Correctly accounting for misclassification when linking latent groups with external variables

Cécile Proust-Lima, Maris Dussartre, Viviane Philipps, Paul Gustafson, Pamela Shaw

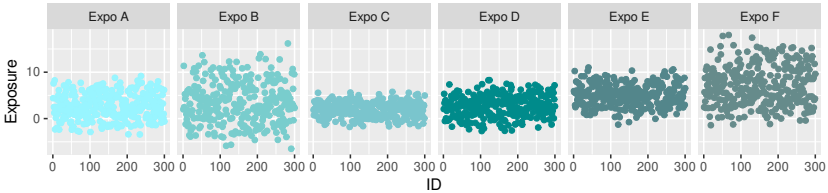
for TG4 of the STRATOS initiative

INSERM U1219, Bordeaux Population Health Research Center, Bordeaux, France
Univ. Bordeaux, ISPED, Bordeaux, France
cecile.proust-lima@inserm.fr

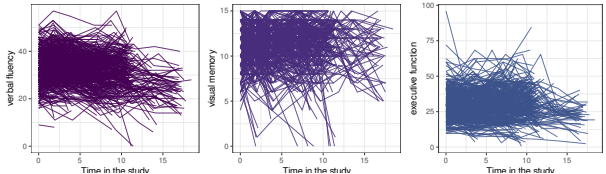
19th International Conference on Applied Statistics
Koper/Capodistria, Slovenia - September, 2023

Context

- Latent class models used to summarize complex information in prospective cohorts:
  - multi-dimensional exposures at baseline: e.g., cardiometabolic health (obesity, activity, glycemia, blood pressure, cholesterol)

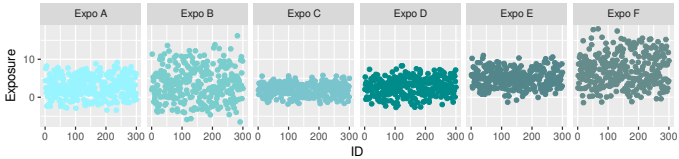


  - trajectories of variables over time: e.g., BMI, cognition, alcool consumption



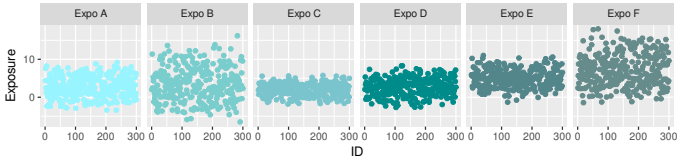
Latent class Strategy

1. Estimate a latent class model on the exposure data:

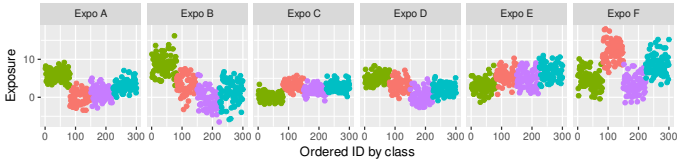


Latent class Strategy

1. Estimate a latent class model on the exposure data:

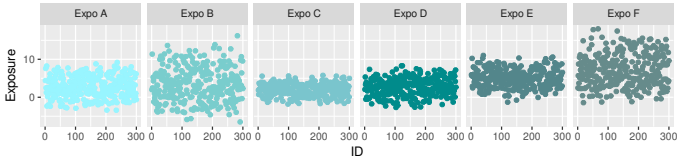


2. Create a classification by assigning each subject to a fitted class:

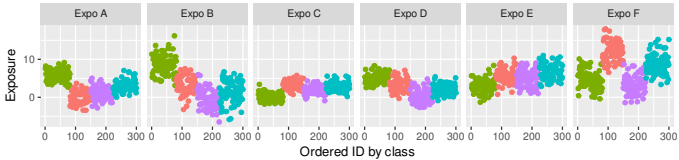


Latent class Strategy

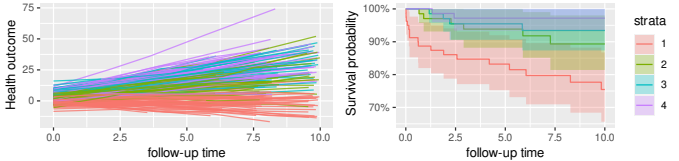
1. Estimate a latent class model on the exposure data:



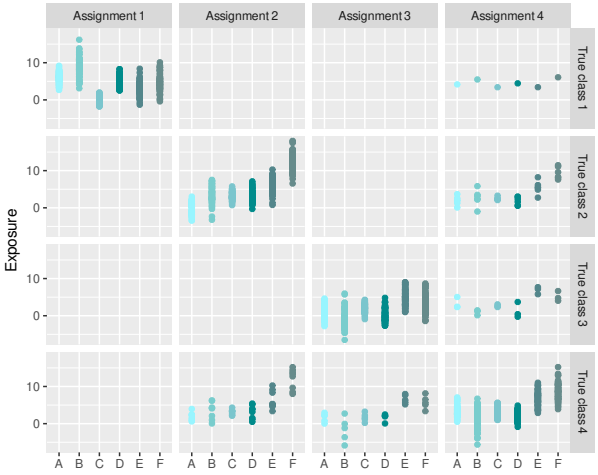
2. Create a classification by assigning each subject to a fitted class:



3. Use this assignment in subsequent analyses



Misclassification issue



⚠ Assignment ≠ Truth

For $k \neq g$,

$$P(\text{assignment} = k \mid \text{true class} = g) \neq 0$$

- Introduces a misclassification / error of classification
- With consequences on the subsequent analyses results

Objective

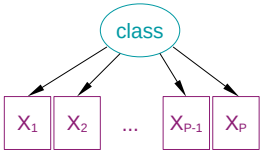
As part of the STRATOS (STRengthening Analytical Thinking for Observational Studies) Topic Group "measurement error and misclassification":

Identify correct statistical strategies to use a latent class structure in subsequent regressions of health outcomes or predictors

- Review the methods in the literature
- Evaluate their performances (i.e., Which ones provide correct/incorrect conclusions?)
- Provide recommendations

Latent class models

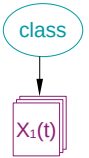
- Latent class $c_i = g$ if subject i belongs to latent class g among G classes
  - Probability of latent class membership : $\pi_{ig} = P(c_i = g)$



- Distribution of the exposures $\boldsymbol{X_i} = (X_{i1}, ..., X_{iP})^\top$ in each latent class g:

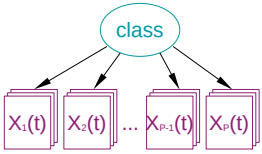
$$\boldsymbol{X}_i | c_i = g \sim \mathscr{D}(\mu_{ig}, V_g)$$

  - Depends on the nature of the data: linear model for continuous data, logistic regression for binary data, mixed models for repeated data, etc.



- Estimation by Maximum Likelihood with individual contribution:

$$\mathscr{L}_i(\theta_G) = \sum_{g=1}^{G} P(c_i = g; \theta_G) P_{\mathscr{D}}(\boldsymbol{X}_i | c_i = g; \theta_G)$$



Posterior Classification

- In which class should be each individual?
  - Posterior class membership:

$$\hat{\pi}_{ig} = P(c_i = g \mid \boldsymbol{X}_i; \hat{\theta}_G)$$

  - Modal assignment: Subject assigned to the class in which (s)he has the highest probability to belong

    $\hat{c}_i = k$ if $P(c_i = k|\boldsymbol{X}_i; \hat{\theta}_G)$ is the highest

  - Proportional assignment: Subject contributes to each class with probability $\hat{\pi}_{ik}$

    $\forall k \in 1, ..., G$: $\hat{c}_i = k$ with weight $\hat{\pi}_{ik}$

ID	$\hat{\pi}_{i1}$	$\hat{\pi}_{i2}$	assigned modal class
Paul	0.75	0.25	1
Maris	0.95	0.05	1
Pamela	0.1	0.9	2
Viviane	0.55	0.45	1
Cécile	0.40	0.60	2

- Discrimination of the classification linked to the separability (Entropy measure)

Methods in the literature