Latent class mixed models for longitudinal data

(Growth mixture models)

Cécile Proust-Lima & Hélène Jacqmin-Gadda

Department of Biostatistics, INSERM U897, University of Bordeaux II

INSERM workshop 205 - june 2010 - Saint Raphael

Analysis of change over time

Linear mixed model (LMM) for describing change over time :

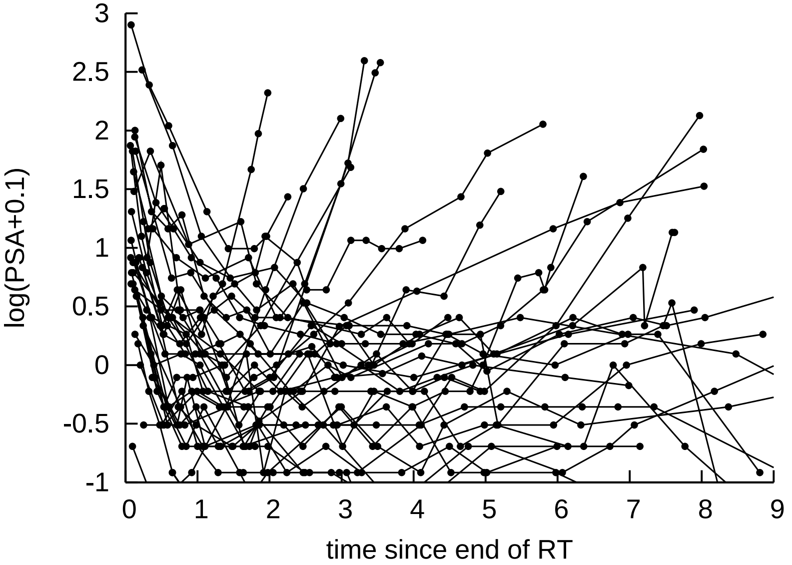
- Correlation between repeated measures (random-effects)
- Single mean profile of trajectory
  - $\rightarrow$ *Restricted to homogeneous populations*

Yet, frequently populations are heterogeneous :

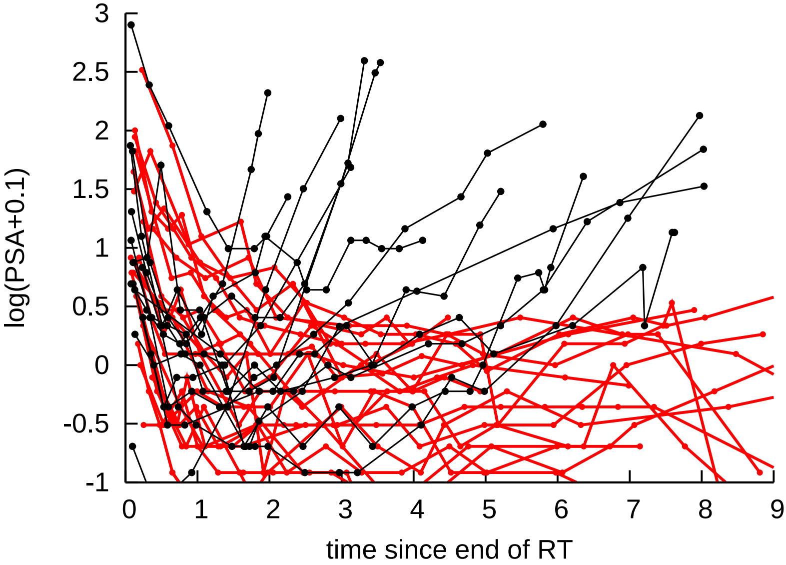
$\rightarrow$ *Latent group structure linked to a behavior, a disease, ...*

- Trajectories of disability before death (Gill, 2010)
- Trajectories of alcohol use in young adults (Muthén, 1999)
- Cognitive declines in the elderly (Proust-Lima, 2009a)
- Progression of prostate cancer after treatment (Proust-Lima, 2009b)

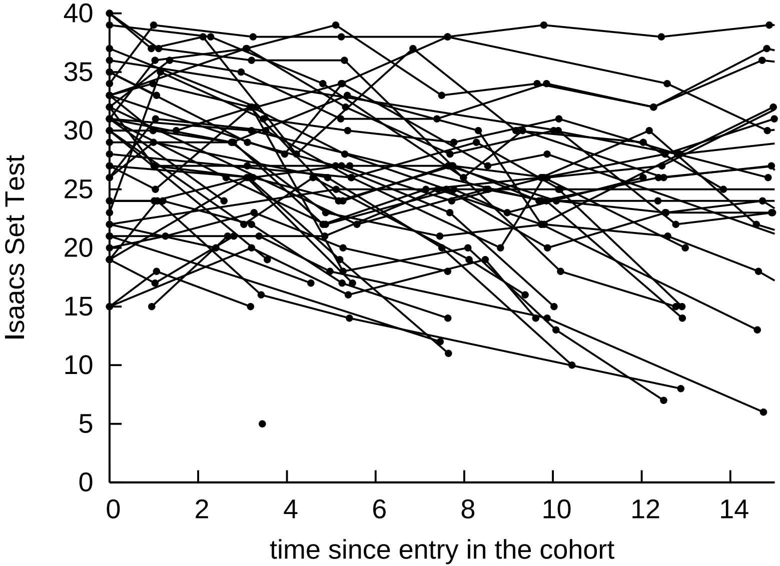
PSA trajectories after radiation therapy (1)



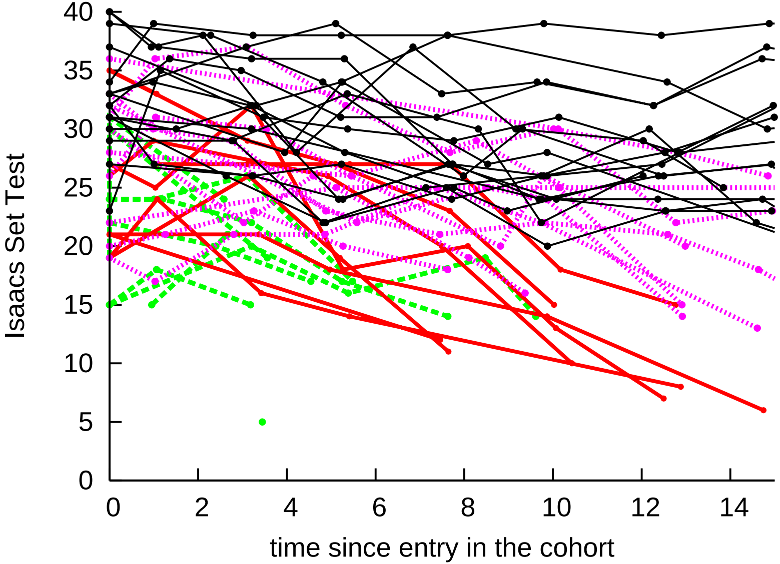
PSA trajectories after radiation therapy (2)



Trajectories of verbal fluency in the elderly (1)



Trajectories of verbal fluency in the elderly (2)



Latent class mixed models / growth mixture models / heterogeneous mixed models

Accounts for both individual variability and latent group structure (Verbeke,1996 ; Muthén, 1999)

$\rightarrow$ Extension of LMM to account for heterogeneity

$\rightarrow$ Extension of LCGA to account for individual variability

Two submodels :

- Probability of latent class membership
- Class-specific trajectory

$\rightarrow$ *both according to covariates/predictors*

Outline of the talk

Methodology

- Latent class linear mixed model
- Estimation methods
- Posterior classification
- Goodness of fit evaluation

Application

- PAQUID cohort of cognitive aging
- Heterogeneous profiles of verbal fluency and predictors

Discussion

Probability of latent class membership

Population of N subjects (subscript i, $i = 1, ..., N$)

G latent homogeneous classes (subscript g, $g = 1, ..., G$)

Discrete latent variable c_i for the latent group structure :

$c_i = g$ if subject i belongs to class g ($g = 1, ..., G$)

$\rightarrow$ *Every subject belongs to only one latent class*

Probability of latent class membership explained according to covariates X_{1i} *(multinomial logistic regression)* :

$$\pi_{ig} = P(c_i = g|X_{1i}) = \frac{e^{\xi_{0g}+X'_{1i}\xi_{1g}}}{\sum_{l=1}^{G} e^{\xi_{0l}+X'_{1i}\xi_{1l}}}$$

with $\xi_{0G} = 0$ and $\xi_{1g} = 0$ i.e. class G = reference class

Class-specific LMM : notations and simple example

Change over time of a longitudinal outcome Y :

- Y_{ij} repeated measure for subject i at occasion j, $j = 1, ..., n_i$
- t_{ij} time of measurement at occasion j, $j = 1, ..., n_i$

  Number & times of measurements can differ across subjects

Linear change over time (without adjustment for covariates) :

$$Y_{ij}|_{c_i=g} = u_{0ig} + u_{1ig} \times t_{ij} + \epsilon_{ij}$$

Class-specific random-effects $(u_{0ig}, u_{1ig})' \sim N((\mu_{0g}, \mu_{1g})', B_g)$

- μ_{0g} and μ_{1g} class-specific mean intercept and slope
- B_g class-specific variance-covariance (usually $B_g = B$ or $B_g = w_g^2 B$)

Independent errors of measurement $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$

Class-specific LMM : General formulation

$$Y_{ij}|_{c_i=g} = Z'_{ij}u_{ig} + X'_{2ij}\beta + X'_{3ij}\gamma_g + \epsilon_{ij}$$

Z_{ij}, X_{2ij}, X_{3ij} : 3 different vectors of covariates without overlap

- $\rightarrow$ Z_{ij} vector of time functions :

  $Z_{ij} = (1, t_{ij}, t_{ij}^2, t_{ij}^3, ...)$ for polynomial shapes

  $Z_{ij} = (B_1(t_{ij}), ..., B_K(t_{ij}))$ for shapes approximated by splines

  $Z_{ij} = (f_1(t_{ij}), ..., f_K(t_{ij}))$ for shapes defined by a set of K parametric functions

- $\rightarrow$ X_{2ij} set of covariates with common effects over classes β

- $\rightarrow$ X_{3ij} set of covariates with class-specific effects γ_g

Estimation of the parameters

Estimation of θ_G for a fixed number of latent classes G :

$\rightarrow$ in a Bayesian framework (Komarek, 2009 ; Elliott, 2005)

$\rightarrow$ in a maximum likelihood framework (Verbeke, 1996 ; Muthén, 1999, 2004 ; Proust, 2005)

$$L(\theta_G) = \sum_{i=1}^{N} \ln \left(\sum_{g=1}^{G} P(c_i = g | X_{1i}, \theta_G) \times \phi_{ig}(Y_i | c_i = g; X_{2i}, X_{3i}, Z_i, \theta_G) \right)$$

with

X_{2i}, X_{3i}, Z_i : matrices of n_i row vectors X_{2ij}, X_{3ij}, Z_{ij} resp.

ϕ_{ig} pdf of $MVN(X_{2i}\beta + X_{3i}\gamma_g + Z_i\mu_g, Z_i B_g Z_i' + \sigma_\epsilon^2 I_{n_i})$

Estimation of LCLMM

Estimation of $\hat{\theta}_G$ for a fixed G

★ *Multiple possible maxima* $\Rightarrow$ *grid of initial values* ★

Selection of the optimal number of latent classes :

- Bayesian Information Criterion (BIC) (Bauer, 2003)
- Deviance Information Criterion (DIC, DIC$_3$,etc) (Celeux, 2006)
- Other possible tests (Lo, 2001 ; Nylund, 2007)

Programs available :

- Mplus (Muthén, 2001)
- R function GLMM_MCMC in mixAK package (Komarek, 2009)
- R function hlme in lcmm package (Proust-Lima, 2010)
- GLLAMM in Stata (Rabe-Hesketh, 2005)

Posterior classification

Posterior probability of class membership

For a subject i in latent class g :

$$\hat{\pi}_{ig} = P(c_i = g | X_i, Y_i, \hat{\theta})$$
$$= \frac{P(c_i = g \mid X_{1i}, \hat{\theta}) \phi_{ig}(Y_i | c_i = g, \theta)}{\sum_{l=1}^{G} P(c_i = l \mid X_{1i}, \hat{\theta}) \phi_{il}(Y_i | c_i = l, X_{2i}, X_{3i}, Z_i, \hat{\theta})}$$

Posterior classification : $\hat{c}_i = \text{argmax}_g(\hat{\pi}_{ig})$
$\rightarrow$ *Class in which the subject has the highest posterior probability*

Goodness-of-fit 1 : Classification

Is the classification very discriminative ? Or is it ambiguous ?

Table of posterior classification :

Final class	$\sharp$	Mean of the probabilities of belonging to each class 1	…	g	…	G
1	N_1	$\frac{1}{N_1}\sum_{i=1}^{N_1}\hat{\pi}_{i1}$	…	$\frac{1}{N_1}\sum_{i=1}^{N_1}\hat{\pi}_{ig}$	…	$\frac{1}{N_1}\sum_{i=1}^{N_1}\hat{\pi}_{iG}$
$\vdots$		$\vdots$	$\ddots$			$\vdots$
g	N_g	$\frac{1}{N_g}\sum_{i=1}^{N_g}\hat{\pi}_{i1}$	…	$\frac{1}{N_g}\sum_{i=1}^{N_g}\hat{\pi}_{ig}$	…	$\frac{1}{N_g}\sum_{i=1}^{N_g}\hat{\pi}_{iG}$
$\vdots$		$\vdots$			$\ddots$	$\vdots$
G	N_G	$\frac{1}{N_G}\sum_{i=1}^{N_G}\hat{\pi}_{i1}$	…	$\frac{1}{N_G}\sum_{i=1}^{N_G}\hat{\pi}_{ig}$	…	$\frac{1}{N_G}\sum_{i=1}^{N_G}\hat{\pi}_{iG}$

Goodness-of-fit 2 : Longitudinal predictions

Does the longitudinal model correctly fit the data ?

Class-specific marginal (M) & subject-specific (SS) predictions :

- $\hat{Y}_{ijg}^{(M)} = X_{2ij}^T \hat{\beta} + X_{3ij}^T \hat{\gamma}_g + Z_{ij}^T \hat{\mu}_g$
- $\hat{Y}_{ijg}^{(SS)} = X_{2ij}^T \hat{\beta} + X_{3ij}^T \hat{\gamma}_g + Z_{ij}^T \hat{\mu}_g + Z_{ij}^T \hat{u}_{ig}$

  with bayes estimates $\hat{u}_{ig} = \omega_g^2 B Z_i^T V_{ig}^{-1} (Y_i - X_{2i} \hat{\beta} + X_{3i} \hat{\gamma}_g + Z_i \hat{\mu}_g)$

Weighted average over classes (and corresponding residuals) :

$$\hat{Y}_{ij}^{(M)} = \sum_{g=1}^{G} \hat{\pi}_{ig} \hat{Y}_{ijg}^{(M)} \qquad \hat{R}_{ij}^{(M)} = Y_{ij} - \hat{Y}_{ij}^{(M)}$$

$$\hat{Y}_{ij}^{(SS)} = \sum_{g=1}^{G} \hat{\pi}_{ig} \hat{Y}_{ijg}^{(SS)} \qquad \hat{R}_{ij}^{(SS)} = Y_{ij} - \hat{Y}_{ij}^{(SS)}$$

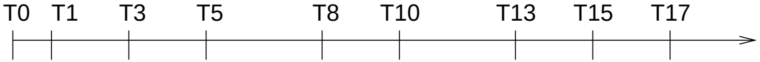
4 kinds of possible analyses in LCLMM

1. Exploration of unconditioned and unadjusted trajectories
   - no covariates in the LMM & the class-membership model
     $\rightarrow$ *raw heterogeneity*
2. Exploration of unconditioned adjusted trajectories
   - covariates in the LMM
     $\rightarrow$ *residual heterogeneity after adjustment for known factors of change over time*
3. Exploration of conditioned unadjusted trajectories
   - covariates in the class-membership model
     $\rightarrow$ *heterogeneity explained by 'targeted' factors*
4. Exploration of conditioned and adjusted trajectories
   - covariates in the class-membership model & the LMM
     $\rightarrow$ *residual heterogeneity explained by 'targeted' factors*

PAQUID cohort

Population-based prospective cohort of cognitive aging

- 3777 subjects of 65 years and older in South West France
  (random selection from electoral rolls)
- Follow-up every 2-3 years :



At each visit :

- Neuropsychological battery
- Two phase diagnosis of dementia
- & Information about health, activities, etc

Verbal fluency trajectories

Verbal fluency impaired early in pathological aging (Amieva, 2008)

$\rightarrow$ *description of heterogeneous trajectories*

- Verbal fluency measured by IST (Isaacs Set Test)
- Quadratic trajectory according to time from entry
- Patients included :
  - Not initially demented
  - At least one measure at IST in T0 - T17
- Covariates of interest :
  - First evaluation effect *(adjustment in the LMM)*
  - Gender, education, age at entry *(classes predictors)*

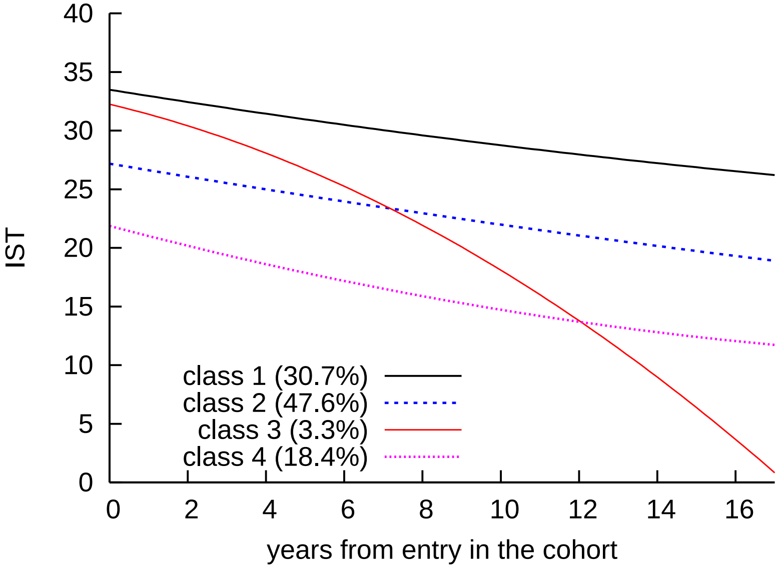
Heterogeneity predicted by gender, education and age

Estimation for a varying number of latent classes :

G	p*	L	BIC	Frequency of the latent classes (%)				
				1	2	3	4	5
1	11	-40651.3	81392.4	100				
2	18	-40104.8	80356.6	51.3	48.7			
3	25	-40015.7	80235.6	29.6	50.6	19.8		
4	**32**	**-39941.3**	**80144.0**	**30.7**	**47.6**	**3.3**	**18.4**	
5	39	-39922.2	80163.0	31.4	1.8	31.4	20.7	14.7

* number of parameters

Trajectories of verbal fluency in the elderly



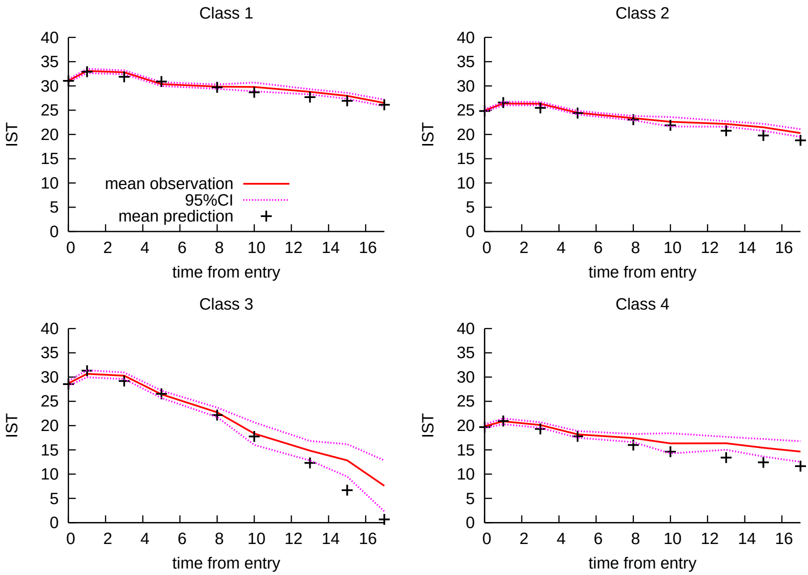
Predictors of class-membership

predictor	class	estimate	SE	p-value
male	1	0.051	0.219	0.817
male	2	0.115	0.197	0.561
male	3	0.530	0.291	0.068
male	4	0		
education+	1	5.150	0.587	$<$0.0001
education+	2	1.742	0.298	$<$0.0001
education+	3	4.379	1.158	0.0001
education+	4	0		
age at entry	1	-0.447	0.046	$<$0.0001
age at entry	2	-0.215	0.027	$<$0.0001
age at entry	3	-0.173	0.040	$<$0.0001
age at entry	4	0		

Posterior classification table

Final classif.	Number of subjects (%)	Mean of the class-membership probabilities in class (in %) : 1	2	3	4
1	1083 (30.7%)	**81.1**	14.4	4.5	<0.1
2	1679 (47.6%)	9.7	**74.0**	5.5	10.8
3	117 (3.3%)	10.8	20.4	**67.7**	1.1
4	648 (18.4%)	<0.2	19.2	0.7	**80.1**

Weighted marginal predictions and observations



Advantages of LCLMM

- Accounts for 2 sources of variability
  - individual variability through random-effects
    $\rightarrow$ *inference possible*
  - latent group structure
    $\rightarrow$ *mean profiles of trajectory (different from LCGA)*
- MAR assumption for missing data and dropout
- Individually varying time (age / exact follow-up)
- Includes flexibly covariates :
  $\rightarrow$ *different questions addressed*

Limits of LCLMM

- Starting values & local solutions (Hipp, 2006)
  - vary the starting values **extensively**
  - compare various solutions to determine the stability of the model
  - assess the frequency of the solution

- Interpretation of the latent classes (Bauer, 2003 + discutants)

  Flexible model that can fit better homogeneous populations

  $\rightarrow$ *relevant assumption of latent groups*

  $\rightarrow$ *evaluation of goodness-of-fit*

- Linear models for Gaussian outcomes only

  $\rightarrow$ *same extension for nonlinear mixed models*

  $\rightarrow$ *same extension for multivariate mixed models*

References

- Amieva H, Le Goff M, Millet X, et al. (2008). *Annals of Neurology*, 64, 492 :8
- Bauer DJ, Curran PJ (2003). *Psychol Meth*,8(3), 338 :63 (+ discutants 364 :93)
- Elliott MR, Ten Have TR, Gallo J, et al. (2005). *Biostatistics*, 6, 119 :43.
- Gill TM, Gahbauer EA, Han L, Allore HG (2010). *The New England journal of medicine*, 362, 1173 :80
- Hipp JR, Bauer DJ (2006). *Psychological methods*, 11(1), 36 :53
- Komarek A (2009). *Computational Statistics and Data Analysis*, 53(12), 3932 :47
- Lo Y, Mendell NR, Rubin DB (2001). *Biometrika*,88(3),767 :78
- Muthén B0, Shedden K (1999). *Biometrics*,55(2),463 :9
- Muthén LK, Muthén B0 (2004). *Mplus user's guide*. `www.statmodel.com`
- Nylund KL, Asparouhov T, Muthén BO (2007). *Structural Equation Modeling*,14,535 :69
- Proust C, Jacqmin-Gadda H (2005). *Computer Methods and Programs in Biomedicine*,78,165 :73
- Proust-Lima C, Joly P, Jacqmin-Gadda H (2009a). *Computational Statistics and Data Analysis*,53,1142 :54
- Proust-Lima C, Taylor J (2009b). *Biostatistics*,10,535 :49
- Rabe-Hesketh S, Skrondal A, Pickles A (2004). *Psychometrika*, 69 , 167 :90
- Verbeke G, Lesaffre E (1996). *Journal of the American Statistical Association*,91(433),217 :21