Developmental trajectories:

the random effect and latent class approach

Cécile Proust-Lima

INSERM U897, Epidemiologie et Biostatistique, Bordeaux, France
Univ. Bordeaux, ISPED, Bordeaux, France

cecile.proust-lima@inserm.fr

Longitudinal analysis in cohort studies

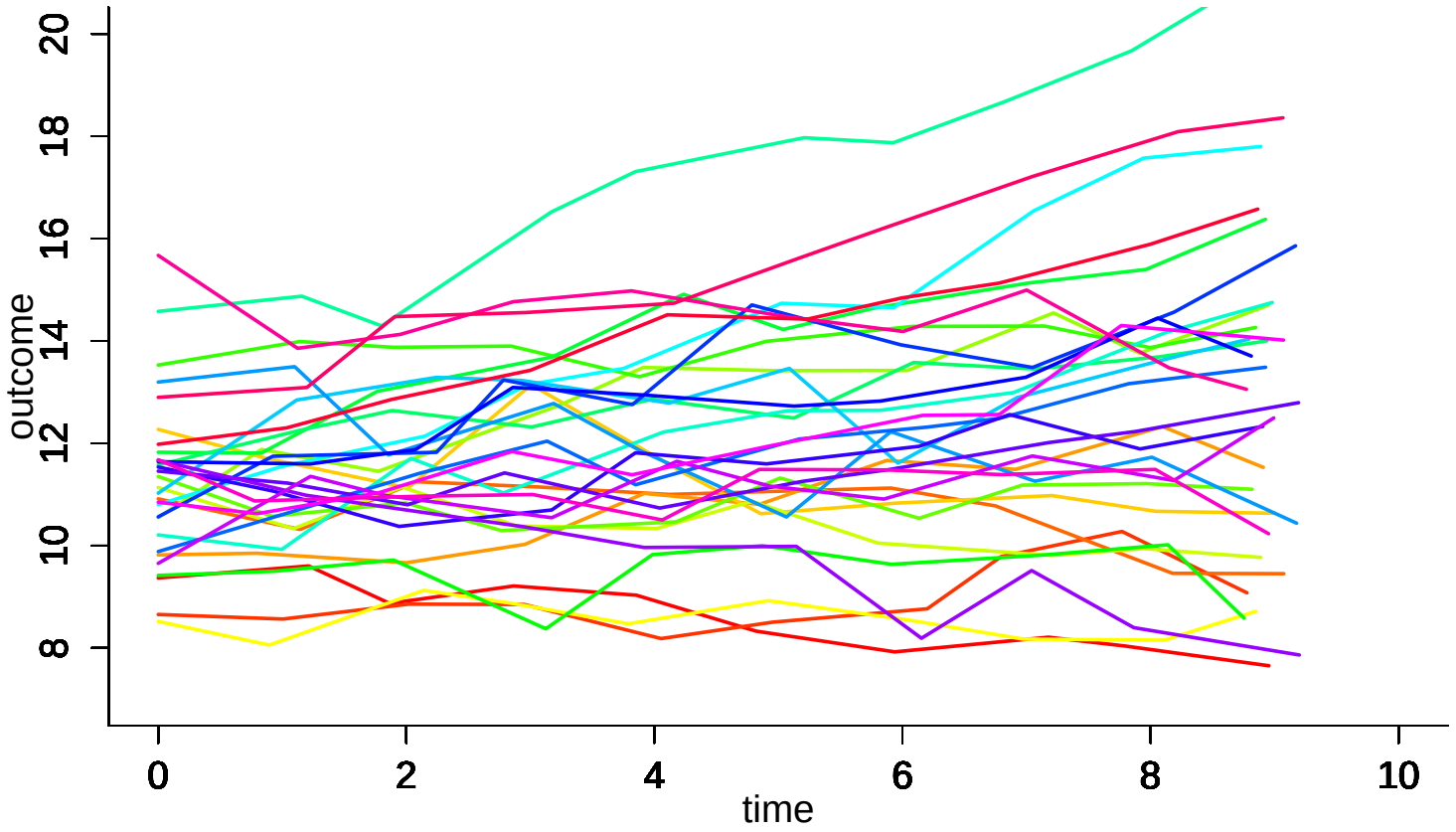
Objective :

- describe the developmental trajectory of a biological or psychological process over time (at the population level)
- assess predictors of this developmental trajectory (at the population level)
- make predictions (at the individual level)

Type of data :

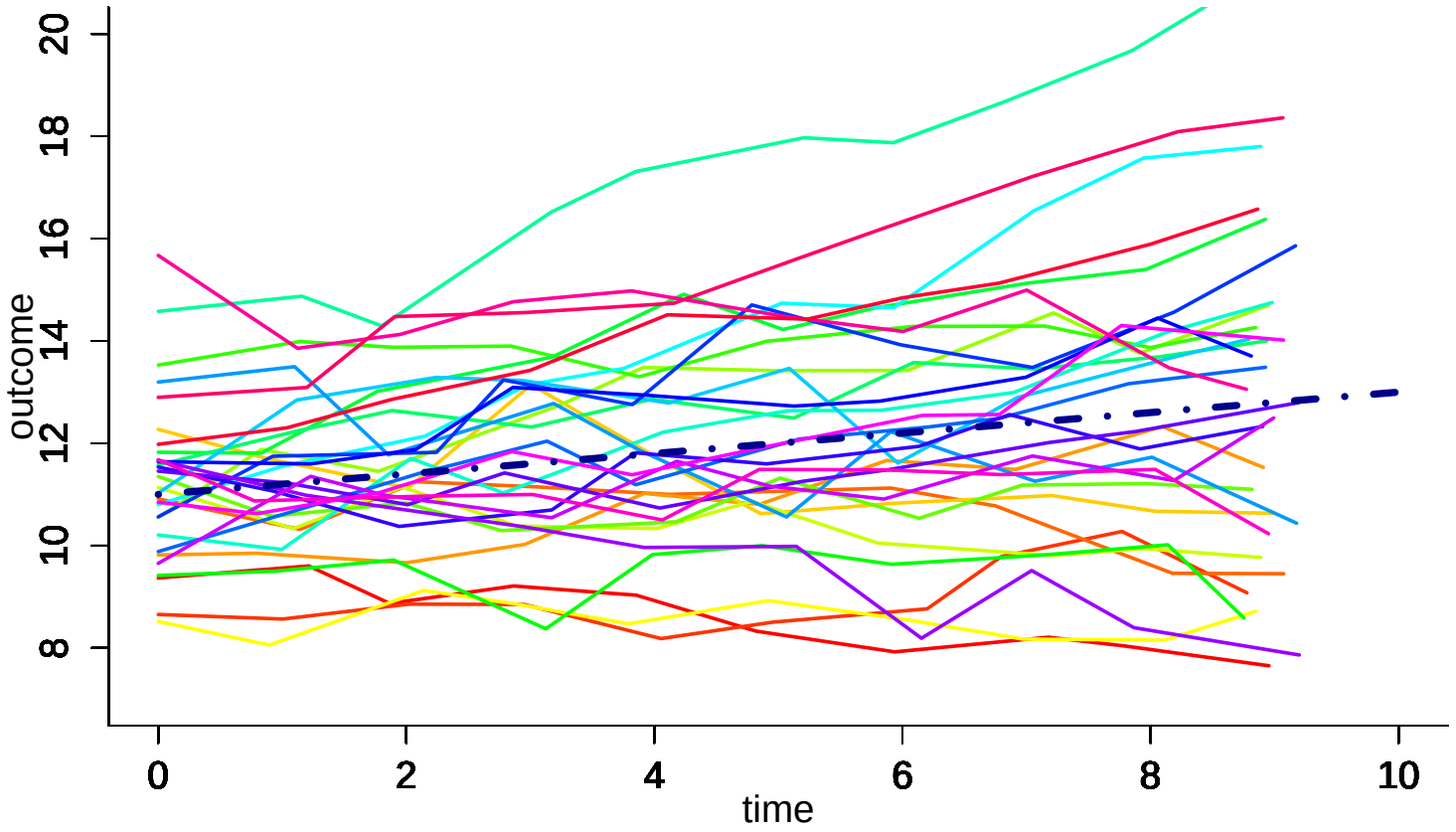
- sample of N subjects
- repeated measures of marker Y over time for each subject (with varying times over subjects)

Toy example



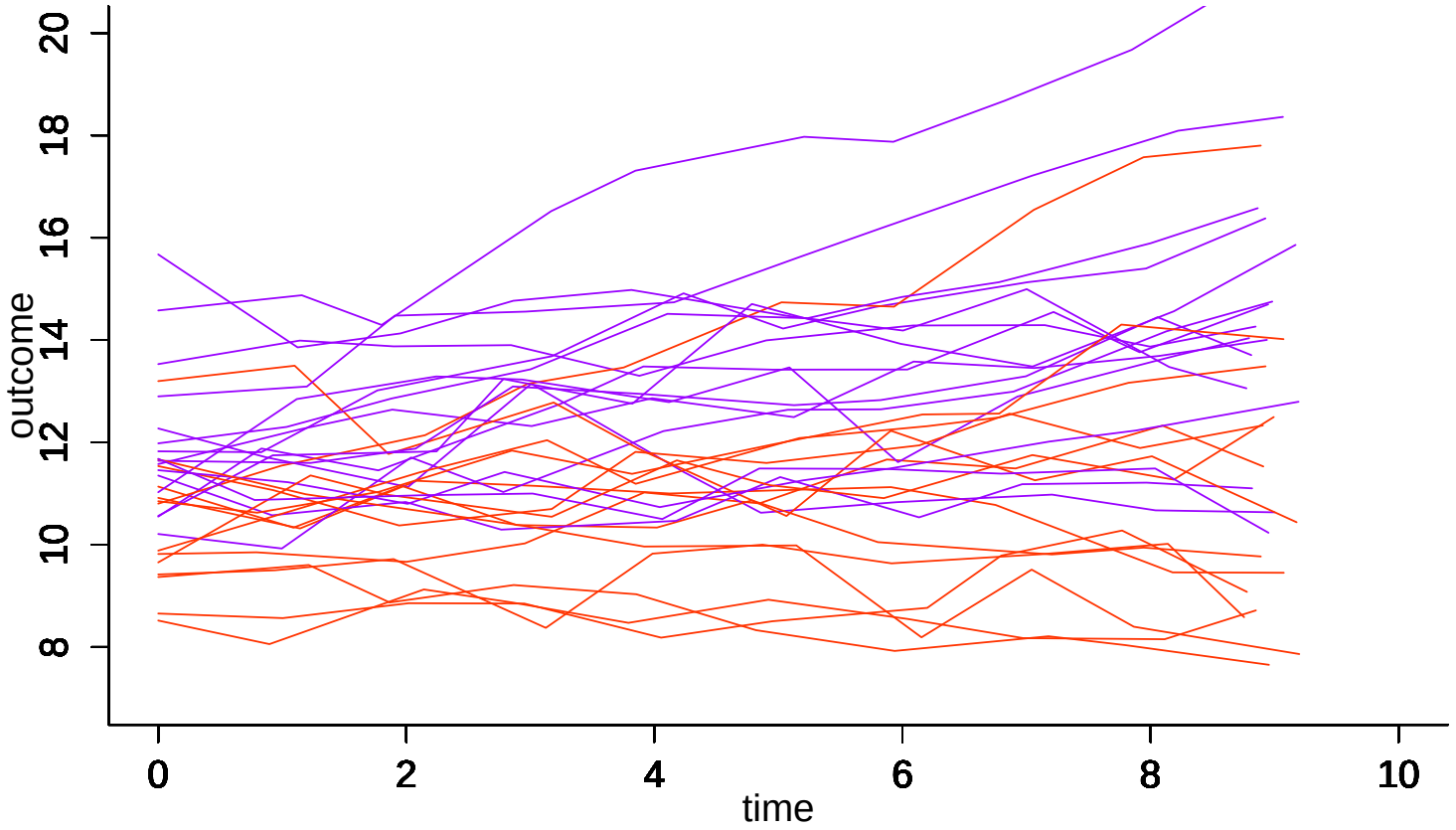
From individual trajectories

Toy example



From individual trajectories we want to estimate the mean trajectory in the sample

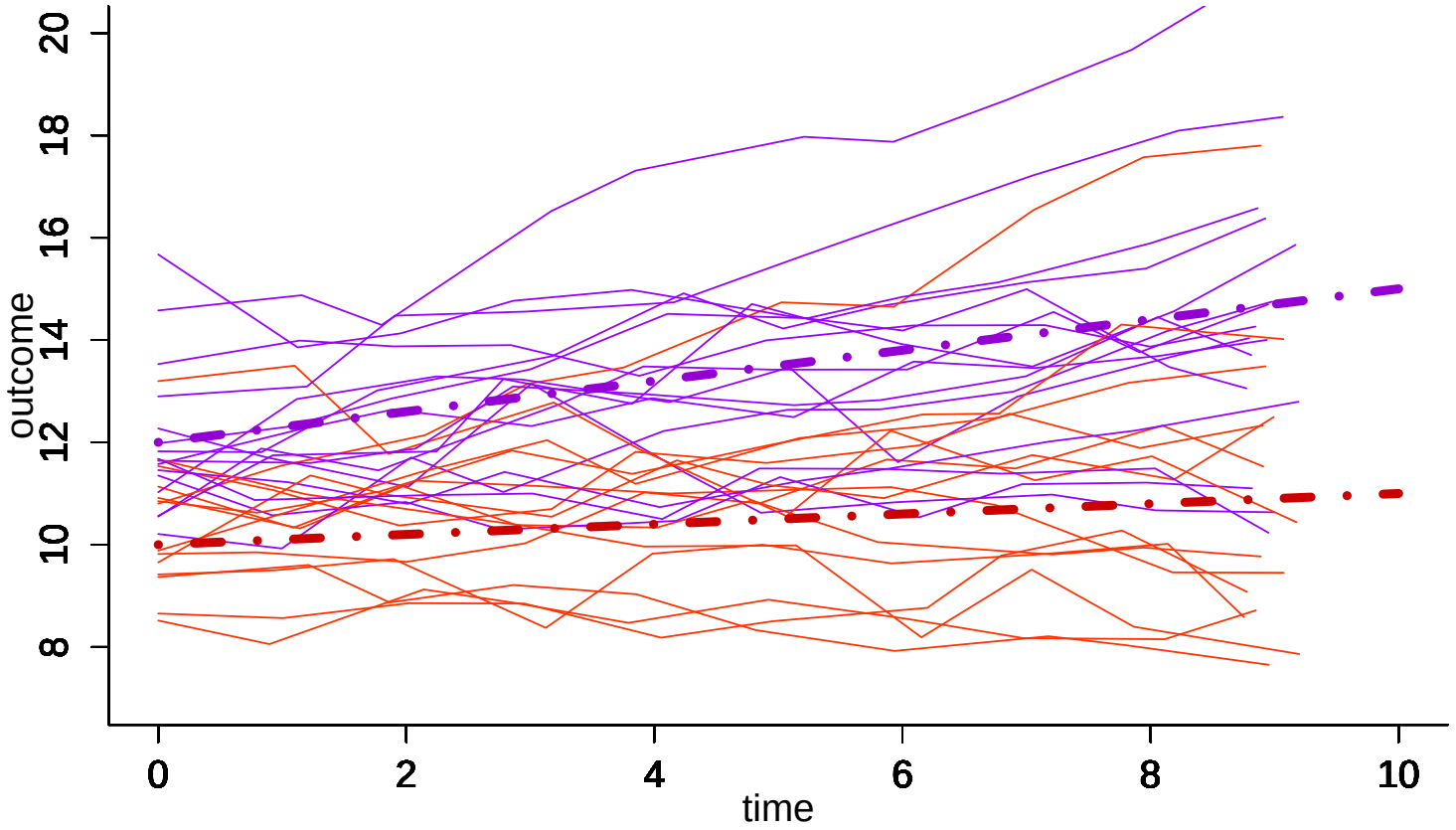
Toy example (cont'd)



Heterogeneity can be explained by covariates :

treatment, exposure, socio-demographic characteristics, etc.

Toy example (cont'd)



Heterogeneity can be explained by covariates :

treatment, exposure, socio-demographic characteristics, etc.

estimate the mean trajectory for each level of the covariate

Longitudinal analysis in cohort studies

Objective :

- describe the developmental trajectory of a biological or psychological process over time (at the population level)
- assess predictors of this developmental trajectory (at the population level)
- make predictions (at the individual level)

Type of data :

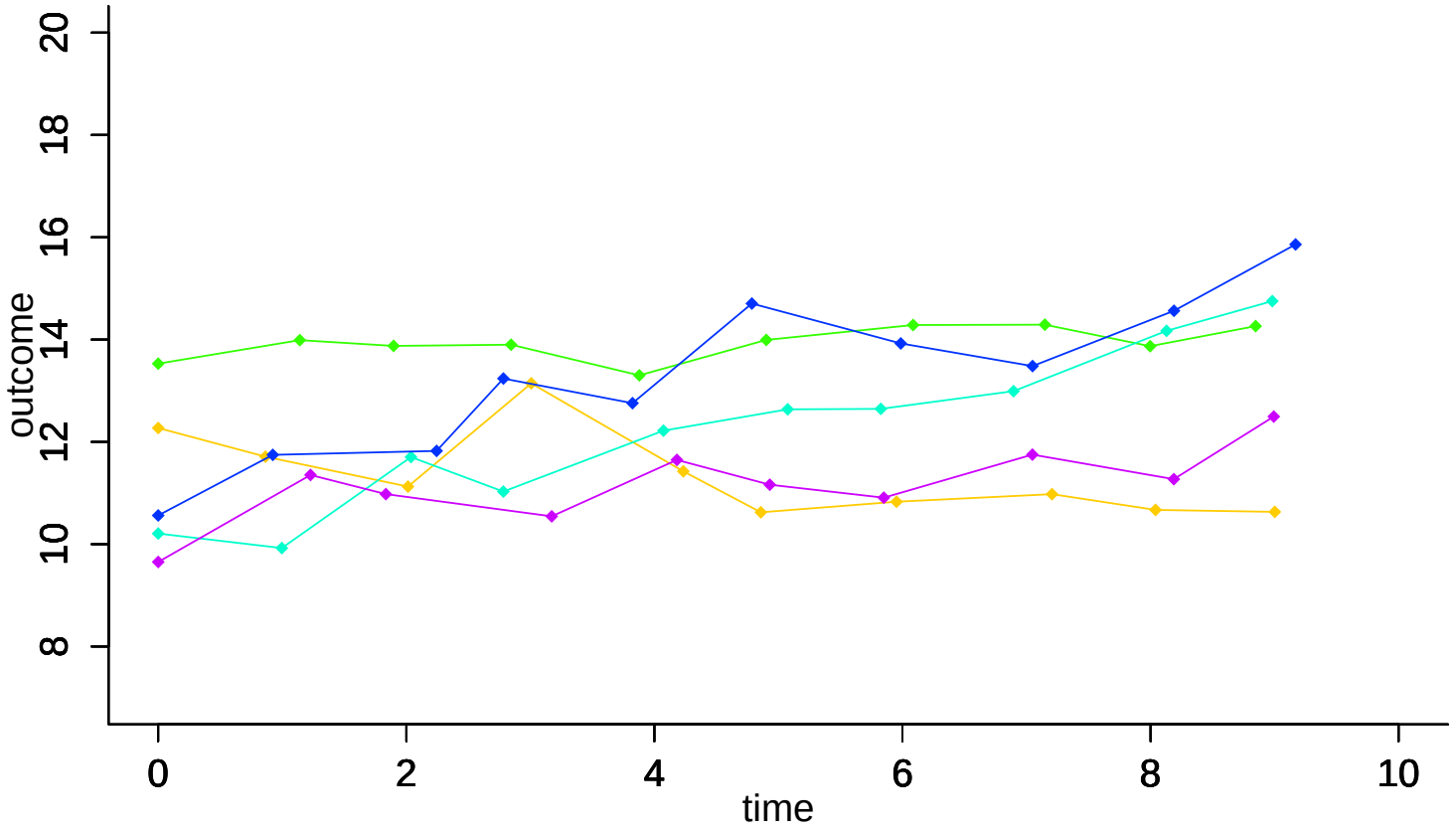
- sample of N subjects
- repeated measures of marker Y over time for each subject (with varying times over subjects)

Challenges :

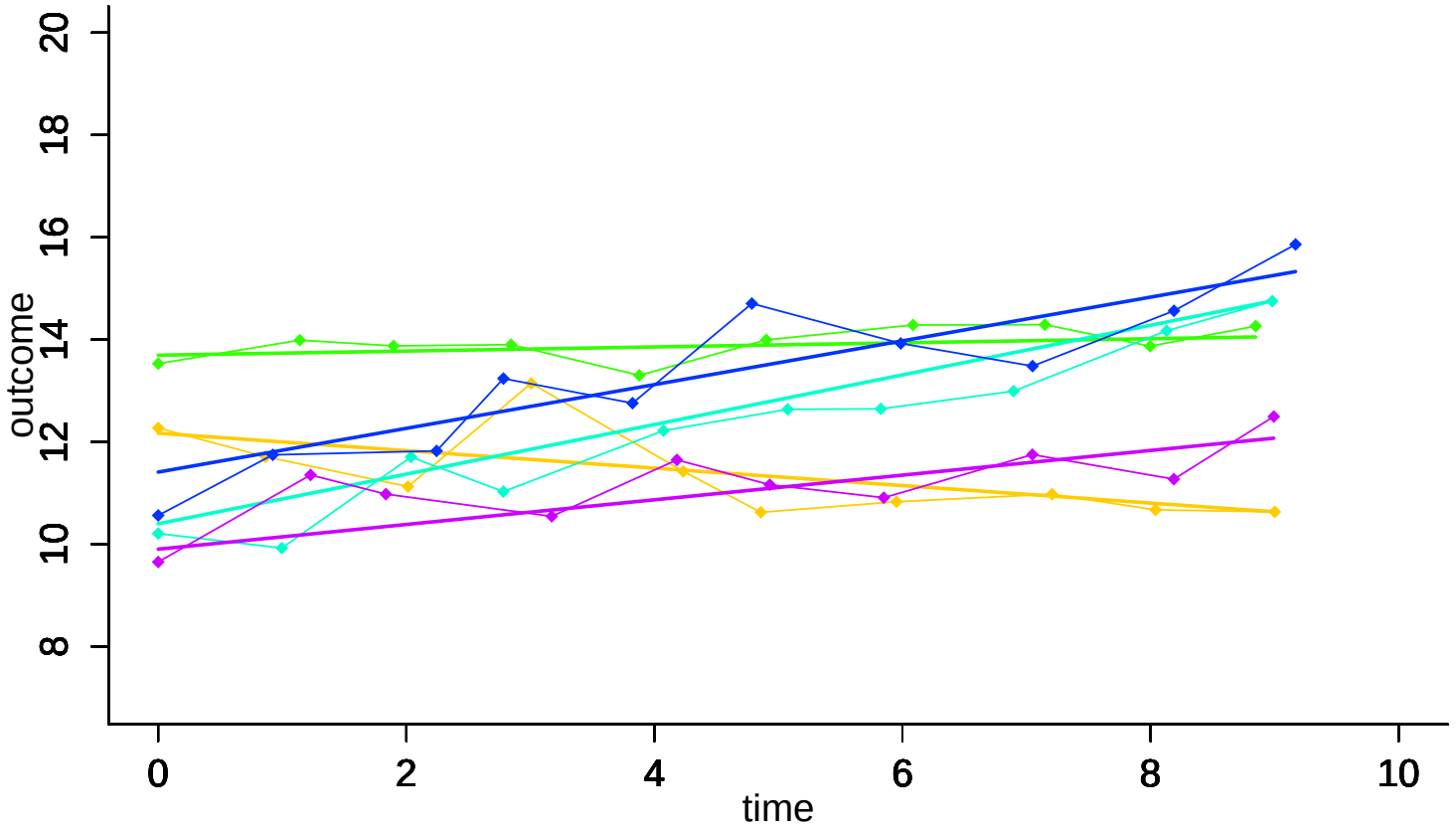
- take into account the correlation within each subject
- take into account the heterogeneity between subjects
- obtain estimations at the population level and at the individual level

$\rightarrow$ the linear mixed model (LMM) theory, the random-effect models

Focus on 5 subjects



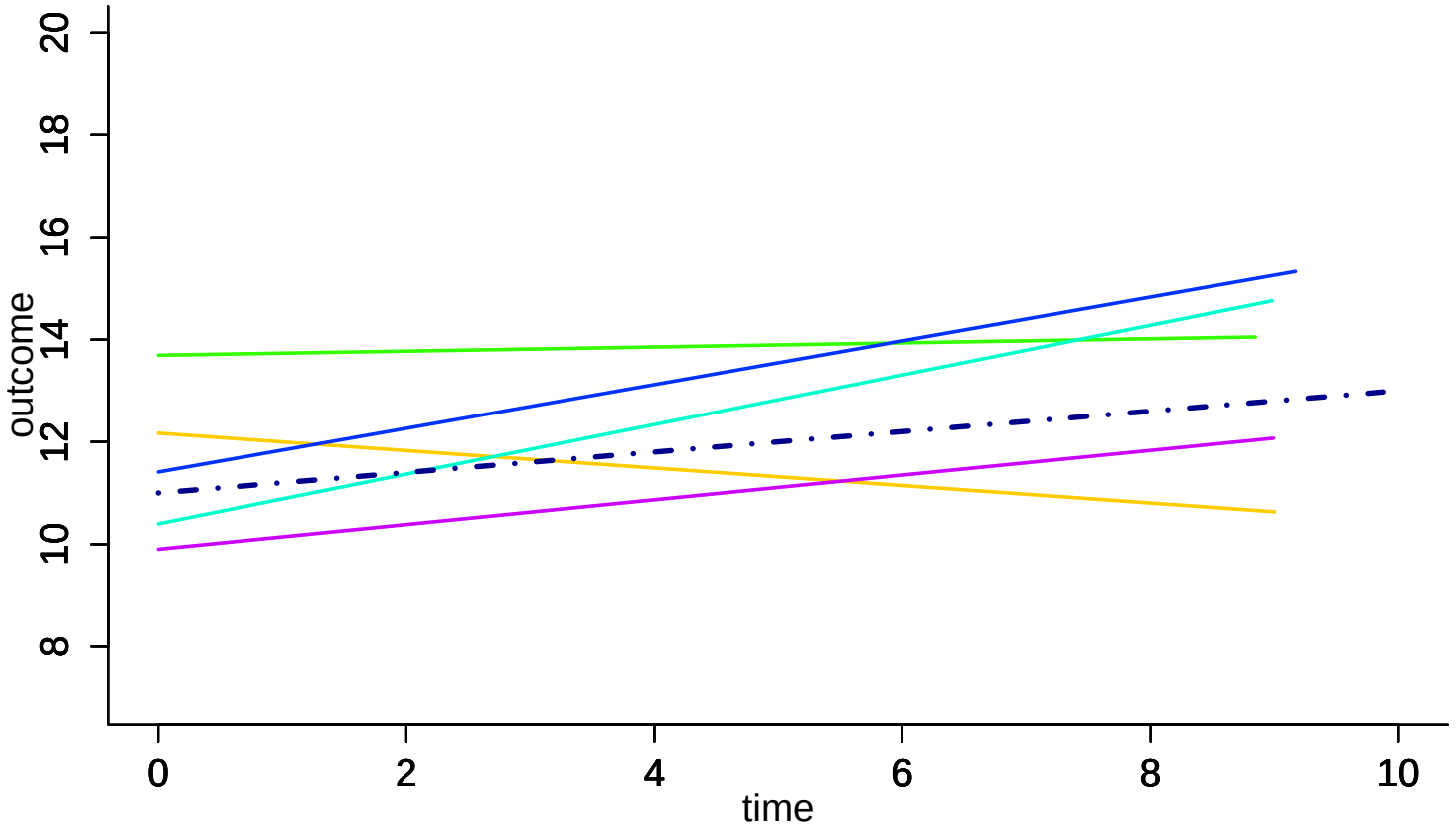
Focus on 5 subjects



Two levels of interest :

- the individual level with the individual trajectory around the noisy measures

Focus on 5 subjects



Two levels of interest :

- the individual level with the individual trajectory around the noisy measures
- the population level with the mean trajectory around the individual deviations

The linear mixed model definition

For subject i at occasion j :

$$
\begin{aligned}
Y_{ij} = Y_i(t_{ij}) &= \beta_0 + \beta_1 \times t_{ij} && \text{at the population level} \\
&+ u_{0i} + u_{1i} \times t_{ij} && \text{at the individual level} \\
&+ \epsilon_{ij}
\end{aligned}
$$

with $u_i \sim \mathcal{N}(0, B)$ and $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$

The linear mixed model definition

For subject i at occasion j (and the binary covariate C) :

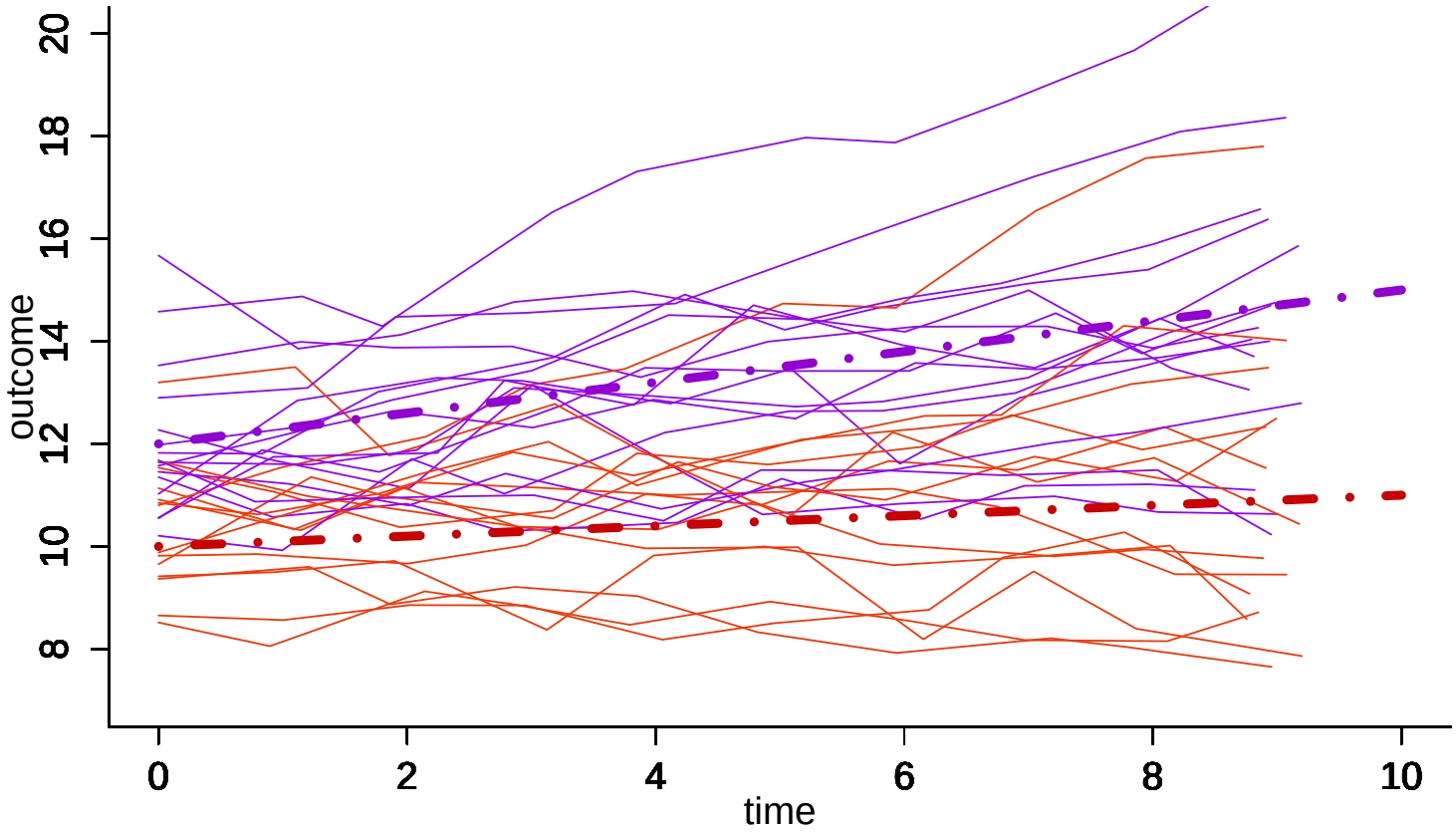
$$\begin{aligned} Y_{ij} = Y_i(t_{ij}) &= \beta_0 + \beta_1 \times t_{ij} + \beta_2 C_i + \beta_3 C_i \times t_{ij} && \text{at the population level} \\ &+ u_{0i} + u_{1i} \times t_{ij} && \text{at the individual level} \\ &+ \epsilon_{ij} \end{aligned}$$

with $u_i \sim \mathcal{N}(0, B)$ and $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$

Covariate-specific mean trajectory :

$$\begin{aligned} E(Y_i(t)|C=0) &= \beta_0 + \beta_1 \times t \\ E(Y_i(t)|C=1) &= (\beta_0 + \beta_2) + (\beta_1 + \beta_3) \times t \end{aligned}$$

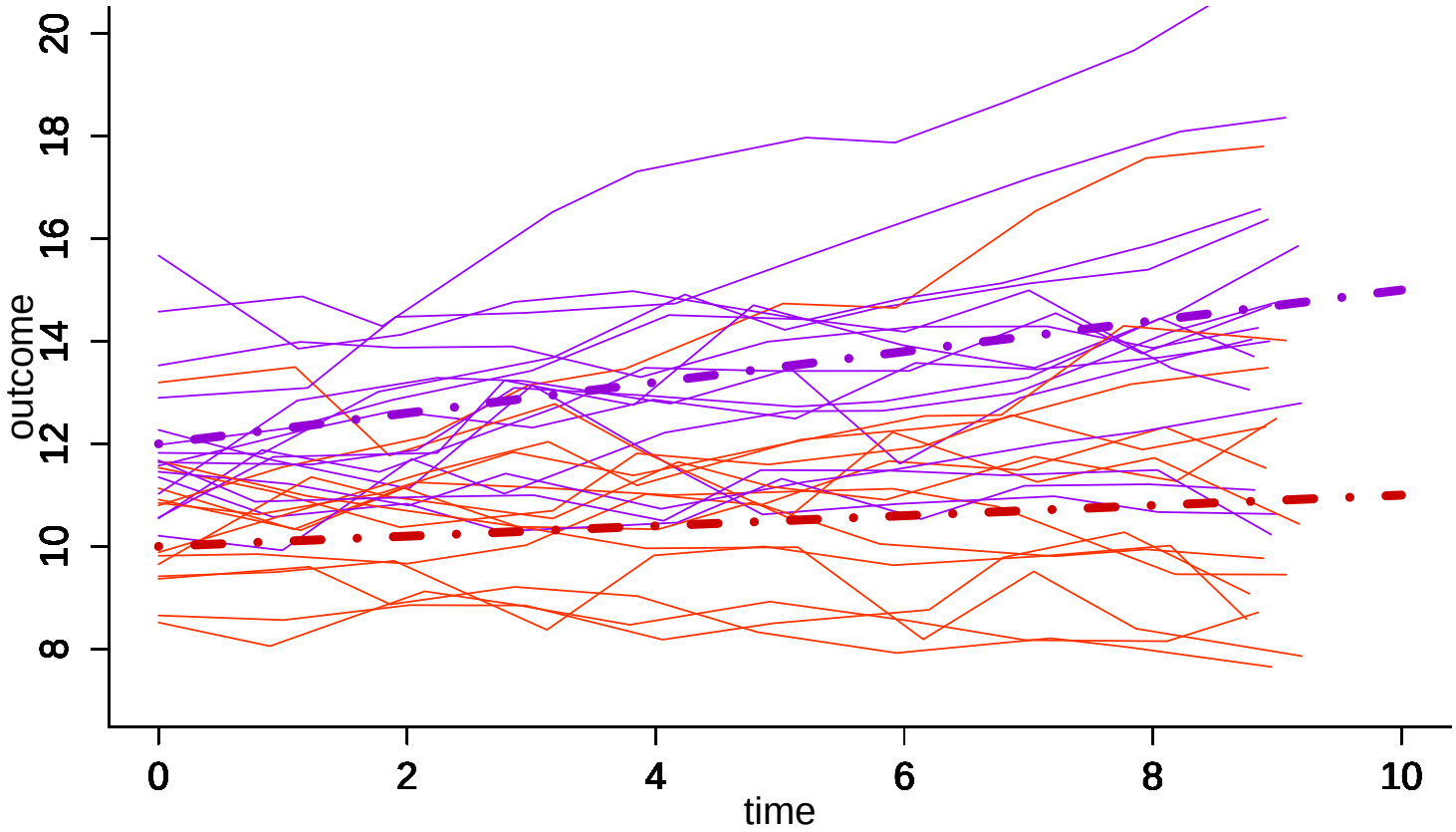
What happens when C not observed ?



Heterogeneity suspected but characteristics not observed :

underlying disease, specific behavior, genetic profile, etc.

What happens when C not observed ?



Heterogeneity suspected but characteristics not observed :

underlying disease, specific behavior, genetic profile, etc.

$\rightarrow$ C becomes a latent variable, **a latent class**

The latent class mixed model (LCMM) or growth mixture model

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

- Y_{ij} repeated measure of the outcome for subject i at occasion j, $j = 1, ..., n_i$
- t_{ij} time of measurement at occasion j, $j = 1, ..., n_i$
- X_i vector of time-independent covariates

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

- c_i discrete latent variable 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*

Two submodels :

- Probability of latent class membership
- Class-specific trajectory of the marker
  both according to observed covariates/predictors

Example of LCMM specification

Probability of latent class membership explained according to covariates X_i :

$\rightarrow$ *multinomial logistic regression*

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

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

Example of LCMM specification

Probability of latent class membership explained according to covariates X_i :

$\rightarrow$ *multinomial logistic regression*

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

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

Class-specific trajectory : linear trajectory example

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

with $u_{ig} = u_i|_{c_i=g} = (u_{0ig}, u_{1ig})' \sim \mathcal{N}((\mu_{0g}, \mu_{1g})', B_g)$ class-specific RE
and $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$, $\epsilon_{ij} \perp u_{ig}$

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

Example of LCMM specification

Probability of latent class membership explained according to covariates X_i :

$\rightarrow$ *multinomial logistic regression*

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

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

Class-specific trajectory : linear trajectory example with observed covariates :

$$Y_{ij}|_{c_i=g} = u_{0ig} + u_{1ig}t_{ij} + \beta_1 X_i + \beta_2 X_i t_{ij} + \epsilon_{ij}$$

with $u_{ig} = u_i|_{c_i=g} = (u_{0ig}, u_{1ig})' \sim \mathcal{N}((\mu_{0g}, \mu_{1g})', B_g)$ class-specific RE
and $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$, $\epsilon_{ij} \perp u_{ig}$

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

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

$u_{ig} = u_i|_{c_i=g} \sim \mathcal{N}(\mu_g, B_g)$ and $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$, $\epsilon_{ij} \perp u_{ig}$

Estimation of LCMM

- Estimation for a fixed number of latent classes
- Mostly estimated by Maximum Likelihood ($\hat{\theta}_G$)

- Number of latent classes chosen from :
  - the fit of the model :
    BIC = -2$\times$log(Likelihood) + #parameters$\times$log(#subjects)
  - the posterior class-membership probabilities :
    $P(c_i = g|Y_i, X_i, \hat{\theta}_G)$ computed by Bayes theorem
  - and the associated posterior classification :
    $\hat{c}_i = \text{argmax}_g \; P(c_i = g|Y_i, X_i, \hat{\theta}_G)$

- Some programs available :
  - Mplus
  - GLLAMM in Stata
  - R functions hlme, lcmm, etc in lcmm package

Example : prostate cancer progression after treatment

- Context :
  - monitoring of prostate cancer progression after radiation therapy
  - prostate specific antigen (PSA), biomarker of progression collected in routine
  - at diagnosis, X_i = initial log PSA, Gleason score and T-stage

- Objective : exploring the trajectories of PSA after radiation therapy
  - non-adjusted class-specific linear mixed model :

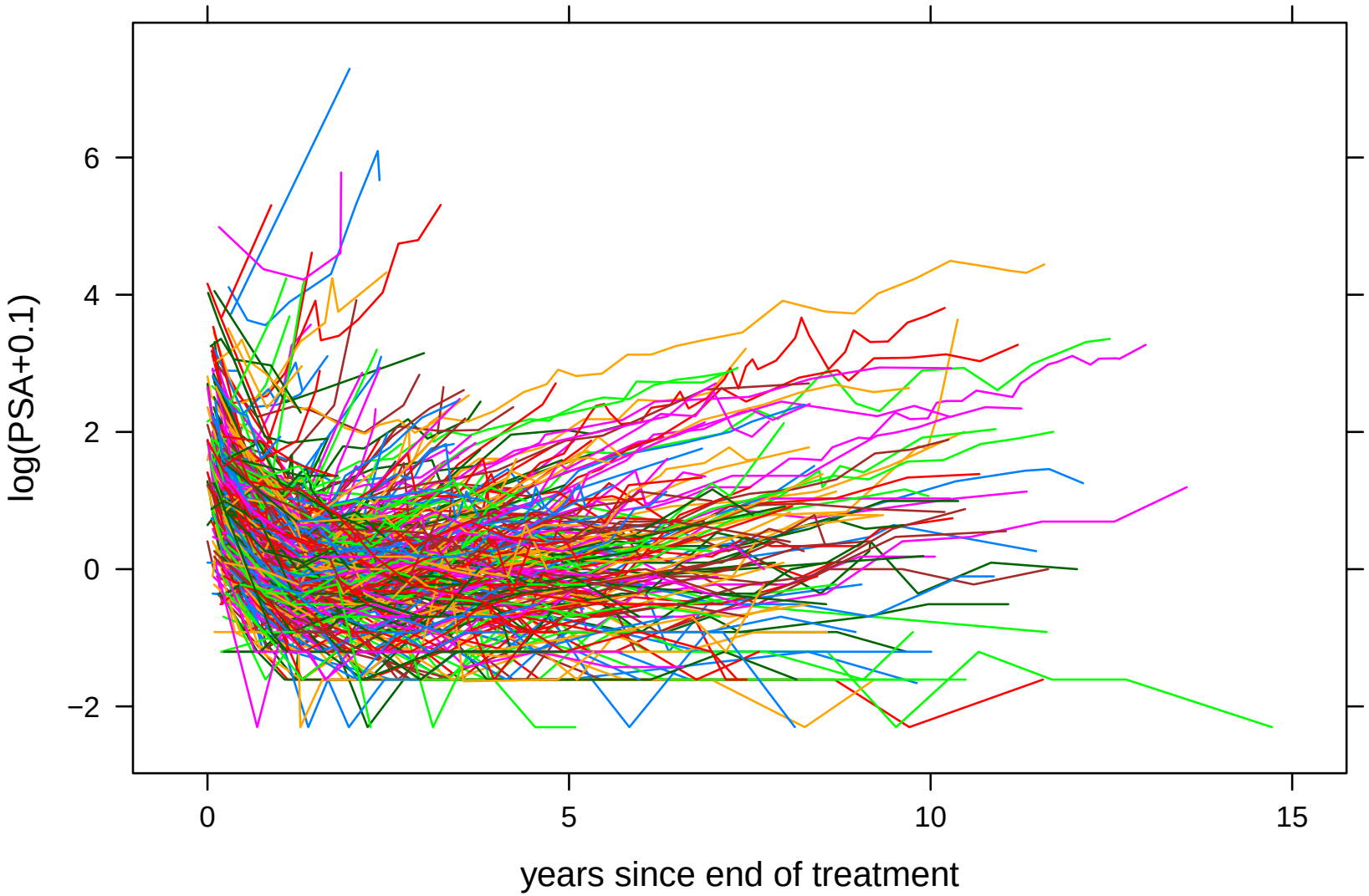
$$\log(PSA(t_{ij}+0.1)) \mid_{c_i=g} = u_{0ig} + u_{1ig} f(t_{ij}) + u_{2ig} t_{ij} + \epsilon_{ij} \qquad \text{and} \quad u_{ig} \sim \mathcal{N}\left(\mu_g, \omega_g^2 B\right)$$

  - Class-membership explained according to prognostic factors :

$$P(c_i = g|X_i) = \frac{\exp(\xi_g X_i)}{\sum_{l=1}^{G} \exp(\xi_l X_i)} \qquad \text{and} \quad \xi_G = 0$$

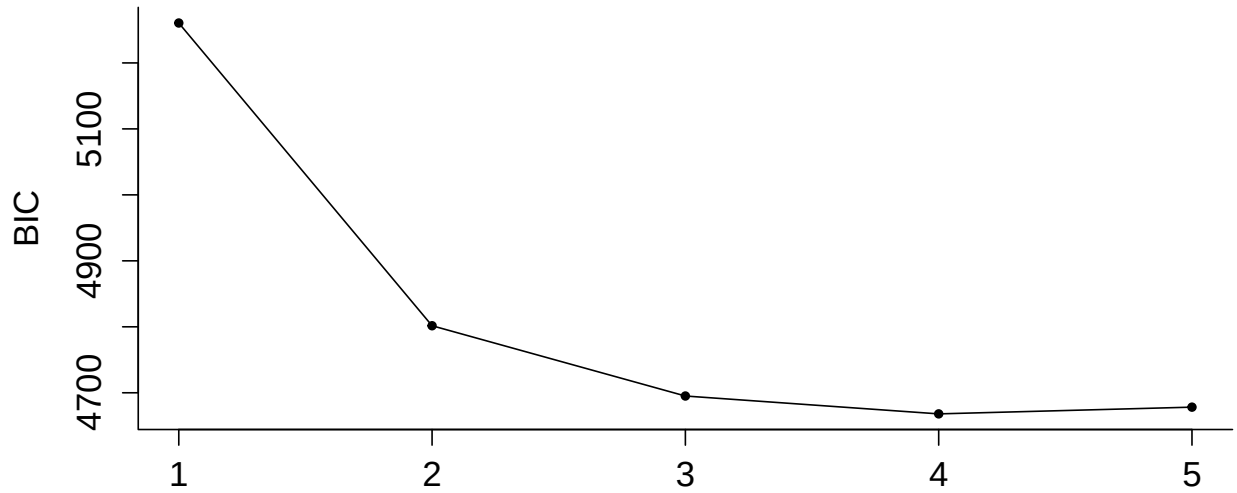
- Data : University of Michigan hospital cohort
  - $N = 459$ patients with 8 (IQR=[5,12]) repeated measures

Individual PSA trajectories after radiation therapy



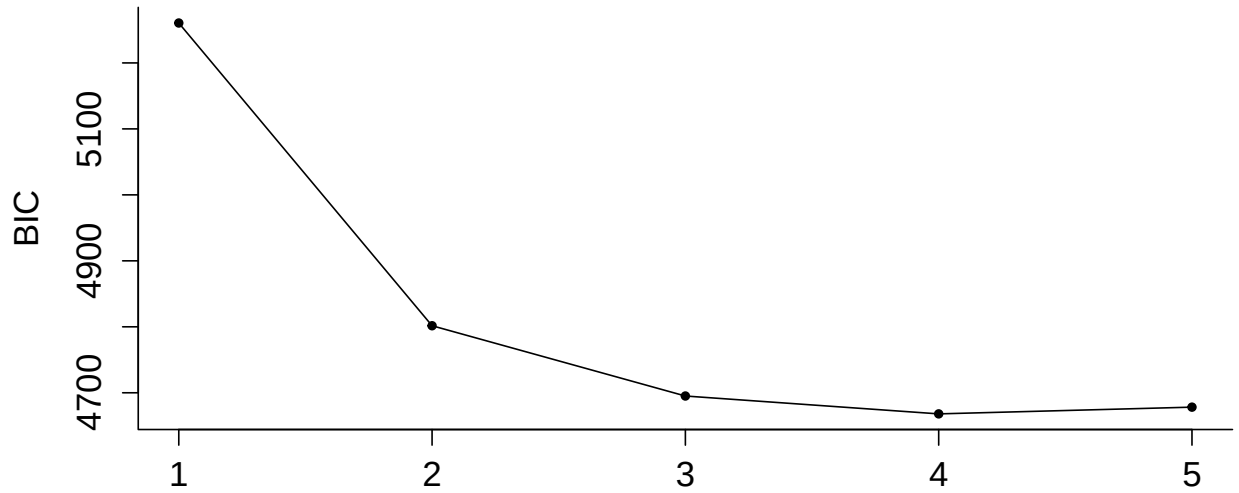
Summary of the estimation process

G	# parms	Log-lik.	BIC	Frequency of the latent classes (%)				
				1	2	3	4	5
1	10	-2599.6	5260.5	100				
2	19	-2342.6	4801.6	85.2	14.8			
3	28	-2261.7	4695.1	66.7	28.1	5.2		
4	37	-2220.6	4668.0	53.4	28.3	15.5	2.8	
5	46	-2198.2	4678.2	47.3	33.8	11.3	5.0	2.6

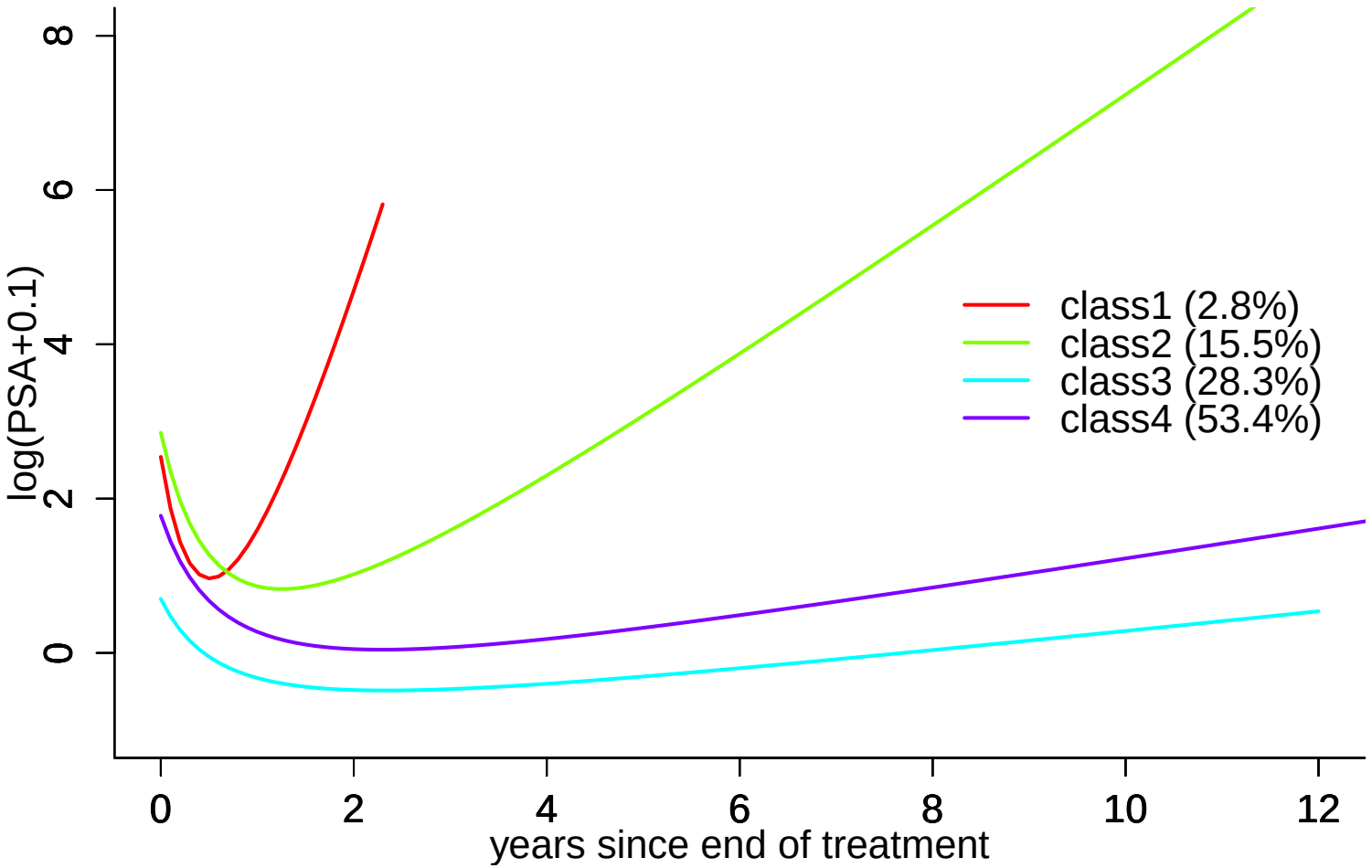


Summary of the estimation process

G	# parms	Log-lik.	BIC	Frequency of the latent classes (%)				
				1	2	3	4	5
1	10	-2599.6	5260.5	100				
2	19	-2342.6	4801.6	85.2	14.8			
3	28	-2261.7	4695.1	66.7	28.1	5.2		
4	**37**	**-2220.6**	**4668.0**	**53.4**	**28.3**	**15.5**	**2.8**	
5	46	-2198.2	4678.2	47.3	33.8	11.3	5.0	2.6



Class-specific predicted trajectories of PSA



Covariate impact on class-membership

Odds-Ratios for class g compared to class G

Covariate	Posterior class :				p
	1	2	3	4	
Gleason <7	1	1	1	1	
Gleason $=7$	>1000	1.58	0.676	1	0.613
Gleason >7	>1000	0.487	0.256	1	0.691
iPSA	0.384	1.113***	0.660***	1	<0.001
Tstage 1-2	1	1	1	1	
Tstage 3-4	2.896***	2.598	0.645	1	0.002

usual interpretation of a multinomial logistic regression

Description of the posterior classification

Covariate	Posterior class : 1 (N=13)	2 (N=71)	3 (N=130)	4 (N=245)	p
gleason <7	0 (0%)	31 (43.7%)	87 (66.9%)	134 (54.7%)	<0.0001
gleason $=7$	9 (69.2%)	30 (42.3%)	42 (32.3%)	92 (37.6%)	
gleason >7	4 (30.8%)	10 (14.1%)	1 (0.8%)	19 (7.8%)	
Tstage 1-2	8 (61.5%)	44 (62.0%)	128 (98.5%)	238 (97.1%)	<0.0001
Tstage 3-4	5 (38.5%)	27 (38.0%)	2 (1.54 %)	7 (2.9%)	
Recurrence No	1 (7.7%)	33 (46.5%)	125 (96.2%)	226 (92.2%)	<0.0001
Recurrence Yes	12 (92.3%)	38 (53.5%)	5 (3.8%)	19 (7.8%)	
iPSA	1.94 (1.07)	3.43 (0.80)	1.37 (0.69)	2.25 (0.46)	<0.0001
Risk of recurrence	91.7 ***	9.6 ***	0.5 $^{(NS)}$	1	<0.0001

Count (frequency) & Chi-square test for qualitative covariates
Mean (standard error) and Kruskal-Wallis test for quantitative covariates

Hazard ratios from a Cox model for the risk of recurrence

Classification assessment

Posterior classification table

Final classif.	Number of subjects (%)	Mean of the class-membership probabilities in class (in %) : 1	2	3	4
1	13 (2.8%)	**93.9**	3.3	2.6	0.2
2	71 (15.5%)	0.7	**90.7**	0.3	8.3
3	130 (28.3%)	$<$0.1	0.2	**85.0**	14.8
4	245 (53.4%)	$<$0.1	6.2	11.1	**82.7**

Percentage of subjects classified with $\pi_{ig}^{(y)} > \eta$:

$\eta \backslash$G	1	2	3	4
$\eta = 0.9$	76.9	70.4	50.8	46.5
$\eta = 0.8$	84.6	80.3	67.7	64.1
$\eta = 0.7$	92.3	87.3	77.7	76.7

Extensions from the latent class approach

Latent class approach for developmental trajectories not limited to the standard LMM.

Practicable with most extensions of LMM (and implemented in **`lcmm`**) :

- refinement of the within-subject correlation with Gaussian processes (**`hlme`**)

- other types of data : ordinal/binary data, curvilinear outcomes, ... (**`lcmm`**)
  - ex : trajectories of functional dependency measured by a 4-level variable

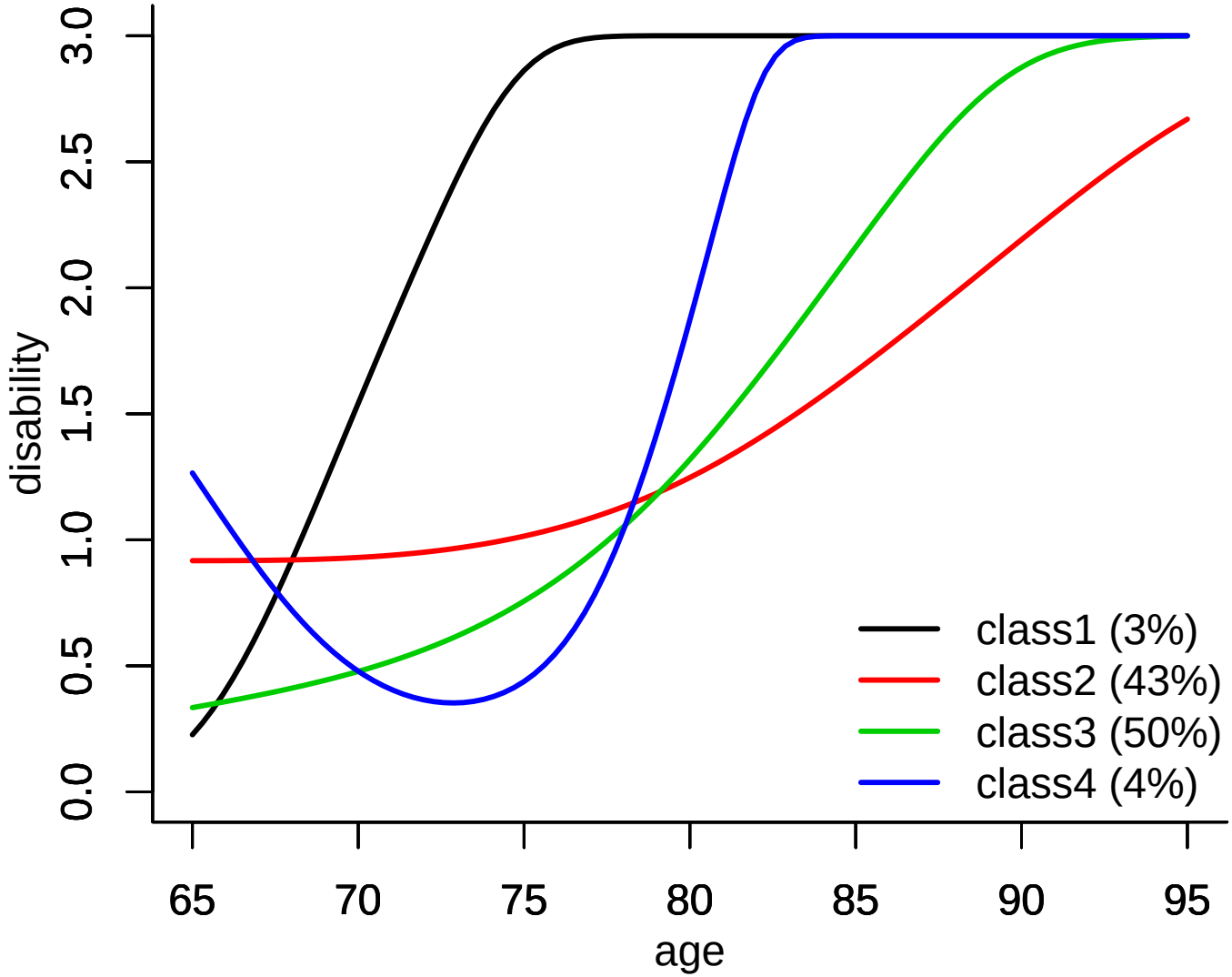
Example : trajectories of disability in the elderly

- Objective : exploring the trajectories of disability according to age
- Disability measured by a 4-level scale :
  - 0= None
  - 1= Mild (mobility only)
  - 2= Moderate (mobility + instrumental activities of daily living (ADL))
  - 3= Severe (mobility + instrumental ADL + ADL)
- Cumulative probit model for ordinal repeated data :

$$Y_{ij}|_{c_i=g} = m \Leftrightarrow \eta_m \leq \Lambda_i(t_{ij})|_{c_i=g} + \epsilon_{ij} < \eta_{(m+1)} \text{ for } m \in \{0,3\}$$
$$\Lambda_i(t_{ij})|_{c_i=g} = X_i(t_{ij})\beta_g + Z_i(t_{ij})u_{ig}$$

  - $Z_i(t_{ij}) =$ a few regression splines ; $X_i(t_{ij}) = 0$; $u_{ig} \sim \mathcal{N}(\mu_g, B)$; $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$
- (Random) sample from PAQUID :
  - PAQUID= prospective study on cerebral aging
  - $N = 500$, *median of 4 (IQR=[2,7]) repeated measures*

Predicted trajectories of disability according to age



Description of the posterior classification

Covariate	Posterior class : 1 (N=15)	2 (N=215)	3 (N=251)	4 (N=19)	p-value
Women	4 (26.7%)	129 (60.0%)	143 (57.0%)	12 (63.2%)	0.083
EL+	12 (80.0 %)	156 (72.6%)	172 (68.5%)	15 (79.0%)	0.537
Incident dementia	3 (20.0%)	33 (15.4%)	77 (30.7%)	15 (79.0%)	$<$0.0001
Death	14 (93.3%)	165 (76.7%)	196 (78.1%)	13 (68.4%)	0.364
Age at entry	70.8 (3.5)	74.0 (6.4)	74.8 (6.5)	72.2 (6.2)	0.030
MMSE at entry	26.7 (3.0)	27.1 (2.4)	26.9 (2.7)	27.1 (2.8)	0.978
IST at entry	26.5 (7.0)	28.4 (5.9)	28.1 (5.8)	27.5 (6.1)	0.737

Count (frequency) & Chi-square test for qualitative covariates

Mean (standard error) and Kruskal-Wallis test for quantitative covariates

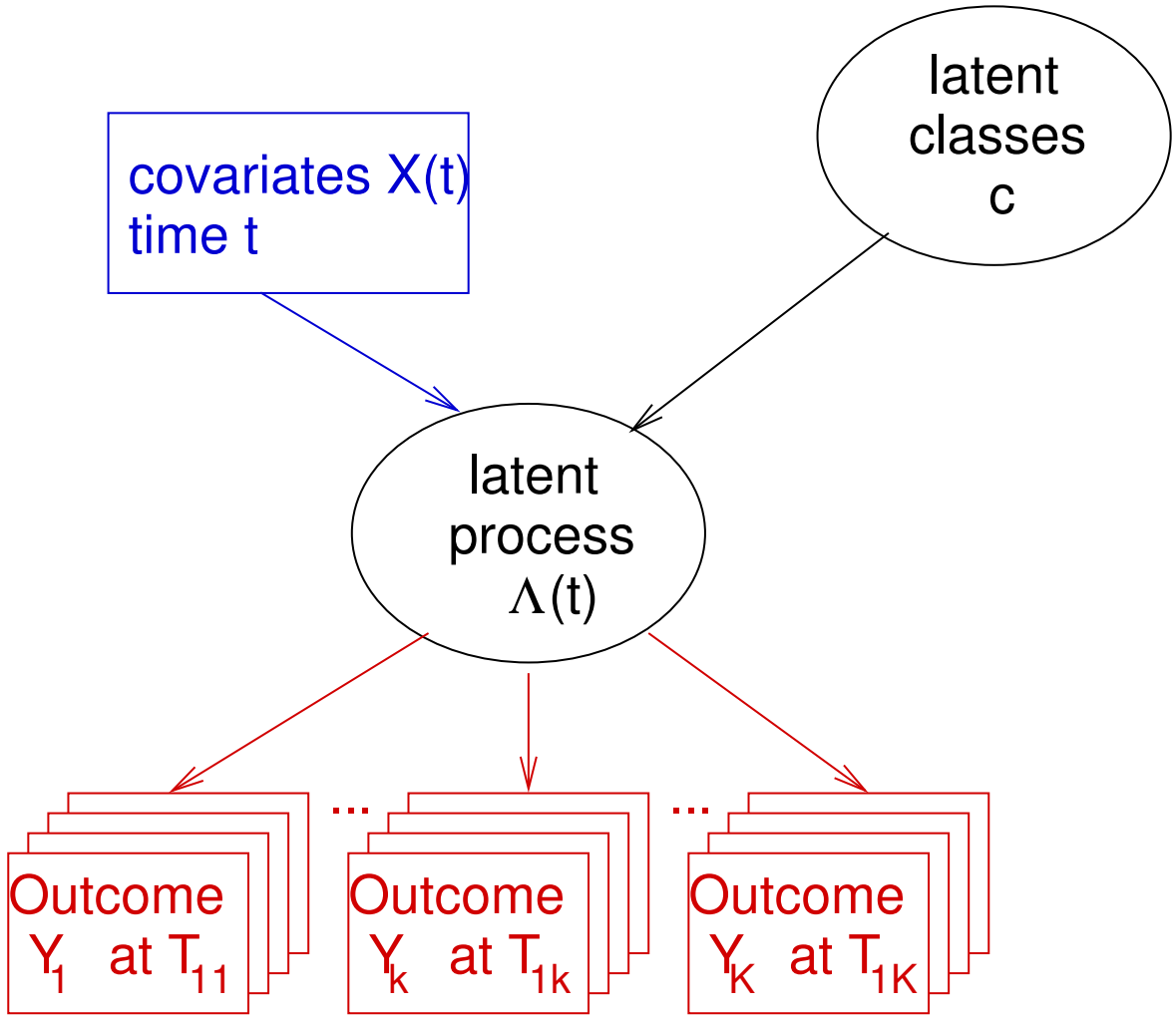
Extensions from the latent class approach (cont'd)

Latent class approach for developmental trajectories not limited to the standard LMM.

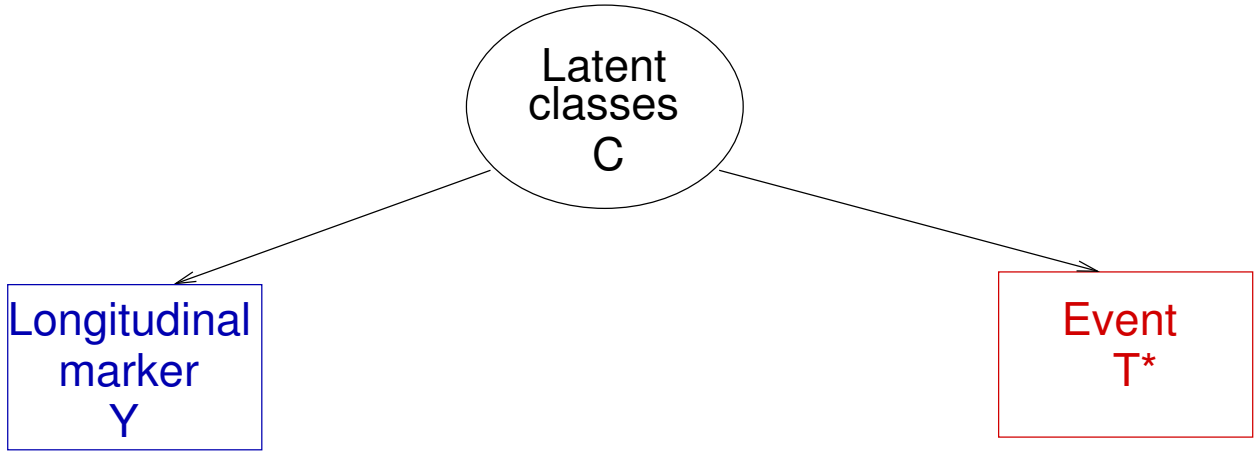
Practicable with most extensions of LMM (and implemented in **`lcmm`**) :

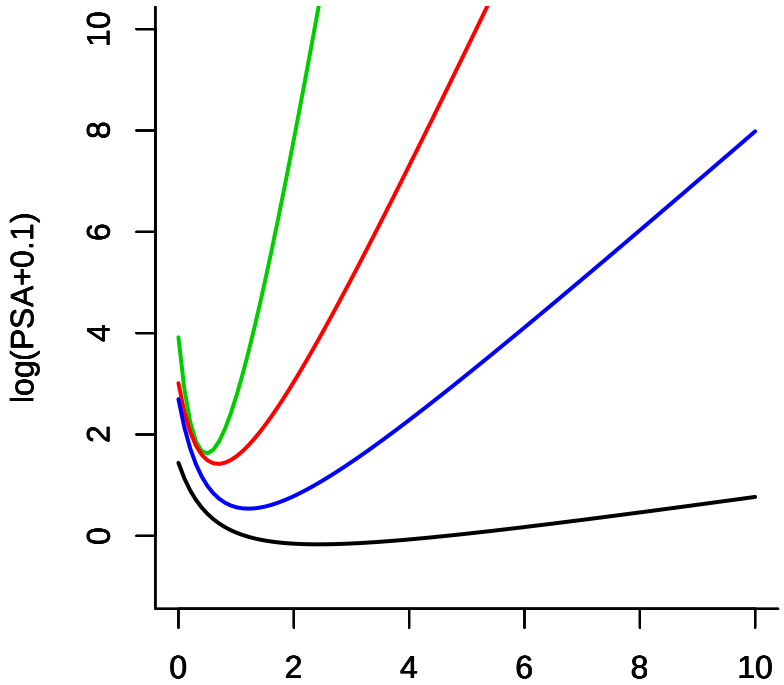
- refinement of the within-subject correlation with Gaussian processes (**`hlme`**)
- other types of data : ordinal/binary data, curvilinear outcomes, ... (**`lcmm`**)
  - ex : trajectories of functional dependency measured by a 4-level variable
- multivariate longitudinal data (**`multlcmm`**)
  - ex : trajectories of cognition measured by multiple psychometric tests
- correlation with clinical events (**`Jointlcmm`**)
  - ex : trajectories of PSA associated with the risk prostate cancer recurrence

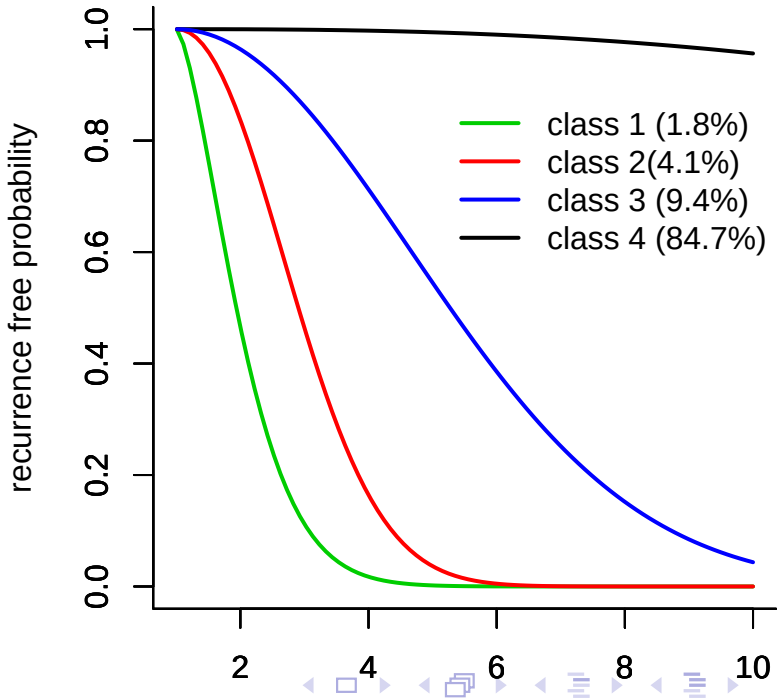
Extension : latent process model for multivariate longitudinal markers



Extension : joint modeling of longitudinal markers and risk of clinical event







Conclusion : LCMM is a powerful statistical tool ...

Applies to any type of longitudinal (multivariate) data

Addresses very different questions - *multiple inclusion of covariate*

- raw exploration of the data
- summary of between-individual heterogeneity (possibly according to covariates)
- identification of disease gravity/diagnosis after adjustment for risk factors
- research of different impacts on profiles of trajectories (e.g. responders/non-responders)

Enjoys the mixed model theory assets

- MAR assumption for missing data and dropout
- individually varying time (age / exact follow-up)
- same inference

Takes into account 2 sources of variability

- individual variability through random-effects - *inference possible*
- latent group structure - *mean profiles of trajectory*

Conclusion : ... to use with caution

With the estimation process

- starting values & local solutions
- not to restrict to the exploratory Nagin's approach (proc Traj) :
  - ⋆ intra-individual correlation neglected (no random-effects)
  - ⋆ number of latent classes overestimated
  - ⋆ inference regarding covariates possibly biased

With the interpretation of the latent classes

- flexible model that may better fit homogeneous populations
  - ⋆ non normal random-effects
  - ⋆ interesting to obtain valid covariate effects (sensitivity analysis)
- identification of latent subgroups / population heterogeneity :
  - ⋆ need of relevant assumption of latent groups
  - ⋆ strict evaluation of goodness-of-fit, discrimination

With the clinical question of interest !

References and aknowledgements

- Aknowledgements :
  - Grants ANR Mobidyq 2009, INCa PREDYC 2010, IRESP Multiple 2013
  - Viviane Philipps for `lcmm` maintainance
  - PAQUID and UM investigators
  - Colleagues from Bordeaux and elsewhere

- A few references :

  - **Bauer, Curran (2003). *Psychol Meth*, 8(3), 338-63 (+ discutants 364-93)**
  - Hipp, Bauer (2006). *Psychological methods*, **11(1)**, 36-53
  - Muthén, Shedden (1999). *Biometrics*, **55(2)**,463-9
  - Muthén, Asparouhov (2009). *In Longitudinal Data Analysis* ed. by Fitzmaurice et al.
  - Proust, Jacqmin-Gadda (2005). *Computer Methods and Programs in Biomedicine*, **78**,165-73
  - Proust-Lima, Joly et al. (2009). *CSDA*, **53**, 1142-54
  - Proust-Lima, Sène et al. (2014). *SMMR*, **23**,74-90
  - Verbeke, Lesaffre (1996). *JASA*, **91**,217-21
  - Xu, Hedeker (2001). *Journal of biopharmaceutical statistics*, **11**,253-73