Comment modéliser les trajectoires de qualité de vie d'un point de vue statistique ?

Cécile Proust-Lima

INSERM U1219, Bordeaux Population Health Research Center
Univ. Bordeaux, ISPED, Bordeaux, France

cecile.proust-lima@inserm.fr

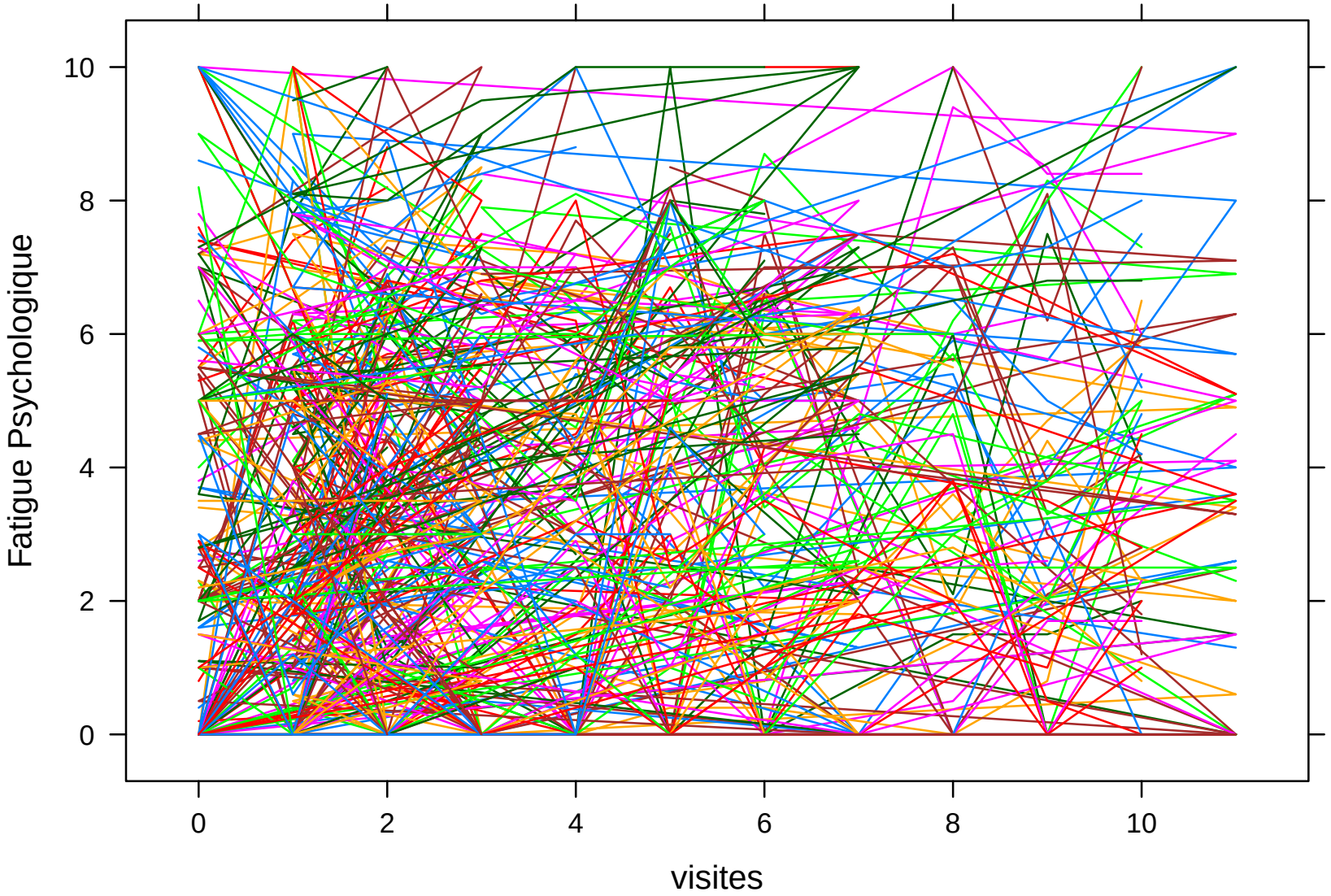
Séminaire Qualité de Vie - Septembre 2018 - Montpellier, France

Plateforme Nationale Qualité de Vie et Cancer, Institut du Cancer de Montpellier, Ligue contre le Cancer

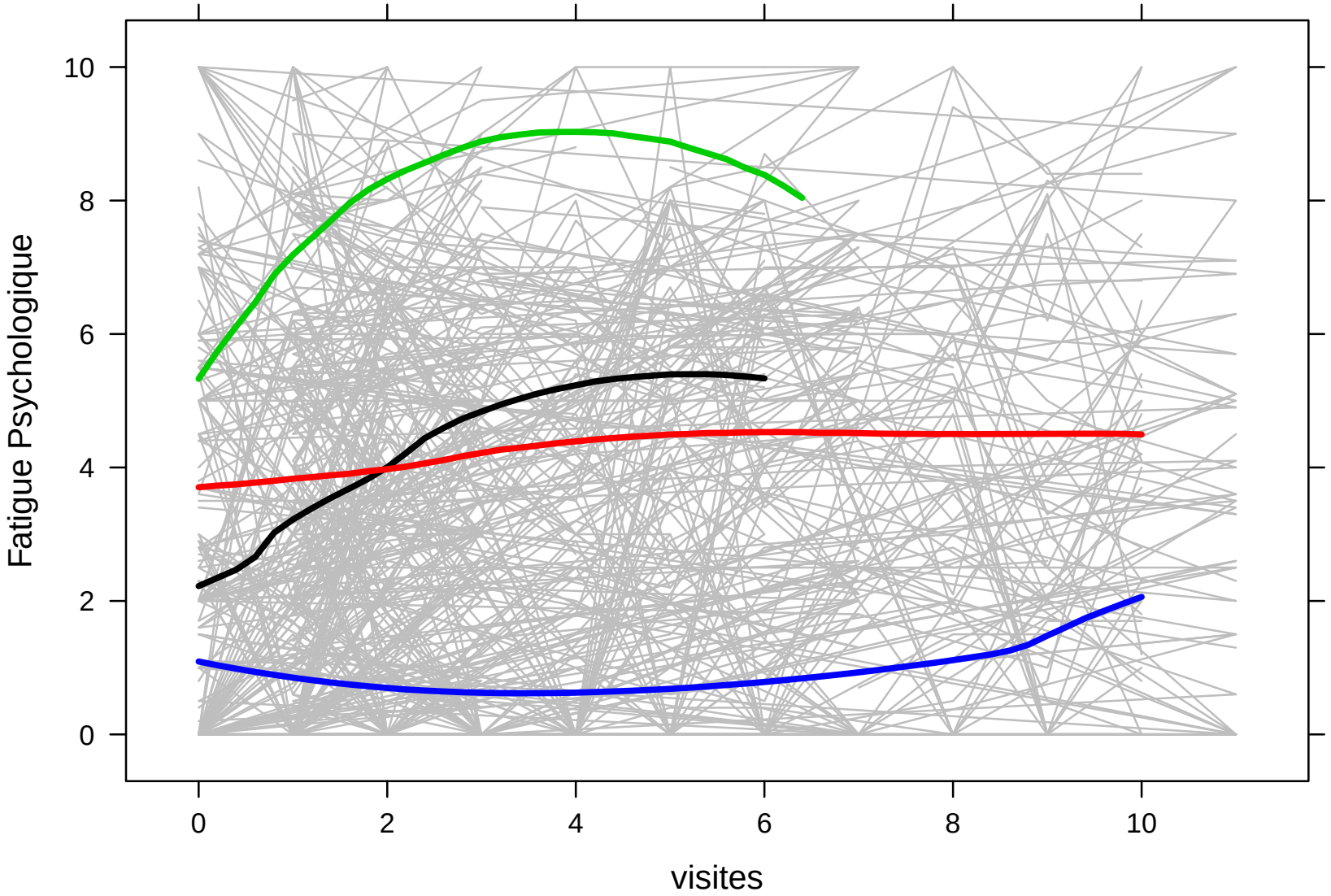
Contexte

- Intérêt pour la dynamique/progression des cancers
  - biomarqueurs classiques (e.g., PSA, taille tumorale)
  - ”nouveaux” marqueurs : la qualité de vie
- Objectifs :
  - décrire la forme de la progression moyenne au cours du temps
  - évaluer des déterminants d'une progression
  - identifier des profils de progression
  - faire des prédictions de la progression au niveau individuel
- Données :
  - Echantillon de N sujets
  - Mesures répétées d'un score Y à plusieurs visites pour chaque sujet

Observations ...



Observations ... à résumer en trajectoires moyennes



Deux contextes d'étude

1. On connait la nature des profils à déterminer
   - selon un traitement,
   - selon des caractéristiques du cancer,
   - selon des caractéristiques du patient ...

2. On ne connait pas a priori la nature des profils
   - on suppose simplement qu'il existe des formes de progression possiblement distinctes

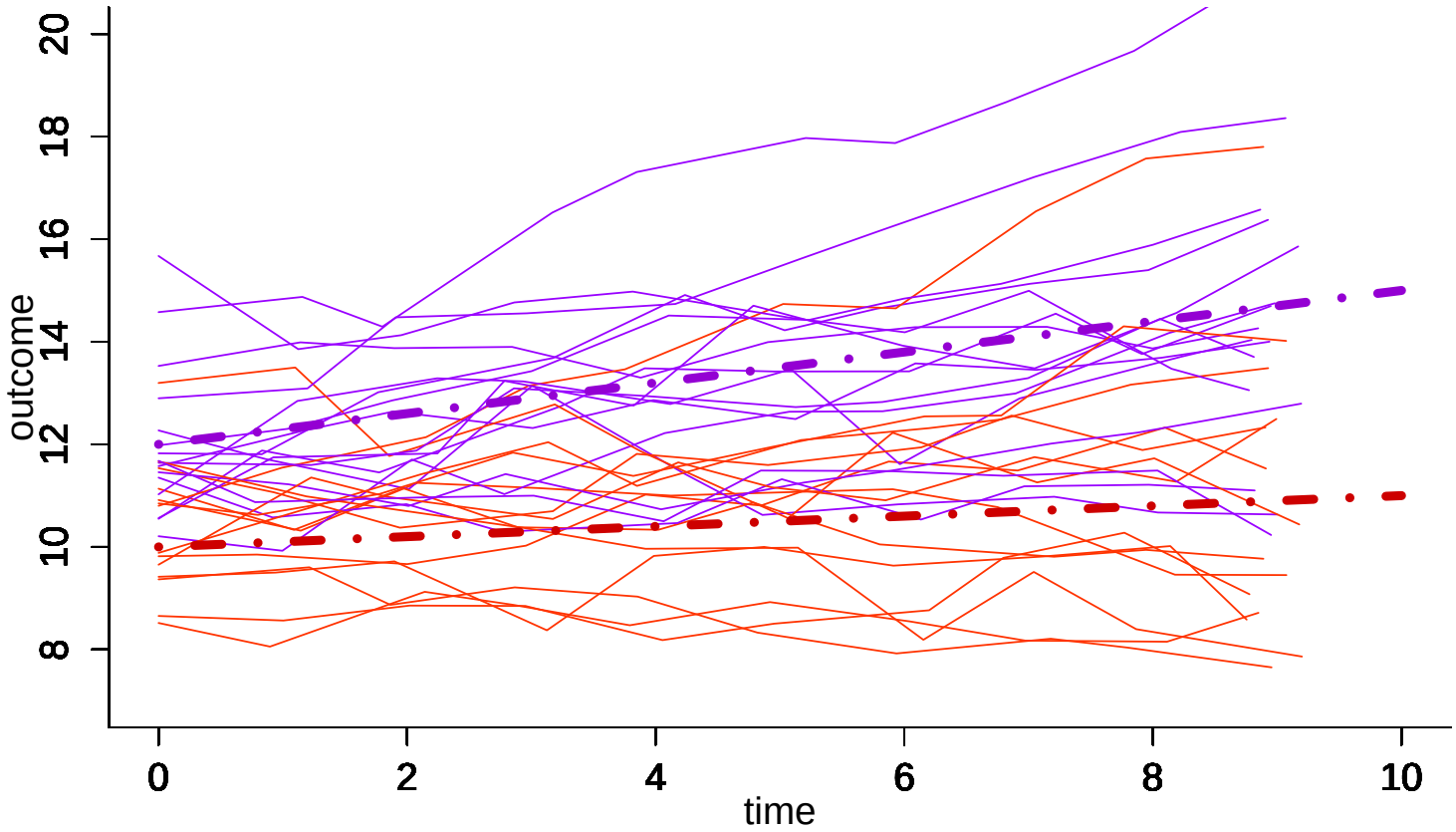
Deux contextes d'étude

1. On connait la nature des profils à déterminer
   - selon un traitement,
   - selon des caractéristiques du cancer,
   - selon des caractéristiques du patient ...
   - modèles mixtes ajustés sur variables explicatives

2. On ne connait pas a priori la nature des profils
   - on suppose simplement qu'il existe des formes de progression possiblement distinctes
   - modèles mixtes à classes latentes

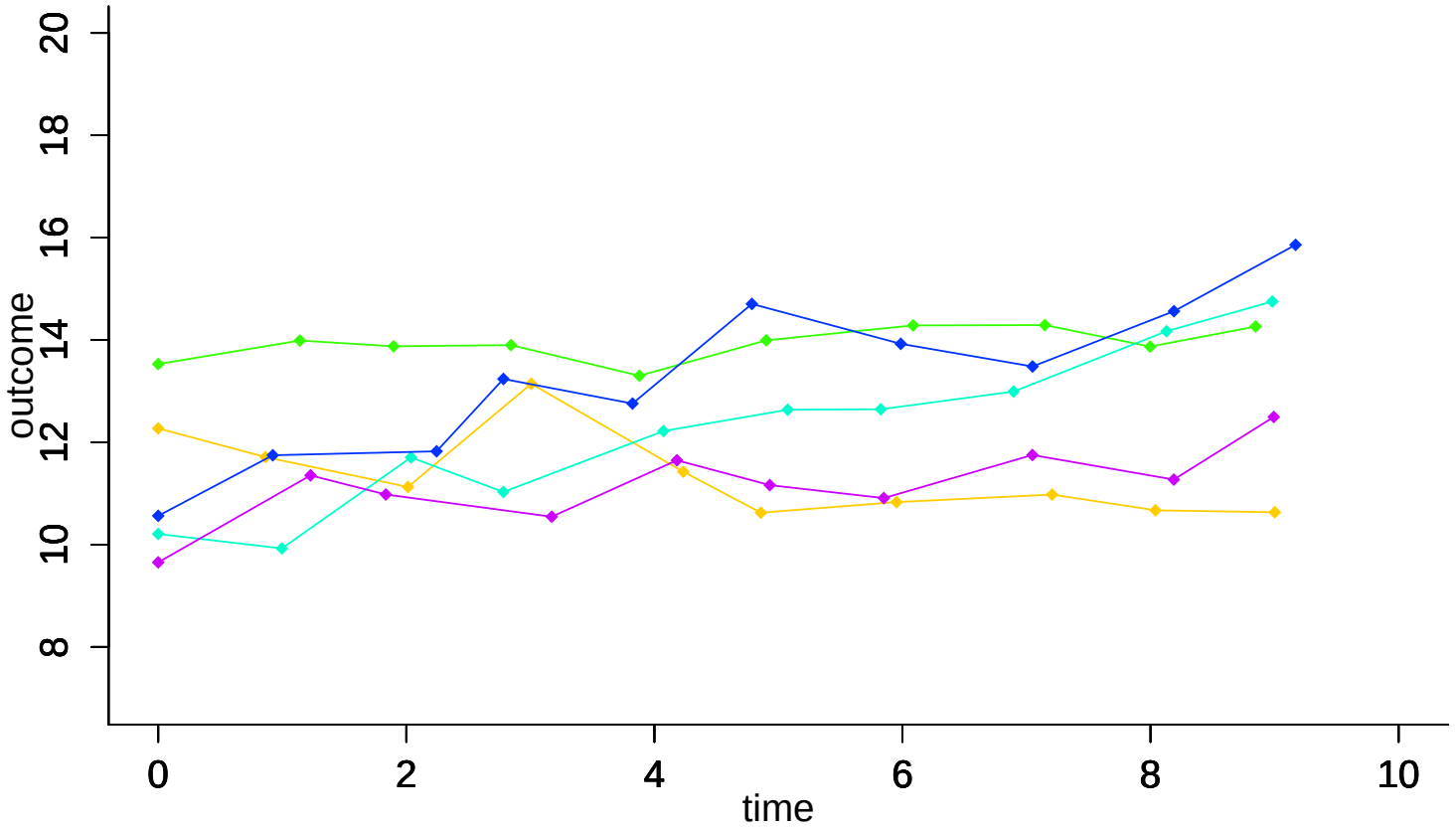
$\rightarrow$ Dans les deux cas, corrélation entre mesures répétées
   - doivent être prises en compte pour assurer une inférence correcte
   - se fait par des modèles mixtes (aka "(linear) mixed models", "random effect models")

Exemple Simulé

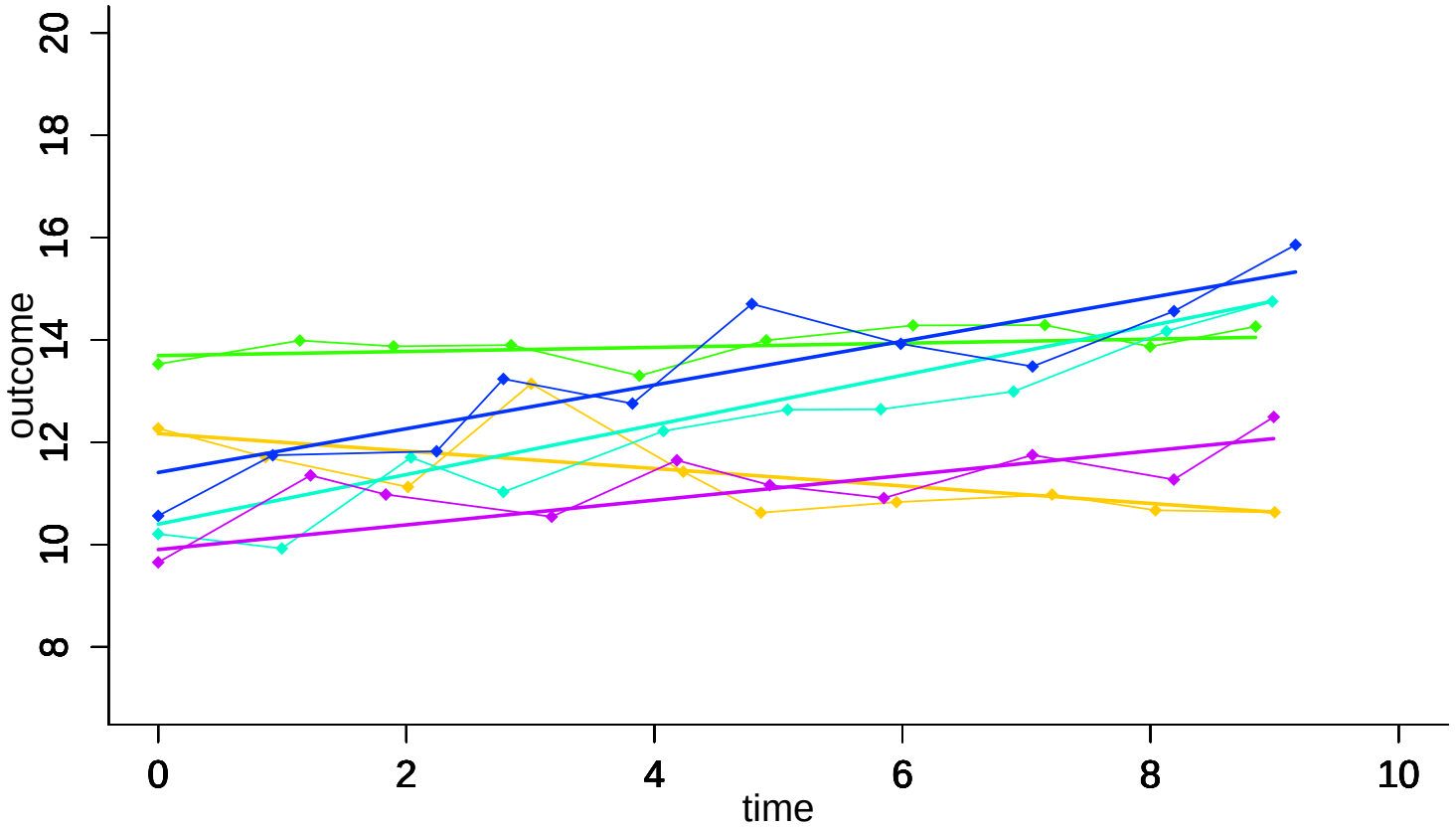


A partir des trajectoires individuelles, trajectoires moyennes dans la population pour chaque niveau d'une variable explicative C (binaire)

Considérons 5 sujets uniquement



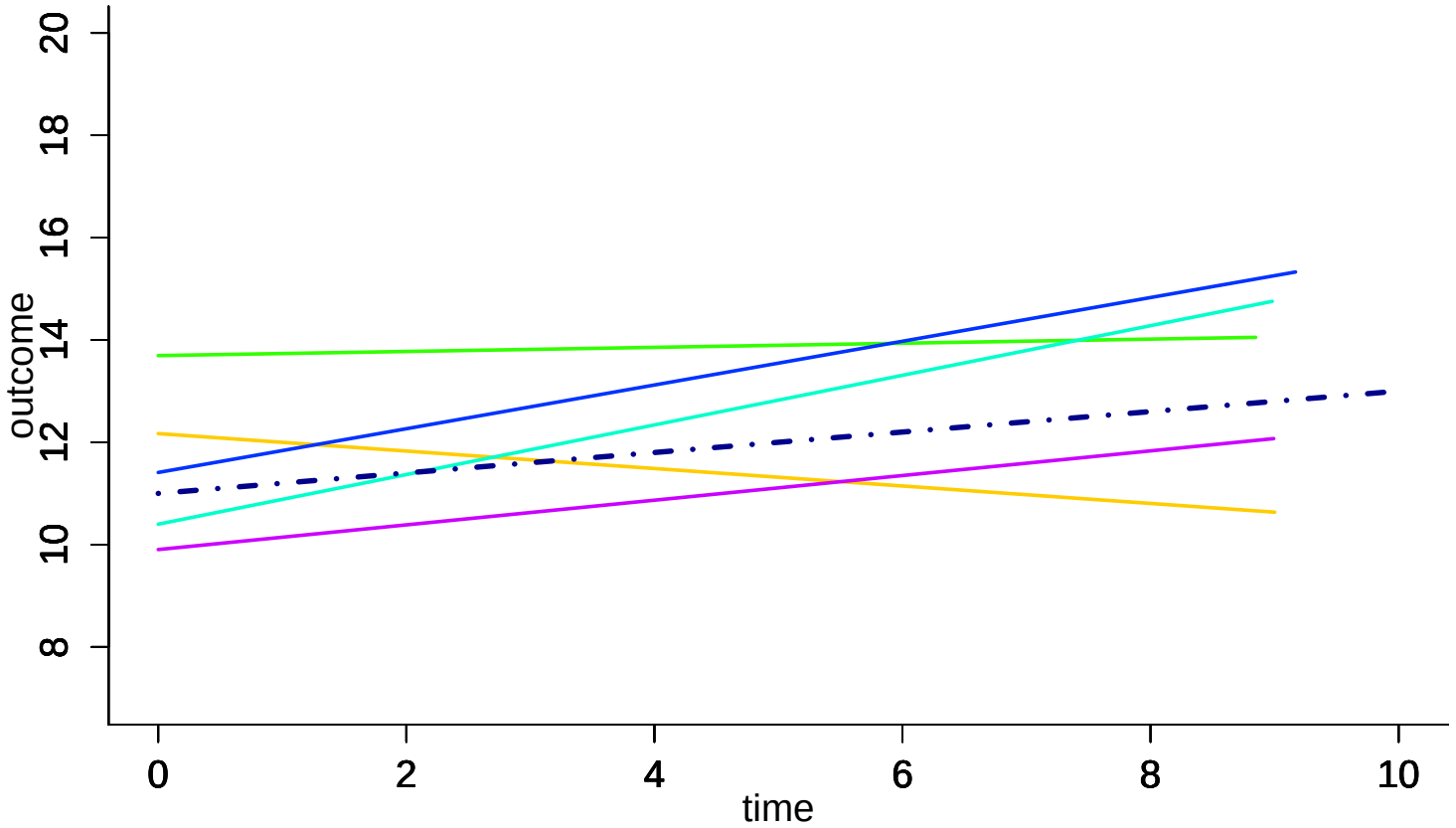
Considérons 5 sujets uniquement



2 niveaux d'intérêt :

- le niveau individuel avec une trajectoire individuelle autour des mesures bruitées

Considérons 5 sujets uniquement



2 niveaux d'intérêt :

- le niveau individuel avec une trajectoire individuelle autour des mesures bruitées
- le niveau populationnel avec une trajectoires moyenne autour des déviations individuelles

Définition du modèle linéaire mixte

- Population de N sujets (indice i, $i = 1, ..., N$)
  - Y_{ij} mesure répétée de la variable pour le sujet i à la visite j, $j = 1, ..., n_i$
  - t_{ij} temps de mesure en j, $j = 1, ..., n_i$

- Pour le sujet i à la répétition j :

$$
\begin{aligned}
Y_{ij} = Y_i(t_{ij}) &= \beta_0 + \beta_1 \times t_{ij} && \text{au niveau de la population} \\
&+ u_{0i} + u_{1i} \times t_{ij} && \text{au niveau de l'individu} \\
&+ \epsilon_{ij}
\end{aligned}
$$

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

Définition du modèle linéaire mixte

- Population de N sujets (indice i, $i = 1, ..., N$)
  - Y_{ij} mesure répétée de la variable pour le sujet i à la visite j, $j = 1, ..., n_i$
  - t_{ij} temps de mesure en j, $j = 1, ..., n_i$

- Pour le sujet i à la répétition j (et la variable explicative binaire C) :

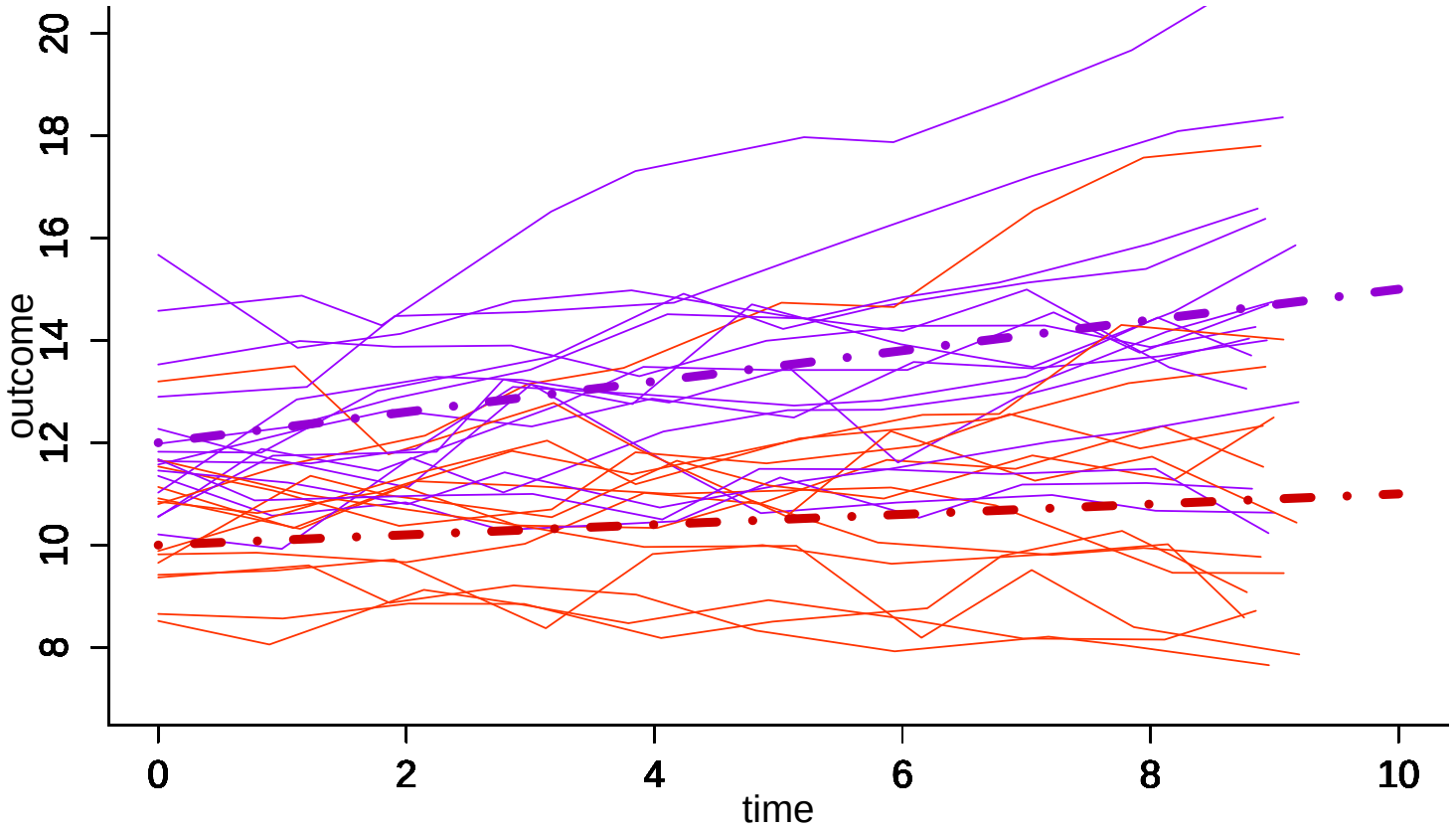
$$\begin{aligned} Y_{ij} = Y_i(t_{ij}) = & \quad \beta_0 + \beta_1 \times t_{ij} + \beta_2 C_i + \beta_3 C_i \times t_{ij} && \text{au niveau de la population} \\ & + \quad u_{0i} + u_{1i} \times t_{ij} && \text{au niveau de l'individu} \\ & + \quad \epsilon_{ij} \end{aligned}$$

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

- Trajectoire moyenne prédite par niveau de C :

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

Si C est inconnu, non observé ?



Hétérogénéité suspectée mais pas observée directement

$\rightarrow$ C devient une variable latente, **une classe latente**

$\rightarrow$ Modèles mixtes à classes latentes (aka "growth mixture models", "latent class mixed models")

Le modèle mixte à classes latentes (LCMM)

Population de N sujets (indice i, $i = 1, ..., N$)

- Y_{ij} mesure répétée de la variable pour le sujet i à la visite j, $j = 1, ..., n_i$
- t_{ij} temps de mesure en j, $j = 1, ..., n_i$
- X_i vecteur de variables explicatives (indépendantes du temps ici pour simplifier)

G Classes homogènes de sujets (indice g, $g = 1, ..., G$)

- c_i variable latente discrète :
  $c_i = g$ si sujet i appartient à la classe g ($g = 1, ..., G$)
  $\rightarrow$ *chaque sujet appartient à une et une seule classe*

Deux sous-modèles :

- Probabilité d'appartenir à une classe latente
- trajectoire du marqueur dans chaque classe
  selon des variables explicatives observées et le temps

Exemple de spécification de LCMM

- Probabilité d'appartenance aux classes expliquée par des variables explicatives X_i :

  $\rightarrow$ *régression logistique multinomiale*

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

avec $\xi_{0G} = 0$ et $\xi_{1G} = 0$ i.e. classe G = référence

Exemple de spécification de LCMM

- Probabilité d'appartenance aux classes expliquée par des variables explicatives X_i :

  $\rightarrow$ *régression logistique multinomiale*

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

avec $\xi_{0G} = 0$ et $\xi_{1G} = 0$ i.e. classe G = référence

- Trajectoires spécifique à la classe : exemple de trajectoire linéaire

$$\begin{aligned} Y_{ij}|_{c_i=g} &= \mu_{0g} + \mu_{1g} \times t_{ij} && \text{au niveau de la population} \\ &+ u_{0ig} + u_{1ig} \times t_{ij} && \text{au niveau de l'individu} \\ &+ \epsilon_{ij} \end{aligned}$$

  - μ_{0g} et μ_{1g} intercept et pente moyennes spécifiques à la classe
  - $u_{ig} = u_i|_{c_i=g} = (u_{0ig}, u_{1ig})' \sim \mathcal{N}((0,0)', B_g)$ class-specific RE
    - ⋆ B_g variance-covariance spécifique à la classe (souvent $B_g = B$ ou $B_g = w_g^2 B$)
  - $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$, $\epsilon_{ij} \perp u_{ig}$

Exemple de spécification de LCMM

- Probabilité d'appartenance aux classes expliquée par des variables explicativesX_i :

  $\rightarrow$ *régression logistique multinomiale*

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

avec $\xi_{0G} = 0$ et $\xi_{1G} = 0$ i.e. classe G = référence

- Trajectoires spécifique à la classe : exemple de trajectoire linéaire avec des variables explicatives :

$$\begin{aligned} Y_{ij}|_{c_i=g} &= \mu_{0g} + \mu_{1g} \times t_{ij} + \beta_2 X_i + \beta_3 X_i \times t_{ij} && \text{au niveau de la population} \\ &+ u_{0ig} + u_{1ig} \times t_{ij} && \text{au niveau de l'individu} \\ &+ \epsilon_{ij} \end{aligned}$$

  - μ_{0g} et μ_{1g} intercept et pente moyennes spécifiques à la classe
  - $u_{ig} = u_i|_{c_i=g} = (u_{0ig}, u_{1ig})' \sim \mathcal{N}((0,0)', B_g)$ class-specific RE
    - ⋆ B_g variance-covariance spécifique à la classe (souvent $B_g = B$ ou $B_g = w_g^2 B$)
  - $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$, $\epsilon_{ij} \perp u_{ig}$

Modèle linéaire mixte : formulation générale

$$
\begin{aligned}
Y_{ij}|_{c_i=g} = \quad & Z'_{ij}\mu_g + X'_{2ij}\beta + X'_{3ij}\gamma_g && \text{au niveau de la population} \\
+ \quad & Z'_{ij}u_{ig} && \text{au niveau de l'individu} \\
+ \quad & \epsilon_{ij}
\end{aligned}
$$

Z_{ij}, X_{2ij}, X_{3ij} : 3 vecteurs de variables explicatives sans overlap

- Z_{ij} vecteur de fonctions du temps :
  $Z_{ij} = (1, t_{ij}, t_{ij}^2, t_{ij}^3, ...)$ pour les formes polynomiales
  $Z_{ij} = (B_1(t_{ij}), ..., B_K(t_{ij}))$ pour les formes approchées par des splines
  $Z_{ij} = (f_1(t_{ij}), ..., f_K(t_{ij}))$ pour des formes définies par K fonctions param
- X_{2ij} variables explicatives avec effets communs sur les classes β
- X_{3ij} variables explicatives avec effets spécifiques sur les classes γ_g

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

Estimation du LCMM

- Estimation pour un nombre de classes latentes fixe
- Le plus souvent par Maximum de Vraisemblance ($\hat{\theta}_G$)

- Nombre de classes latentes choisi a posteriori selon :
  - Le fit du modèle avec le BIC, Size-Ajusted BIC (SABIC), ...
    - ⋆ BIC = -2$\times$log(Likelihood) + #parameters$\times$log(#subjects)
  - Le pouvoir discriminant de la classification a posteriori
    - ⋆ Probas a posteriori $P(c_i = g|Y_i, X_i, \hat{\theta}_G)$ calculées par le théorème de Bayes
    - ⋆ Classification $\hat{c}_i = \text{argmax}_g \; P(c_i = g|Y_i, X_i, \hat{\theta}_G)$
    - ⋆ Mesure d'entropie (le plus proche de 1 possible pour une discrimination parfaite)
    - ⋆ Table de classification (probabilités moyennes d'affectation dans chaque classe)
  - Autres critères : la taille des classes, l'interprétation clinique, etc

- Quelques programmes
  - Mplus, GLLAMM in Stata
  - **R functions hlme, lcmm, etc in lcmm package**

Exemple : Trajectoires de fatigue psychologique en cancer colorectal métastatique

- Etude prospective dans le cancer colorectal métastatique
  - N= 169 patients initiant un nouveau cycle de chimiothérapie
  - Evaluation tous les 15jours de la fatigue (physique et psychologique) par une échelle visuelle analogique
- Etude réalisée dans la thèse de Louise Baussard (Institut du Cancer de Montpellier)
  - détaillée dans une présentation demain
- Modèles mixtes à classes latentes :
  - évolution au cours du temps approchée par un polynome quadratique
  - pas d'ajustement sur des variables explicatives

Résumé d'estimation

Estimation des modèles de 1 à 5 classes

						Proportion des classes (%)				
G	logL*	p*	BIC	SABIC	entropy	1	2	3	4	5
1	-1766.7	16	3615.4	3564.8	1.00	100.0				
2	-1744.1	22	3601.1	3531.4	0.75	58.0	42.0			
3	-1734.5	28	3612.7	3524.0	0.69	38.5	42.6	18.9		
4	**-1725.7**	**33**	**3620.6**	**3516.1**	**0.74**	**12.4**	**39.6**	**5.9**	**42.0**	
5	-1721.5	40	3648.1	3521.5	0.69	36.1	29.0	6.5	11.2	17.2

* logL = log-vraisemblance ; p = nombre de paramètres à estimer

Résumé d'estimation

Estimation des modèles de 1 à 5 classes

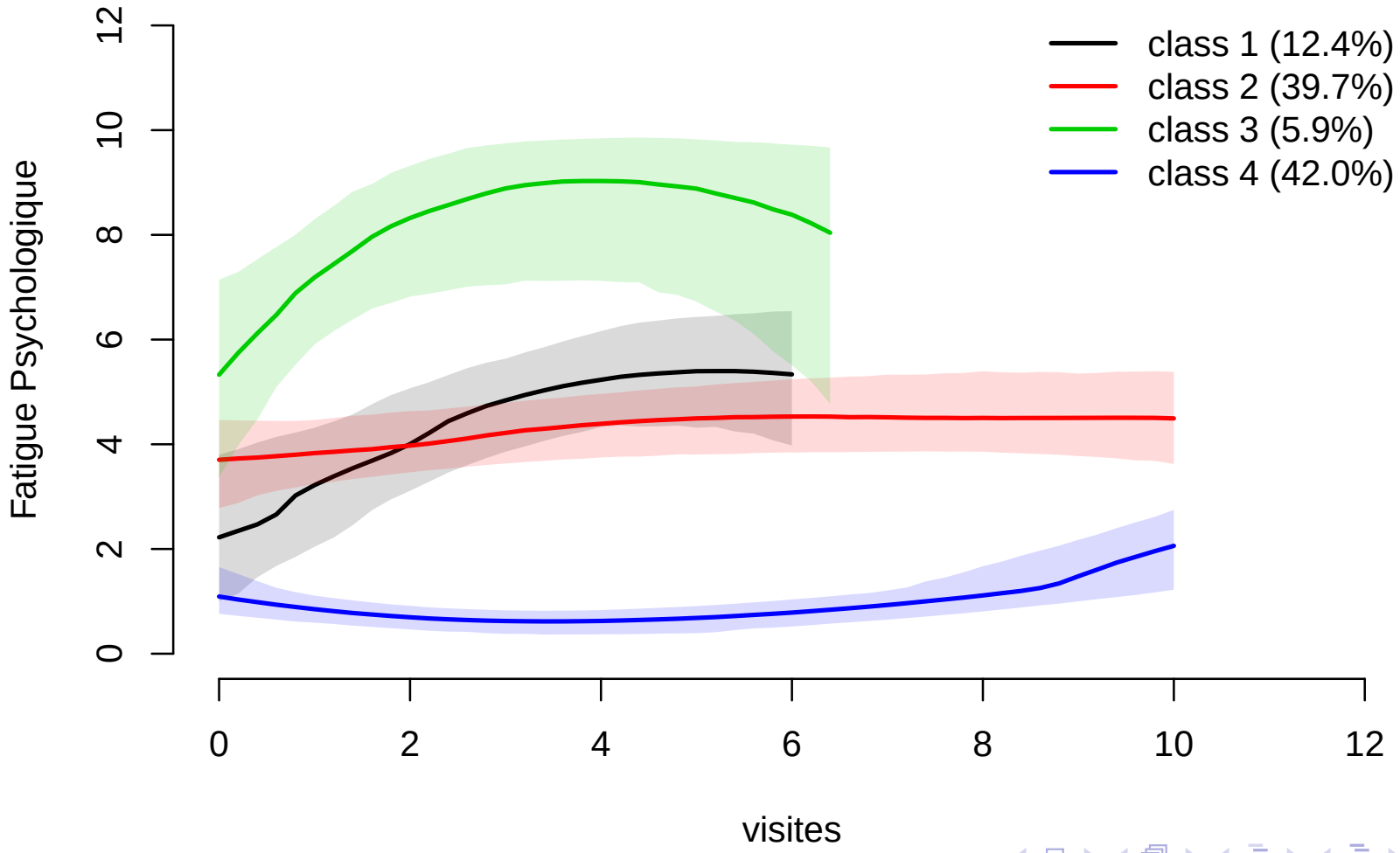
						Proportion des classes (%)				
G	logL*	p*	BIC	SABIC	entropy	1	2	3	4	5
1	-1766.7	16	3615.4	3564.8	1.00	100.0				
2	-1744.1	22	3601.1	3531.4	0.75	58.0	42.0			
3	-1734.5	28	3612.7	3524.0	0.69	38.5	42.6	18.9		
4	**-1725.7**	**33**	**3620.6**	**3516.1**	**0.74**	**12.4**	**39.6**	**5.9**	**42.0**	
5	-1721.5	40	3648.1	3521.5	0.69	36.1	29.0	6.5	11.2	17.2

* logL = log-vraisemblance ; p = nombre de paramètres à estimer

Table de classification a posteriori

classif. finale	Nombre de sujets (%)	Moyenne des probabilités d'appartenir aux classes (%)			
		1	2	3	4
1	21 (12.4%)	**81.6**	13.3	2.4	2.7
2	67 (39.6%)	14.7	**77.3**	1.4	6.5
3	10 (5.9%)	8.5	1.2	**90.3**	0.0
4	71 (42.0%)	1.9	6.6	$<$0.1	**91.4**

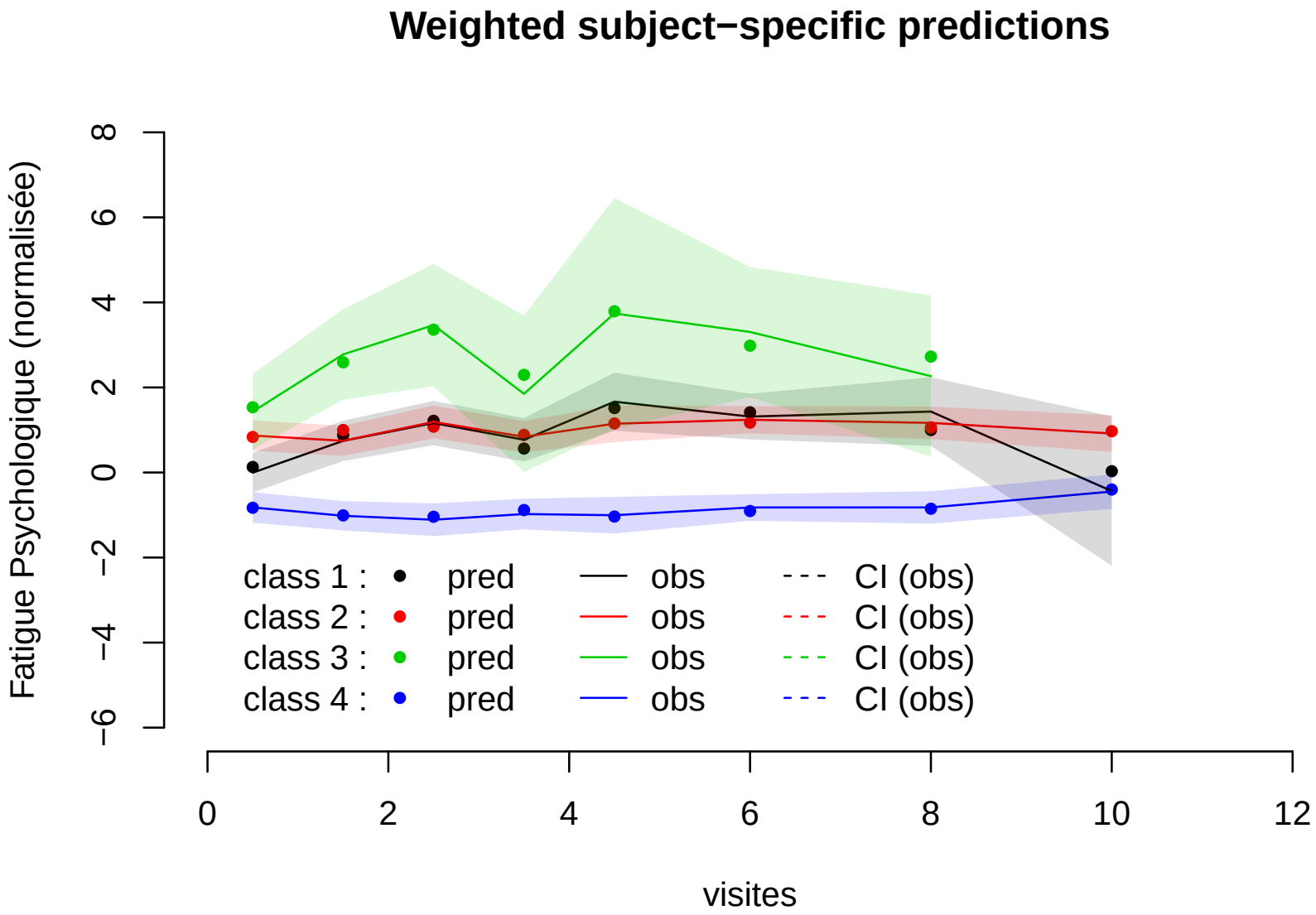
Trajectoires moyennes prédites de fatigue (avec intervalles de confiance à 95%)



Description de la classification a posteriori

- On dispose de la classe a posteriori de chaque sujet
  - avec la proba d'appartenance associée
- On peut décrire la classification :
  - en croisant avec des variables explicatives d'intérêt + tests statistiques
  - par une régression multinomiale pondérée pour l'appartenance aux classes pondération importante pour éviter des biais

Evaluation du fit du modèle



Extensions du modèle à classes latentes

prise en compte des spécificités des données de Qualité de Vie en Cancer

1. Aspect non Gaussien des scores de qualité de vie

Solution changer le modèle pour $Y|c$

- normaliser les données par une transformation continue (fonction R `lcmm`)
- considérer un modèle probit pour données ordinales (fonction R `lcmm`)

Extensions du modèle à classes latentes

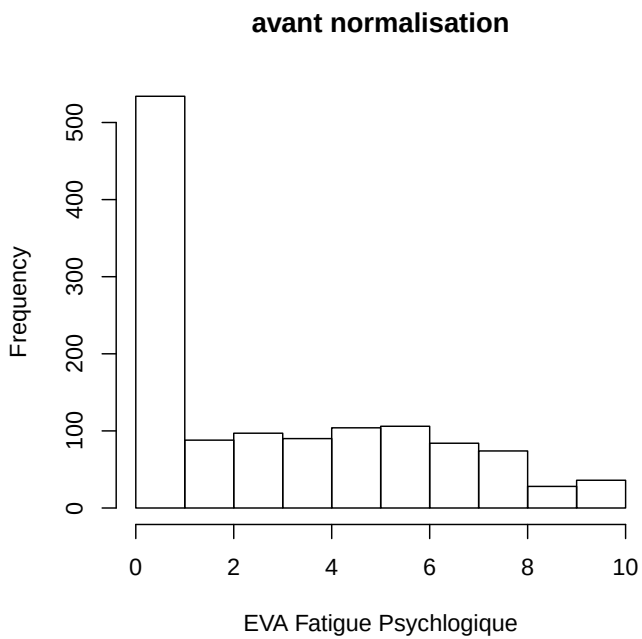
prise en compte des spécificités des données de Qualité de Vie en Cancer

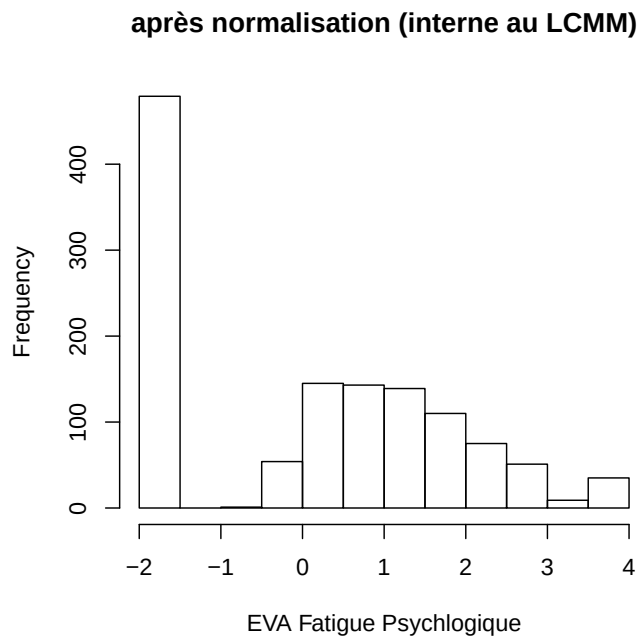
1. Aspect non Gaussien des scores de qualité de vie

Solution changer le modèle pour $Y|c$

- normaliser les données par une transformation continue (fonction R `lcmm`)
- considérer un modèle probit pour données ordinales (fonction R `lcmm`)

Exemble normalisation de la fatigue par des splines dans l'exemple précédent





Extensions du modèle à classes latentes (suite)

2. Association avec des événements cliniques

Solution modèle conjoint à classes latentes

- pour événements possiblement en compétition (fonction R `Jointlcmm`)

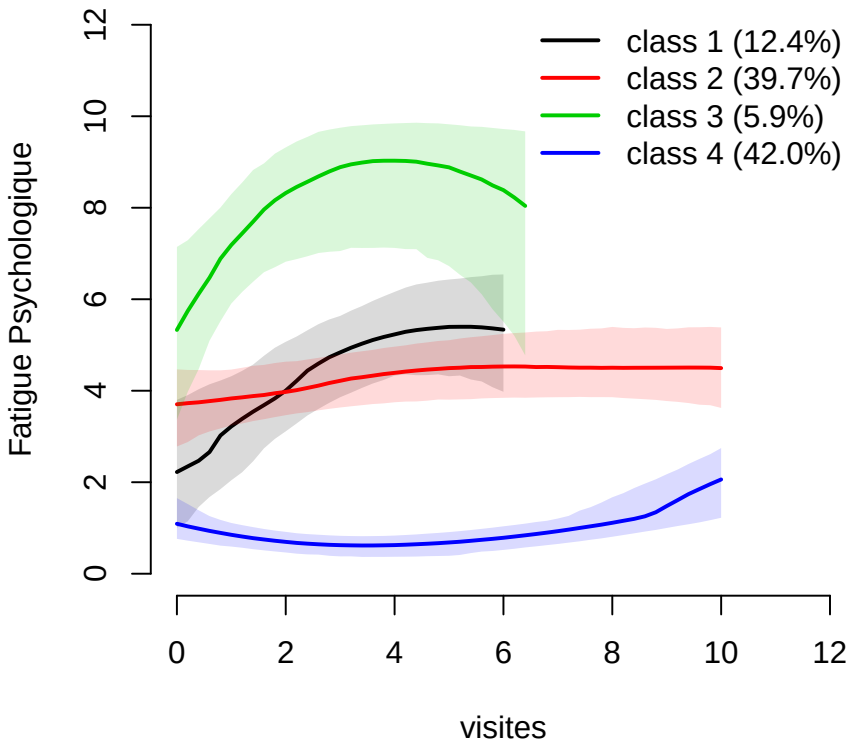
Extensions du modèle à classes latentes (suite)

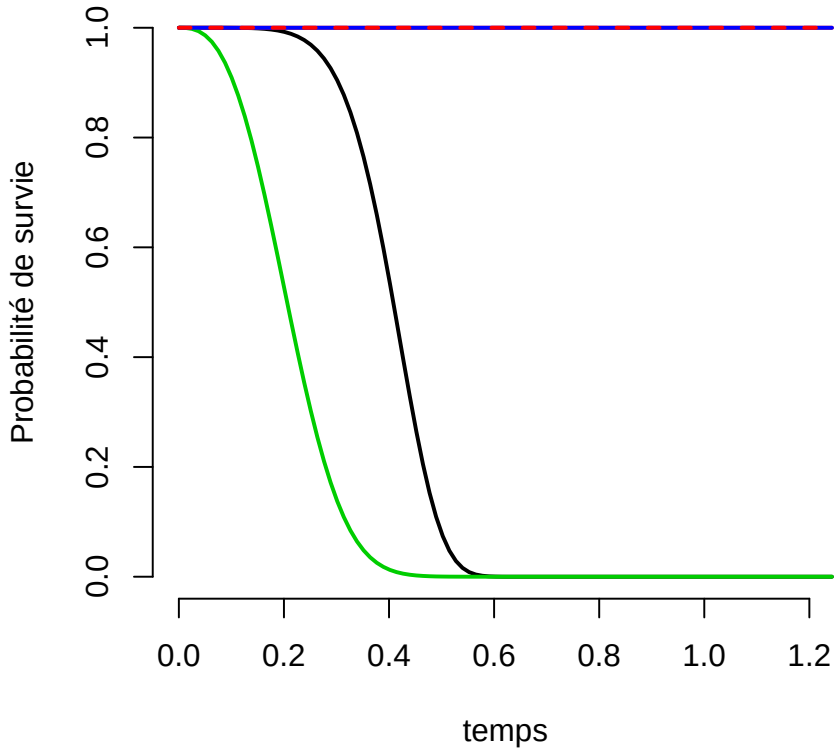
2. Association avec des événements cliniques

Solution modèle conjoint à classes latentes

- pour événements possiblement en compétition (fonction R `Jointlcmm`)

Exemple trajectoires de fatigue avec prise en compte du décès





Extensions du modèle à classes latentes (suite)

3. Aspect multivarié des échelles (unidimensionnelles)

Solution supposer un modèle type IRT à classes latentes

- avec des items continus (possiblement non Gaussiens) (R `multlcmm`)
- avec des items ordinaux (programme disponible en `Fortran`, à venir en R)

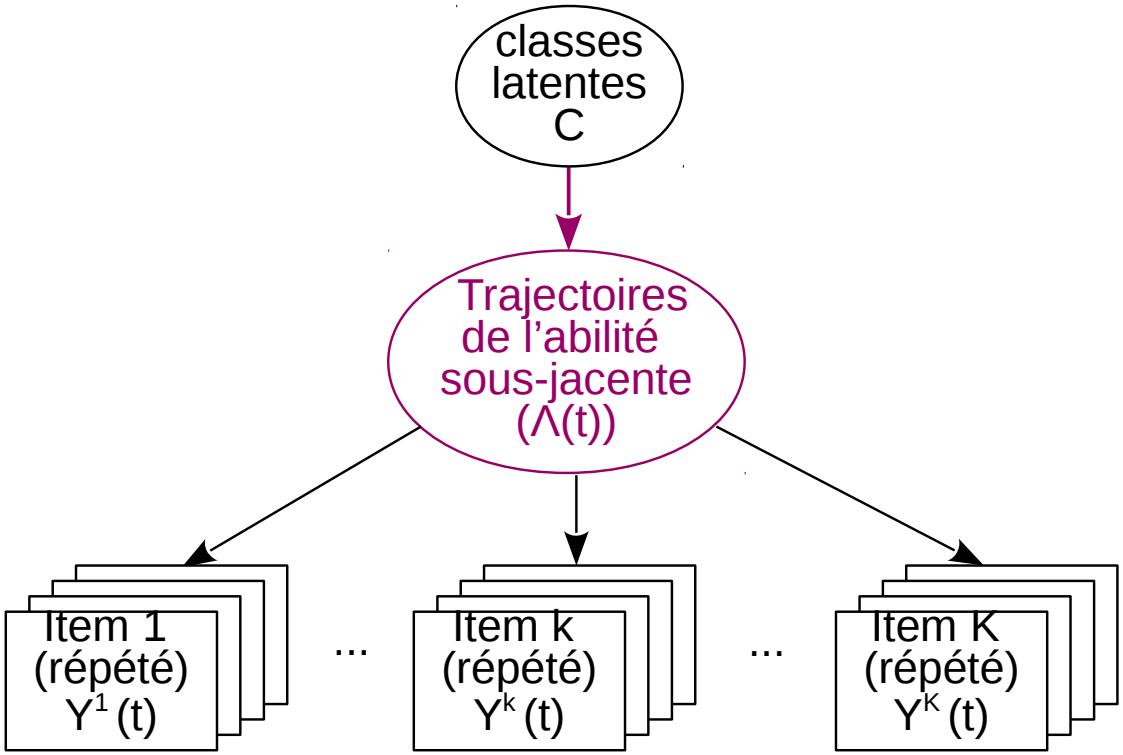
Extensions du modèle à classes latentes (suite)

3. Aspect multivarié des échelles (unidimensionnelles)

Solution supposer un modèle type IRT à classes latentes

- avec des items continus (possiblement non Gaussiens) (R `multlcmm`)
- avec des items ordinaux (programme disponible en `Fortran`, à venir en R)

Exemple Dépendance chez les personnes âgées (Edjolo et al., technical report)



Extensions du modèle à classes latentes (fin)

4. Dimensions multiples

Solution modèle conjoint à classes latentes

- pour plusieurs marqueurs répétés (fonction R actuellement en test)

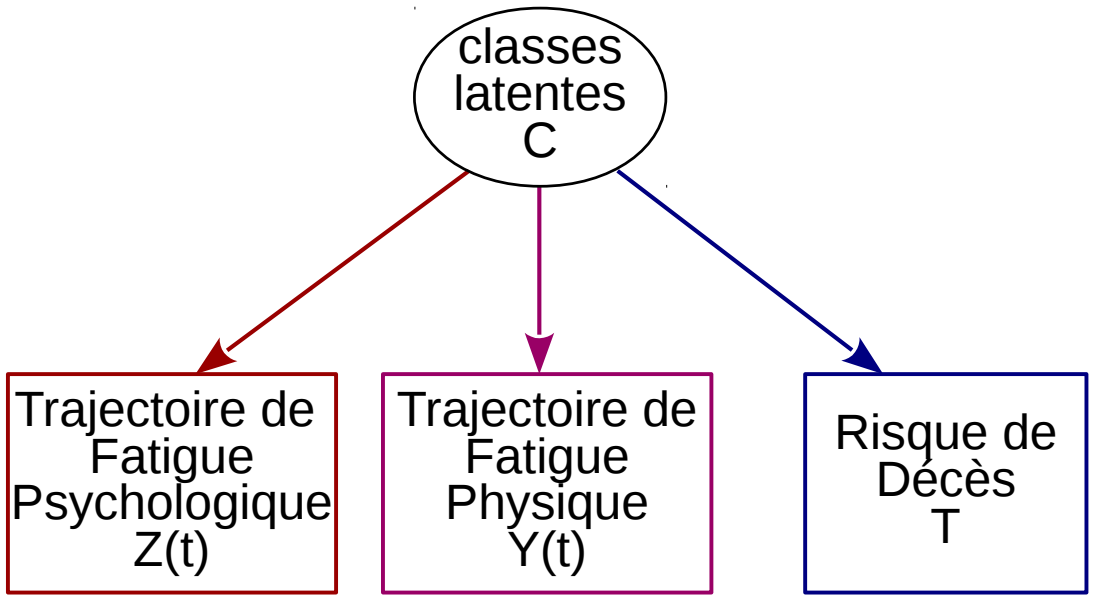
Extensions du modèle à classes latentes (fin)

4. Dimensions multiples

Solution modèle conjoint à classes latentes

- pour plusieurs marqueurs répétés (fonction R actuellement en test)

Exemple trajectoires de fatigue physique et psychologique (avec prise en compte du décès)



Conclusion

LCMM est un outil statistique puissant ...

- s'applique à tout type de données répétées
- répond à des questions diverses sur l'hétérogénéité
- profite des bonnes propriétés statistiques des modèles mixtes
- sépare bien les 2 sources of variabilité (individuel / groupe)
- implémenté dans des logiciels

Conclusion

LCMM est un outil statistique puissant ...

- s'applique à tout type de données répétées
- répond à des questions diverses sur l'hétérogénéité
- profite des bonnes propriétés statistiques des modèles mixtes
- sépare bien les 2 sources of variabilité (individuel / groupe)
- implémenté dans des logiciels

... à utiliser avec une extrême précaution

- Processus d'estimation (e.g. maxima locaux, choix du nombre de classes)
- Ne doit pas se réduire à l'approche exploratoire de Nagin (proc Traj)
  - ⋆ néglige la corrélation intra-individuelle
  - ⋆ surestime le nombre de classes
  - ⋆ biais dans les inférences
- Evaluation de la discrimination et du fit
- Interprétation prudente des classes latentes en lien avec
  - ⋆ le pouvoir discriminant
  - ⋆ hypothèses pertinentes sur l'existence de groupes latents

Remerciements et références

- Remerciements :
  - Viviane Philipps (statisticienne sur `lcmm`)
  - Louise Baussard, Florence Cousson-Gélie (Grant INCA FATIGUE-TR)

- Quelques références :

  - **Bauer, Curran (2003). *Psychol Meth*, 8(3), 338-63 (+ discutants 364-93)**
  - Hipp, Bauer (2006). *Psychological methods*, **11(1)**, 36-53
  - Muthén, Asparouhov (2009). *In Longitudinal Data Analysis* ed. by Fitzmaurice et al.
  - **Proust-Lima, Philipps, Liquet (2017). *Journal of Statistical Software* 78, 1-56.**
  - **van de Schoot, Sijbrandij, Winter et al. (2016) : The GRoLTS-Checklist, Structural Equation Modeling : A Multidisciplinary Journal**