

Journées 2007 du GDR Statistique et Santé

**Modèles par mélange de schémas d'observation
et modèles à classes latentes
pour le traitement des sorties d'étude
informatives**

**Etienne Dantan, Cécile Proust-Lima
& Hélène Jacqmin-Gadda**

INSERM U875, Bordeaux, France

Introduction

- Beaucoup de données manquantes (sorties d'étude) dans cohortes observationnelles ⇔ biais
- Typologie des données manquantes (Little & Rubin, 1987)

$y_i = (y_{o,i}, y_{m,i})$ où $y_{o,i}$ correspond aux données observées, $y_{m,i}$ aux manquantes et R_i au vecteur de variables indicatrices d'une donnée manquante

- Manquantes complètement aléatoirement (MCAR)

$$P(R_i = r | y_i) = P(R_i = r)$$

- Manquantes aléatoirement (MAR)

$$P(R_i = r | y_i) = P(R_i = r | y_{o,i})$$

- Manquantes non aléatoirement (MNAR) ou informatives

$$P(R_i = r | y_i) = P(R_i = r | y_{o,i}, y_{m,i})$$

Introduction

- Estimateurs du maximum de vraisemblance consistant si hypothèse de données MAR: probabilité d'une donnée manquante indépendante des données non observées
- Hypothèse MAR invérifiable \Rightarrow évaluation de la sensibilité à cette hypothèse

Introduction (Méthodes pour données MNAR)

⇒ Modélisation conjointe de l'évolution du marqueur et du processus de sortie d'étude

- Modèle de sélection

- Outcome dependent selection model

$$f(y_{o,i}, R_i | X_i) = \int f(y_{o,i}, y_{m,i} | X_i, \theta) f(R_i | y_{o,i}, y_{m,i}, X_i, \psi) dy_{m,i}$$

- Random-effect dependent selection models

$$f(y_{o,i}, R_i | X_i) = \int f(y_{o,i} | X_i, \alpha_i, \theta) f(R_i | X_i, \alpha_i, \psi) f(\alpha_i) d\alpha_i$$

- Principales limites:

- * $\hat{\theta}$ sensible à une mauvaise spécification des distributions
- * Difficultés numériques

Introduction (Méthodes pour données MNAR)

- Modèle par mélange de schémas d'observation (PMM)

$$f(y_{o,i}, R_i | X_i) = f(y_{o,i} | R_i, X_i, \theta) f(R_i)$$

– Principales limites:

- * nombre important de schémas d'observation
- * 2 sujets avec même date de sortie ont un même profil d'évolution

- Modèle à classes latentes (LCM)

– Simple LCM

$$f(y_{o,i}, R_i | X_i) = \sum_{g=1}^G f(y_{o,i} | c_{ig} = 1, X_i, \theta) f(c_{ig} = 1 | R_i, X_i) f(R_i | X_i)$$

– Joint LCM

$$f(y_{o,i}, R_i | X_i) = \sum_{g=1}^G f(y_{o,i} | c_{ig} = 1, X_i, \theta) f(R_i | c_{ig} = 1, X_i) f(c_{ig} = 1 | X_i)$$

Objectif

- Comparer 2 types d'approches sur un exemple:
 - Pattern Mixture Model
 - Modèle à classes latentes: en analyse simple et conjointe
- En effectuant une analyse de sensibilité aux sorties d'étude des résultats de l'influence de facteurs de risque de déclin cognitif

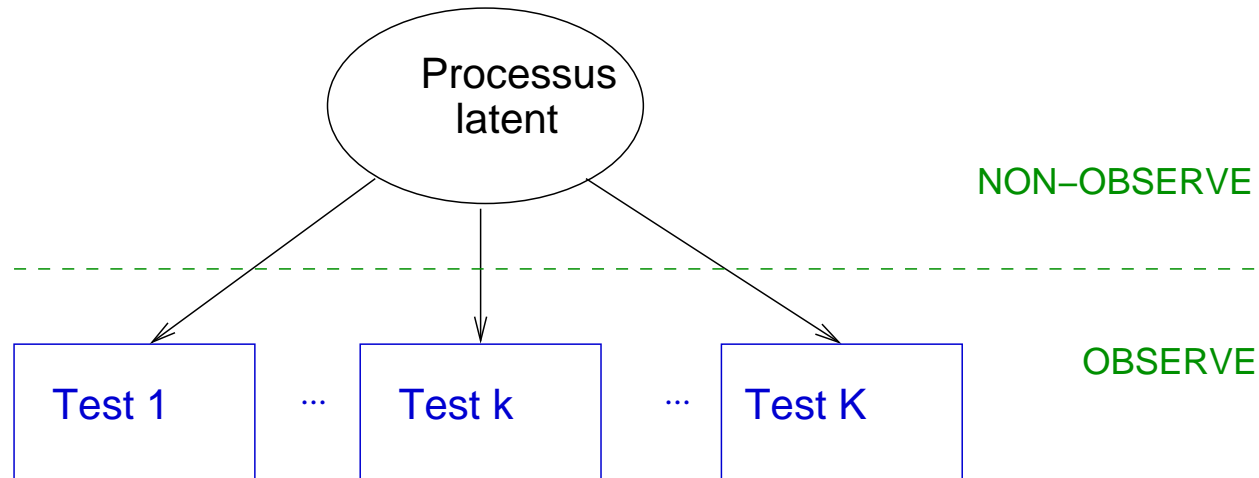
Plan

- Données
- Modèle pour données MAR: Modèle pour données longitudinales multivariées
- Modèles pour données MNAR
 - Pattern Mixture Model
 - Simple Latent Class Model
 - Joint Latent Class Model
- Résultats
- Conclusion

Données de l'application: Données longitudinales multivariées

- Cohorte PAQUID : étude du vieillissement normal et pathologique au cours d'un suivi de 7 visites (V0, V1, V3, V5, V8, V10, V13)
- Extrait étudié :
 - 1640 sujets ayant participé à une sous-étude de V3 à V13
 - sujets non déments à V3
 - 3 tests psychométriques : Benton, Isaacs, MMSE
- Sortie d'étude à t : aucun des 3 tests à t et jusqu'à la fin du suivi
⇒ Sortie à V5, V8, V10, V13 et pas de sortie

1. Modèle à processus latent pour données longitudinales multivariées: hypothèse de données MAR



$$\Lambda_i(t) = \beta_0 X_{1i}^T + \beta_1 X_{1i}^T t + u_{0i} + u_{1i} t$$

$$g(Y_{ijk}, \eta_k) = \Lambda_i(t_{ijk}) + \alpha_{ik} + \epsilon_{ijk}$$

$$u_i \sim \mathcal{N}(0, B) \text{ et } \alpha_{ik} \sim \mathcal{N}(0, \sigma_{\alpha_k}^2)$$

Proust et al., Biometrics 2006

2. PMM pour données longitudinales multivariées

Principe: évolution cognitive conditionnelle à la date de sortie

- Analyse stratifiée : paramètres tous spécifiques aux profils de sorties
 - beaucoup de paramètres
 - sous-groupes de tailles et de nombres de mesures variables
 - problème d'identifiabilité: sortie à V5 et V8 ensemble
- Analyse ajustée :

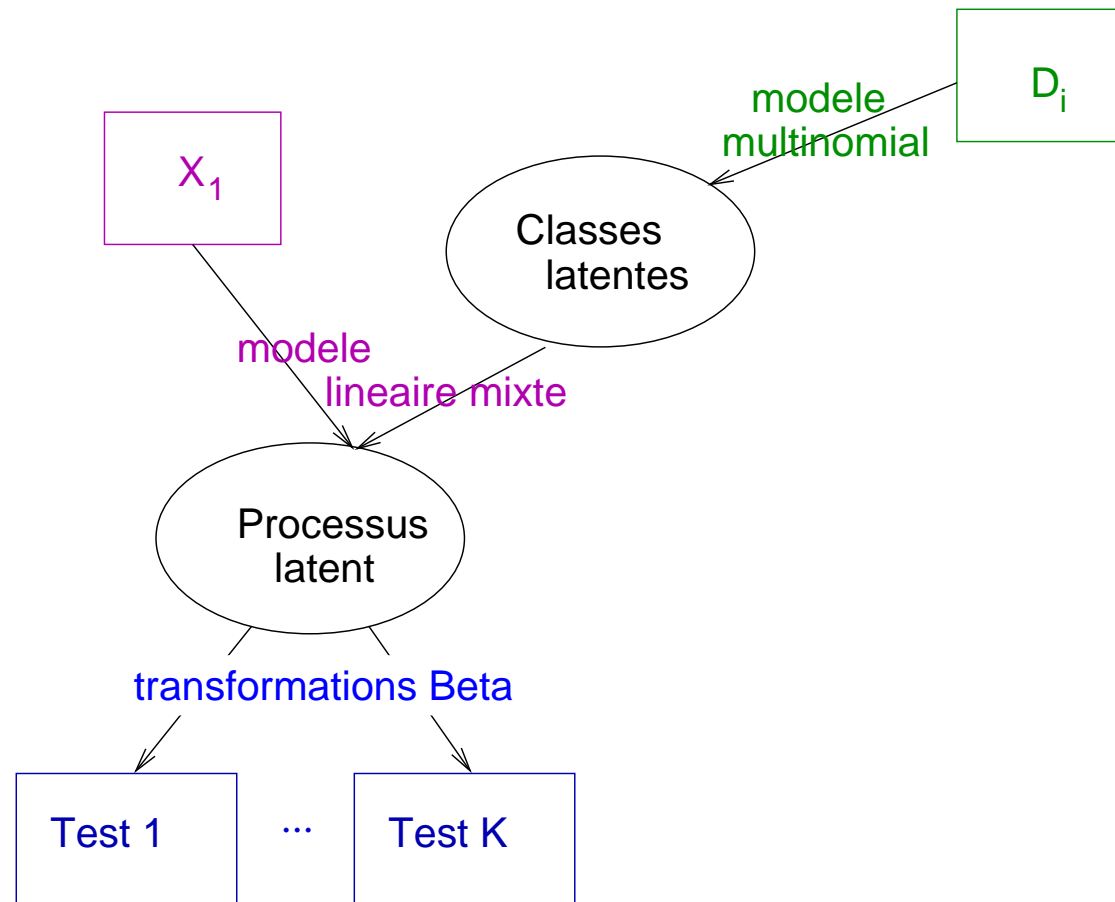
$$\Lambda_i(t) = \beta_0 X_{1i}^T + \beta_1 X_{1i}^T t + \sum_j D_{ij} (u_{0ij} + u_{1ij} t) \quad (1)$$

$$(u_{0ij}, u_{1ij})^T \sim \mathcal{N}((\beta_{0j}, \beta_{1j})^T, B_j) \text{ et } D_{ij} = \mathbb{1}\{\text{Sortie à } V_j\}$$

$$j \in \{5/8, 10, 13\}$$

3. Simple LCM pour données longitudinale multivariées

Principe: évolution cognitive conditionnelle aux G classes latentes liées au processus de sortie d'étude



3. Joint LCM pour données longitudinales multivariées

- Cognition latente spécifique à classe g

$$\Lambda_i(t) |_{c_{ig}=1} = \beta_0 X_{1i}^T + \beta_1 X_{1i}^T t + u_{0ig} + u_{1ig} t, \quad t \geq 0 \quad (2)$$

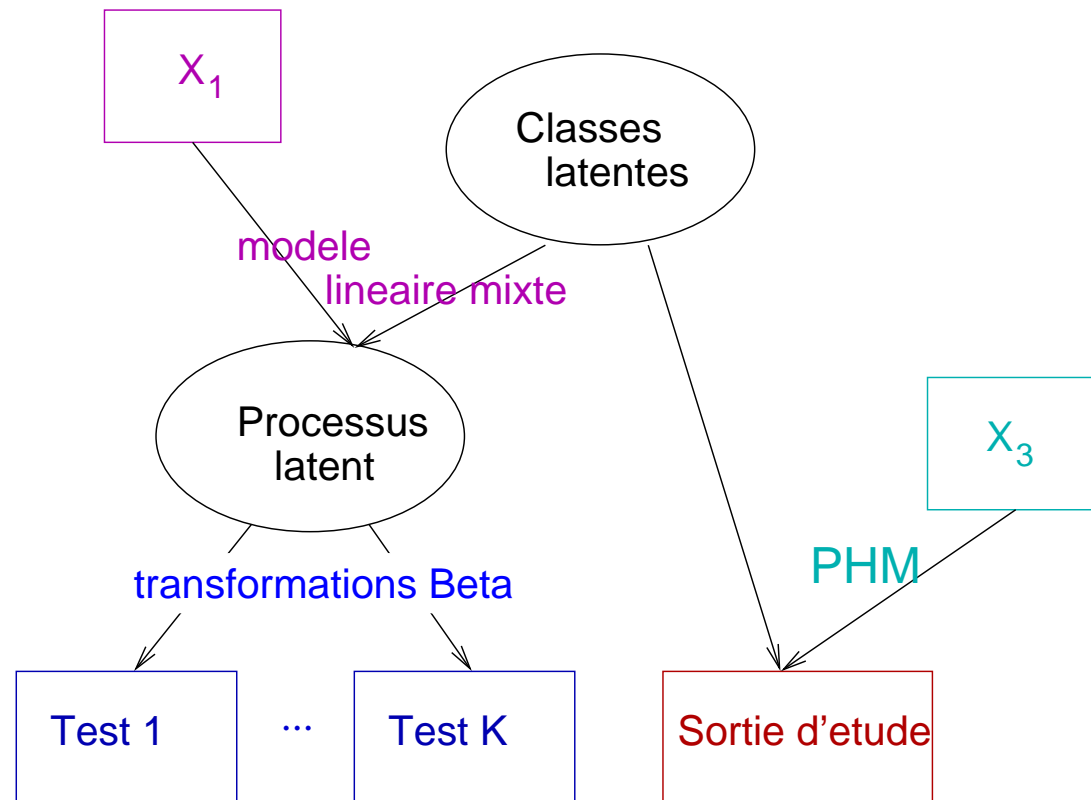
$$u_{ig} = (u_{0ig}, u_{1ig}) \sim \mathcal{N}(\mu_g, \omega_g^2 B) \text{ avec } \sum_{g=1}^G \mu_g = 0$$

- Probabilité appartenance aux classes par regression logistique multinomiale

$$\pi_{ig} = P(c_{ig} = 1 | D_i) = \frac{e^{\xi_{0g} + D_i^T \xi_{1g}}}{\sum_{l=1}^G e^{\xi_{0l} + D_i^T \xi_{1l}}} \quad (3)$$

$$D_i^T = (D_{i5}, D_{i8}, D_{i10}, D_{i13})^T \text{ avec } D_{ij} = \mathbb{1}\{\text{sortie à } V_j\}$$

4. Traitement des sorties d'étude: Modèle Conjoint à Classes Latentes



4. Traitement des sorties d'étude: Modèle Conjoint à Classes Latentes

- Probabilité appartenance aux classes

$$\pi_{ig} = P(c_{ig} = 1) = \frac{e^{\xi_{0g}}}{\sum_{l=1}^G e^{\xi_{0l}}} \quad (4)$$

- Modélisation conjointe du risque de sortie d'étude par un PHM :

$$\lambda(t \mid c_{ig} = 1, X_{2i}; \gamma_{0g}, \gamma_{1g}) = \lambda_0(t) e^{\gamma_{0g} + X_{2i} \gamma_1} \text{ avec } \gamma_{01} = 0 \quad (5)$$

- Risque de base $\lambda_0(t)$: fonction en escalier

Estimation

- Algorithme d'optimisation du type Newton-Raphson
 - Programme développé en Fortran
 - LCM estimé pour G classes
- ⇒ Selection par $BIC = -2 * \log(\text{Vraisemblance}) + k * \log(n)$
- où k =nombre de paramètres du modèle et n =nombre de sujets

Classification a posteriori

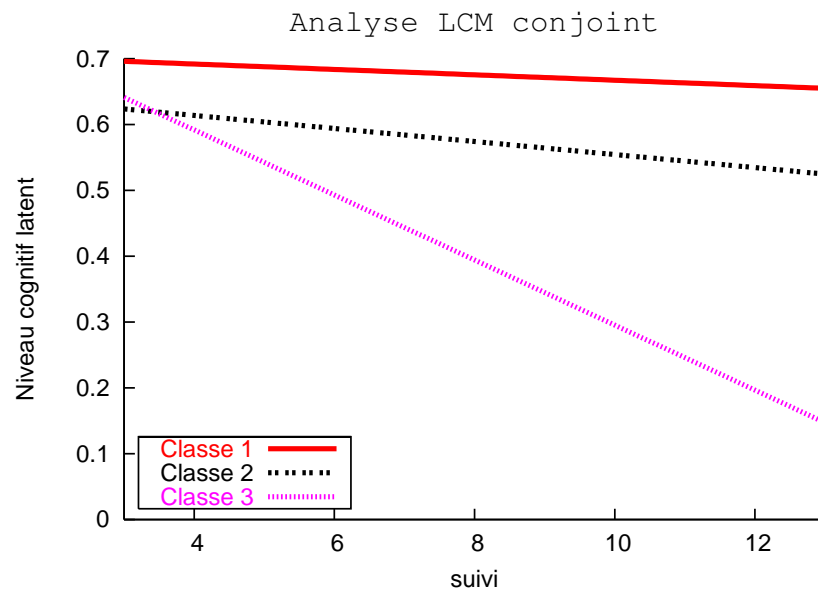
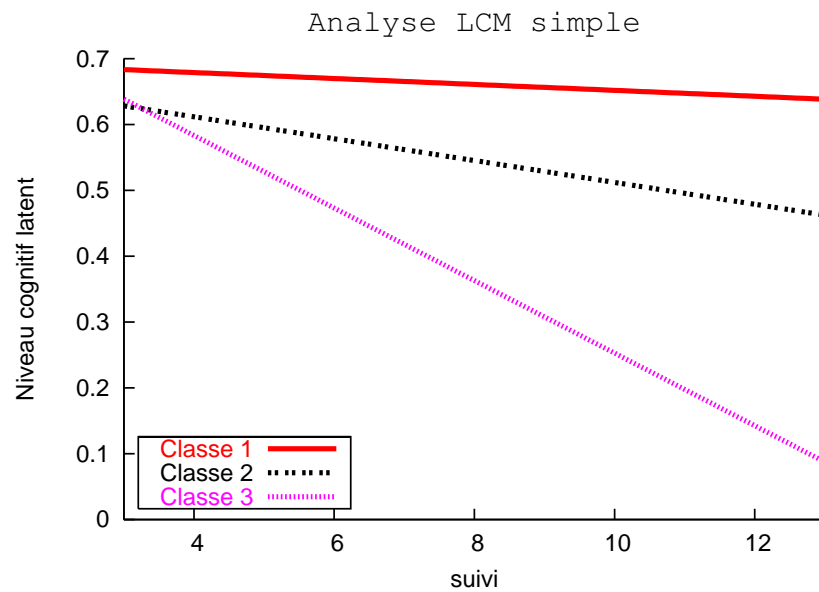
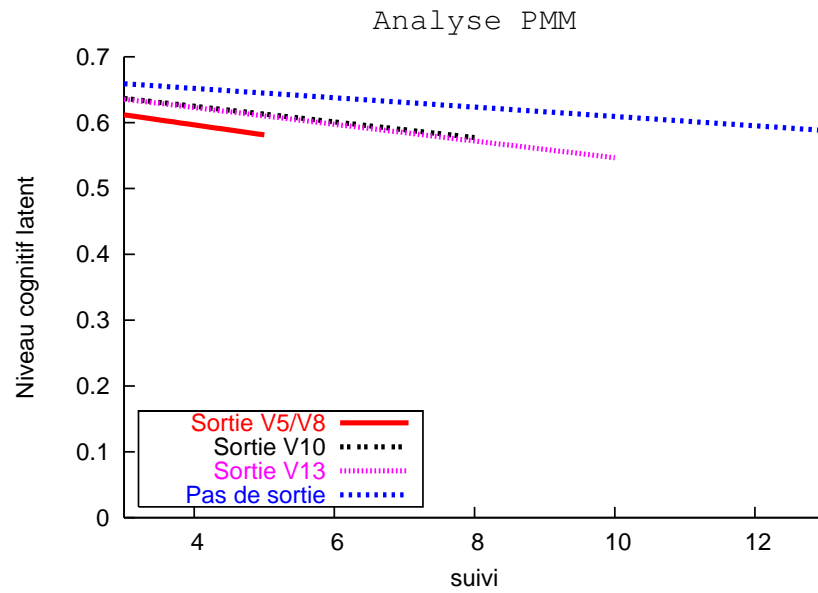
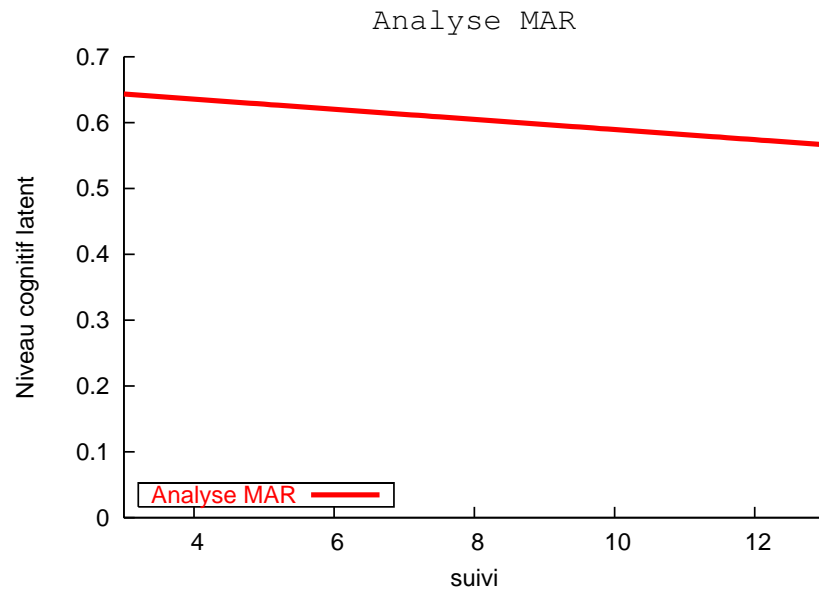
- $P(c_{ig} | Y_{ijk}, D_i)$ obtenu à l'aide du théorème de Bayes

Résultats

Table 1: Estimations du modèle pour données longitudinales multivariées pour l'analyse MAR, PMM, LCM simple et LCM conjoint

	MAR		PMM		LCM simple		LCM conjoint	
	β	p	β	p	β	p	β	p
Homme	0.008	0.07	0.012	0.01	0.008	0.06	0.004	0.33
Homme * t	0.021	0.01	0.024	<0.01	0.012	0.05	0.008	0.19
Pas de diplôme	-0.104	<0.01	-0.103	<0.01	-0.096	<0.01	-0.096	<0.01
Pas de diplôme * t	-0.006	0.56	-0.006	0.49	-0.016	0.03	-0.018	0.02

Résultats



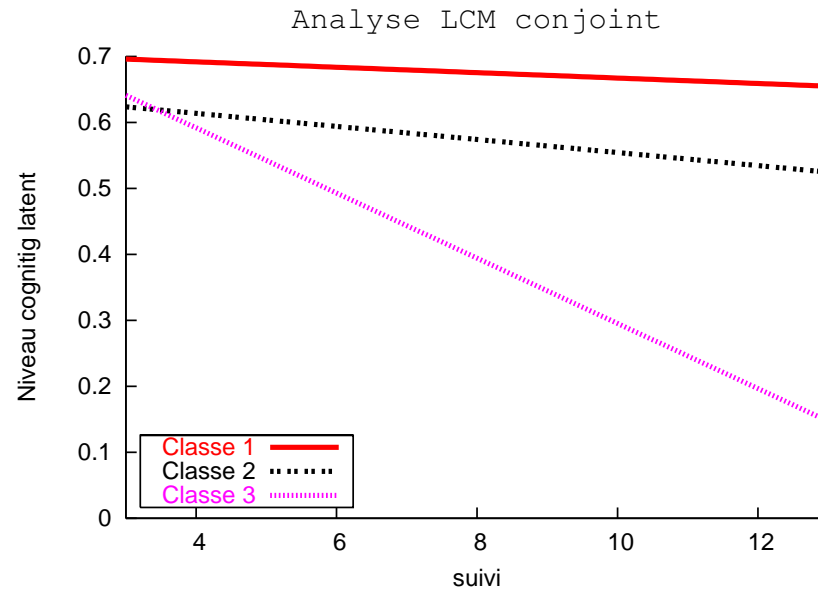
Résultats

Table 2: Description des caractéristique des sujets selon les classes latentes a posteriori pour le LCM simple (%)

	Classe 1 (N=1032)	Classe 2 (N=548)	Classe 3 (N=60)	p-value
Sortie à :				< 0.0001
V5	0.0	43.0	0.0	
V8	11.7	21.2	31.7	
V10	10.7	10.6	8.3	
V13	14.5	16.6	33.3	
Pas de sortie	63.1	8.6	26.7	
Déments	7.4	10.6	56.7	< 0.0001
Femmes	58.4	54.7	75.0	
Diplomés	76.6	70.3	81.7	0.0094

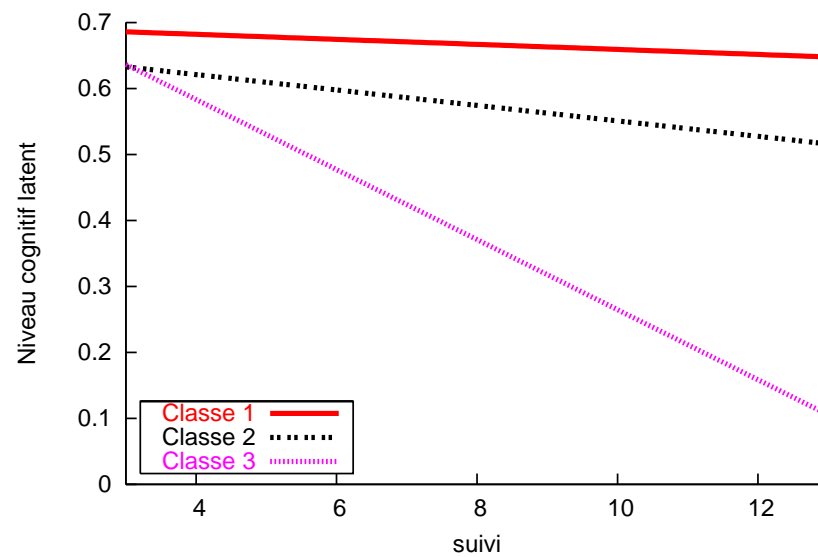
Résultats

$$\lambda(t) = \lambda_0(t)e^{\gamma_0 g + X_{2i}\gamma_1}$$



LCM avec indépendance entre classes latentes et sortie d'étude

$$\lambda(t) = \lambda_0(t)e^{X_{2i}\gamma_1}$$



Conclusions

- Différences d'interprétations de l'ajustement pour PMM et LCM
 - PMM: ajustement sur sortie d'étude
 - LCM: ajustement sur classes

⇒ analyses complémentaires pour mieux interpréter les classes latentes
- intérêt principal: interprétation des paramètres marginaux
 - calculées à partir des paramètres ajustés
 - formulation marginale directe^{1,2}
- PMM: mise en oeuvre et interprétation simple, mais hypothèses fortes
- LCM: plus de flexibilité, mais
 - complexité des procédures d'estimation
 - difficulté d'interprétation

¹Roy et al., *Biostatistic* 2007

²Wilkins et al., *online in* 2007

References

- [1] Little RJA, Rubin DB (1987). Statistical Analysis with Missing Data. New York: Wiley.
- [2] Little RJA (1993). Pattern-Mixture Models for Multivariate Incomplete Data. Journal of the American Statistical Association, 88, 125-134.
- [3] Lin H, McCulloch CE, Rosenheck RA (2004). Latent pattern mixture models for informative intermittent missing data in longitudinal studies. Biometrics, 60(2):295-305.
- [4] Proust C, Jacqmin-Gadda H, Taylor J, Ganiayre J, Commenges D (2006). A nonlinear model with latent process for cognitive evolution using multivariate longitudinal data. Biometrics 2006, 62, 1014-1024
- [5] Roy J (2003). Modeling longitudinal data with nonignorable dropouts using a latent dropout class model. Biometrics ;59(4):829-36.