Estimation des paramètres du modèle à classes latentes pour des données longitudinales

Ndiogou Seck, Mounir Mesbah

Pub. Inst. Stat. Univ. Paris
62, fasc.1-2, 2018, 33-60

Estimation des paramètres du modèle à classes latentes pour des données longitudinales

Ndiogou Seck [1] et Mounir Mesbah[2]
L.S.T.A., Université Pierre et Marie Curie, Paris 6.
4, Place Jussieu, Paris, France.
[1] seck.ndiogou@gmail.com et [2] mounir.mesbah@upmc.fr

12 mars 2018

Résumé
Dans ce travail, nous présentons le modèle à classes latentes dans le cas longitudinal. Ce modèle est une extension du modèle à classes latentes. Nous ne traitons que le cas où le temps est considéré comme étant discret. Dans ce modèle, avec des items polytomiques, nous avons trois paramètres à estimer : la probabilité qu'un individu soit dans une classe latente à l'étape initiale, la probabilité de transition et enfin, la probabilité qu'un individu d'une classe quelconque réponde à un item à un instant donné. Enfin, nous appliquons la méthode à des données réelles en qualité de vie. Pour cela, nous utiliserons la procédure proc LTA de SAS pour obtenir numériquement l'estimation des paramètres.
Mots-clés : classes latentes, probabilité de transition, longitudinal, algorithme EM, proc LTA.
Abstract
In this work, we introduce the latent class model in the longitudinal case. This model is an extension of the classical latent class model. We consider only the discrete time case. In this model with polytomous items, we have three parameters to estimate : the first one is the probability that an individual is in a latent class at the initial stage, the second is the transition probability and the third is the probability that at a given time an individual in a class to answer an item. Finally, we apply the method to a real data Quality of life data set. We use the SAS proc LTA to get numerical estimation of these parameters.
Keywords : Latent class, transition probability, longitudinal, EM algorithm, proc LTA.

1 Introduction

Le modèle des variables latentes est un modèle de mélange introduit par Lazarsfeld et Henry (1968). D'autres auteurs se sont également intéressé à ce ce modèle, comme Andersen(1982) pour l'estimation des paramètres du modèle à classes latentes. Gibson(1955) a étendu les solutions solutions proposées par Anderson (1954) et Goodman(1974) a traité le cas multidimensionnel. Ce modèle est important et utile pour l'analyse de données multivariées. Il suppose l'existence de variables inobservées (latentes) dont on peut quantifier l'effet. Connaître les variables latentes permettrait donc de diminuer les corrélations entre variables observées. Les variables observées sont supposées indépendantes conditionnellement aux variables latentes. Le modèle des classes latentes caractérise souvent des variables discrètes latentes comme l'ont présenté Bartholomew et Knott(1999). Le but de ce travail est de présenter le modèle des classes latentes dans le cas longitudinal. Les données longitudinales, constituent un domaine important de la statistique. On entend par données longitudinales des données telles que, pour chaque individu considéré, on dispose d'observations à différents instants, autrement dit répétées dans le temps. Le principal domaine d'application de ce type de données est la santé. La difficulté majeure dans le traitement statistique de ces données provient de ce qu'il n'est en général pas réaliste de supposer que les observations réalisées sur un même individu, au cours du temps, sont indépendantes. Il est donc nécessaire d'introduire une structure de covariance, pour les variables aléatoires associées à chaque individu, expliquant ces corrélations. Par ailleurs, il est fréquent dans les modèles pour données répétées de considérer, en plus des facteurs à effets fixes que l'on souhaite étudier dans le modèle, des effets aléatoires associés aux individus. On aura pour cela recours à des modèles mixtes.
Dans ce qui suit, nous allons considérer les cas suivants :
-Le modèle marginal, qui se concentre sur le changement de distribution d'une variable pendant le temps. Dans ce modèle, les individus sont supposés appartenir à la même classe latente au cour du temps.
-Le modèle conditionnel ou transitoire par classe, qui étudie les changements entre les classes consécutifs au cour du temps. Dans ce modèle l'appartenance d'un individu à une classe peut varier en fonction du temps. Dans ce cas, nous allons définir le modèle de Markov pour les classes latentes ou modèle de transition développé par des auteurs comme Collins et Wugalter (1992) et Vermunt et Magidson (2003). Dans la suite de notre étude nous considérons que le temps est discret et que les covariables sont indépendantes du temps. Diggle et al. (2014), ont développé des approches alternatives pour l'analyse des données longitudinales.
Dans la section 2, nous précisons nos notations et présentons les observations et modèles, en particulier le modèle marginal dans lequel l'appartenance à une classe latente ne change pas dans le temps, et le modèle transitoire, dans lequel les indi-

vidus peuvent changer de classe latente.
Dans la Section 3, nous présenterons l'algorithme de EM (Expectation-Maximisation) pour l'estimation des paramètres. Cet algorithme a été proposé par Dempster et Rubin (1977) dans un contexte plus général de données incomplètes (correspondant dans notre cas, aux variables observées), que complètent les variables latentes inobservées.
Dans la Section 4, les données réelles d'une étude en qualité de vie sont décrites. La procédure de SAS Proc LTA, qui va nous permettre d'obtenir les estimations numériques est présentée succinctement. Ces paramètres sont définis ainsi : la probabilité qu'un individu soit dans une classe latente à l'étape initiale, la probabilité de transition et la probabilité qu'un individu d'une classe quelconque réponde à une modalité d'un item sont interprétés. La Section 5 est consacrée à la conclusion.

2 Observations et modèles

2.1 Observations

On observe les réponses de n individus à J variables. Soit $Y_1^{(i)}(t), Y_2^{(i)}(t), ..., Y_J^{(i)}(t)$ J variables aléatoires appelées items ou questions à l'instant t. Notons $Y_j^{(i)}(t)$ le j-vecteur aléatoire à l'instant t constitué par les variables aléatoires qui prennent les valeurs $l \in \{1, 2, ..., L_j\}$, l est appelé modalité de réponse, L_j le nombre de modalités de réponses de la variable j. Les composantes du vecteur sont : $Y_j^{(i)}(t) = (Y_{j,1}^{(i)}(t), Y_{j,2}^{(i)}(t), ..., Y_{j,L_j}^{(i)}(t))$ pour tout, $t \in \{1, 2, ..., T\}$. La réponse de l'individu $i \in \{1, 2, ..., n\}$ à la question $j \in \{1, 2, ..., J\}$ à l'instant $t \in \{1, 2, ..., T\}$ est notée par $y_j^{(i)}(t) = (y_{j,1}^{(i)}(t), y_{j,2}^{(i)}(t), ..., y_{j,L_j}^{(i)}(t)) \in \{0, 1\}^{L_j}$

2.2 Les modèles

Les individus sont répartis en K classes inobservées ou latentes. Nous allons présenter deux modèles : le modèle marginal et le modèle de transition.

2.2.1 Le modèle marginal

Le modèle marginal suppose un changement de distribution au cours du temps. Néanmoins, pour ce modèle chaque individu est supposé rester dans la même classe latente au cours du temps. Les classes latentes sont inchangées. Le nombre de classes latentes ne varie pas avec le temps.
Soit k_t, la classe d'un individu à l'instant t, pour le modèle marginal, l'individu i est supposé appartenir à la même classe latente au cours du temps et dans ce cas k_t ne dépendra pas de t dans la suite on posera $k_t = k$.

Notons $\Theta^i(t)$ la variable aléatoire associée aux classes latentes au temps t pour un individu i avec $\Theta^i(t) \in \{1, 2, ..., K\}$.
Dans le cas du modèle marginal les individus sont supposés appartenir à la même classe durant tout le temps. Dans ce type de modèle le nombre de classes latentes reste constant dans ce cas, la variable aléatoire $\Theta^i(t)$ ne dépendra pas du temps. On peut poser $\Theta^i(t) = \Theta^i = 1, 2, ..., K$.
Soit $P(\Theta^i = k)$ la probabilité, indépendante de t, qu'un individu i soit dans la classe k à l'instant t.
On notera par

$$\lambda_{j,l}(t|k) = P[Y_{j,l}^{(i)}(t) = 1/\Theta^i = k] \tag{1}$$

$i \in \{1, 2, ..., n\}$, $j \in \{1, 2, ..., J\}$, $k \in \{1, 2, ..., K\}$ et $l \in \{1, 2, ..., L_j\}$

la probabilité qu'un individu i de classe k donne la réponse l à l'item j à l'instant t. Cette probabilité $\lambda_{j,l}(t|k)$ vérifie les contraintes suivantes : $\lambda_{j,l}(t|k) \geq 0$ et $\sum_{l=1}^{L_j} \lambda_{j,l}(t|k) = 1$, car pour chaque item, les individus choisissent une et seule modalité de réponse proposée.
La probabilité marginale de réponse aux items pour un individu (i), définie par, $P(i) = P(Y_1^{(i)}(t) = y_1^{(i)}(t), Y_2^{(i)}(t) = y_2^{(i)}(t), ..., Y_J^{(i)}(t) = y_J^{(i)}(t))$, est égale à :

$$P(i) = \sum_{k=1}^{K} \prod_{t=1}^{T} P(\Theta^i = k) \prod_{j=1}^{J} \prod_{l=1}^{L_j} P(Y_{j,l}^{(i)}(t) = y_{j,l}^{(i)}(t)/\Theta^i = k) \tag{2}$$

$$i = 1, ..., n, j = 1, ..., J, t = 1, ..., T, k = 1, ..., K, k = 1, 2, ...K, k_t = 1, 2, ...K$$

Cette expression est obtenue en utilisant la propriétée d'indépendance locale des items (indépendance des items conditionellement à la classe latente). Posons $\alpha_k = P(\Theta^i = k)$, la probabilité d'appartenir à la classe latente k. Les réponses des individus sont indépendantes, la vraisemblance du modèle est donnée par

$$L(Y) = \prod_{i=1}^{n} \sum_{k=1}^{K} \prod_{t=1}^{T} \alpha_k \prod_{j=1}^{J} \prod_{l=1}^{L_j} \left(\lambda_{j,l}(t|k) \right)^{y_{j,l}^{(i)}(t)} \tag{3}$$

Avec $\lambda_{j,l}(t|k)$ et α_k, les paramètres à estimer.
La probabilité qu'un individu i de réponse $y^i(t)$ à un temps t appartienne à une classe k appelée probabilité à posteriori est définie par :

$$P(\Theta^i = k)/Y^{(i)}(t) = y^{(i)}(t)) = \frac{\alpha_k P(y^{(i)}(t)/\Theta^i = k)}{\sum_{r=1}^{K} \alpha_r P(y^{(i)}(t)/\Theta^i = r)} \tag{4}$$

2.2.2 Le modèle conditionnel ou transitoire

Pour décrire le modèle conditionnel ou transitoire, nous allons nous servir du modèle de transition qui utilise les chaînes de Markov. Le modèle de transition de Markov, les chaînes et processus de Markov sont souvent utilisés en Biologie, Finance, Physiques etc... Les chaînes de Markov modélisent le passage d'une classe à une autre à un instant t donné. Un des éléments clés de ce modèle est que les transitions qui se produisent entre classes latentes au fil du temps sont modélisées à l'aide du processus de premier ordre de Markov (Wiggins (1973)). Nous définissons également le modèle à classes latentes de Markov plus connu sous le modèle caché de Markov (voir les travaux de Baum (1970)) sur les modèles de changement ou ceux aussi Hamilton et Raj (2002).
Dans ce qui suit nous allons considérer le temps t comme étant discret, et l'existence de K classes latentes au cours du temps t.
Le modèle de Markov suppose une dépendance particulière entre l'appartenance d'un individu à une classe au temps t et sa présence dans une autre classe au temps $t+1$.
La probabilité de transition est la probabilité conditionnelle qu'un individu soit dans la classe k_{t+1} au temps $t+1$ sachant qu'il était dans la classe k_t à l'instant t, avec $k_t \in \{1, 2, ..., K\}$ et $k_{t+1} \in \{1, 2, ..., K\}$. On observe les réponses de n individus $i(i = 1, ...n)$ appartenant à K classes. Soit $\Theta^i(t)$ la variable aléatoire associée à la classe latente k_t au temps t pour un individu i avec $\Theta^i(t) \in \{1, 2, ..., K\}$ et $\Theta^i = (\Theta^i(1), \Theta^i(2), ..., \Theta^i(T))$.
Pour tout $i \in \{1, 2, ..., n\}$, les $\Theta^i(t); t \in \{1, 2, ..., K\}$ sont des chaines de Markov homogènes i.i.d. dont la matrice de transition commune $\Gamma = (\gamma_{(u,v)})_{1 \leq (u,v) \leq K}$ est définie pour $t \geq 1$ par

$$\gamma_{(u,v)} = P(\Theta^i(t) = v / \Theta^i(t-1) = u) \tag{5}$$

et obéissent à un modèle de régression logistique (voir les travaux de Baum (1970)) sur les modèles de changement ou ceux aussi Hamilton et Raj (2002).
La probabilité marginale qu'un individu i soit dans une classe k_t à l'instant t est définie par l'équation :

$$P(\Theta^i(t) = k_t) = \sum_{k_t} P(\Theta^i(t-1) = k_{t-1}) P(\Theta^i(t) = k_t / \Theta^i(t-1) = k_{t-1}) \tag{6}$$

En posant $\gamma_{(k_t, k_{t-1})} = P(\Theta^i(t) = k_t / \Theta^i(t-1) = k_{t-1})$ la probabilité de transition, la probabilité marginale qu'un individu soit dans une classe k_t devient :

$$P(\Theta^i(t) = k_t) = \sum_{k_t=1}^{K} P(\Theta^i(t-1) = k_{t-1}) \gamma_{(k_t, k_{t-1})}, k_t = 1, ..., K \tag{7}$$

La probabilité de transition $\gamma_{(k_t,k_{t-1})}$ vérifie les contraintes suivantes :

$\gamma_{(k_t,k_{t-1})} \geq 0$ et $\sum_{k_t=1}^{K} \gamma_{(k_t,k_{t-1})} = 1, i = 1, ..., n; t = 1, ...T.$

Les individus sont supposés appartenir à des classes latentes différentes selon l'évolution du temps.

Le modèle logit associé au modèle de transition pour des réponses nominales est :

$$\log\left(\frac{p(\Theta^i(t) = k_t/\Theta^i(t-1) = k_{(t-1)})}{p(\Theta^i(t) = 1/\Theta^i(t-1) = k_{(t-1)})}\right) = \beta_k(t), \tag{8}$$

Avec $k_t = 2, ..., K, k_{t-1} = 1, ..., K$ et $\beta_k(t)$ le coefficient de régression.

La première classe latente est choisie arbitrairement comme la classe de référence.

Dans ce qui suit nous allons utiliser les chaînes de Markov pour décrire le modèle à classes latentes dans le cas où les données sont longitudinales.

Soit $\Theta^i(t)$ la variable aléatoire associée aux classes latentes au temps t pour un individu i avec $\Theta^i(t) = 1, 2, ..., K$,

et $\Theta^i = (\Theta^i(1), \Theta^i(2), ..., \Theta^i(T))'$, le vecteur des classes latentes pour un individu i durant toute la période T.

Contrairement au modèle marginale, dans le cas du modèle de transition, la classe latente d'un individu varie en fonction du temps. Par définition la probabilité marginale de réponse d'un individu i est :

$$P(Y^{(i)} = l) = \sum_{\Theta^i(1)=1}^{K} \sum_{(\Theta^i(2)=1}^{K} ... \sum_{(\Theta^i(T)=1}^{K} P((\Theta^i(1) = k_1, (\Theta^i(2) = k_2, ..., (\Theta^i(T) = k_T) \times P(Y^{(i)} = l/\Theta^i(1) = k_1, \Theta^i(2) = k_2, ..., \Theta^i(T) = k_T) \tag{9}$$

Avec $l \in \{1, 2, ..., L_j\}$ et pour tout temps t $k_t \in \{1, 2, ..., K\}$.

On suppose qu'à chaque instant t un individu i est supposé appartenir à un classe k_t. On définit par $P(\Theta^i(1) = k_1, \Theta^i(2) = k_2, ..., \Theta^i(T) = k_T)$, la probabilité jointe qu'un individu i soit dans une classe k à chaque étape.

Cette probabilité jointe est définie comme suite :

$$P(\Theta^i(1) = k_1, \Theta^i(2) = k_2, ..., \Theta^i(T) = k_T) = \prod_{t=1}^{T} P(\Theta^i(1) = k_1) P(\Theta^i(t) = k_t/\Theta^i(t-1) = k_{t-1}) \tag{10}$$

Comme $P(\Theta^i(1) = k_1)$ est indépendante du temps nous avons

$$P(\Theta^i(1) = k_1, \Theta^i(2) = k_2, ..., \Theta^i(T) = k_T) = P(\Theta^i(1) = k_1) \prod_{t=2}^{T} P(\Theta^i(t) = k_t/\Theta^i(t-1) = k_{t-1}) \tag{11}$$

Avec $P(\Theta^i(1) = k_1)$ la probabilité qu'un individu i soit dans la classe k_1 à l'instant initial, et $P(\Theta^i(t) = k_t/\Theta^i(t-1) = k_{t-1})$ la probabilité de transition entre les instants t et $t+1$.

Si à chaque instant t, un individu i est supposé appartenir à une classe, la probabilité que cet individu réponde à la modalité l aux items est définie par

$$P(Y^{(i)} = l/\Theta^i(1) = k_1, \Theta^i(2) = k_2, ..., \Theta^i(T) = k_T).$$

Cette probabilité est égale à :

$$\begin{aligned} P(Y^{(i)} = l/\Theta^i(1) = k_1, \Theta^i(2) = k_2, ..., \Theta^i(T) = k_T) &= \prod_{t=1}^{T} P(Y^{(i)}(t) = l/\Theta^i(t) = k_t) \quad (12) \\ &= \prod_{t=1}^{T}\prod_{j=1}^{J}\prod_{l=1}^{L_j} P(Y_{j,l}^{(i)}(t) = 1/\Theta^i(t) = k_t) \end{aligned}$$

Finalement la probabilité marginale de réponse aux items est égale à :

$$\begin{aligned} P(Y^{(i)} = l) = &\sum_{\Theta^i(1)=1}^{K}\sum_{\Theta^i(2)=1}^{K} ... \sum_{\Theta^i(T)=1}^{K} P(\Theta^i(1) = k_1) \\ &\times \prod_{t=2}^{T} P(\Theta^i(t) = k_t/\Theta^i(t-1) = k_{t-1}) \prod_{t=1}^{T}\prod_{j=1}^{J}\prod_{l=1}^{L_j} P(Y_{j,l}^{(i)}(t) = 1/\Theta^i(t) = k_t) \end{aligned} \quad (13)$$

$$i = 1, ..., n, j = 1, ..., J, t = 1, ..., T, k = 1, ..., K, k_{t+1} = 1, 2, ...K, k_t = 1, 2, ...K$$

Soit α_{k_1} la probabilité d'appartenir à la classe latente initiale, c'est à dire la probabilité qu'un individu i soit dans une classe k à l'instant $t = 1$
Avec

$$\alpha_{k_1} = P(\Theta^i(1) = k_1), i = 1, ..., n, k_1 = 1, ..., K. \quad (14)$$

la probabilité α_1 vérifie les contraintes suivantes :

$$\alpha_{k_1} \geq 0 \text{ et } \sum_{k_1=1}^{K} \alpha_{k_1} = 1, i = 1, ..., t = 1, ...T,$$

En remplaçant la probabilité de transition $P(\Theta^i(t) = k_t/\Theta^i(t-1) = k_{t-1})$ par

$\gamma_{(k_t,k_{t-1})}$ et $P(Y_{j,l}^{(i)}(t) = 1/\Theta^i(t) = k_t)$ par $\lambda_{j,l}(t|k)$ on obtient :

$$P(Y^{(i)} = l) = \sum_{k_1=1}^{K} \sum_{\theta_{i_2}=1}^{K} ... \sum_{\theta_{i_T}=1}^{K} \alpha_{k_1} \times \prod_{t=1}^{T} \gamma_{(k_t,k_{t-1})} \prod_{t=1}^{T} \prod_{j=1}^{J} \prod_{l=1}^{L_j} (\lambda_{j,l}(t|k)) \tag{15}$$

$i = 1, ..., n, j = 1, ..., J, t = 1, ..., T, k = 1, ..., K, k_{t+1} = 1, 2, ...K, k_t = 1, 2, ...K$

Les réponses des individus sont indépendantes, la vraisemblance du modèle est donnée par

$$L(Y) = \prod_{i=1}^{n} P(Y^{(i)} = l).$$

En remplaçant $P(Y^{(i)} = l)$ par son expression, on obtient :

$$L(Y) = \prod_{i=1}^{n} \sum_{k_1=1}^{K} \sum_{\theta_{i_2}=1}^{K} ... \sum_{\theta_{i_T}=1}^{K} \alpha_{k_1} \prod_{t=1}^{T} \gamma_{(k_t,k_{t-1})} \prod_{t=1}^{T} \prod_{j=1}^{J} \prod_{l=1}^{L_j} (\lambda_{j,l}(t|k))^{y_{j,l}^i(t)}. \tag{16}$$

$$i = 1, ..., n, j = 1, ..., J, t = 1, ..., T, z_{it} = 1, ..., K$$

$\gamma_{(k_t,k_{t-1})}$, $\lambda_{j,l}(t|k)$ et α_{k_1}, représentent les paramètres à estimer.
Les paramètres du modèle de Markov à classes latentes peuvent être estimés à l'aide d'un maximum vraisemblance ML. Ces paramètres à estimer peuvent être décrits ainsi :

$$\alpha_{k_1} = \frac{\exp(\beta_{k_1})}{\sum_{k_1=1}^{K} \exp(\beta_{k_1})}, k_1 = 1, ..., K \tag{17}$$

$$\gamma_{(k_t,k_{t-1})} = \frac{\exp(\beta_k^t)}{\sum_{k=1}^{K} \exp(\beta_k^t)}, k = 1, ..., K, t = 1, ..., T \tag{18}$$

$$\lambda_{j,l}(t|k) = \frac{\exp(\beta_{kl})}{\sum_{l=1}^{L_j} \exp(\beta_{kl})}, k_1 = 1, ..., K, l = 1, ..., L_j \tag{19}$$

β_{k_1} peut être interprété comme le log odds de la classe k par rapport à la classe 1, β_k^t le log odds de transition par rapport aux classes et β_{k_l} le log odds d'un individu de la classe k et qui répond l à l'item j.

3 Estimation des paramètres par la méthode de L'algorithme EM (Expectation-Maximisation)

L'algorithme EM (Expectation-Maximisation) est une méthode de calcul du maximum de vraisemblance et des estimations des paramètres dans des situations où certaines données sont manquantes. L'algorithme EM donne des estimations de paramètres qui maximisent la vraisemblance des données observées à l'aide de calculs qui impliquent la probabilité de l'ensemble des données. C'est une méthode itérative pour la maximisation de la vraisemblance dans le cas de données manquantes.
Cet algorithme a été proposé par Dempster et Rubin (1977) avec de nombreux exemples. Orchard et Woodbury (1972) ont développé une idée similaire qu'ils ont appelé le principe de l'information manquante appliqué à l'algorithme EM. D'autres auteurs comme Little et Rubin (1983) ont développé des applications en utilisant cet algorithme pour estimer des paramètres. Par contre Becker (1997) a aussi utilisé cet algorithme en supposant que les variables latentes sont manquantes pour estimer les paramètres dans le cas du modèle à variables latentes.
Le principe de l'algorithme EM est suivant :
-La première étape c'est le calcul de l'espérance de log-vraisemblance (c'est à-dire en supposant les classes non observées connues).
-La deuxième étape on estime le maximum de vraisemblance des paramètres en maximisant la vraisemblance trouvé à la première étape.
Nous allons conserver les mêmes définitions c'est-à-dire $i \in \{1, 2, ..., n\}$, $j \in \{1, 2, ..., J\}$, $k \in \{1, 2, ..., K\}$ et $l \in \{1, 2, ..., L_j\}$.
Les composantes du vecteur sont : $Y_j^{(i)}(t) = (Y_{j,1}^{(i)}(t), Y_{j,2}^{(i)}(t), ..., Y_{j,L_j}^{(i)}(t))$ pour tout, $t \in \{1, 2, ..., T\}$.
La réponse de l'individu $i \in \{1, 2, ..., n\}$ à la question $j \in \{1, 2, ..., J\}$ à l'instant $t \in \{1, 2, ..., T\}$ est notée par $y_j^{(i)}(t) = (y_{j,1}^{(i)}(t), y_{j,2}^{(i)}(t), ..., y_{j,L_j}^{(i)}(t)) \in \{0, 1\}^{L_j}$.
$y_{j,l}^{(i)}(t)$ la réponse l d'un individu i à l'item j y_j prend les valeurs l, l est appelé modalité de réponse, L_j le nombre de modalités de réponses de la variable j.
La réponse de l'individu i à la question j est notée y_j^i.
Les individus sont répartis en K classes. Notons $\theta = (\theta(1), ..., \theta(K))$ le vecteur des k classes latentes, $\theta^i = (\theta^i(1), ..., \theta^i(K))$ et α_k, le proportion d'individus dans la classe k.
On suppose que les θ^i suivent une distribution telle que $f(\theta^i) = \prod_{k=1}^{K} \alpha_k^{\theta^i(1)}$ avec

$\sum_{k=1}^{K} \alpha_k = 1, \alpha_k \geq 0.$

La distribution conditionnelle de y^i par rapport à θ^i est notée $f(y^i/\theta^i)$.

$$f(y^i/\theta^i) = \prod_{k=1}^{K} (P(y^i = l/\theta^i = k))^{\theta^i(k)}. \tag{20}$$

L'indépendance conditionnelle des variables réponses par rapport aux classes latentes entraîne que $P(y^i = l/\theta^i = k) = \prod_{j=1}^{J} \prod_{l=1}^{L} (P(y^i_{jl} = 1/\theta^i = k))^{y^i_{jl}}$

En posant $\lambda_{j,l}(k) = P(y^i_{jl} = 1/\theta^i = k)$, la distribution conditionnelle de y^i par rapport à θ^i devient :

$$f(y_i/\theta_i) = \prod_{k=1}^{K} \left(\prod_{j=1}^{J} \prod_{l=1}^{L_j} (\lambda_{j,l}(k))^{y^i_{jl}} \right)^{\theta^i(k)} \tag{21}$$

La distribution jointe de y et θ est :
$f(y, \theta) = P(y, \theta) = f(y/\theta) f(\theta)$
On a dans ce cas :

$$\begin{aligned}
f(y^i, \theta^i) &= f(y^i/\theta^i) f(\theta^i) \\
&= \prod_{k=1}^{K} \left(\prod_{j=1}^{J} \prod_{l=1}^{L_j} (\lambda_{j,l}(k))^{y^i_{jl}} \right)^{\theta^i(k)} \cdot \prod_{k=1}^{K} \alpha_k^{\theta^i(k)} \\
&= \prod_{k=1}^{K} \left(\prod_{j=1}^{J} \prod_{l=1}^{L_j} (\lambda_{j,l}(k))^{y^i_{jl}} \right)^{\theta^i(k)} \alpha_k^{\theta^i(k)} \\
&= \prod_{k=1}^{K} \left(\prod_{j=1}^{J} \prod_{l=1}^{L_j} ((\lambda_{j,l}(k))^{y^i_{jl}} \alpha_k)^{\theta^i(k)} \right)
\end{aligned}$$

$$i = 1, ..., n, j = 1, ..., J, l = 1, ..., L, k = 1, ..., K$$

Finalement la distribution jointe de y et θ est donnée par :

$$f(y, \theta) = \prod_{i=1}^{n} \prod_{k=1}^{K} \left(\prod_{j=1}^{J} \prod_{l=1}^{L_j} ((\lambda_{j,l}(k))^{y^i_{jl}} \alpha_k)^{\theta^i(k)} \right) \tag{22}$$

Le log vraisemblance est $L_{EM}f(y,\theta)$:

$$\begin{aligned} L_{EM}f(y,\theta) &= \log f(y,\theta) \\ &= \log\left(\prod_{i=1}^{n}\prod_{k=1}^{K}\left(\prod_{j=1}^{J}\prod_{l=1}^{L_j}((\lambda_{j,l}(k))^{y^i_{jl}}\alpha_k)^{\theta^i(k)}\right)\right) \\ &= \sum_{i=1}^{n}\sum_{k=1}^{K}\theta^i(k)\log\left(\alpha_k\prod_{j=1}^{J}\prod_{l=1}^{L}(\lambda_{j,l}(k))^{y^i_{jl}}\right) \end{aligned} \tag{23}$$

Comme $\theta = (\theta^1, ..., \theta^n)$ n'est pas observé, l'algorithme procède de manière itérative :
Par définition on note $f(\theta^i(k)/y^i)$ la probabilité qu'un individu appartienne à la classe k sachant qu'il a répondu à l'item j. Alors on a :

$$\begin{aligned} f(\theta^i(k)/y^i) &= \frac{f(y^i,\theta^i(k))}{f(y^i)} \\ &= \frac{f(y^i/\theta^i(k))f(\theta^i(k))}{\sum_{t=1}^{K} f(y^i/\theta^i(k))f(\theta^i(k))}, i = 1, ..., n, t = 1, ..., K, k = 1, ..., K \end{aligned}$$

En utilisant le principe de Bayes on a :

$$f(\theta^i(k)/y^i) = \frac{f(y^i/\theta^i(k))f(\theta^i(k))}{\sum_{s=1}^{K} f(y^i/\theta^i(s))f(\theta^i(s))} = \frac{f(y^i,\theta^i(k))}{\sum_{s=1}^{K} f(y^i,\theta^i(s))} \tag{24}$$

$$i = 1, ..., n, s = 1, ..., K, k = 1, ..., K.$$

En posant $\theta^i(k)$ le nombre d'individus qui se trouve dans la classe k et $r^i_{jl}(k) = y^i_{jl}\theta^i(k)$ le nombre d'individus de la classe k et qui donne la réponse l à l'item j, la distribution jointe de y et θ est :

$$f(y,\theta) = \prod_{i=1}^{n}\prod_{k=1}^{K}\left(\prod_{j=1}^{J}\prod_{l=1}^{L_j}((\lambda_{j,l}(k)^{r^i_{jl}(k)})\alpha_k^{\theta^i(k)}\right) \tag{25}$$

Le log vraisemblance est $L_{EM}f(y,\theta)$:

$$
\begin{aligned}
L_{EM} f(y,\theta) &= \log f(y,\theta) \\
&= \log\left(\prod_{i=1}^{n}\prod_{k=1}^{K}\left(\prod_{j=1}^{J}\prod_{l=1}^{L_j}((\lambda_{j,l}(k)^{r^i_{jl}(k)})\alpha_k^{\theta^i(k)}\right)\right) \\
&= \sum_{i=1}^{n}\sum_{k=1}^{K}\theta^i(k)\log\left[\alpha_k\prod_{j=1}^{J}\prod_{l=1}^{L_j}(\lambda_{j,l}(k)^{r^i_{jl}(k)}\right] \\
&= \sum_{i=1}^{n}\sum_{k=1}^{K}\theta^i(k)\left[\log\alpha_k + \sum_{j=1}^{J}\sum_{l=1}^{L_j} r^i_{jl}(k)\log(\lambda_{j,l}(k))\right] \qquad (26)
\end{aligned}
$$

Les valeurs attendues $\hat{\theta}^s(k)$ et $\hat{r}^s_{jl}(k)$ respectivement pour $\theta(k)$ et $r_{jl}(k)$ sont obtenues par itération pour $s = 0, 1, \ldots$.
$\alpha^s_k = f(\hat{\theta}^s(k))$ la proportion d'individus dans la classe k à l'itération s.
On peut séparer l'algorithme EM en deux étapes, l'étape Estimation et l'étape Maximisation.
L'étape Estimation :

$$
\begin{aligned}
\hat{\theta}^s(k) &= E(\theta(k)/y^i) \\
&= \sum_{i=1}^{n}\frac{\prod_{j=1}^{J}\prod_{l=1}^{L_j}(\lambda_{j,l}(k))^{y^i_{jl}}\alpha^s_k}{\sum_{t=1}^{K}\left(\prod_{j=1}^{J}\prod_{l=1}^{L_j}(\lambda_{j,l}(t))^{y^i_{jl}}\right)\alpha^s_t} \\
&= \sum_{i=1}^{n}\frac{\alpha^s_k\prod_{j=1}^{J}\prod_{l=1}^{L_j}(\lambda_{j,l}(k))^{y^i_{jl}}}{\sum_{t=1}^{K}\alpha^s_t\left(\prod_{j=1}^{J}\prod_{l=1}^{L_j}((\lambda_{j,l}(t))^{y^i_{jl}}\right)} \qquad (27)
\end{aligned}
$$

$i = 1, \ldots, n, k = 1, \ldots, K$
De la même manière $\hat{r}^s_{jl}(k)$ est définie comme suite :

$$\hat{r}^s_{jl}(k) = E(\hat{r}_{jl}(k)/y) = E(y^i_{jl}\theta(k)/y) = \sum_{i=1}^{n} y^i_{jl} f(\theta(k)/y^i)$$

$i = 1, \ldots, n, k = 1, \ldots, K; j = 1, \ldots, J.$

On obtient :

$$\hat{r}_{jl}^{s}(k) = \sum_{i=1}^{n} \frac{y_{jl}^{i} f(y^{i}/\theta(k)) f(\hat{\theta}(k)^{s})}{\sum_{t=1}^{k} f(y^{i}/x(t)) f(\hat{\theta}(t)^{s})}$$

$$= \sum_{i=1}^{n} \frac{y_{jl}^{i} \alpha_{k}^{s} \prod_{j=1}^{J} \prod_{l=1}^{L_j} (\lambda_{j,l}(k))^{y_{jl}^{i}}}{\sum_{t=1}^{K} \alpha_{t}^{s} \left(\prod_{j=1}^{J} \prod_{l=1}^{L_j} (\lambda_{j,l}(t))^{y_{jl}^{i}} \right)}$$

$i = 1, ..., n, k = 1, ..., K, j = 1, ..., J.$

L'étape Maximisation :

Cette étape consiste à substituer dans le log vraisemblance $\theta(k)$ et $r_{jl}(k)$ par les valeurs attendues dans l'étape E. Les valeurs α_k et $\lambda_{j,l}(k)$ sont trouvées en maximisant le log-vraisemblance.

En utilisant $\hat{\theta}^{s}(k)$ et $\hat{r}_{jl}^{s}(k)$ trouvés dans l'étape E, le log-vraisemblance devient :

$$L_{EM} f(y, \theta) = \sum_{i=1}^{n} \sum_{k=1}^{K} \hat{\theta}^{s(i)}(k) \left[\log \alpha_k + \sum_{j=1}^{J} \sum_{l=1}^{L_j} \hat{r}_{jl}^{s}(k) \log(\lambda_{j,l}(k)) \right] \tag{28}$$

Pour trouver la valeur d'un paramètre qui maximiser le log vraisemblance il faut dériver le log-vraisemblance par rapport à ce paramètre. Ce rapport de dérivée partielle est égale à zéro afin de trouver les paramètres.

Par définition le paramètre α_k^{s+1} obtenu à l'itération s+1 est :

$$\alpha_k^{s+1} = \frac{\sum_{i=1}^{n} \hat{\theta}^{s(i)}(k)}{n}, i = 1, ..., n, k = 1, ...K, s = 0, 1, ... \tag{29}$$

Avec α_k^s le paramètre obtenue à l'itération s. Pour ce qui est du paramètre $\lambda_{j,l}^{s+1}(k)$, il est obtenu es résolvant cette équation :

$$\frac{\partial \log f(y, \theta)}{\partial \lambda_{j,l}(k)} = 0, i = 1, ..., n, k = 1, ...K \tag{30}$$

Nous allons pas donner les détails pour la résolution de cette équation. Nous pouvons dire qu'à iteration s+1 $\lambda_{j,l}^{s+1}(k)$ est la solution de l'équation $\frac{\partial \log f(y,\theta)}{\partial \lambda_{j,l}(k)} = 0$. c'est-à-dire la solution de

$$\sum_{i=1}^{n} \sum_{k=1}^{K} \hat{\theta}^{s(i)}(k) \sum_{j=1}^{J} \sum_{l=1}^{L_j} \frac{\hat{r}_{jl}^{s}(k)}{\lambda_{j,l}(k)} = 0 \tag{31}$$

$$i = 1, ..., n, k = 1, ...K.$$

Nous allons élargir cette méthode d'estimation dans le cas où les données sont longitudinales. Dans ce cas les paramètres à estimer deviennes $\lambda_{j,l}(t|k)$ et α_{k_t} et $\gamma_{(k_t,k_{t-1})}$. En posant $\theta^i(k|t)$ la classe k d'un individu i à un instant t donné et $r_{jl}(k|t) = y^i_{jl}(t)\theta^i_{k|t}$ le nombre d'individus de la classe k et qui donne la réponse l à l'item j à un instant t.

Les valeurs attendues $\hat{\theta}^{s(i)}(k|t)$ et $\hat{r}^s_{jl}(k|t)$ respectivement pour $\hat{\theta}^i(k|t)$ et $\hat{r}^s_{jl}(k|t)$ sont obtenues par itération pour $s = 0, 1, ...$.
$\alpha^s_{k_t} = f(\hat{\theta}^{s(i)}(k|t)$ la proportion d'individus dans la classe k à un instant t et à l'itération s.
On considère $y^i(t)$ le vecteur réponse d'un individu à un instant t et $\theta^i(k|t)$ la classe latente d'un individu i à un instant t. En utilisant le principe de Bayes à un instant t, on a :

$$f(\theta^i(k|t)/y^i(t)) = \frac{f(y^i(t)/\theta^i(k|t))f(\theta^i(k|t))}{\sum_{z=1}^{K} f(y^i(t)/\theta^i(z|t))f(\theta^i(z|t))} = \frac{f(y^i(t), \theta^i(k|t))}{\sum_{z=1}^{K} f(y^i(t), \theta^i(z|t))} \quad (32)$$

On suppose que les $\theta^i(t))$ suivent une distribution$f(\theta^i(t))) = \prod_{k=1}^{K} \alpha_{k_t}^{\theta^i(k|t)}$ avec $\sum_{k=1}^{K} \alpha_{k_t} = 1, \alpha_{k_t} \geq 0,$
La distribution conditionnelle de $y^i(t)$ par rapport à $\theta^i(t)$ est notée $f(y^i(t)/\theta^i(t))$.

$$f(y^i/\theta^i) = \prod_{k=1}^{K}\prod_{t=1}^{T}(P(y^i(t) = l/\theta^i(t) = k_t))^{\theta^i(k|t)}, i = 1, ..., n, l = 1, ..., L, k = 1, ..., K \quad (33)$$

L'indépendance conditionnelle des variables réponses par rapport aux classes latentes entraîne que $p(y^i = l/\theta^i = k) = \prod_{t=1}^{T}\prod_{j=1}^{J}\prod_{l=1}^{L}(P(y^i_{j,l}(t) = 1/\theta^i(k|t) = k_t))^{y^i_{j,l}(t)}$
En posant $\lambda_{j,l}(t|k) = P(y^i_{j,l}(t) = 1/\theta^i(k|t) = k_t)$, la distribution conditionnelle de y^i par rapport à θ^i devient :

$$f(y^i/\theta^i) = \prod_{k=1}^{K}\prod_{t=1}^{T}\left(\prod_{j=1}^{J}\prod_{l=1}^{L_j}(\lambda_{j,l}(t|k))^{y^i_{j,l}(t)}\right)^{\theta^i(k|t)} \quad (34)$$

L'étape Estimation :

$$\begin{aligned}\hat{\theta}^s(k|t) &= E(\theta(k|t)/y^i(t)) \\ &= \sum_{i=1}^{n} \frac{\prod_{j=1}^{J}\prod_{l=1}^{L_j}(\lambda_{j,l}(t|k))^{y^i_{j,l}(t)}\alpha^s_{kt}}{\sum_{z=1}^{K}\left(\prod_{j=1}^{J}\prod_{l=1}^{L_j}(\lambda_{j,l}(t|z))^{y^i_{j,l}(t)}\right)\alpha^s_{zt}} \\ &= \sum_{i=1}^{n} \frac{\alpha^s_{kt}\prod_{j=1}^{J}\prod_{l=1}^{L_j}(\lambda_{j,l}(t|k))^{y^i_{j,l}(t)}}{\sum_{z=1}^{K}\left(\alpha^s_{zt}\prod_{j=1}^{J}\prod_{l=1}^{L_j}(\lambda_{j,l}(t|z))^{y^i_{j,l}(t)}\right)} \end{aligned} \quad (35)$$

$i = 1, ..., n, k = 1, ..., K$

De la même manière $\hat{r}^s_{jl}(k|t)$ est définie comme suite :

$$\hat{r}^s_{jl}(k|t) = E(r_{jl}(k|t)/y^i(t)) = E(y^i_{j,l}(t)\theta(k|t)/y^i(t)) = \sum_{i=1}^{n} y^i_{j,l}(t) f(\theta(k|t)/y^i(t))$$

$i = 1, ..., n, k = 1, ..., K, j = 1, ..., J$

On obtient :

$$\begin{aligned}\hat{r}^s_{jl}(k|t) &= \sum_{i=1}^{n} \frac{y^i_{j,l}(t) f(y_{it}/\theta(k|t) f(\hat{\theta}^s_{kt})}{\sum_{z=1}^{K} f(y^i(t)/\theta(z|t) f(\hat{\theta}^s_{zt})} \\ &= \sum_{i=1}^{n} \frac{y^i_{j,l}(t)\alpha^s_{kt}\prod_{j=1}^{J}\prod_{l=1}^{L_j}(\lambda_{j,l}(t|k))^{y^i_{j,l}(t)}}{\sum_{z=1}^{K}\alpha^s_{zt}\left(\prod_{j=1}^{J}\prod_{l=1}^{L_j}(\lambda_{j,l}(t|z))^{y^i_{j,l}(t)}\right)} \end{aligned}$$

$i = 1, ..., n, k = 1, ..., K, j = 1, ..., J.$

Ces paramètres ont été définis dans le modèle conditionnel. Le paramètre α^{s+1}_{kt} obtenu par itération est :

$$\alpha^{s+1}_{kt} = \frac{\sum_{i=1}^{n} \hat{\theta}^{s(i)}(k|t)}{n}, i = 1, ..., n, k = 1, ...K, s = 0, 1, ... \quad (36)$$

et $\lambda_{j,l}(t|k)$ vérifie aussi

$$\sum_{i=1}^{n}\sum_{k=1}^{K}\hat{\theta}^{s(i)}(k|t)\sum_{j=1}^{J}\sum_{l=1}^{L_j}\frac{\hat{r}^{s}_{jl}(k|t)}{\lambda_{j,l}(t|k)}=0 \tag{37}$$

$$i=1,...,n, k=1,...K.$$

Puis que le modèle transitoire étudie aussi le changement de classe en fonction du temps, nous allons également étudier l'estimation du paramètre de transition. Considérons k_t la classe latente d'un individu i à l'instant t avec $k_t = 1, 2, ..., K$ on définira $\gamma_{(k_t,k_{t-1})}$ l'équation :

$$\gamma_{(k_t,k_{t-1})}=\frac{\exp(\beta_k(t))}{\sum_{k=1}^{K}\exp(\beta_k(t))}, k=1,...,K, t=1,...,T. \tag{38}$$

Cela revient à estimer les paramètres β_k^t qui est obtenu par :

$$\log\left(\frac{p(\hat{\theta}^{s(i)}(k)=k_t/\hat{\theta}^{s(i)}(k)=k_{t-1})}{p(\hat{\theta}^{s(i)}(k)=1/\hat{\theta}^{s(i)}(k)=k_{t-1})}\right)=\beta_k(t), \tag{39}$$

Il existe aussi d'autres méthodes d'estimation des paramètres dans le cas du modèle à classes latentes pour des données longitudinales exemple celle proposée par Baum (1970).
La méthode de Baum utilise les indépendances conditionnelles implicites par le modèle afin de contourner le problème de calcul à cause des valeurs élevées de T. Cette méthode a connu une extension par Paas et al (2007) afin de faire face à de multiples observations.
L'algorithme EM est souvent utilisé dans d'autres circonstances. Par exemple, il a été utilisé par Dempster (1977) et Chen (1981) dans le cas ou les variables réponses sont continues.

4 Application sur des jeux de données simulées

4.1 Outils et méthodes

Les procédure LCA et LTA sont développée depuis la version 9.1 de SAS par Lanza et al. (2007). La procedure LTA est adaptée aux modèles de transition latentes . Dans le cas de notre application l'utilisation de la procédure LTA nous

permettra de modéliser trois types de paramètres : le premier est la probabilité d'appartenance d'un individu (patient) à une classe donnée à l'état initial, le deuxième est la probabilité de transition entre les classes au cours du temps et la troisième est la probabilité conditionnelle qu'un individu d'une classe latente réponde à un item.

4.2 Utilisation de la procédure LTA à des données simulées

L'objectif de cette partie est d'évaluer l'approche de la procédure LTA qui utilise l'algorithme EM. Pour se faire nous comparons la qualité des estimations fournies en utilisant l'écart quadratique moyen entre les paramètres estimés et la moyenne. Sur la table 1 nous définissons par $\alpha_{k_1}, \alpha_{k_2}, \alpha_{k_3}$et α_{k_4} les probabilités qu'un individu soit dans les classes latentes k_1, k_2, k_3, et k_4. $\bar{\alpha}_k$ définit la moyenne déterminée à l'issue de la simulation.
Nous remarquons que l'écart quadratique moyen entre les paramètres est très insignifiant voir la table 1. Le macro et la procédure LTA de SAS sont présentés en Annexe.

TABLE 1 – **Distribution des classes latentes à chaque instant**

simulation 1	$\frac{(\alpha_{k_1}-\bar{\alpha}_k)^2}{1000}$	$\frac{(\alpha_{k_2}-\bar{\alpha}_k)^2}{1000}$	$\frac{(\alpha_{k_3}-\bar{\alpha}_k)^2}{1000}$	$\frac{(\alpha_{k_4}-\bar{\alpha}_k)^2}{1000}$
temps 1	$1,90.10^{-05}$	$5,06.10^{-06}$	$8,43.10^{-05}$	2.10^{-04}
temps 2	$1,17.10^{-05}$	$2,05.10^{-05}$	10^{-04}	10^{-05}
temps 3	$1,8.10^{-05}$	$7,65.10^{-06}$	10^{-04}	10^{-05}
simulation 2				
temps 1	$1,15.10^{-05}$	$4,39.10^{-05}$	$6,90.10^{-05}$	$2,94.10^{-05}$
temps 2	$8,91.10^{-05}$	$3,22.10^{-05}$	10^{-04}	$6,51.10^{-06}$
temps 3	$8,49.10^{-05}$	$4,42.10^{-05}$	2.10^{-04}	$1,45.10^{-06}$
simulation 3				
temps 1	10^{-04}	$6,30.10^{-06}$	2.10^{-04}	$1,66.10^{-05}$
temps 2	10^{-04}	10^{-06}	10^{-04}	$3,86.10^{-06}$
temps 3	10^{-04}	$1,2310^{-06}$	2.10^{-04}	$4,83.10^{-06}$

4.3 Description des données réelles

Nous venons d'appliquer la procédure LTA de SAS pour des données simulées. L'écart quadratique moyen entre les paramètres estimés et la moyenne reste insignifiant. Nous pouvons utiliser ce modèle pour des données réelles en utilisant la

procédure sur SAS.
Les données utilisées dans cette étude proviennent d'une enquête qualité de la vie. L'échantillon est composé de 339 individus. Ces individus ont répondu à sept questions. Les modalités de réponses varient d'une question à une autre. Les questions posées sont détaillés dans la table 2.

TABLE 2 – **Description des questions posées**

Variables	Libellé.
Question 1	Efforts physiques modérés (déplacer une table, passer aspirateur) 1='oui, beaucoup gêné' 2='oui, un peu gêné' 3='Non, pas du tout gêné';.
Question 2	Au cours des quatre dernières semaines, en fonction de votre état physique, vous faites moins de choses que ce que vous auriez souhaité 1='Oui, un peu gêné' 2='Non, pas du tout gêné'.
Question 3	Au cours des quatre dernières semaines, en fonction de votre état émotionnel, fait moins de choses que ce que vous auriez souhaité vous faites moins de choses que ce que vous auriez souhaité 1='Oui' 2='Non'.
Question 4	Au cours des quatre dernières semaines, en fonction de votre état émotionnel, avez vous eu des difficultés à faire ce que vous auriez à faire avec autant de soin 1='Oui' 2='Non'.
Question 5	Au cours des quatre dernières semaines, êtes-vous senti calme et détendu ? 1='En permanence' 2='Très souvent' 3='Souvent' 4= 'quelquefois' 5='rarement' 6= 'jamais'.
Question 6	Au cours des quatre dernières semaines, êtes-vous senti débordant d'énergie ? 1='En permanence' 2='Très souvent' 3='Souvent' 4= 'quelquefois' 5='rarement' 6= 'jamais'.
Question 7	Au cours des quatre dernières semaines, êtes-vous senti triste d'énergie ? 1='En permanence' 2='Très souvent' 3='Souvent' 4= 'quelquefois' 5='rarement' 6= 'jamais'.

4.4 Interprétation des résultats

Comme nous l'avions mentionné, avec la procédure LTA de SAS (Lanza et al. (2007)), nous obtenons trois types de paramètres : la probabilité qu'un individu soit dans une classe latente à un instant t donné, la probabilité de transition et la probabilité qu'un individu d'une classe latente réponde à une modalité d'un item. Dans la partie application, nous avons choisi de fixer arbitrairement le nombre de classes latentes à 4.

4.4.1 Les probabilités conditionnelles

Nous allons représenter les probabilités qu'un individu d'une classe latente réponde à une modalité d'un item. Dans le cas de notre étude les individus ont répondu à 7 items dont les modalités varient en fonction des items. Les individus sont répartis en quatre classes latentes numérotées de 1 à 4. Les résultats sont présentés sur la table 3. Cette table permet d'interpréter les classes latente en fonction du questionnaire. C'est ce que nous allons détailler ici.

-**Question 1**(Efforts physiques modérés (déplacer une table, passer aspirateur) Au vu des résultats sur la table 3, nous pouvons dire que plus de la moitié des individus interrogés ne sont pas du tout gêné quand ils exercent un effort physique modéré exemple déplacer une table. C'est ainsi, 54.48% des individus de la classe 1, 89.56% de la classe 2, 52.63% de la classe 3 et 94.32% de la classe 4 ne se sentent pas du tout gênés quand ils font un effort physique. Par contre 12.97% des individus de la classe 1, 1.93% de la classe 2, 1.35% de la classe 3 et 1.12% de la classe 4 se sentent gênés quand ils font un effort physique.
-**Question 2**(Au cours des quatre dernières semaines, en fonction de votre état physique, vous faites moins de choses que ce que vous auriez souhaité ?)
Nous avons 95.06% des individus de la classe latente 1, 12.37% de la classe 2, 79.62% de la classe 3 et 3.35% de la classe 4 qui ont affirmés qu'au cours des quatre dernières semaines, en fonction de leur état physique, ils font moins de choses que ce qu'ils auraient souhaité. Par contre nous constatons que 4.94% des individus de la classe 1, 87.63% de la classe 2 et 20.38% de la classe 3 et 96.65% de la classe 4 réussissent à faire des choses qu'ils auraient souhaité faire
-**Question 3**(Au cours des quatre dernières semaines, en fonction de votre état physique, avez vous du arrêter de faire certaines choses ?)
Pour la question 3, 74.17% des individus de la classe latente 1, 3% de la classe 2, 53.94% de la classe 3 et 2.93% de la classe 4, au cours des quatre dernières semaines, en fonction de leur état émotionnel, font moins de choses que ce que ils auraient souhaité. Cependant 25.83% des individus de la classe latente 1, 97% de la classe 2, 46.06% de la classe 3 et 97.07% de la classe 4 ont réussi à faire des choses qu'ils auraient souhaité faire
-**Question 4**(Au cours des quatre dernières semaines, en fonction de votre état émotionnel, avez-vous eu des difficultés à faire ce que vous auriez à faire avec autant de soin ?)
Au cours des quatre dernières semaines, en fonction de leur état émotionnel, 62.8% des individus de la classe latente 1, 17.82% de la classe 2, 84.3% de la classe 3 et 2.67% de la classe 4 ont eu des difficultés à faire ce qu'ils auraient à faire avec autant de soin.
-**Question 5**(Au cours des quatre dernières semaines, êtes-vous senti calme et dé-

tendu ?)
Sur la table 3, nous pouvons lire qu'au cours des quatre dernières semaines, 23.71% des individus de la classe latente 4 sont en permanence clame et détendu. C'est dans cette même classe que l'on retrouve aussi la majorité des personnes qui au cours des quatre dernières semaines sont très souvent clame et détendu avec un pourcentage de 55.06%. On constate également qu'il y a peu d'individus appartenant au quatre classes latentes qui au cours des quatre dernières semaines ne sont jamais calmes et détendus.
-**Question 6**(Au cours des quatre dernières semaines, êtes-vous senti débordant d'énergie ?)
Les résultats sont représentés dans la table. Au cours des quatre dernières semaines, nous constatons que la majorité des individus de la classe latente 4 se sentent très souvent (44.36%) ou souvent (25.56%) débordant d'énergie. La majorité des individus de la classe latente 3 se sentent rarement (41.49%) ou jamais (31.02%) débordant d'énergie.
-**Question 7**(Au cours des quatre dernières semaines, êtes-vous senti triste et abattu ?)
Pour cette question, nous constatons que 96% des individus de classe latente 4 se sentent rarement ou jamais triste d'énergie. Cela s'explique aussi par le fait que ces individus sont débordant d'énergie d'après les résultats de la question 5. C'est la même remarque pour les individus supposés appartenir à classe latente 3 qui en majorité se sentent très souvent ou souvent tristes.

4.4.2 Les probabilités d'appartenir à une classe à un instant t

Sur la table 4, nous avons les probabilités qu'un individu appartienne à une classe latente à temps t donné. Nous rappelons que les individus sont repartis en quatre classes latentes et chaque individu ne peut être que dans une seule classe. Á l'instant t=1, la classe la plus représentative est la classe latente 2 avec 28.75% des individus. La classe latente 1 contient 16.9% des individus, c'est la classe la moins représentative. Á l'instant t= 2, nous constatons que 38.26% des individus sont supposés appartenir à classe latente 3 qui contient plus d'individus. La classe latente 1 est la classe la plus faiblement représentée avec 13.24% des individus. Au temps t=3, c'est la classe latente 2 qui contient le plus d'individus et la classe latente 1 reste la classe la moins représentée avec 12.26% des individus. Á l'instant t=4 c'est toujours la classe latente 2 qui contient plus d'individus (39.04%) et classe latente 1 demeure la moins représentative avec 9.44% des individus. Nous avons les mêmes constats à l'instant t=5. La remarque en est que le nombre d'individus appartenant à une classe latente varie en fonction du temps. Exemple la classe latente 1 a perdu environs 3% des individus entre les temps t=1 et t=2. C'est la

TABLE 3 – **Probabilités conditionnelles**

	1	2	3	4
items 1				
$p(x_{11} = 1/k)$	0.1227	0.0193	0.1135	0.0112
$p(x_{11} = 2/k)$	0.3326	0.0848	0.3602	0.0456
$p(x_{11} = 3/k)$	0.5448	0.8959	0.5263	0.9432
items 2				
$p(x_{21} = 1/k)$	0.9506	0.1237	0.7962	0.0335
$p(x_{21} = 2/k)$	0.0494	0.8763	0.2038	0.9665
items 3				
$p(x_{31} = 1/k)$	0.7417	0.0300	0.5394	0.0293
$p(x_{31} = 2/k)$	0.2583	0.9700	0.4606	0.9707
items 4				
$p(x_{41} = 1/k)$	0.6280	0.1782	0.8430	0.0267
$p(x_{41} = 2/k)$	0.3720	0.8218	0.1570	0.9733
items 5				
$p(x_{51} = 1/k)$	0.0526	0.0011	0.0228	0.2240
$p(x_{51} = 2/k)$	0.2240	0.1184	0.0000	0.5011
$p(x_{51} = 3/k)$	0.3552	0.4649	0.0140	0.2128
$p(x_{51} = 4/k)$	0.3379	0.3132	0.3969	0.0346
$p(x_{51} = 5/k)$	0.0146	0.0995	0.4371	0.0039
$p(x_{51} = 6/k)$	0.0157	0.0029	0.1292	0.0236
items 6				
$p(x_{61} = 1/k)$	0.0154	0.0000	0.0089	0.1300
$p(x_{61} = 2/k)$	0.0698	0.0527	0.0113	0.4203
$p(x_{61} = 3/k)$	0.2649	0.2749	0.0074	0.2750
$p(x_{61} = 4/k)$	0.3689	0.4486	0.2135	0.1229
$p(x_{61} = 5/k)$	0.2570	0.1986	0.4246	0.0303
$p(x_{61} = 6/k)$	0.0240	0.0251	0.3343	0.0214
items 7				
$p(x_{71} = 1/k)$	0.0000	0.0000	0.0942	0.0027
$p(x_{71} = 2/k)$	0.0453	0.0127	0.2209	0.0039
$p(x_{71} = 3/k)$	0.0413	0.0487	0.3484	0.0035
$p(x_{71} = 4/k)$	0.2812	0.4170	0.2721	0.0287
$p(x_{71} = 5/k)$	0.3659	0.4126	0.0591	0.2818
$p(x_{71} = 6/k)$	0.2664	0.1090	0.0052	0.6793

même remarque pour la classe latente 2 qui a vu le nombre d'individus diminuer de 5% entre les instants t=4 et t=5. Des individus ont changé de classe latente entre deux périodes consécutives. Ce qui nous amène à expliquer les probabilités de transition entre les classes latentes (voir la table 5).

TABLE 4 – **Distribution des classes latentes à chaque instant**

	1	2	3	4
temps 1	0.1698	0.2875	0.2736	0.2692
temps 2	0.1324	0.2843	0.2008	0.3826
temps 3	0.1226	0.3580	0.1695	0.3499
temps 4	0.0944	0.3904	0.1927	0.3225
temps 5	0.1048	0.3385	0.2107	0.3460

4.4.3 Les probabilités de transition

Durant les différentes étapes de l'enquête (les instants t=1, t=2, t=3, t=4 et t=5), les individus sont répartis en quatre classes latentes. Dans chaque étape le nombre de classes reste le même. L'appartenance d'un individu à une classe latente peut varier entre deux instants consécutifs. Ce qui nous ramène à présenter les probabilités de transition entre deux classes latentes (table 5).
Les probabilités de transition entre les instants t=1 et t=2 : entre les temps t=1 et temps t=2, 37.33% des individus de la classe latente 1 ne vont pas changer de classe, ils resterons toujours dans la classe latente 1. Cependant d'autres individus vont changer de classe latente. 46.45% des individus vont passer de la classe latente 1 à la classe latente 4, et 8.68% à la classe 3. Toujours entre les instants t=1 et t=2, la classe latente 4 semble la plus stable car 77.48% de ces individus ne vont pas changer de classe, ils resteront dans la classe 4.
Entre les instants t=2 et t=3, la classe latente 2 demeure la plus stable car 91.81% des individus resteront dans cette même classe latente à l'instant t=2. Aucun individu de la classe latente 3 n'ira rejoindre la classe latente 1. C'est la même remarque entre les classes latentes 4 et 3. Cependant la classe latente 2 accueillera 17.37% des individus venant de la classe 1 entre les instants t=2 et t=3.
Pour ce qui est de la transition entre les instants t=3 et t=4, c'est la classe latente 3 qui semble être la classe la plus stable car 88.56% des individus de cette classe latente resterons dans la même classe. Cependant 29.39% des individus qui étaient dans la classe latente 1 sont susceptibles de se retrouver à la classe latente 2 à

TABLE 5 – **Probabilités de transition entre les classes latentes**

	1	2	3	4
Temps 1 et 2				
classe latente 1	0.3733	0.0754	0.0868	0.4645
classe latente 2	0.0000	0.5975	0.1277	0.2749
classe latente 3	0.1214	0.2990	0.5205	0.0591
classe latente 4	0.1328	0.0667	0.0257	0.7748
Temps 2 et 3				
classe latente 1	0.7565	0.1737	0.0000	0.0699
classe latente 2	0.0212	0.9181	0.0029	0.0578
classe latente 3	0.0000	0.1599	0.8401	0.0000
classe latente 4	0.0429	0.1095	0.0000	0.8476
Temps 3 et 4				
classe latente 1	0.5593	0.2939	0.0000	0.1469
classe latente 2	0.0000	0.8811	0.1189	0.0000
classe latente 3	0.0000	0.1144	0.8856	0.0000
classe latente 4	0.0740	0.0560	0.0000	0.8701
Temps 4 et 5				
classe latente 1	0.7963	0.0000	0.2034	0.0003
classe latente 2	0.0757	0.8671	0.0052	0.0520
classe latente 3	0.0000	0.0000	0.9566	0.0434
classe latente 4	0.0000	0.0000	0.0159	0.9841

l'instant t=4. Autre enseignement, aucun individu appartenant à la classe latente 1 à l'instant t=3 ne retrouvera les classes latentes 3 à l'instant t=4.
En fin la classe latente 4 restera stable entre les instants t=4 et t=5 98.41% de ses individus resterons dans cette même classe. Entre les instants t=4 et t=5, nous constatons également une stabilité pour la classe 3. Cependant mise à part les 4.34% de ses individus qui vont rejoindre la classe 4 aucun autre individu n'est susceptible de passer de la classe latente 3 vers les classes 1 et 2 entre les instants t= 4 et t=5.

5 Conclusion

Nous venons de présenter le modèle à classes latentes dans le cas où les données sont longitudinales. Nous avions fait appel au modèle du chaînes de Markov afin de pouvoir étudier la transition entre les classes latentes. C'est ainsi, on a introduit

un paramètre à savoir la probabilité de transition qui nous permet de voir comment sont répartis les individus durant le temps. Le modèle à classes latentes pour les données longitudinales requiert beaucoup de paramètres. Pour la réduction du nombre de paramètres à étudier, nous sommes limités aux paramètres :
$\alpha_{z_{i1}}$ qui est la probabilité qu'un individu appartienne à une classe à l'instant initial.
$\beta_{z_{it}/z_{i(t-1)}}$ la probabilité transition entre classes.
λ_{kjlt} la probabilité qu'un individu de classe k répond à la modalité l à l'item j à l'instant t.
Dans la partie application, nous avions appliqué ce modèle à des données issues d'une étude sur la qualité de vie. Dans cette partie, nous nous avons utilisé une récente procédure de SAS mise au point par une équipe d'universitaires américains, la procédure LTA afin de pouvoir estimer les paramètres cités ci-dessus.
Dans le cadre cette étude, nous n'avions pas tenu compte du paramètre de transition entre les modalités au cours du temps car l'objectif principal était de voir la répartition des individus au cours du temps.
Dans les perspectives, il serait nécessaire de voir comment l'appartenance d'un individu à une classe pourrait influencer la variation de la modalité de réponse au cours du temps surtout lorsque les variables réponses sont polytomiques. Dans le futur il serait possible de comparer le modèle à classes latentes pour des données longitudinale et celui de grade of membership à trajectoire pour lequel les individus peuvent appartenir à plusieurs classes latentes différentes.

6 Annexes

```
%macro ndiogou;/*création de la macro sans paramatètres pour
 simuler 3 variables 1000 individus durant
3 temps 1 dimension */
data table1;
do i= 1 to  1000;
%do t=1 %to 3;
x1_&t=RANTBL(344555,0.2208, 0.1585,0.1635);
x2_&t=RANTBL(344555,0.13, 0.1459,0.1842);
x3_&t=RANTBL(344555,0.0757,0.1199, 0.188);
%end;
output;
end;
run;
%mend;
%ndiogou;/*éxécution de la macro sans paramètres*/
%macro ndiogou;/*création de la macro sans paramatètres
```

```
 pour simuler 3 variables 1000 individus
durant 3 temps 1 dimension */
data table2;
do i= 1 to  1000;
%do t=1 %to 3;
x1_&t=RANTBL(344555,0.1939,0.1298,0.1380);
x2_&t=RANTBL(344555,0.2279,0.1592,0.1378);
x3_&t=RANTBL(344555,0.3888,0.0771,0.0728);
%end;
output;
end;
run;
%mend;
%ndiogou;/*éxécution de la macro sans paramètres*/
%macro ndiogou;/*création de la macro sans paramatètres
pour simuler 3 variables 1000 individus durant
3 temps 1 dimension */
data table3;
do i= 1 to  1000;
%do t=1 %to 3;
x1_&t=RANTBL(344555,0.2789,0.1066,0.1295);
x2_&t=RANTBL(344555,0.0262,0.0716,0.1670);
x3_&t=RANTBL(344555,0.2372,0.1624,0.1529);
%end;
output;
end;
run;
%mend;
%ndiogou;/*éxécution de la macro sans paramètres*/
/*Utilisation de la procédure lta pour estimer
les paramètres de la table1*/
proc lta DATA=table1 ;
   nstatus 4;
  ntimes 3;
  ITEMS x1_1 x2_1 x3_1
  x1_2 x2_2 x3_2
  x1_3 x2_3 x3_3 ;
  CATEGORIES 4 4 4 ;
 measurement times;
  seed 941623;
```

```
RUN;
/*Utilisation de la procédure lta pour estimer
 les paramètres de la table2*/
proc lta DATA=table2 ;
   nstatus 4;
  ntimes 3;
  ITEMS x1_1 x2_1 x3_1
  x1_2 x2_2 x3_2
  x1_3 x2_3 x3_3 ;
  CATEGORIES 4 4 4 ;
 measurement times;
  seed 941623;
  run;
/*Utilisation de la procédure lta pour estimer
les paramètres de la table3*/
RUN;proc lta DATA=table3 ;
   nstatus 4;
  ntimes 3;
  ITEMS x1_1 x2_1 x3_1
  x1_2 x2_2 x3_2
  x1_3 x2_3 x3_3 ;
  CATEGORIES 4 4 4 ;
 measurement times;
  seed 941623;
RUN;
```

Références

[1] Andersen, E. B. (1954). Latent Structure Analysis : A Survey. Scandinavian Journal of Statistics, 9, 1-12.

[2] Anderson, T.W. (1954). On estimation of parameters in latent structure analysis. Psychometrika, 19, 1-10.

[3] Bartholomew, D.J. (1999). Latent Variable Models and Factor Analysis. 2nd ed, Wiley, London.

[4] Baum, L.E. Petrie, T. Soules, G. and Weiss, N. (1970). A maximization technique occuring in the statistical analysis of probabilistic functions of Markov chains. Annals of MathematicalStatistics. Annals of Mathematical Statistics, 41, 164-171.

[5] Becker, M.P., Yang, I. and Lange, K.(1997), Em algorithms without missing data. Statistical Methods in medical Research, 6 :37–53.

[6] Chen, C.F. (1981), The em approach to the multiple indicators and multiple causes model via the estimation of the latent variable. Journal of American Statistical Association, 76 :704–708.

[7] Collins, L.M. and Wugalter, S.E. (1992). Latent class models for stage-sequential dynamic latent variables. Multivariate Behavioral Research, 27, 131-157.

[8] Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977), Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Association Society, Series B, 39 :1–38.

[9] Diggle, P.J., Liang, K.Y. and Zeger, S.L. (1994). Analysis of longitudinal data. Oxford University Press, Oxford.

[10] Gibson, W.A. (1955). An of latent Anderson's Solution for the latent structure équations. Psychometrika, 20, 69-73.

[11] Goodman, L. A. (1974) Exploratory Latent Structure Analysis Using Both Identifiable and Unidentifiable Models. Biometrika, 61, 215-231.

[12] Hamilton, J.D. and Raj B.(2002). Advances in Markov-switching models. Springer, Berlin.

[13] Lazarsfeld, P.F. and Henry N.W.(1968). Latent structure analysis. Houghton Mifflin, Boston.

[14] Lanza S.T., Lemmon D., Schafer J.L. and Collins L.M. (2007). PROC LCA and PROC LTA User's Guide Version 113 beta. University Park, PA : The Pennsylvania State University, The Methodology Center.

[15] Liang, K.Y., and Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. Biometrika, 73, 13-22.

[16] Little, R. J. A. and Rubin, D. B. (1983), Models for nonresponse in sample surveys. The American Statistician, 37 :218–220.

[17] Orchard, T. and Woodbury, M. A. (1972), Missing information principe : Theory and applications. Proceeding of 6th Berkeley Symposium on mathematical Statistics and Probability. Berkeley, CA : University of California, pages 697– 715.

[18] Paas, L.J. and Vermunt, J.K. and Bijmolt,T.H.A. (2007). Discrete time, discrete state latent Markov modelling for assessing and predicting household acquisitions of financial products. Journal of the Royal Statistical Society A. Séries A, 170, 955-974.

[19] Van de Pol, F. and Langeheine, R. (1990). Mixed Markov latent class models. Sociological Methodology, 213-247.

[20] Vermunt, J.K. and Magidson, J. (2003). Latent class models for classification. Computational Statistics, 41, 531-537.

[21] Wiggins, L.M. (1973). Panel analysis : Latent Probability Models for Attitude and Behavior Processes, Elsevier, Amsterdam.