**Université de Montréal**

# Modèle de mélange de lois multinormales appliqué à l'analyse de comportements et d'habiletés cognitives d'enfants

par

**Charles-Édouard Giguère**

Département de mathématiques et de statistique
Faculté des arts et des sciences

Mémoire présenté à la Faculté des études supérieures
en vue de l'obtention du grade de
Maître ès sciences (M.Sc.)
en Statistiques

décembre 2011

**Université de Montréal**

Faculté des études supérieures

Ce mémoire intitulé

# Modèle de mélange de lois multinormales appliqué à l'analyse de comportements et d'habiletés cognitives d'enfants

présenté par

**Charles-Édouard Giguère**

a été évalué par un jury composé des personnes suivantes :

*Jean-François Angers*

(président-rapporteur)

*Martin Bilodeau*

(directeur de recherche)

*Jean R. Séguin*

(co-directeur)

*David Haziza*

(membre du jury)

Mémoire accepté le:

*24 novembre 2011*

# SOMMAIRE

Cette étude aborde le thème de l'utilisation des modèles de mélange de lois pour analyser des données de comportements et d'habiletés cognitives mesurées à plusieurs moments au cours du développement des enfants. L'estimation des mélanges de lois multinormales en utilisant l'algorithme EM est expliquée en détail. Cet algorithme simplifie beaucoup les calculs, car il permet d'estimer les paramètres de chaque groupe séparément, permettant ainsi de modéliser plus facilement la covariance des observations à travers le temps. Ce dernier point est souvent mis de côté dans les analyses de mélanges. Cette étude porte sur les conséquences d'une mauvaise spécification de la covariance sur l'estimation du nombre de groupes formant un mélange. La conséquence principale est la surestimation du nombre de groupes, c'est-à-dire qu'on estime des groupes qui n'existent pas. En particulier, l'hypothèse d'indépendance des observations à travers le temps lorsque ces dernières étaient corrélées résultait en l'estimation de plusieurs groupes qui n'existaient pas. Cette surestimation du nombre de groupes entraîne aussi une surparamétrisation, c'est-à-dire qu'on utilise plus de paramètres qu'il n'est nécessaire pour modéliser les données. Finalement, des modèles de mélanges ont été estimés sur des données de comportements et d'habiletés cognitives. Nous avons estimé les mélanges en supposant d'abord une structure de covariance puis l'indépendance. On se rend compte que dans la plupart des cas l'ajout d'une structure de covariance a pour conséquence d'estimer moins de groupes et les résultats sont plus simples et plus clairs à interpréter.

**Mots-clés :** modèle de mélanges de lois multinormales, analyse de trajectoires, développement de l'enfant, cognition, comportement.

# SUMMARY

This study is about the use of mixture to model behavioral and cognitive data measured repeatedly across development in children. Estimation of multinormal mixture models using the EM algorithm is explained in detail. This algorithm simplifies computation of mixture models because the parameters in each group are estimated separately, allowing to model covariance across time more easily. This last point is often disregarded when estimating mixture models. This study focused on the consequences of a misspecified covariance matrix when estimating the number of groups in a mixture. The main consequence is an overestimation of the number of groups, *i.e.* we estimate groups that do not exist. In particular, the independence assumption of the observations across time when they were in fact correlated resulted in estimating many non existing groups. This overestimation of the number of groups also resulted in an overfit of the model, *i.e.* we used more parameters than necessary. Finally mixture models were fitted to behavioral and cognitive data. We fitted the data first assuming a covariance structure, then assuming independence. In most cases, the analyses conducted assuming a covariance structure ended up having fewer groups and the results were simpler and clearer to interpret.

**Keywords :** multinormal mixture model, trajectory analysis, child development, cognition, behavior.

# TABLE DES MATIÈRES

# LISTE DES FIGURES

# LISTE DES TABLEAUX

# LISTE DES SIGLES ET ABRÉVIATIONS EN FRANÇAIS

**AR1 :** autorégressive d'ordre 1

**CS :** symétrie composée

**EM :** espérance-maximisation

**IND :** indépendance

**UAR1 :** autorégressive d'ordre 1 avec variance hétérogène

**UCS :** symétrie composée avec variance hétérogène

**UIND :** indépendance avec variance hétérogène

# LISTE DES SIGLES ET ABRÉVIATIONS EN ANGLAIS

**BIC :** Bayesian information criterion

# REMERCIEMENTS

Je remercie mon directeur Martin Bilodeau d'avoir facilité mon entrée à la maîtrise et d'avoir pris intérêt dans mon projet ainsi que mon co-directeur Jean R. Séguin pour son support constant et sa flexibilité qui m'ont permis de mener ce projet à terme. Finalement, merci à ma femme Lucie pour son support et à notre fille Alycia qui illumine nos vies.

# INTRODUCTION

Les modèles de mélange de lois sont utilisés depuis le tout début du développement des statistiques. À la fin du XIX$^{e}$ siècle, Pearson (1894, 1895) utilisait cette méthode afin estimer la proportion de deux espèces de crabe. Il estimait la proportion des deux espèces en se basant sur des observations morphologiques. À cette époque, Pearson utilisait la méthode des moments pour estimer son modèle. En l'absence d'ordinateur, ses calculs étaient effectués à la main sur des grilles de calculs. L'estimation d'un seul modèle prenait énormément de temps. Cette méthode a donc été délaissée par les contemporains de Pearson étant donnée la difficulté à estimer ce type de modèle.

Presque 100 ans plus tard, Dempster et coll. (1977) ont publié un article décrivant un nouvel algorithme qui sert à estimer des modèles statistiques pour lesquels une partie des données n'est pas observée, l'algorithme EM (Espérance-Maximisation). Dempster et coll. utilisent les modèles de mélanges de lois pour illustrer l'application de l'algorithme EM. Comme cet algorithme a été utilisé dans toutes nos analyses et qu'*a priori* il semble très abstrait, nous démontrons au chapitre 1 comment il peut être appliqué à des modèles de mélanges de lois multinormales et comment il permet d'estimer les paramètres de chaque groupe séparément contrairement à d'autres algorithmes.

Au tournant du millénaire, Muthén et Shedden (1999), Muthén et Muthén (2000) et Nagin (1999) ont développé des bibliothèques permettant d'estimer des modèles de mélanges de lois appliqués à des données longitudinales. En rendant cette méthode un peu plus accessible, de nombreux chercheurs en sciences humaines ont utilisé cette méthode afin d'identifier des sous-groupes

d'individus ayant des patrons de développement similaires dans leurs échantillons. Utilisés dans ce contexte, les mélanges de lois sont souvent désignés comme des analyses de trajectoires développementales. Bauer (2007) démontre à quel point cette méthode n'a cessé de gagner en popularité depuis l'année 2000 en soulignant le nombre de citations de ces trois articles entre 2000 et 2006. Dans l'introduction de son article, Bauer parle de l'enthousiasme qui l'a amené initialement à étudier ce type de modèle. Il soulève ensuite les problèmes liés au fait que des mauvaises spécifications de modèles peuvent mener à de sérieux problèmes d'estimation du nombre de groupes. Dans la plupart des cas, il conclue que des mauvaises spécifications de modèles mènent à une surestimation du nombre de groupes.

Une des erreurs de spécification mentionnée par Bauer concerne la matrice de covariance à l'intérieur des groupes. Cette problématique est discutée au chapitre 2 à l'aide de simulations. Nous nous intéressons plus particulièrement à vérifier l'hypothèse d'indépendance en présence de données corrélées. Ce point est important car certaines bibliothèques, par exemple Proc Traj de Jones et coll. (2001), ne permettent pas de supposer autre chose que l'indépendance des observations d'un individu à travers le temps.

Finalement nous appliquerons les modèles de mélanges à des données longitudinales de comportements et de fonctions cognitives provenant de l'étude "En 2001 ... j'avais 5 ans", une étude longitudinale portant sur le développement des enfants du Québec. Nous testerons le modèle en supposant une structure de covariance entre les observations d'un individu d'abord puis en supposant l'indépendance par la suite. Ces modèles seront présentés au chapitre 3.

# Chapitre 1

# MÉLANGE DE LOIS MULTINORMALES

## 1.1. DÉFINITION D'UN MÉLANGE

Un mélange de lois est une loi statistique dont la densité est une combinaison convexe de plusieurs densités

$$f(\mathbf{y}) = \sum_{j=1}^{G} \pi_j f_j(\mathbf{y}), \tag{1.1.1}$$

où $\mathbf{y}$ représente une observation possiblement multivariée, G représente le nombre de classes ou de groupes formant le mélange, $\pi_j$ représente la proportion d'individus dans la classe j et finalement $f_j(\mathbf{y})$, la densité de $\mathbf{y}$ dans la classe j.

Les mélanges s'appliquent lorsqu'on peut scinder la population à l'étude en plusieurs classes ou sous-populations. Chaque classe possède des caractéristiques (paramètres) qui peuvent être combinées afin de former ce que l'on désigne comme un mélange de lois.

Le célèbre jeu de données sur les iris analysé par Fisher (1936) donne un bon exemple de mélange. Cet exemple est utilisé dans plusieurs livres traitant de classification. Pour 3 variétés d'iris, Setosa, Versicolor et Virginica, on a mesuré la longueur et la largeur d'un sépale et d'un pétale pour 50 spécimens donnant un total de 150 spécimens. Chacune des variétés possède ses propres caractéristiques. La figure 1.1 montre que la distribution de la longueur d'un sépale diffère selon sa localisation et sa dispersion pour chacune des variétés d'iris. Le graphique du bas montre la distribution théorique du mélange sur la

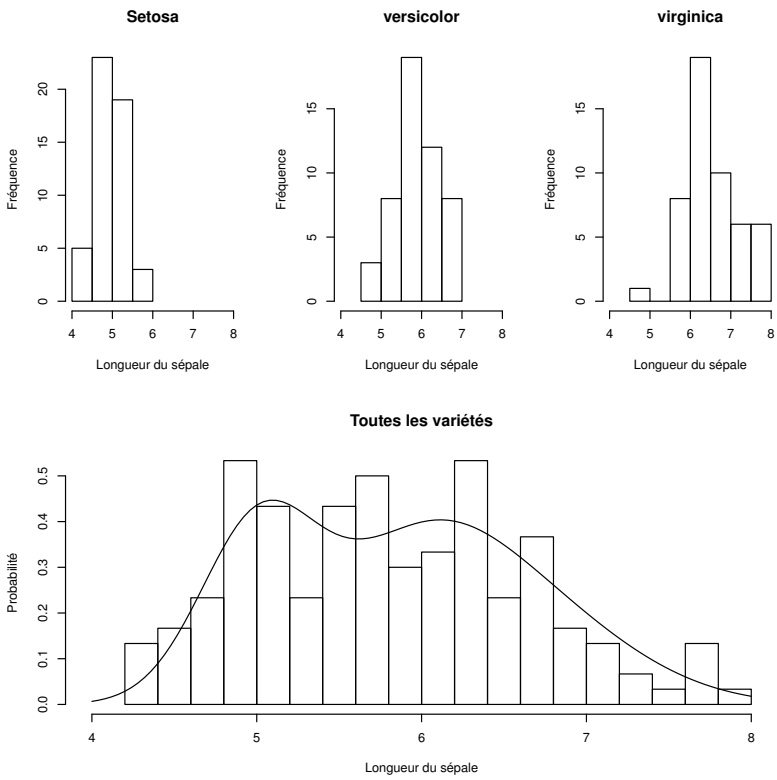



FIGURE 1.1. Histogramme de la longueur d'un sépale pour 3 variétés d'iris. Cette figure a été généré à partir du jeu de données Iris disponible dans R.

distribution empirique de tous les iris sans distinction de variétés. La courbe théorique d'un mélange de lois normales ajuste bien cette distribution. Cependant, on sait que la distribution n'est pas normale et contient plus d'un mode parce que les différentes variétés d'iris sont connues à l'avance, mais ce n'est pas toujours le cas. Nous pouvons parfois soupçonner l'existence de plus d'une classe dans une population sans toutefois connaître les variables identifiant ces classes. Dans ce cas, on parle de classes latentes ou de classes non observées. On distingue donc les classes latentes, qui ne sont pas observées, des classes pour lesquelles nous avons de l'information, comme l'espèce dans le cas des iris. Ce mémoire traite plus spécifiquement des classes latentes.

Jusqu'à présent la distribution dans chaque classe de l'équation (1.1.1) n'a pas été spécifiée. Dans ce mémoire, nous supposons que le mélange est formé

de lois multinormales. La variable $\mathbf{y} = (y_1, \ldots, y_p)'$ à l'équation (1.1.1) représente un vecteur de p variables. Les variables représentent un phénomène mesuré plusieurs fois dans le temps. Pour cette raison, dans les modèles de mélange appliqués à cette fin on parle souvent d'analyse de trajectoires. Sans perte de généralité, la densité dans la classe j est définie comme suit

$$f_j(\mathbf{y}) = (2\pi)^{-\frac{p}{2}} |\mathbf{\Sigma}_j|^{-\frac{1}{2}} exp\left[-\frac{1}{2}(\mathbf{y} - \boldsymbol{\mu}_j)' \mathbf{\Sigma}_j^{-1} (\mathbf{y} - \boldsymbol{\mu}_j)\right], \tag{1.1.2}$$

où p représente la dimension du vecteur $\mathbf{y}$, $\boldsymbol{\mu}_j$ et $\mathbf{\Sigma}_j$ représentent, respectivement, le vecteur de moyenne et la matrice de covariance spécifique au groupe j.

## 1.2. ESTIMATION DES PARAMÈTRES

Lorsque les groupes d'un mélange sont connus à l'avance, il est relativement facile d'en obtenir les paramètres. Il est toutefois plus difficile d'estimer ces mêmes paramètres lorsque les groupes ne sont pas connus. Il est possible de trouver le vecteur de paramètres qui maximisera la fonction de vraisemblance, mais cette méthode de calcul engendre des calculs numériques d'une très grande complexité rendant l'estimation très ardue et très longue. La difficulté principale tient au fait que les classes définissant le mélange ne sont pas connues. L'algorithme EM de Dempster et coll. (1977) a été développé dans les années 1970 spécifiquement pour régler ce type de problème. L'algorithme se compose de deux étapes, E pour espérance et M pour maximisation. On débute l'algorithme en supposant une valeur de départ à la partie de notre jeu de données qui est manquante. Dans le cas présent, l'appartenance aux différentes classes est manquante. On alterne ensuite entre les étapes E et M jusqu'à ce qu'on atteigne la convergence. À l'étape E, on calcule l'espérance de la fonction de log-vraisemblance étant donné les variables observées et en utilisant les paramètres estimés à l'itération précédente. On utilise ensuite cette fonction à l'étape M afin de trouver un nouveau vecteur de paramètres qui maximise cette fonction. L'application de cette méthode sera décrite en détail dans la prochaine section.

## 1.3. ALGORITHME EM

Afin d'utiliser l'algorithme EM, on doit avoir un jeu de données incomplet, c'est-à-dire qu'une partie des données est observée et une autre partie des données n'est pas observée. On doit donc supposer qu'un jeu de données complet existe. Dans le cas d'un mélange, Dempster et coll. (1977) de même que McLachlan et Peel (1998) et McLachlan et Peel (2000) proposent d'utiliser le jeu de données complet

$$(\mathbf{y}_i, \mathbf{z}_i),\ i = 1, \ldots, N.$$

Ici, les données observées de l'individu $i$ sont

$$\mathbf{y}_i = (y_{i1}, \ldots, y_{ip_i})',$$

où la dimension $p_i$ peut varier avec l'individu. Cette dimension $p_i$ représente le nombre de temps où l'individu $i$ a été observé. Par exemple, $y_{it}$ est une variable telle que l'agressivité ou la mémoire à court terme d'un enfant mesurée à l'âge $t$. Il y a aussi les données manquantes de l'individu $i$ notées

$$\mathbf{z}_i = (z_{i1}, \ldots, z_{iG})',$$

où $z_{ij}$ vaut 1 ou 0 selon que l'individu $i$ appartient ou non au groupe $j$. On suppose que $\mathbf{y}_i$ et $\mathbf{z}_i$ sont indépendantes et que

$$\mathbf{z}_i \sim \text{multinomiale}(1; \pi_{i1}, \ldots, \pi_{iG}).$$

Les probabilités $\pi_{ij}$ peuvent dépendre de variables auxiliaires comme le sexe de l'enfant. Ces probabilités sont paramétrées selon une fonction logistique généralisée

$$\pi_{ij} = \frac{e^{\mathbf{r}_i'\boldsymbol{\lambda}_j}}{\sum_{j=1}^{G} e^{\mathbf{r}_i'\boldsymbol{\lambda}_j}}, \tag{1.3.1}$$

où $\mathbf{r}_i$ est un vecteur de variables auxiliaires de l'individu $i$ et $\boldsymbol{\lambda}_j$ est un vecteur de paramètres associé au groupe $j$. On pose sans perte de généralité $\boldsymbol{\lambda}_1 = (0, \ldots, 0)'$ afin de rendre ces paramètres identifiables.

Si on n'utilise pas le cadre conceptuel EM, la fonction de densité du mélange est définie comme suit

$$f(\mathbf{y}_i) = \sum_{j=1}^{G} \pi_{ij} f_j(y_i) \tag{1.3.2}$$

et

$$f_j(\mathbf{y}_i) = \frac{1}{(2\pi)^{p_i/2}|\mathbf{\Sigma}_{ij}|^{1/2}} exp\left[-\frac{1}{2}(\mathbf{y}_i - \mathbf{X}_{ij}\boldsymbol{\beta}_j)'\mathbf{\Sigma}_{ij}^{-1}(\mathbf{y}_i - \mathbf{X}_{ij}\boldsymbol{\beta}_j)\right]. \tag{1.3.3}$$

La matrice $\mathbf{X}_{ij}$ est une matrice de plan servant, par exemple, à une modélisation des observations de l'individu $i$ par une régression polynomiale dans le temps. Il est à noter ici que la matrice de plan $\mathbf{X}_{ij}$, la matrice de covariance $\mathbf{\Sigma}_{ij}$ ainsi que la dimension $p_i$ de $\mathbf{y}_i$ sont indicées pour chaque sujet car les temps pour lesquels nous avons des observations peuvent varier d'un sujet à l'autre.

Dans le cas où $p_i = p$, c'est-à-dire pour un jeu de données où chaque individu est observé à chacun des temps, on peut réécrire le modèle de la façon suivante

$$f_j(\mathbf{y}_i) = \frac{1}{(2\pi)^{p/2}|\mathbf{\Sigma}_j|^{1/2}} exp\left[-\frac{1}{2}(\mathbf{y}_i - \mathbf{X}_j\boldsymbol{\beta}_j)'\mathbf{\Sigma}_j^{-1}(\mathbf{y}_i - \mathbf{X}_j\boldsymbol{\beta}_j)\right]. \tag{1.3.4}$$

Dans le cadre de l'algorithme EM, en supposant un jeu de données complet, le groupe duquel est issue l'observation $\mathbf{y}_i$ serait connue ; la fonction de densité est donc exprimée seulement en fonction des paramètres dans le groupe. On peut écrire la fonction de densité complète de la façon suivante

$$f(\mathbf{y}_i, \mathbf{z}_i) = \prod_{j=1}^{G} [\pi_{ij} f_j(\mathbf{y}_i)]^{z_{ij}}. \tag{1.3.5}$$

Maintenant que la fonction de densité est établie, voici la fonction de vraisemblance et de log-vraisemblance complètes pour le vecteur de paramètres

$$\boldsymbol{\theta} = (\boldsymbol{\lambda}_1', \cdots, \boldsymbol{\lambda}_G', \boldsymbol{\beta}_1', \cdots, \boldsymbol{\beta}_G', \boldsymbol{\xi}_1', \cdots, \boldsymbol{\xi}_G')',$$

où $\boldsymbol{\xi}_j$ est le vecteur de valeurs distinctes pour la paramétrisation des $\mathbf{\Sigma}_{ij}$,

$$L_c(\boldsymbol{\theta}) = \prod_{i=1}^{N}\prod_{j=1}^{G} [\pi_{ij} f_j(\mathbf{y}_i)]^{z_{ij}} \tag{1.3.6}$$

et

$$\log L_c(\theta) = \sum_{i=1}^{N} \sum_{j=1}^{G} z_{ij} \left[\log \pi_{ij} + \log f_j(\mathbf{y}_i)\right]. \tag{1.3.7}$$

Par exemple, en supposant l'indépendance et l'égalité des variances entre les observations d'un individu du groupe $j$, le vecteur $\xi_j$ serait constitué uniquement de la variance commune dans ce groupe. D'autres hypothèses seront toutefois considérées par la suite.

### 1.3.1. Étape E

À l'étape E (Espérance) de l'algorithme EM on doit calculer l'espérance conditionnelle de la log-vraisemblance étant donné les données observées, $\mathbf{Y}$, et l'estimateur de $\theta$ à l'itération $k-1$, à savoir $\hat{\theta}^{(k-1)}$. À cette fin, on utilise

$$\begin{aligned}
Q(\theta, \hat{\theta}^{(k-1)}) &= E\left[\log L_c(\theta) | \mathbf{Y}, \hat{\theta}^{(k-1)}\right] \\
&= E\left\{\sum_{i=1}^{N} \sum_{j=1}^{G} z_{ij} \left[\log \pi_{ij} + \log f_j(\mathbf{y}_i)\right] | \mathbf{Y}, \hat{\theta}^{(k-1)}\right\} \\
&= \sum_{i=1}^{N} \sum_{j=1}^{G} E\left[z_{ij} | \mathbf{Y}, \hat{\theta}^{(k-1)}\right] \left[\log \pi_{ij} + \log f_j(\mathbf{y}_i)\right] \\
&= \sum_{i=1}^{N} \sum_{j=1}^{G} \tau_j(\mathbf{y}_i, \hat{\theta}^{(k-1)}) \left[\log \pi_{ij} + \log f_j(\mathbf{y}_i)\right],
\end{aligned} \tag{1.3.8}$$

où

$$\begin{aligned}
\tau_j(\mathbf{y}_i, \hat{\theta}^{(k-1)}) &= E[z_{ij} | \mathbf{Y}, \hat{\theta}^{(k-1)}] \\
&= 0 \cdot \frac{f(z_{ij} = 0, \mathbf{y}_i | \hat{\theta}^{(k-1)})}{f(\mathbf{y}_i | \hat{\theta}^{(k-1)})} + 1 \cdot \frac{f(z_{ij} = 1, \mathbf{y}_i | \hat{\theta}^{(k-1)})}{f(\mathbf{y}_i | \hat{\theta}^{(k-1)})} \\
&= \frac{f(z_{ij} = 1, \mathbf{y}_i | \hat{\theta}^{(k-1)})}{f(\mathbf{y}_i | \hat{\theta}^{(k-1)})} \\
&= \frac{\hat{\pi}_{ij}^{(k-1)} f_j(\mathbf{y}_i | \hat{\theta}^{(k-1)})}{f(\mathbf{y}_i | \hat{\theta}^{(k-1)})} \\
&= \frac{\hat{\pi}_{ij}^{(k-1)} f_j(\mathbf{y}_i | \hat{\theta}^{(k-1)})}{\sum_{j=1}^{G} \hat{\pi}_{ij}^{(k-1)} f_j(\mathbf{y}_i | \hat{\theta}^{(k-1)})}.
\end{aligned} \tag{1.3.9}$$

Le terme $\tau_j(\mathbf{y}_i, \hat{\theta}^{(k-1)})$ est la probabilité *a posteriori* que le sujet $i$ appartienne au groupe $j$. Afin d'alléger la suite du texte, cette probabilité *a posteriori* sera dénotée $\tau_{ij}$.

### 1.3.2. Étape M

À l'étape M (Maximisation) de l'itération $k$, on veut maximiser la fonction $Q(\theta, \hat{\theta}^{(k-1)})$. McLachlan et Peel (1998) et McLachlan et Peel (2000) donnent certaines solutions pour des cas particuliers du modèle précédemment cité. Une solution du modèle plus générale est maintenant présentée afin de couvrir tous les cas possibles.

#### 1.3.2.1. *Estimation des proportions*

Si les variables $z_{ij}$ étaient des données observées, on utiliserait ces valeurs pour estimer les paramètres de proportion. Par contre, comme nous n'observons pas ces valeurs, nous estimons les paramètres à partir de la probabilité *a posteriori* $\tau_{ij}$ qui se substitue au valeur de $z_{ij}$. L'estimation des paramètres de proportion est effectuée selon la méthode de Fisher. La méthode est décrite en détail par Dobson (1990).

On commence par estimer le gradient de la fonction de log-vraisemblance pour chaque valeur de $\hat{\boldsymbol{\lambda}}_j^{(k-1)}$ qui est

$$\begin{aligned}
\mathbf{g}\left(\hat{\boldsymbol{\lambda}}_j^{(k-1)}\right) &= \left.\frac{\partial\left[\sum_{i=1}^{N}\sum_{j=1}^{G}\tau_{ij}\log\pi_{ij}\right]}{\partial\boldsymbol{\lambda}_j}\right|_{\boldsymbol{\lambda}_j=\hat{\boldsymbol{\lambda}}_j^{(k-1)}} \\
&= \left.\sum_{i=1}^{N}\sum_{j=1}^{G}\frac{\tau_{ij}}{\pi_{ij}}\frac{\partial\pi_{ij}}{\partial\boldsymbol{\lambda}_j}\right|_{\boldsymbol{\lambda}_j=\hat{\boldsymbol{\lambda}}_j^{(k-1)}} \\
&= \sum_{i=1}^{N}\mathbf{r}_i.\left[\tau_{ij}-\hat{\pi}_{ij}^{(k-1)}\right], \qquad (1.3.10)
\end{aligned}$$

car $\frac{\partial\pi_{ij}}{\partial\boldsymbol{\lambda}_j} = \mathbf{r}_i.\pi_{ij}(1-\pi_{ij})$ et $\frac{\partial\pi_{ik}}{\partial\boldsymbol{\lambda}_j} = -\mathbf{r}_i.\pi_{ij}\pi_{ik}$ lorsque $j \neq k$. De façon similaire, on obtient la matrice hessienne de la fonction de log-vraisemblance :

$$\mathbf{H}\left(\hat{\boldsymbol{\lambda}}_j^{(k-1)}\right) = -\sum_{i=1}^{N}\mathbf{r}_i\mathbf{r}_i'.\left[\hat{\pi}_{ij}^{(k-1)}(1-\hat{\pi}_{ij}^{(k-1)})\right]. \qquad (1.3.11)$$

On met ensuite à jour l'estimation des vecteurs de paramètres $\boldsymbol{\lambda}_j$ en se rappelant que nous avons fixé $\boldsymbol{\lambda}_1 = (0, \ldots, 0)'$ pour que les paramètres soient identifiables

$$\hat{\boldsymbol{\lambda}}_j^{(k)} = \hat{\boldsymbol{\lambda}}_j^{(k-1)} - \mathbf{H}\left(\hat{\boldsymbol{\lambda}}_j^{(k-1)}\right)^{-1} \mathbf{g}\left(\hat{\boldsymbol{\lambda}}_j^{(k-1)}\right). \tag{1.3.12}$$

Finalement, on met à jour les probabilités d'appartenance à chacun des groupes

$$\hat{\pi}_{ij}^{(k)} = \frac{e^{\mathbf{r}_i' \hat{\boldsymbol{\lambda}}_j^{(k)}}}{\sum_{j=1}^{G} e^{\mathbf{r}_i' \hat{\boldsymbol{\lambda}}_j^{(k)}}}. \tag{1.3.13}$$

On arrête l'algorithme lorsque les valeurs obtenues à l'itération $k$ ne diffèrent pas de celles obtenues à l'itération $k-1$ au delà d'une tolérance donnée. Par exemple, on peut arrêter lorsque $|\hat{\boldsymbol{\lambda}}_j^{(k)} - \hat{\boldsymbol{\lambda}}_j^{(k-1)}|_\infty < 10^{-6}$, $j = 2, \ldots, G$, où $|\mathbf{x}|_\infty = \max\{|x_1|, \ldots, |x_p|\}$ est la norme sup du vecteur $\mathbf{x}$.

Dans le cas où $r_i = 1$, pour tout $i$, c'est-à-dire qu'aucune variable auxiliaire n'est présente dans le modèle, l'estimateur de $\pi_{ij} = \pi_j$ est la moyenne des probabilités *a posteriori* :

$$\hat{\pi}_j = \sum_{i=1}^{N} \frac{\tau_{ij}}{N}. \tag{1.3.14}$$

1.3.2.2. *Estimation des paramètres de localisation et de covariance*

Nous voulons maintenant estimer les paramètres de localisation $\boldsymbol{\beta}_j$ et de covariance $\boldsymbol{\Sigma}_j$ pour chacun des groupes $j = 1, \ldots, G$. Retournons à l'expression générale de l'équation (1.3.8). On peut maintenant développer la fonction de densité $\log f_j(\mathbf{y}_i)$ et l'appliquer à la loi multinormale

$$\begin{aligned}
Q(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{(k-1)}) &= \sum_{i=1}^{N} \sum_{j=1}^{G} \tau_{ij} \left[\log \pi_{ij} + \log f_j(\mathbf{y}_i)\right] \\
&= \sum_{i=1}^{N} \sum_{j=1}^{G} \tau_{ij} \log \pi_{ij} - \frac{p_i}{2} \log(2\pi) \\
&\quad + \sum_{i=1}^{N} \sum_{j=1, j\neq l}^{G} \tau_{ij} \left[-\frac{1}{2} \log |\boldsymbol{\Sigma}_{ij}| - \frac{1}{2} (\mathbf{y}_i - \mathbf{X}_{ij}\boldsymbol{\beta}_j)' \boldsymbol{\Sigma}_{ij}^{-1} (\mathbf{y}_i - \mathbf{X}_{ij}\boldsymbol{\beta}_j)\right] \\
&\quad + \sum_{i=1}^{N} \tau_{il} \left[-\frac{1}{2} \log |\boldsymbol{\Sigma}_{il}| - \frac{1}{2} (\mathbf{y}_i - \mathbf{X}_{ij}\boldsymbol{\beta}_l)' \boldsymbol{\Sigma}_{il}^{-1} (\mathbf{y}_i - \mathbf{X}_{ij}\boldsymbol{\beta}_l)\right].
\end{aligned} \tag{1.3.15}$$

On peut noter que seulement le dernier terme de l'équation (1.3.15) dépend de $\boldsymbol{\beta}_l$ et $\boldsymbol{\Sigma}_l$. On peut donc maximiser seulement cette partie de l'équation. En multipliant ce terme par (-2), la nouvelle fonction objective à minimiser s'écrit comme

$$\begin{aligned} Q^*(\theta, \hat{\theta}^{(k-1)}) &= \sum_{i=1}^{N} \tau_{il} \left[ \log |\boldsymbol{\Sigma}_{il}| + (\mathbf{y}_i - \mathbf{X}_{il}\boldsymbol{\beta}_l)' \boldsymbol{\Sigma}_{il}^{-1} (\mathbf{y}_i - \mathbf{X}_{il}\boldsymbol{\beta}_l) \right] \\ &= \sum_{i=1}^{N} \tau_{il} \log |\boldsymbol{\Sigma}_{il}| + \sum_{i=1}^{N} \tau_{il} (\mathbf{y}_i - \mathbf{X}_{il}\boldsymbol{\beta}_l)' \boldsymbol{\Sigma}_{il}^{-1} (\mathbf{y}_i - \mathbf{X}_{il}\boldsymbol{\beta}_l). \end{aligned} \tag{1.3.16}$$

Lindstrom et Bates (1988) et Pinheiro et Bates (1996) proposent des solutions afin de réduire le problème de l'estimation en utilisant une fonction profilée de façon à réduire le nombre de paramètres estimés à chaque itération. Tout d'abord, on peut sortir la valeur $\sigma_l^2$ de la matrice de covariance et poser $\boldsymbol{\Sigma}_{il} = \sigma_l^2 \boldsymbol{\Lambda}_{il}$. Cette décomposition comporte plusieurs avantages. Tout d'abord, cela nous permet d'éliminer le paramètre d'échelle $\sigma_l^2$ de la fonction objective. De plus, Lindstrom et Bates (1988) indiquent que cette nouvelle fonction objective, définit un peu plus loin, nécessite moins d'itérations avant d'atteindre la convergence et que son gradient et sa matrice hessienne sont plus simples à calculer. Ayant fait la décomposition, la fonction à minimiser devient

$$\begin{aligned} Q^*(\theta, \hat{\theta}^{(k-1)}) &= \sum_{i=1}^{N} \tau_{il} \log |\boldsymbol{\Lambda}_{il}| + \sum_{i=1}^{N} \tau_{il} p_i \log \sigma_l^2 \\ &\quad + \sum_{i=1}^{N} \tau_{il} \sigma_l^{-2} (\mathbf{y}_i - \mathbf{X}_{il}\boldsymbol{\beta}_l)' \boldsymbol{\Lambda}_{il}^{-1} (\mathbf{y}_i - \mathbf{X}_{il}\boldsymbol{\beta}_l). \end{aligned} \tag{1.3.17}$$

On peut aisément montrer que si l'on connaît $\boldsymbol{\Lambda}_{il}$ on peut trouver les estimateurs de $\boldsymbol{\beta}_l$ et $\sigma_l^2$ qui maximisent l'équation (1.3.17). Ces estimateurs sont

$$\hat{\boldsymbol{\beta}}_l = \left( \sum_{i=1}^{N} \tau_{il} \mathbf{X}_{il}' \boldsymbol{\Lambda}_{il}^{-1} \mathbf{X}_{il} \right)^{-1} \sum_{i=1}^{N} \tau_{il} \mathbf{X}_{il}' \boldsymbol{\Lambda}_{il}^{-1} \mathbf{y}_i \tag{1.3.18}$$

$$\hat{\sigma}_l^2 = \frac{\sum_{i=1}^{N} \tau_{il} (\mathbf{y}_i - \mathbf{X}_{il}\hat{\boldsymbol{\beta}}_l)' \boldsymbol{\Lambda}_{il}^{-1} (\mathbf{y}_i - \mathbf{X}_{il}\hat{\boldsymbol{\beta}}_l)}{N_l}, \tag{1.3.19}$$

où

$$N_l = \sum_{i=1}^{N} \tau_{il} p_i. \tag{1.3.20}$$

Un exemple de paramétrisation de la matrice de covariance est celui où $\mathbf{\Lambda}_l$ (la matrice pour les participants qui ont des mesures à tous les temps de mesure) est simplement définie positive et symétrique. On peut alors la décomposer selon la méthode de Cholesky, c'est-à-dire $\mathbf{\Lambda}_l = \mathbf{U}'\mathbf{U}$, où $\mathbf{U}$ est une matrice triangulaire supérieure. Comme on a extrait $\sigma_l^2$ de la matrice $\mathbf{\Sigma}_l$ on peut donc éliminer un paramètre de $\mathbf{\Lambda}_l$ en posant un des paramètres à la valeur 1. On choisit l'élément à la première colonne et la première ligne ($u_{11} = 1$). Par rapport à un vecteur de paramètre $\xi_l = (\xi_{l,1}, \xi_{l,2}, \ldots, \xi_{l,\frac{p^2+p}{2}-1})'$, la matrice $\mathbf{U}$ a donc la forme suivante

$$\mathbf{U}(\xi_l) = \begin{pmatrix} 1 & \xi_{l,1} & \cdots & \xi_{l,p-1} \\ 0 & \xi_{l,p} & \cdots & \xi_{l,2p-2} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \xi_{l,\frac{p^2+p}{2}-1} \end{pmatrix}. \tag{1.3.21}$$

Cette paramétrisation correspond à une matrice de covariance générale et nous garantit que $\mathbf{\Lambda}_l$ sera définie positive. De plus, en contraignant les éléments de la diagonale à être strictement positif, on s'assure que le vecteur de paramètres $\xi_l$ est unique et donc identifiable.

D'autres types de covariances structurées peuvent être utilisés afin de réduire la dimension du vecteur $\xi_l$. Finalement, on peut faire une dernière réduction dans l'équation (1.3.17) en substituant leurs estimateurs aux paramètres $\beta_l$ et $\sigma_l^2$ pour obtenir la fonction suivante

$$pQ^*(\theta, \hat{\theta}^{(k-1)}) = \sum_{i=1}^{N} \tau_{il} \log|\mathbf{\Lambda}_{il}| + N_l \log \sum_{i=1}^{N} \tau_{il} (\mathbf{y}_i - \mathbf{X}_{il}\hat{\boldsymbol{\beta}}_l)' \mathbf{\Lambda}_{il}^{-1} (\mathbf{y}_i - \mathbf{X}_{il}\hat{\boldsymbol{\beta}}_l). \tag{1.3.22}$$

Pour faire l'estimation du vecteur $\xi_l$ on peut utiliser un algorithme de type Newton-Raphson. Lindstrom et Bates (1988) décrivent la façon de calculer le gradient et la matrice hessienne de la fonction $pQ^*$ par rapport au vecteur $\xi_l$.

## 1.4. ESTIMATION DU NOMBRE DE GROUPES

La section précédente montre comment estimer un modèle de mélange pour un nombre de groupes donné. Toutefois, en pratique, nous ne connaissons pas le nombre de groupes. Afin de trouver le nombre de groupes optimal, nous utilisons le *Bayesian information criterion* (BIC). Le BIC est la mesure de log-vraisemblance ajustée pour le nombre de paramètres et la taille échantillonnale. Plus spécifiquement, $\mathrm{BIC} = -2\mathrm{log}\mathrm{L}(\boldsymbol{\theta}|\mathbf{Y}) + \mathrm{df}_{\boldsymbol{\theta}}\mathrm{log}(\mathrm{N})$ où $\mathrm{df}_{\boldsymbol{\theta}}$ est le nombre de paramètres dans le modèle. Définie de cette façon, on favorise le modèle avec la plus petite valeur du BIC.

On commence donc par estimer un modèle à un groupe et on évalue le BIC. On réévalue ensuite des nouveaux modèles en augmentant le nombre de groupes jusqu'à ce que la valeur du BIC augmente. Cette façon de procéder, décrite par McLachlan et Peel (2000), est la plus utilisée.

# Chapitre 2

## SIMULATIONS

Comme il a été brièvement mentionné dans l'introduction, Bauer (2007) propose que les modèles de mélange de lois génèrent souvent plus de groupes qu'il n'est nécessaire. Une cause potentielle identifiée par Bauer serait une mauvaise modélisation de la matrice de covariance dans les groupes formant le mélange. En révisant des applications de modèles de mélanges, il se rendait compte que les matrices de covariance semblaient souvent mal spécifiées. Nous nous sommes intéressés particulièrement à trois questions interreliées :

(1) Est-ce que l'estimation du nombre de groupes, en se basant sur le BIC, est robuste à une mauvaise spécification de la matrice de covariance ?

(2) Est-ce que la supposition d'indépendance lorsqu'on est en présence d'un jeu de données où les observations d'un individu dans le temps sont corrélées est raisonnable ?

(3) Est-ce qu'une mauvaise spécification de la matrice de covariance mène à une surparamétrisation ?

Bauer souligne que ces questions sont importantes puisque l'intérêt de ce type d'analyses réside dans l'interprétation des groupes formant le mélange. Si les groupes trouvés ne servent qu'à compenser pour une mauvaise spécification de la matrice de covariance, ils perdent tout leurs sens. Ainsi nous choisissons d'examiner l'hypothèse d'indépendance et la distinguons des autres cas car il existe au moins un logiciel servant à estimer des modèles de trajectoires, *Proc Traj* dans *SAS*, qui suppose toujours l'indépendance. Comme ce logiciel

est très utilisé en recherche, les interprétations d'analyses basées sur *Proc Traj* risquent d'être erronées.

Avant d'entrer dans les détails de la simulation, voici une illustration de ce qui peut se passer lorsqu'on suppose l'indépendance en présence de données corrélées. Ce cas particulier est aussi un exemple d'une mauvaise spécification de la matrice de covariance.

Soient 50 individus fictifs, pour lesquels nous avons des observations aux temps $t = 1, 2, 3$ générées selon le modèle suivant :

$$\begin{aligned} y_{it} &= \beta_{0i} + 3t + \epsilon_{it} \\ \beta_{0i} &= 2 + \zeta_i \end{aligned} \tag{2.0.1}$$

où $(\epsilon_{i1}, \epsilon_{i2}, \epsilon_{i3})' \sim N(\mathbf{0}, (1-\rho)\mathbf{I}_3)$, $\zeta_i \sim N(0, \rho)$ et $\rho = 0,6$. Ce modèle correspond à une régression linéaire où l'ordonnée à l'origine est aléatoire. Des données ont été générées avec la version 2.13.0 du logiciel R selon le modèle décrit à l'équation (2.0.1). Le processus de génération aléatoire est décrit un peu plus loin dans ce chapitre. Ces données sont présentées dans le tableau 2.1.

Tout d'abord, les données ont été analysées selon le modèle spécifié à l'équation (2.0.1). Il s'agit en fait d'un modèle à un seul groupe dont la moyenne est une régression linéaire donnant lieu à deux paramètres de régression et dont la covariance a la structure à symétrie composée (CS), définie plus loin, donnant lieu à deux autres paramètres : la variance et la corrélation. Nous avons ensuite estimé trois modèles de mélanges de lois multinormales à un, deux et trois groupes en supposant l'indépendance. Les mélanges suivaient le modèle suivant :

$$\begin{aligned} E(y_{it}|g = j) &= \beta_{0j} + \beta_{1j}t \\ \text{Cov}[(y_{i1}, y_{i2}, y_{i3})'|g = j] &= \sigma_j^2 \mathbf{I}_3. \end{aligned} \tag{2.0.2}$$

Les BIC des quatre modèles sont présentés au tableau 2.2. On peut voir que le mélange à deux groupes est le meilleur modèle parmi les modèles supposant l'indépendance. Ce modèle nécessite sept paramètres (quatre de régression, deux de variance et un de probabilité de mélange) alors que le modèle

TABLEAU 2.1. Données fictives pour 50 individus, générées aléatoirement selon le modèle de l'équation (2.0.1)

| i | $y_{i1}$ | $y_{i2}$ | $y_{i3}$ | i | $y_{i1}$ | $y_{i2}$ | $y_{i3}$ | i | $y_{i1}$ | $y_{i2}$ | $y_{i3}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | -0,08 | 4,33 | 6,83 | 18 | 2,16 | 4,40 | 7,25 | 35 | 1,28 | 2,69 | 7,32 |
| 2 | 1,85 | 4,91 | 8,28 | 19 | 2,19 | 4,78 | 7,15 | 36 | 1,54 | 3,86 | 8,35 |
| 3 | 2,14 | 5,88 | 9,88 | 20 | 4,08 | 6,85 | 10,80 | 37 | 0,66 | 3,11 | 7,05 |
| 4 | -0,46 | 2,87 | 5,63 | 21 | 2,11 | 4,98 | 9,18 | 38 | 0,71 | 2,73 | 6,64 |
| 5 | 2,23 | 4,81 | 8,41 | 22 | 1,33 | 4,13 | 7,65 | 39 | 1,65 | 5,35 | 8,22 |
| 6 | 2,75 | 5,50 | 9,25 | 23 | 1,43 | 5,96 | 7,45 | 40 | 0,88 | 4,25 | 7,51 |
| 7 | 2,60 | 3,99 | 7,41 | 24 | 2,77 | 5,83 | 7,20 | 41 | 3,09 | 5,91 | 8,78 |
| 8 | 1,09 | 4,68 | 6,91 | 25 | 2,77 | 5,62 | 8,36 | 42 | 1,33 | 5,05 | 5,37 |
| 9 | 2,58 | 4,79 | 7,01 | 26 | 0,78 | 3,93 | 6,35 | 43 | 2,42 | 4,74 | 6,84 |
| 10 | 0,58 | 4,28 | 7,06 | 27 | 1,57 | 5,05 | 7,73 | 44 | 2,42 | 4,71 | 8,09 |
| 11 | 2,05 | 4,51 | 7,09 | 28 | 0,75 | 3,25 | 9,13 | 45 | 0,92 | 4,55 | 8,60 |
| 12 | 2,84 | 3,82 | 7,06 | 29 | 2,15 | 4,59 | 8,14 | 46 | 1,47 | 4,50 | 7,57 |
| 13 | 1,38 | 3,70 | 7,14 | 30 | 1,07 | 4,42 | 7,25 | 47 | 0,42 | 5,19 | 7,53 |
| 14 | 1,63 | 5,59 | 7,93 | 31 | 2,74 | 6,49 | 7,13 | 48 | 1,59 | 4,20 | 6,42 |
| 15 | 2,74 | 5,76 | 9,00 | 32 | 1,52 | 4,79 | 7,57 | 49 | 2,21 | 4,91 | 7,70 |
| 16 | 3,04 | 5,44 | 8,31 | 33 | 0,58 | 3,71 | 8,07 | 50 | 2,96 | 4,83 | 6,30 |
| 17 | 0,88 | 3,82 | 8,67 | 34 | 1,50 | 5,03 | 7,87 | | | | |

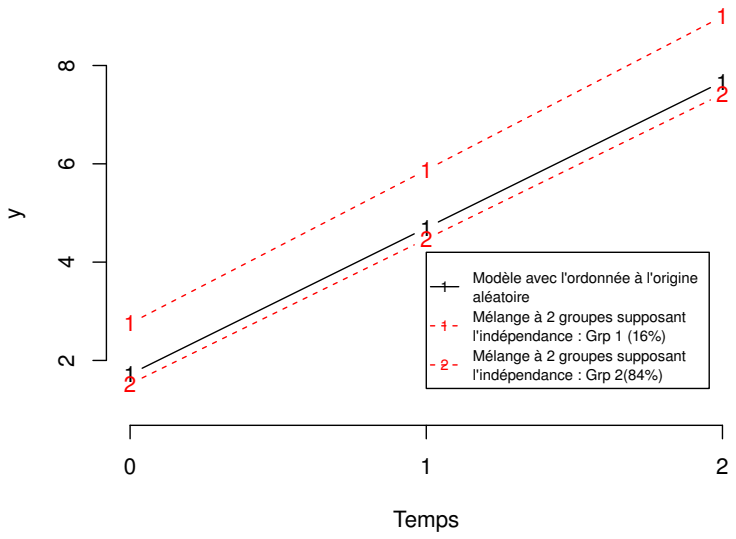
de régression avec l'ordonnée à l'origine aléatoire ne nécessite que quatre paramètres. De plus, le modèle (2.0.1) possède un BIC plus petit que le mélange à deux groupes supposant l'indépendance (390,74 *vs* 414,24). On peut constater à la figure 2.1, qu'en fait, l'analyse supposant l'indépendance estime un mélange à deux groupes avec des ordonnées à l'origine différentes afin de modéliser la variance de l'ordonnée à l'origine du vrai modèle.

Afin de tester les hypothèses de façon plus formelle, une série de simulations ont été faites.

TABLEAU 2.2. BIC d'un modèle selon l'équation (2.0.1) et de trois modèles supposant l'indépendance.

| | Modèles | BIC | Nombre de paramètres |
|---|---|---|---|
| 1 : | Modèle de régression linéaire avec ordonnée à l'origine aléatoire | 390,74 | 4 |
| 2 : | Mélange de régression linéaire à un groupe | 420,20 | 3 |
| 3 : | Mélange de régression linéaire à deux groupes | 414,24 | 7 |
| 4 : | Mélange de régression linéaire à trois groupes | 414,26 | 11 |

FIGURE 2.1. Illustration d'une mauvaise spécification de modèle




## 2.1. DESCRIPTION DES SIMULATIONS

Dans toutes les simulations, on génère les données selon le modèle suivant :

$$y_{it} = 2 + 3t + \epsilon_{it}. \tag{2.1.1}$$

Les facteurs suivants sont pris en compte :

(1) le nombre de temps de mesure,

(2) le véritable type de covariance,

(3) la taille échantillonnale,

(4) le degré de corrélation,

(5) le type de covariance supposée.

### 2.1.1. Nombre de temps de mesure

On considère deux valeurs du nombre de temps de mesure $p$, soient trois et cinq. Tout d'abord on génère des données pour les temps $t = 1, 2$ et $3$, c'est-à-dire $p = 3$, puis une autre série de données pour les temps $t = 1, 2, 3, 4$ et $5$, c'est-à-dire $p = 5$.

### 2.1.2. Véritable type de covariance

Quatre structures de covariance ont été utilisées. La première structure de covariance est la structure à symétrie composée (CS[1]) :

$$\text{Cov}\begin{pmatrix} \epsilon_{i1} \\ \vdots \\ \epsilon_{ip} \end{pmatrix} = \begin{pmatrix} \sigma^2 & \sigma^2\rho & \dots & \sigma^2\rho \\ \sigma^2\rho & \sigma^2 & \dots & \sigma^2\rho \\ \vdots & \vdots & \ddots & \vdots \\ \sigma^2\rho & \sigma^2\rho & \dots & \sigma^2 \end{pmatrix}, \tag{2.1.2}$$

$$\text{Cov}(\epsilon_{ij}, \epsilon_{ik}) = \sigma^2 \left[\rho + (1-\rho)\delta_{jk}\right],$$

où $\sigma^2 > 0$, $0 < \rho < 1$ et $\delta_{jk}$ est le delta de Kronecker qui vaut 1 lorsque $j = k$ et 0 sinon. Définie de cette façon et en fixant $\rho = 0,6$, la structure CS utilisée dans le modèle de l'équation (2.1.1) est équivalente à la structure de la covariance du modèle avec ordonnée à l'origine aléatoire de l'équation (2.0.1). Dans toutes les simulations, nous avons posé $\sigma^2 = 1$.

1. On utilise l'acronyme CS pour cette structure car, en anglais, elle s'appelle *Compound Symmetry*

La deuxième structure de covariance est une structure à symétrie composée avec variance hétérogène (UCS) :

$$\text{Cov}\begin{pmatrix} \epsilon_{i1} \\ \vdots \\ \epsilon_{ip} \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho & \dots & \sigma_1\sigma_p\rho \\ \sigma_1\sigma_2\rho & \sigma_2^2 & \dots & \sigma_2\sigma_p\rho \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_1\sigma_p\rho & \sigma_2\sigma_p\rho & \dots & \sigma_p^2 \end{pmatrix}, \tag{2.1.3}$$

$$\text{Cov}(\epsilon_{ij}, \epsilon_{ik}) = \sigma_j\sigma_k\left[\rho + (1-\rho)\delta_{jk}\right],$$

où pour tout $i = 1, \dots, p$, $\sigma_i > 0$ et $0 < \rho < 1$. Dans toutes les simulations, nous avons posé $\sigma_i^2 = i$.

La troisième structure de covariance est une structure autorégressive d'ordre 1 (AR1) :

$$\text{Cov}\begin{pmatrix} \epsilon_{i1} \\ \vdots \\ \epsilon_{ip} \end{pmatrix} = \begin{pmatrix} \sigma^2 & \sigma^2\rho^1 & \dots & \sigma^2\rho^{p-1} \\ \sigma^2\rho^1 & \sigma^2 & \dots & \sigma^2\rho^{p-2} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma^2\rho^{p-1} & \sigma^2\rho^{p-2} & \dots & \sigma^2 \end{pmatrix}, \tag{2.1.4}$$

$$\text{Cov}(\epsilon_{ij}, \epsilon_{ik}) = \sigma^2\rho^{|j-k|},$$

où $\sigma^2 > 0$ et $0 < \rho < 1$. Dans toutes les simulations nous avons posé $\sigma^2 = 1$.

La quatrième structure de covariance est une structure autorégressive d'ordre 1 avec variance hétérogène (UAR1) :

$$\text{Cov}\begin{pmatrix} \epsilon_{i1} \\ \vdots \\ \epsilon_{ip} \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho^1 & \dots & \sigma_1\sigma_p\rho^{p-1} \\ \sigma_1\sigma_2\rho^1 & \sigma_2^2 & \dots & \sigma_2\sigma_p\rho^{p-2} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_1\sigma_p\rho^{p-1} & \sigma_2\sigma_p\rho^{p-2} & \dots & \sigma_p^2 \end{pmatrix}, \tag{2.1.5}$$

$$\text{Cov}(\epsilon_{ij}, \epsilon_{ik}) = \sigma_j\sigma_k\rho^{|j-k|},$$

où pour tout $i = 1, \dots, p$, $\sigma_i > 0$ et $0 < \rho < 1$. Dans toutes les simulations, nous avons posé $\sigma_i^2 = i$.

### 2.1.3. Taille échantillonnale

Les tailles échantillonnales N considérées sont $N = 200$, 400 et 600.

### 2.1.4. Degré de corrélation

Le degré de corrélation $\rho$ est un autre facteur. Les données ont été générées avec $\rho = 0, 0, 3$ et $0, 6$. Comme les données sont générées selon une loi multinormale, le cas où $\rho = 0$ correspond à l'indépendance des données à travers le temps.

### 2.1.5. Type de covariance supposée

Les quatre structures de covariance définies précédemment, CS, UCS, AR1 et UAR1 ont été utilisées. Deux autres structures supposant l'indépendance, IND et UIND, ont aussi été utilisées dans les analyses. La structure IND correspond aux cas particuliers des structures CS et AR1 où le paramètre $\rho$ est fixé à 0. La structure UIND correspond aux cas particuliers des structures UCS et UAR1 où le paramètre $\rho$ est fixé à 0.

### 2.1.6. Génération aléatoire des données

Pour générer les données aléatoires nous avons utilisé la bibliothèque *mvtnorm* du logiciel *R (V2.13.0)*. Cette bibliothèque utilise l'algorithme 1 pour générer $N$ observations de dimension $p$. On peut voir que les $\mathbf{y}_i$ suivent bien la

**Algorithme 1** Générer aléatoirement $N$ observations selon la loi $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

Générer $Np$ variables $z_1, \ldots, z_{Np} \overset{iid}{\sim} N(0, 1)$.
$\mathbf{Z} := ((z_1, \ldots, z_N)', \ldots, (z_{N(p-1)+1}, \ldots, z_{Np})')$
Trouver $\boldsymbol{\Gamma}$ et $\boldsymbol{\lambda}$ tel que $\boldsymbol{\Sigma} = \boldsymbol{\Gamma}\mathbf{D}_{\boldsymbol{\lambda}}\boldsymbol{\Gamma}'$, $\boldsymbol{\Gamma}\boldsymbol{\Gamma}' = \mathbf{I}_p$ et $\mathbf{D}_{\boldsymbol{\lambda}} = diag((\lambda_1, \ldots, \lambda_p)')$
$\mathbf{D}^{\frac{1}{2}} := diag(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_p})$
**pour** $i = 1 \rightarrow 10$ **faire**
  $\mathbf{y}_i := \boldsymbol{\Sigma}^{\frac{1}{2}}\mathbf{Z}_i' + \boldsymbol{\mu}$
**fin pour**
$\mathbf{Y} := (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N)'$
Retourner $\mathbf{Y}$

loi multinormale $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ car

$$\begin{aligned} E(\mathbf{y}_i) &= E(\boldsymbol{\Sigma}^{\frac{1}{2}}\mathbf{z}_i + \boldsymbol{\mu}) \\ &= \boldsymbol{\Sigma}^{\frac{1}{2}}E(\mathbf{z}_i) + \boldsymbol{\mu} \\ &= \boldsymbol{\mu} \end{aligned} \tag{2.1.6}$$

et

$$\begin{aligned} \text{Cov}(\mathbf{y}_i) &= \text{Cov}(\boldsymbol{\Sigma}^{\frac{1}{2}}\mathbf{z}_i + \boldsymbol{\mu}) \\ &= \boldsymbol{\Sigma}^{\frac{1}{2}}\text{Cov}(\mathbf{z}_i)\boldsymbol{\Sigma}^{\frac{1}{2}} \\ &= \boldsymbol{\Sigma}. \end{aligned} \tag{2.1.7}$$

### 2.1.7. Logiciels utilisés

Toutes les analyses de mélanges ont été effectuées avec le logiciel Mplus (V6.11), Muthén et Muthén (2000) et Muthén et coll. (2002). Ce logiciel estime les mélanges de lois à l'aide de l'algorithme EM tel que décrit au chapitre 1.

Un des problèmes de l'algorithme EM et de l'estimation de modèles complexes en général est qu'il arrive que l'algorithme converge vers une solution correspondant à un maximum local de la fonction de log-vraisemblance. Pour s'assurer de ne pas rapporter ce type de solution, nous avons effectué au minimum 200 estimations avec des valeurs de départs différentes. Pour les estimations à plus de 2 groupes nous utilisons 500 estimations au minimum et finalement 800 estimations pour les modèles à plus de 6 groupes. Afin d'automatiser les analyses dans *Mplus* et de récupérer les résultats dans *R*, la bibliothèque *MplusAutomation* a été utilisée. Nous établissons le nombre de groupes selon la méthode décrite à la section 1.4.

Il est à noter que la bibliothèque *MMELN* Giguère (2011), créé dans le cadre de cette recherche, permet d'estimer des modèles de mélange avec une covariance de type CS ou IND. Cette bibliothèque est disponible gratuitement dans le logiciel *R* et continuera d'être développé dans le futur permettant ainsi d'utiliser d'autres types de covariance.

### 2.1.8. Déroulement de la simulation

Pour toutes les combinaisons des facteurs manipulés, 20 jeux de données ont été générés indépendamment et analysés en utilisant des modèles de mélange. Au total, $2 \times 4 \times 3 \times 3 \times 6 \times 20 = 8640$ jeux de données ont donc été générés. Pour chaque jeu de données, un mélange de lois multinormales a été estimé. Le nombre de groupes du mélange a été estimé en prenant le modèle dont le BIC était minimal. Cette façon de procéder est décrite par McLachlan et Peel (2000). Les résultats obtenus dans les simulations étaient donc le nombre de groupes des mélanges. On se rappellera que tous les jeux de données proviennent d'un mélange à un seul groupe. Les données ont été ajustées selon le modèle suivant :

$$\begin{aligned} E(y_{it}|g = j) &= \beta_{0j} + \beta_{1j}t + \epsilon_{it}, \\ \text{Cov}[(\epsilon_{i1}, \ldots, \epsilon_{ip})'|g = j] &= \mathbf{\Sigma}_j, \end{aligned} \tag{2.1.8}$$

où $\mathbf{\Sigma}_j$ suit une des six structures décrites précédemment. Les tableaux complets des résultats des simulations sont présentés à l'annexe A. Le sommaire des résultats est présenté dans la prochaine section.

## 2.2. Sommaire des résultats

### 2.2.1. Robustesse de l'estimation du nombre de groupes

La première question de recherche était : est-ce que l'estimation du nombre de groupes, en se basant sur le BIC, est robuste à une mauvaise spécification de la matrice de covariance ?

Lorsque la structure de covariance supposée était appropriée, c'est-à-dire que les données étaient générées et analysées en utilisant la même structure de covariance, la plupart des analyses n'estimaient pas plus d'un groupe. Les données étant générées selon un modèle à un seul groupe, on ne devrait pas en obtenir plus. Parmi 1440 analyses faites en supposant une matrice de covariance appropriée, 1377 (95,6%) estimaient un mélange à un groupe, 62 (4,3%)

estimaient un mélange à deux groupes et finalement 1 (<,01%) estimait un modèle à trois groupes. On peut donc voir que si la structure de covariance est bien spécifiée, dans environ 95 % des cas, on n'estime pas de groupes supplémentaires. On peut voir au tableau 2.3 que les analyses estimant des mélanges à plus d'un groupe tendaient à provenir des jeux de données ayant des niveaux de corrélation ρ plus élevés et des tailles échantillonnales N plus petites.

TABLEAU 2.3. Tableau croisé de p × N pour les mélanges à plus d'un groupe ayant une matrice de covariance appropriée

| | | N | | | |
|---|---|---|---|---|---|
| | | 200 | 400 | 600 | Total |
| ρ | 0 | 3 | 1 | 1 | 5 |
| | 0,3 | 17 | 9 | 2 | 28 |
| | 0,6 | 18 | 6 | 6 | 30 |
| | Total | 38 | 16 | 9 | 63 |

Ensuite, 4320 analyses ont été faites en supposant une structure alternative à celle qui a servi à générer les données. Ces résultats sont présentés dans le tableau 2.4. Les jeux de données analysés en supposant l'indépendance ne sont pas inclus dans ce nombre et seront considérés dans la prochaine sous-section. Parmi ces 4320 analyses, 2133 (49,4 %) estimaient des mélanges à un groupe, 1361 (31,5 %) estimaient des mélanges à 2 groupes, 757 (17,5 %) estimaient des mélanges à 3 groupes et 69 (1,6 %) estimaient des mélanges à quatre groupes. Les modèles de mélanges ne sont donc pas robustes à une mauvaise spécification de la structure de covariance, c'est-à-dire lorsque la matrice de covariance est inappropriée, on surestime le nombre de groupes. Tel qu'illustré dans le tableau 2.5, les analyses qui estimaient des mélanges à plus d'un groupe tendaient à avoir des tailles échantillonnales N et des dimensions p plus grandes ainsi que des niveaux de corrélation ρ plus élevés.

Certaines des erreurs de spécification étaient moins graves que d'autres. Lorsqu'on supposait des variances hétérogènes sur des données générées avec une variance homogène à travers le temps et en spécifiant adéquatement la

TABLEAU 2.4. Nombre de groupes estimés pour chaque combinaison de structure de covariance réelle *vs* supposée

| Covariance | | Nombre de groupes trouvés | | | | |
|---|---|---|---|---|---|---|
| véritable | supposée | 1 | 2 | 3 | 4 | Total |
| AR1 | CS | 199 | 147 | 14 | 0 | 360 |
| | UAR1 | 337 | 23 | 0 | 0 | 360 |
| | UCS | 223 | 128 | 9 | 0 | 360 |
| CS | AR1 | 211 | 80 | 69 | 0 | 360 |
| | UAR1 | 226 | 73 | 50 | 11 | 360 |
| | UCS | 345 | 15 | 0 | 0 | 360 |
| UAR1 | AR1 | 38 | 156 | 166 | 0 | 360 |
| | CS | 30 | 227 | 98 | 5 | 360 |
| | UCS | 241 | 104 | 14 | 1 | 360 |
| UCS | AR1 | 33 | 112 | 194 | 21 | 360 |
| | CS | 33 | 221 | 98 | 8 | 360 |
| | UAR1 | 217 | 75 | 45 | 23 | 360 |
| | Total | 2133 | 1361 | 757 | 69 | 4320 |

structure de corrélation, on obtenait des résultats comparables à une bonne spécification du modèle. Par exemple, si on analysait les jeux de données générées avec les structures AR1 et CS en supposant les structures UAR1 et UCS, seulement 5,3 % des analyses estimaient des mélanges à 2 groupes, les autres analyses estimaient des mélanges à 1 groupe. Ce constat est logique puisque la supposition des variances hétérogènes est plus générale que la supposition d'une variance homogène à travers le temps.

La supposition contraire était toutefois problématique. Lorsqu'on supposait une variance homogène sur des jeux de données générées avec des variances hétérogènes, en considérant une structure de corrélation appropriée, 52,4 % des analyses estimaient des mélanges à deux groupes, 36,7 % estimaient des mélanges à trois groupes et 1 % estimaient des mélanges à quatre groupes.

TABLEAU 2.5. Tableau croisé de p × ρ × N pour les mélanges à plus d'un groupe supposant une structure de covariance mal spécifée

| | | N | | | |
|---|---|---|---|---|---|
| | ρ | 200 | 400 | 600 | Total |
| p = 3 | 0 | 34 | 72 | 81 | 187 |
| | 0,3 | 83 | 94 | 100 | 277 |
| | 0,6 | 123 | 149 | 175 | 447 |
| Total | | 240 | 315 | 356 | 911 |
| p = 5 | 0 | 35 | 71 | 81 | 187 |
| | 0,3 | 132 | 169 | 185 | 486 |
| | 0,6 | 202 | 201 | 200 | 603 |
| Total | | 369 | 441 | 466 | 1276 |
| | Total | 609 | 756 | 822 | 2187 |

Ces cas incluent les analyses où les structures de covariance supposées étaient CS ou AR1 alors que les données étaient générées respectivement avec les structures UCS et UAR1.

Parmi les 4320 analyses précédentes, 1440 analyses ont été faites en supposant une structure de corrélation inappropriée et une variance bien spécifiée. Par exemple, la véritable covariance est CS et la covariance supposée est AR1. Parmi celles-ci, 868 (60,2 %) analyses estimaient des mélanges à un groupe, 406 (28,2 %) analyses estimaient des mélanges à deux groupes, 142 (10,0 %) analyses estimaient des mélanges à trois groupes et 24 (1,6 %) estimaient des mélanges à quatre groupes. La spécification de la structure de corrélation est donc très importante.

Finalement 720 analyses ont été faites en supposant une structure de corrélation inappropriée et en supposant une variance homogène sur des données générées avec des variances hétérogènes. Par exemple, la véritable covariance est UAR1 et la covariance supposée est CS. Dans ce cas, 63 (8,8 %) analyses

estimaient des mélanges à un groupe, 448 (62,2 %) analyses estimaient des mélanges à deux groupes, 196 (27,2 %) analyses estimaient des mélanges à trois groupes et 13 (1,8 %) analyses estimaient des mélanges à quatre groupes. On a donc pu constater que cette combinaison de corrélation inappropriée et de variance inappropriée nous a donné des analyses qui surestimaient le nombre de groupes dans plus de 90 % des cas.

### 2.2.2. Estimation du nombre de groupes en supposant l'indépendance

Examinons maintenant la deuxième question à laquelle on voulait répondre. Est-ce que la supposition d'indépendance lorsqu'on est en présence d'un jeu de données où les observations d'un individu dans le temps sont corrélées est raisonnable ? Afin de répondre à cette question examinons les résultats des analyses lorsque nous avons supposé l'indépendance. Ces résultats sont présentés au tableau 2.6. On peut constater que la supposition d'indépendance introduit un biais énorme dans l'estimation du nombre de groupes lorsqu'on fait des analyses sur des données corrélées. On peut estimer jusqu'à huit groupes dans un mélange dont on sait que les données proviennent d'un seul groupe.

TABLEAU 2.6. Nombre de groupes estimés en supposant l'indépendance des données

| Covariance | | Nombre de groupes trouvés | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| véritable | supposée | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | Total |
| AR1 | IND | 126 | 66 | 64 | 39 | 34 | 18 | 12 | 1 | 360 |
| | UIND | 134 | 93 | 70 | 36 | 27 | 0 | 0 | 0 | 360 |
| CS | IND | 115 | 58 | 83 | 50 | 46 | 8 | 0 | 0 | 360 |
| | UIND | 119 | 86 | 73 | 66 | 16 | 0 | 0 | 0 | 360 |
| UAR1 | IND | 15 | 149 | 97 | 47 | 27 | 19 | 6 | 0 | 360 |
| | UIND | 136 | 104 | 60 | 41 | 17 | 2 | 0 | 0 | 360 |
| UCS | IND | 15 | 123 | 103 | 63 | 48 | 8 | 0 | 0 | 360 |
| | UIND | 122 | 91 | 59 | 73 | 15 | 0 | 0 | 0 | 360 |
| | Total | 782 | 770 | 609 | 415 | 230 | 55 | 18 | 1 | 2880 |

Le niveau de corrélation $\rho$ influence beaucoup ce biais comme on peut le constater au tableau 2.7. En particulier, on peut voir que si la supposition d'indépendance est appropriée, c'est-à-dire $\rho = 0$, seulement 217 (22,6 %) analyses estimaient des mélanges à plus d'un groupe. On obtient un résultat équivalent lorsqu'on suppose une certaine forme de covariance puisque dans ce cas, 19,7 % des analyses estimaient des mélanges à plus d'un groupe. Par contre, avec $\rho = 0,6$ on constate qu'aucune analyse n'a estimé un mélange à un groupe et on estime jusqu'à 8 groupes.

TABLEAU 2.7. Impact du niveau de corrélation $\rho$ si on suppose l'indépendance des données

| | Nombre de groupes trouvés | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| $\rho$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | Total |
| 0 | 743 | 211 | 6 | 0 | 0 | 0 | 0 | 0 | 960 |
| 0,3 | 39 | 546 | 347 | 28 | 0 | 0 | 0 | 0 | 960 |
| 0,6 | 0 | 13 | 256 | 387 | 230 | 55 | 18 | 1 | 960 |
| Total | 782 | 770 | 609 | 415 | 230 | 55 | 18 | 1 | 2880 |

Les analyses effectuées en supposant une structure de covariance UIND tendaient à estimer moins de groupes même lorsque les données étaient générées avec une variance homogène. Les analyses tendaient toutefois à estimer encore moins de groupes même lorsque les variances étaient hétérogènes (UCS et UAR1) et que l'on supposait une structure UIND. Ce résultat est présenté au tableau 2.8.

Parmi les analyses supposant l'indépendance, plus la taille échantillonnale $N$ et le nombre d'observations $p$ augmentaient, plus le nombre de groupes estimés tendait à augmenter. Ce résultat est illustré au tableau 2.9.

### 2.2.3. Surestimation des paramètres

Finalement, examinons la troisième question : est-ce qu'une mauvaise spécification de la matrice de covariance mène à une surparamétrisation des modèles de mélange ? On parle de surparamétrisation lorsqu'on utilise plus de

TABLEAU 2.8. Nombre de groupes selon le type de variance pour les analyses supposant l'indépendance

| Variances | Covariance | Nombre de groupes estimés | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| véritables | supposée | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | Total |
| homogènes | IND | 241 | 124 | 147 | 89 | 80 | 26 | 12 | 1 | 720 |
| | UIND | 253 | 179 | 143 | 102 | 43 | 0 | 0 | 0 | 720 |
| Total | | 494 | 303 | 290 | 191 | 123 | 26 | 12 | 1 | 1440 |
| hétérogènes | IND | 30 | 272 | 200 | 110 | 75 | 27 | 6 | 0 | 720 |
| | UIND | 258 | 195 | 119 | 114 | 32 | 2 | 0 | 0 | 720 |
| Total | | 288 | 467 | 319 | 224 | 107 | 29 | 6 | 0 | 1440 |
| | Total | 782 | 770 | 609 | 415 | 230 | 55 | 18 | 1 | 2880 |

TABLEAU 2.9. Tableau croisé de $p \times N \times g$ pour les groupes supposant l'indépendance

| | | Nombre de groupes | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | N | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | Total |
| $p = 3$ | 200 | 149 | 158 | 126 | 44 | 3 | 0 | 0 | 0 | 480 |
| | 400 | 121 | 166 | 98 | 80 | 13 | 2 | 0 | 0 | 480 |
| | 600 | 120 | 142 | 96 | 96 | 26 | 0 | 0 | 0 | 480 |
| Total | | 390 | 466 | 320 | 220 | 42 | 2 | 0 | 0 | 1440 |
| $p = 5$ | 200 | 149 | 122 | 87 | 83 | 35 | 3 | 1 | 0 | 480 |
| | 400 | 123 | 109 | 88 | 69 | 62 | 25 | 4 | 0 | 480 |
| | 600 | 120 | 73 | 114 | 43 | 91 | 25 | 13 | 1 | 480 |
| Total | | 392 | 304 | 289 | 195 | 188 | 53 | 18 | 1 | 1440 |
| | Total | 782 | 770 | 609 | 415 | 230 | 55 | 18 | 1 | 2880 |

paramètres qu'il n'est nécessaire pour ajuster les données. Selon les résultats que nous avons obtenus aux deux sections précédentes, il semble que les modèles de mélanges dont les matrices de covariance sont mal spécifiées mènent à une surparamétrisation. La surparamétrisation contrevient au principe de parcimonie qui tend à favoriser les modèles les plus simples.

Pour chaque nouveau groupe généré dans un mélange, on ajoute le nombre de paramètres nécessaires afin d'estimer un mélange à un groupe plus le nombre de paramètres nécessaires à modéliser la probabilité d'appartenir au nouveau groupe. Dans les simulations, nous avons supposé que la probabilité que l'individu $i$ appartienne au groupe $j$ ne dépend pas de variables auxiliaires, c'est-à-dire $\pi_{ij} = \pi_j$ suivant la notation au chapitre 1. Si on utilise une structure de covariance CS ou AR1, par exemple, nous devons ajouter cinq paramètres pour chaque groupe ajouté. En effet, il faut deux paramètres pour modéliser la localisation des observations, deux paramètres pour modéliser la structure de covariance et un paramètre pour modéliser la probabilité d'être dans le nouveau groupe. Le nombre de paramètres nécessaires à chaque modèle est présenté au tableau 2.10.

TABLEAU 2.10. Nombre de paramètres nécessaires pour estimer un mélange selon le nombre de groupes et la structure de covariance utilisée

| Structure de | Nombre de groupes | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| covariance | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| CS/AR1 | 4 | 9 | 14 | 19 | 24 | 29 | 34 | 39 |
| UCS/UAR1 | 8 | 17 | 26 | 35 | 44 | 53 | 62 | 71 |
| IND | 3 | 7 | 11 | 15 | 19 | 23 | 27 | 31 |
| UIND | 7 | 15 | 23 | 31 | 39 | 47 | 55 | 63 |

Lorsqu'on regarde le nombre de paramètres nécessaires pour modéliser un mélange à deux groupes, on constate qu'il faut plus de paramètres que pour un modèle à un groupe. Les seules exceptions sont les cas où on utilise la structure IND et qu'on compare avec des mélanges à un groupe comportant des structures de covariance UCS ou UAR1. Toutefois, nous pouvons voir au tableau 2.8, que si les données sont générées selon des structures de covariance UCS ou UAR1 et que l'on suppose l'indépendance, dans 58 % des cas nous estimerons des mélanges à plus d'un groupe. On peut donc conclure qu'une

mauvaise spécification de la matrice de covariance mène, dans la majorité des cas, à une surparamétrisation.

Finalement, afin de quantifier le phénomène de surparamétrisation, nous avons défini une nouvelle variable $\Delta$. Cette variable est définie comme la différence entre le nombre de paramètres du modèle théorique et le nombre de paramètres nécessaires pour ajuster les modèles de mélange. Nous avons séparé les analyses en trois catégories, les analyses avec 1) une structure de covariance bien spécifiée, 2) une structure de covariance mal spécifiée (excluant l'indépendance) et 3) une structure de covariance supposant l'indépendance des observations. Pour ces trois groupes, la valeur moyenne de $\Delta$ est respectivement de 0,39, 4,27 et 7,89. On note que l'hypothèse d'indépendance nous mène à estimer plus de paramètres. Ce phénomène s'explique par le fait qu'une structure de covariance, même mal spécifiée, nous permet de tenir compte de l'interdépendance entre les données à travers le temps. L'hypothèse d'indépendance, toutefois, nous oblige à partitionner les données jusqu'à ce que chaque partition de données soient plus ou moins contenues dans des petites hypersphères.

Le tableau 2.11 indique les valeurs moyennes de $\Delta$ selon le nombre d'observations $p$ d'un individu dans le temps, le niveau de corrélation $\rho$ et la taille échantillonnale $N$. On peut voir que si la matrice de covariance est bien spécifiée, plus la taille échantillonnale augmente plus la moyenne de $\Delta$ diminue. Dans tous les autres cas, plus le nombre d'observations $p$, le niveau de corrélation $\rho$ ou la taille échantillonnale $N$ augmente, plus la moyenne de $\Delta$ augmente.

TABLEAU 2.11. Moyennes de $\Delta$ pour les différents niveaux de $p \times \rho \times N$ selon le type de spécification de la covariance

| | | | Types de spécification de la covariance | | |
|---|---|---|---|---|---|
| p | ρ | N | Bien spécifiée | Mal spécifiée | Indépendance |
| 3 | 0 | 200 | 0,11 | 0,76 | -0,13 |
| | | 400 | 0,11 | 1,52 | 0,05 |
| | | 600 | 0,00 | 1,70 | 0,03 |
| | 0,3 | 200 | 0,68 | 2,15 | 4,8 |
| | | 400 | 0,40 | 2,21 | 5,8 |
| | | 600 | 0,23 | 2,59 | 6,6 |
| | 0,6 | 200 | 0,74 | 3,41 | 12,1 |
| | | 400 | 0,23 | 4,84 | 14,5 |
| | | 600 | 0,34 | 6,44 | 15,8 |
| 5 | 0 | 200 | 0,23 | 1,02 | -0,30 |
| | | 400 | 0,00 | 2,21 | -0,08 |
| | | 600 | 0,11 | 2,56 | 0,05 |
| | 0,3 | 200 | 1,35 | 4,16 | 5,63 |
| | | 400 | 0,56 | 5,51 | 7,83 |
| | | 600 | 0,00 | 6,98 | 9,68 |
| | 0,6 | 200 | 1,24 | 8,15 | 16,18 |
| | | 400 | 0,45 | 9,75 | 20,38 |
| | | 600 | 0,34 | 10,92 | 22,93 |

# Chapitre 3

# APPLICATION DES MÉLANGES À DES DONNÉES COMPORTEMENTALES ET COGNITIVES

## 3.1. MÉTHODE

### 3.1.1. Description de l'étude

Toutes les données analysées proviennent de l'étude "En 2001 ... j'avais 5 ans", décrit dans le rapport de Jetté et coll. (1997), qui était, dans sa phase initiale, l'échantillon pilote de l'enquête longitudinale des enfants du Québec (ÉLDEQ). Pour participer à l'étude, les enfants devaient habiter dans l'une des régions administratives suivantes : Capitale-Nationale, Montréal, Chaudière-Appalaches, Laval, Lanaudière, Laurentides ou Montérégie. 1000 enfants ont été sélectionnés aléatoirement à partir du registre des naissances du Québec de façon à ce que le nombre d'enfants tirés dans chaque région soit proportionnel au total dans la population. De ce nombre, 572 parents ont consenti à participer à l'étude donnant un échantillon total de 279 garçons et 293 filles. La première collecte a eu lieu en 1996, lorsque les enfants de l'étude étaient âgés d'environ 5 mois. Des collectes annuelles ont suivi. Les participants de cette étude sont actuellement âgés de 15 ans.

### 3.1.2. Description des données comportementales

Deux variables de comportement proviennent de l'instrument PBQ (Preschool Behaviour Questionnaire) de Tremblay et coll. (1992). Il s'agit de l'agressivité physique et de l'hyperactivité. Les deux comportements ont été mesurés à 1, 5, 2, 5, 3, 5, 5, 6, 7 et 9 ans. Chaque item est évalué sur une échelle de Likert de 0 à 2 qui représente respectivement "Jamais", "Quelquefois" et "Souvent". Ces deux séries de variables ont été évaluées par la PCM, c'est-à-dire la personne qui connaît le mieux l'enfant. Dans la majorité des cas, cette personne était la mère biologique de l'enfant.

#### 3.1.2.1. *Agressivité physique*

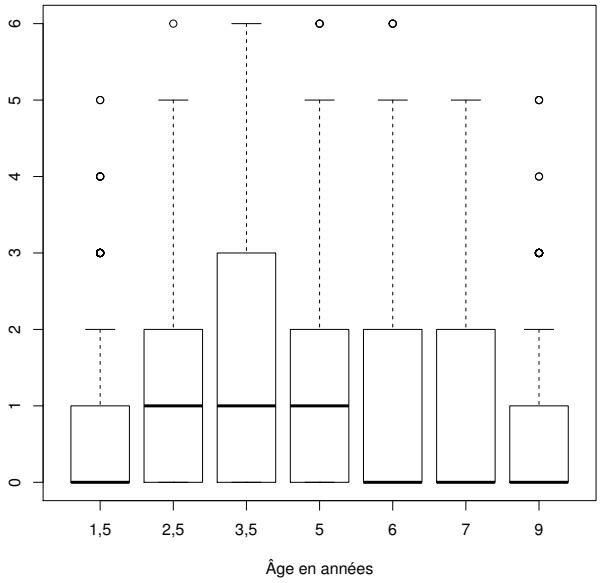



FIGURE 3.1. Agressivité physique de 1,5 à 9 ans

L'agressivité physique était composée de 3 items : "Frappe, mord, donne des coups de pied", "Se bagarre" et "Attaque physiquement les autres". L'échelle d'agressivité est la somme des réponses aux 3 items. Cette somme donne une échelle de 0 à 6. Si une réponse manquait à un des items, l'échelle était calculée en effectuant la moyenne des deux items valides puis en multipliant par 3 pour

TABLEAU 3.1. Statistiques descriptives de l'agressivité physique de 1,5 à 9 ans

| | **1,5 ans** | **2,5 ans** | **3,5 ans** | **5 ans** | **6 ans** | **7 ans** | **9 ans** |
|---|---|---|---|---|---|---|---|
| **N** | 511 | 496 | 474 | 423 | 358 | 244 | 233 |
| **$\alpha$ de Cronbach** | 0,55 | 0,63 | 0,77 | 0,73 | 0,78 | 0,73 | 0,78 |
| **Minimum** | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **Maximum** | 5 | 6 | 6 | 6 | 6 | 5 | 5 |
| **Moyenne** | 0,68 | 1,08 | 1,48 | 1,29 | 0,94 | 0,86 | 0,80 |
| **Écart type** | 0,99 | 1,19 | 1,46 | 1,36 | 1,25 | 1,17 | 1,18 |

ramener l'échelle entre 0 et 6. Le résultat final a été arrondi à l'unité près afin de respecter les propriétés de l'échelle originale. Si plus d'un item étaient manquants, l'échelle n'a pas été calculée. Les statistiques descriptives de l'échelle d'agressivité sont colligées dans le tableau 3.1. La statistique $\alpha$ de Cronbach (1951) est une mesure de cohérence interne définie comme suit

$$\alpha = \frac{n}{n-1}\left(1 - \frac{\sum_{i=1}^{n}\sigma_i^2}{\sigma_T^2}\right),$$

où $n$ représente le nombre d'items dans l'échelle, $\sigma_i^2$ est la variance de l'item $i$ et $\sigma_T^2$ est la variance de la somme des items, *i.e.* l'échelle totale. Une bonne échelle devrait avoir un $\alpha$ de Cronbach s'approchant de 1. Une valeur de $\alpha$ entre 0,7 et 0,8 correspond généralement à une cohérence interne jugée acceptable par Cronbach et Richard (2004). Si tous les items sont positivement corrélés, la statistique $\alpha$ de Cronbach varie entre 0 et 1. La figure 3.1 montre les graphiques en boîtes de l'échelle d'agressivité à travers le temps.

3.1.2.2. *Hyperactivité*

La mesure d'hyperactivité était composée de 5 items : "Ne peut rester en place, agité", "Remue sans cesse", "Est impulsif, agit sans réfléchir","A de la difficulté à attendre son tour dans un jeu" et "A de la difficulté à rester tranquille pour faire quelque chose plus de quelques instants". L'échelle d'hyperactivité est la somme des réponses aux 5 items. Cette somme donne une échelle de 0 à 10. Si un ou deux des items n'étaient pas répondus, l'échelle a été calculée

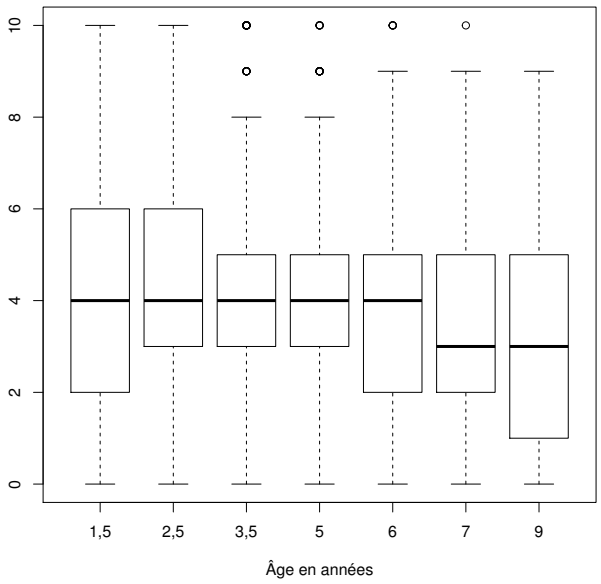



FIGURE 3.2. Hyperactivité de 1,5 à 9 ans

TABLEAU 3.2. Statistiques descriptives de l'hyperactivité de 2,5 à 9 ans

| | **1,5 ans** | **2,5 ans** | **3,5 ans** | **5 ans** | **6 ans** | **7 ans** | **9 ans** |
|---|---|---|---|---|---|---|---|
| **N** | 511 | 497 | 474 | 423 | 358 | 244 | 233 |
| **$\alpha$ de Cronbach** | 0,69 | 0,74 | 0,75 | 0,74 | 0,79 | 0,78 | 0,75 |
| **Minimum** | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **Maximum** | 10 | 10 | 10 | 10 | 10 | 10 | 9 |
| **Moyenne** | 3,98 | 4,50 | 4,17 | 4,18 | 3,78 | 3,50 | 3,33 |
| **Écart type** | 2,4 | 2,4 | 2,3 | 2,2 | 2,2 | 2,2 | 2,2 |

comme la moyenne des items valides multipliée par 5. Le résultat final a été arrondi à l'unité près afin de respecter les propriétés de l'échelle originale. Si plus de deux items étaient manquants, l'échelle n'a pas été calculée. Les statistiques descriptives de l'échelle d'hyperactivité sont colligées dans le tableau 3.2. On peut voir l'évolution de l'hyperactivité dans les graphiques en boîtes pour tous les temps de mesure à la figure 3.2.

### 3.1.3. Description des données cognitives

Plusieurs tâches ont été effectuées par les participants de l'étude afin de mesurer leurs différentes habiletés cognitives. Ces tâches étaient administrées par des assistantes de recherche au domicile de l'enfant.

3.1.3.1. *Mesure de mémoire à court terme*

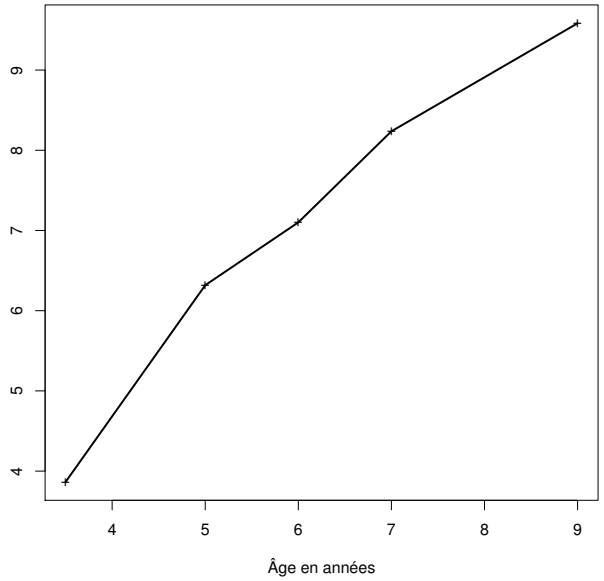



FIGURE 3.3. Niveau de mémoire à court terme de 3,5 à 9 ans

TABLEAU 3.3. Statistiques descriptives du VCR de 3,5 à 9 ans

| | **3,5 ans** | **5 ans** | **6 ans** | **7 ans** | **9 ans** |
|---|---|---|---|---|---|
| **N** | 376 | 388 | 311 | 277 | 231 |
| **Minimum** | 1,0 | 1,0 | 1,0 | 4,1 | 5,0 |
| **Maximum** | 10,0 | 10,0 | 12,0 | 12,0 | 12,0 |
| **Moyenne** | 3,86 | 6,32 | 7,10 | 8,24 | 9,58 |
| **Écart type** | 2,4 | 2,3 | 2,3 | 2,1 | 1,9 |

Le VCR (Visually Cued Recall) de Zelazo et coll. (2002) est une tâche évaluant la mémoire à court terme des participants. Cette tâche a été administrée

aux participants alors qu'ils étaient âgés de 3,5, 5, 6, 7 et 9 ans. À chaque niveau, l'expérimentateur pointe un certain nombre d'images sur un carton. L'expérimentateur retourne ensuite le carton durant une seconde puis demande à l'enfant de lui pointer les mêmes images. L'enfant réussit l'essai si toutes les images sont identifiées correctement. Le nombre d'images augmente d'un niveau à l'autre. La tâche s'arrête lorsque l'enfant a échoué deux niveaux consécutifs. Il y a 12 niveaux à cette tâche.

L'échelle de mémoire à court terme est la somme des proportions de réussite de tous les niveaux. Voici comment sont calculées les proportions de réussite :

$$PR = \frac{N_{BS} - (N_S - N_{BS})}{N_I}, \tag{3.1.1}$$

où PR représente la proportion de réussite, $N_{BS}$ représente le nombre d'images bien identifiées par l'enfant, $N_S$ représente le nombre total d'images montrées par l'enfant (incluant les mauvaises) et $N_I$ représente le nombre d'images pointées par l'expérimentateur. Bien que cette situation ne se présente que très rarement, il peut arriver qu'un sujet ne montre que des mauvaises images dans ce cas PR prend une valeur négative, on attribue alors une valeur de 0 à la proportion de réussite. L'échelle totale est donc une somme de 12 items variant de 0 à 1 résultant en une échelle de 0 à 12.

Les statistiques descriptives de l'échelle de mémoire à court terme sont présentées au tableau 3.3 et on peut voir graphiquement l'évolution de l'échelle à la figure 3.3.

3.1.3.2. *Mesure de vocabulaire*

Le niveau de vocabulaire réceptif (compréhension) du participant a été mesuré à l'aide de l'ÉVIP (échelle de vocabulaire en image de Peabody). L'ÉVIP est la version française du PPVT-R (Peabody Picture Vocabulary Task - Revised edition) de Dunn et Dunn (1981) et Dunn et coll. (1993). Comme il y avait aussi des participants anglophones, la version originale anglaise a été utilisée pour ces participants. Puisqu'une minorité de participants ont répondu à la version

TABLEAU 3.4. Statistiques descriptives de l'ÉVIP de 3,5 à 8 ans

| | **3,5 ans** | **5 ans** | **6 ans** | **7 ans** | **8 ans** |
|---|---|---|---|---|---|
| **N** | 396 | 347 | 275 | 254 | 230 |
| **Minimum** | 0 | 10 | 22 | 58 | 62 |
| **Maximum** | 81 | 108 | 130 | 141 | 146 |
| **Moyenne** | 30,47 | 62,14 | 87,51 | 101,94 | 110,55 |
| **Écart Type** | 14,3 | 18,3 | 14,9 | 14,0 | 15,3 |

anglaise, ils ont été exclus des analyses. Cette tâche a été administrée aux participants à 3, 5, 5, 6, 7 et 8 ans. À chaque étape, l'enfant doit identifier la bonne image parmi quatre images associées à un mot de vocabulaire. L'expérimentateur identifie tout d'abord le niveau plancher du participant. Le plancher correspond au niveau auquel l'enfant est confortable, c'est-à-dire qu'il réussit huit niveaux sans se tromper. Ensuite, en partant de ce niveau, l'expérimentateur continue jusqu'à ce qu'il atteigne le niveau plafond du participant. Le plafond est le niveau précédent une séquence de huit essais contenant au moins six erreurs. L'échelle de vocabulaire correspond au niveau du plafond moins le nombre d'erreurs entre la base et le plafond. Il y a 170 niveaux de difficulté dans l'ÉVIP. Le tableau 3.4 montre les statistiques descriptives pour l'ÉVIP de 3, 5 à 8 ans. La figure 3.4 montre les moyennes de l'ÉVIP à travers le temps.

### 3.1.4. Échantillons des analyses

#### 3.1.4.1. *Données comportementales*

Plusieurs participants n'ont fait que les entrevues qui ont eu lieu lors des premières années de collecte. Typiquement, ces participants manifestaient légèrement plus de problèmes de comportement. Le mécanisme d'attrition n'était donc pas aléatoire. En incluant ces participants seulement dans les premiers temps de collecte, on aurait pu être amené à penser qu'il y a une baisse des problèmes de comportement dans le temps. En fait, ces participants n'auraient pas été inclus dans les calculs des derniers temps et auraient donc abaissé les moyennes. Nous avons donc effectué les analyses sur le sous-échantillon de

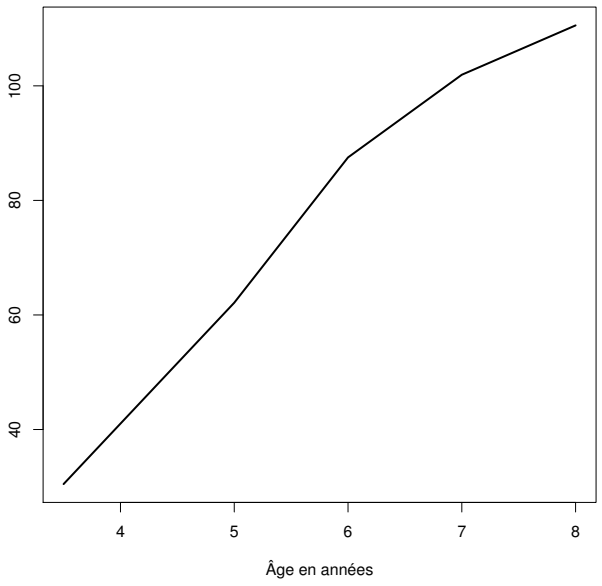



FIGURE 3.4. Niveau de vocabulaire de 3,5 à 8 ans

participants qui ont répondu à au moins deux collectes de données de 1,5 à 5 ans et qui ont répondu à deux autres collectes de 6 à 9 ans. Dans ce nouvel échantillon, le mécanisme de données manquantes est supposé aléatoire. Les résultats ne s'appliquaient donc qu'à ce sous-échantillon et non à la population totale de l'étude. En imposant ces deux conditions, un échantillon de 281 participants a été obtenu. Cette démarche est résumée au tableau 3.5. Les données pouvaient donc être incomplètes pour plusieurs participants mais elles couvraient tout de même la période visée de façon adéquate.

TABLEAU 3.5. Conditions d'inclusion des participants dans l'analyse des comportements

| Âge en années | 1,5 2,5 3,5 5 | 6 7 8* 9 |
|---|---|---|
| Conditions | 1 : $\geq$ 2 présences | 2 : $\geq$ 2 présences |

Note : *Une collecte de données a eu lieu à 8 ans mais les comportements n'ont pas été mesurés.

3.1.4.2. *Données cognitives*

Pour des raisons similaires à celles évoquées pour les comportements, nous avons exclu certains participants aux tâches cognitives des analyses. Nous avons inclu seulement les participants qui ont effectué les tâches, le VCR ou l'ÉVIP, à au moins trois reprises entre 3,5 ans et 9 ans. En imposant cette condition, un échantillon de 320 participants a été obtenu pour la mémoire à court terme et un échantillon de 300 sujets a été obtenu pour le vocabulaire réceptif. On se rappelle que pour le vocabulaire, les participants ayant répondu à la version originale anglaise ont aussi été exclus. Le processus est décrit au tableau 3.6

TABLEAU 3.6. Conditions d'inclusion des participants dans les analyses des données cognitives

| Âge en années | 3,5 5 6 7 8* 9** |
|---|---|
| Conditions | 1 : $\geq$ 3 présences |

Note : *À 8 ans, la mémoire n'a pas été mesurée.

**À 9 ans, le vocabulaire n'a pas été mesuré.

### 3.1.5. Logiciel utilisé

Toutes les analyses ont été effectuées avec le logiciel Mplus (V6.11), Muthén et Muthén (2000) et Muthén et coll. (2002). Ce logiciel estime les mélanges de lois à l'aide de l'algorithme EM tel que décrit au chapitre 1. Un des problèmes de l'algorithme EM est qu'il est possible que l'on converge sur une solution correspondant à un maximum local de la fonction de log-vraisemblance. Pour s'assurer de ne pas rapporter ce type de solution, nous avons effectué au minimum 200 estimations avec des valeurs de départs différentes en s'assurant que la solution soit répliquée au moins deux fois. Pour les estimations à plus de trois groupes, 1000 estimations au minimum ont été effectuées tout en s'assurant que la solution soit répliquée au moins quatre fois. Les précisions des paramètres (erreurs-types) ont été estimées en utilisant la méthode du maximum de vraisemblance (ML). Il est à noter toutefois que Mplus permet d'estimer

des erreurs-types robustes. Comme les méthodes d'estimation d'erreurs-types robustes peuvent varier d'un logiciel à l'autre et que l'inférence sur les paramètres n'était pas le but central de cette recherche, l'approche classique, basée sur l'inversion de la matrice d'information de Fisher, a été utilisée.

## 3.2. ANALYSE

### 3.2.1. Agressivité physique

#### 3.2.1.1. *Modèle statistique*

Dans les graphique en boîtes de la figure 3.1, on a pu constater que plusieurs participants démontraient un niveau d'agressivité très faible à tous les temps de mesure. Nous avons donc considéré qu'une des trajectoires aura une matrice de variance très faible (toutes les valeurs propres de la matrice de covariance seraient très basses) et une moyenne près de 0 à tous les temps. Nous avons laissé les paramètres de localisation des autres groupes s'estimer librement, en supposant une variance commune. Cette décision a été prise étant donnée la faible variation de l'échelle d'agressivité. Les matrices de covariances ont toutes été supposées à symétrie composée (CS). Ce type de covariance ajustait aussi bien les données qu'une matrice de covariance non-structurée lorsqu'on comparait leurs BIC. Nous avons estimé tout d'abord un groupe avec une croissance polynomiale cubique puis nous avons augmenté le nombre de groupes jusqu'à ce que le BIC se détériore. Le même processus a été repris par la suite en supposant l'indépendance des observations à travers le temps, c'est-à-dire que toutes les covariances sont fixées à 0. Voici la

description plus formelle du modèle pour l'agressivité physique :

$$
\begin{aligned}
E(y_{ij}|g=1) &= \beta_{01}, \\
E(y_{ij}|g=k) &= \beta_{0k} + \beta_{1k}t_j + \beta_{2k}t_j^2 + \beta_{3k}t_j^3,\ k>1, \\
\mathrm{Cov}\left[\left.\begin{pmatrix} y_{i1} \\ \vdots \\ y_{i7} \end{pmatrix}\right| g=1\right] &= \begin{bmatrix} \sigma_1^2 & \sigma_1^2\rho_1 & \dots & \sigma_1^2\rho_1 \\ \sigma_1^2\rho_1 & \sigma_1^2 & \dots & \sigma_1^2\rho_1 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_1^2\rho_1 & \sigma_1^2\rho_1 & \dots & \sigma_1^2 \end{bmatrix}, \\
\mathrm{Cov}\left[\left.\begin{pmatrix} y_{i1} \\ \vdots \\ y_{i7} \end{pmatrix}\right| g>1\right] &= \begin{bmatrix} \sigma_2^2 & \sigma_2^2\rho_2 & \dots & \sigma_2^2\rho_2 \\ \sigma_2^2\rho_2 & \sigma_2^2 & \dots & \sigma_2^2\rho_2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_2^2\rho_2 & \sigma_2^2\rho_2 & \dots & \sigma_2^2 \end{bmatrix},
\end{aligned}
\tag{3.2.1}
$$

où $y_{ij}$ représente la mesure d'agressivité du participant $i$ au temps $t_j$, $j = 1, 2, \dots, 7$, et $g$ désigne les groupes du mélange.

Concernant l'estimation des $\pi_{ij}$, c'est-à-dire la probabilité pour le participant $i$ d'appartenir au groupe $j$, le modèle a été établi tout d'abord sans aucune variable explicative. On parle donc des $\pi_j$ puisque les proportions sont les mêmes pour tous les participants. Le modèle s'écrit comme suit :

$$
\pi_j = \frac{e^{\lambda_j}}{\sum_{j=1}^{G} e^{\lambda_j}},\ \text{où } \lambda_1 = 0. \tag{3.2.2}
$$

Lorsque le meilleur modèle a été établi par la méthode décrite un peu plus haut, le modèle a été réajusté afin d'inclure le sexe dans le calcul des probabilités. Le modèle s'écrit donc maintenant comme suit :

$$
\pi_{ij} = \frac{e^{\lambda_{0j}+\lambda_{1j}1_{garcon_i}}}{\sum_{j=1}^{G} e^{\lambda_{0j}+\lambda_{1j}1_{garcon_i}}},\ \text{où } \lambda_{01} = \lambda_{11} = 0, \tag{3.2.3}
$$

et

$$
1_{garcon_i} = \begin{cases} 1, & \text{si } sexe_i = \text{garçon}, \\ 0, & \text{si } sexe_i = \text{fille}. \end{cases}
$$

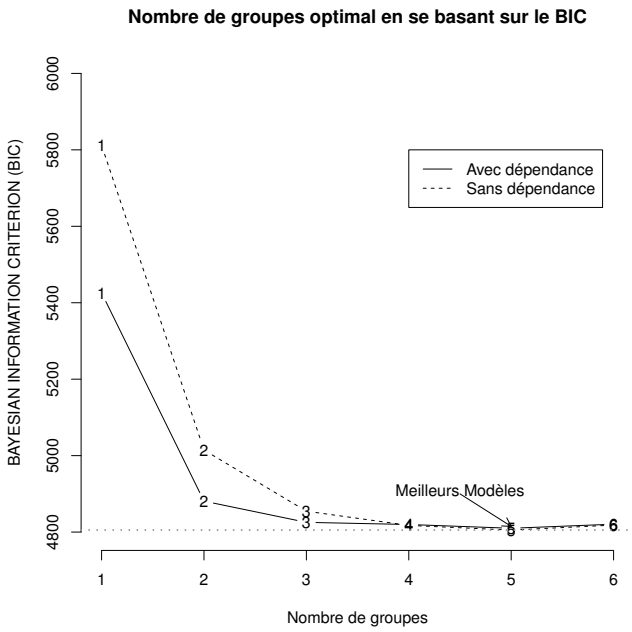



FIGURE 3.5. Sélection du meilleur modèle (selon le BIC)

3.2.1.2. *Résultats*

La figure 3.5 illustre le processus de sélection du meilleur modèle. Les deux séries d'analyses indiquaient, en se basant sur le BIC, que le nombre de groupes optimal pour ajuster les données d'agressivité était un modèle à cinq groupes. Le nombre de groupes ayant été établi, les deux modèles pouvaient s'exprimer en termes de moyennes à travers le temps. La figure 3.6 montre les moyennes prédites d'agressivité pour les cinq trajectoires de 1,5 à 9 ans. À la vue de cette figure, on a pu noter que la supposition de dépendance à l'intérieur des trajectoires n'altérait pas l'allure des courbes. Le BIC du modèle supposant l'indépendance des observations était légèrement plus faible (4805,216 *vs* 4809,951) que celui qui supposait la dépendance, ce modèle a donc été retenu. Le modèle a finalement été réévalué en incluant la variable de sexe du participant, tel que décrit dans la méthode. Nous avons donc obtenu le modèle final, incluant le risque associé de se trouver dans l'une ou l'autre des trajectoires selon que l'on soit un garçon ou une fille. Le tableau 3.7 montre les

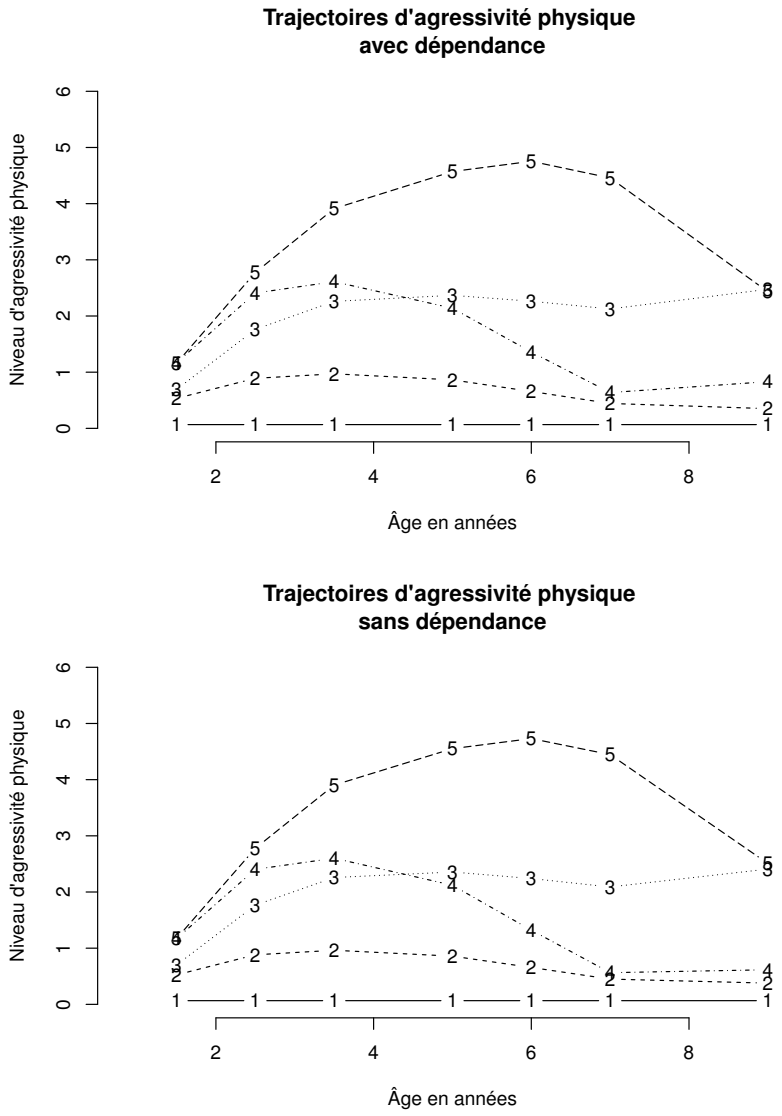



FIGURE 3.6. Moyennes prédites d'agressivité physique de 1,5 à 9 ans

valeurs des paramètres estimés de localisation et de dispersion correspondant à la série d'équations (3.2.1) décrivant le modèle dans la section méthode.

Les résultats ont indiqué, tel que supposé, que le premier groupe, composé de 18,4% des participants avait une variance presque nulle et une moyenne proche de 0. Les groupes 2 et 4 avec respectivement une proportion de 46,6% et de 16,2% avaient des trajectoires qui démarraient avec une moyenne relativement basse lorsque les enfants avaient 1,5 ans, atteignaient leur maximum lorsque les enfants avaient 3,5 ans et décroissaient par la suite pour atteindre

TABLEAU 3.7. Paramètres de localisation et de dispersion des trajectoires d'agressivité physique

| Paramètres | Valeurs Estimés | é. t. | $Z$ | $P(\|Z\| > 0)$ |
|---|---|---|---|---|
| Groupe 1 (18,4%) : Aucune agressivité | | | | |
| $\beta_{01}$ | 0,065 | 0,014 | 4,590 | 0,000 |
| $\sigma_1^2$ | 0,061 | 0,006 | 9,949 | 0,000 |
| Groupe 2 (46,6%) : Agressivité basse | | | | |
| $\beta_{02}$ | 0,524 | 0,088 | 5,936 | 0,000 |
| $\beta_{12}$ | 0,521 | 0,135 | 3,851 | 0,000 |
| $\beta_{22}$ | -0,182 | 0,048 | -3,749 | 0,000 |
| $\beta_{32}$ | 0,015 | 0,004 | 3,423 | 0,001 |
| $\sigma_2^2$ | 0,926 | 0,037 | 24,935 | 0,000 |
| Groupe 3 (16,5%) : Agressivité basse ascendante | | | | |
| $\beta_{03}$ | 0,706 | 0,180 | 3,918 | 0,000 |
| $\beta_{13}$ | 1,415 | 0,231 | 6,113 | 0,000 |
| $\beta_{23}$ | -0,380 | 0,083 | -4,566 | 0,000 |
| $\beta_{33}$ | 0,031 | 0,008 | 3,898 | 0,000 |
| $\sigma_2^2$ | 0,926 | 0,037 | 24,935 | 0,000 |
| Groupe 4 (16,2%) : Agressivité modérée | | | | |
| $\beta_{04}$ | 1,166 | 0,222 | 5,262 | 0,000 |
| $\beta_{14}$ | 1,874 | 0,270 | 6,944 | 0,000 |
| $\beta_{24}$ | -0,699 | 0,099 | -7,086 | 0,000 |
| $\beta_{34}$ | 0,060 | 0,009 | 6,504 | 0,000 |
| $\sigma_2^2$ | 0,926 | 0,037 | 24,935 | 0,000 |
| Groupe 5 (2,2%) : Agressivité élevée | | | | |
| $\beta_{05}$ | 1,148 | 0,379 | 3,028 | 0,002 |
| $\beta_{15}$ | 1,828 | 0,543 | 3,366 | 0,001 |
| $\beta_{25}$ | -0,230 | 0,198 | -1,164 | 0,244 |
| $\beta_{35}$ | 0,000 | 0,019 | -0,021 | 0,983 |
| $\sigma_2^2$ | 0,926 | 0,037 | 24,935 | 0,000 |

Le paramètre $\sigma_2^2$ est commun aux groupes 2 à 5

leur niveau initial d'agressivité à 9 ans. Les participants du groupe 2 avaient un niveau relativement plus faible ($\approx 1$) à 3,5 ans que ceux du groupe 4 ($\approx 2,5$). Le groupe 3, avec une proportion de 16,5%, avait une trajectoire ascendante qui démarrait à un niveau relativement bas puis croissait rapidement entre 1,5 ans et 5 ans et se maintenait au même niveau par la suite jusqu'à l'âge de 9 ans. La dernière trajectoire, avec une proportion de 2,2%, suivait une courbe quadratique qui démarrait à un niveau relativement haut à 1,5 ans, atteignait son maximum à 6 ans ($\approx$4,5), et redescendait par la suite jusqu'à 9 ans ($\approx$2,5). Cette trajectoire est intéressante à étudier car ces participants manifestaient un niveau élevé d'agressivité de 1,5 à 9 ans lorsqu'on les comparait aux participants des autres groupes. Il faut noter par contre que l'étude de cette sous-population dans des analyses ultérieures pourrait s'avérer difficile étant donné son effectif réel d'environ six participants. Dans ce contexte, un modèle à quatre trajectoires aurait peut-être été préférable. Comme le but de cette recherche se limitait à étudier le comportement à travers le temps, le modèle à cinq groupes nous donnait le meilleur modèle. Les trajectoires 2 à 5 partageaient une variance commune proche de 1. Il y avait donc davantage de variation dans ces trajectoires lorsqu'on les comparait à la trajectoire 1. On a pu constater que la supposition de dépendance à l'intérieur des groupes n'était pas nécessaire dans ce contexte puisque que l'ajustement était aussi bon lorsqu'on fixait la covariance à 0. Il est à noter toutefois que le fait d'avoir ajouté les deux paramètres utilisés pour modéliser la covariance nous donnait le même résultat en terme de trajectoire d'agressivité. Ceci ne signifie pas que les données d'agressivité ne soient pas corrélées à travers le temps mais plutôt que la dépendance s'expliquait par l'appartenance à l'une des trajectoires. Ce phénomène est bien illustré à la figure 3.5 puisqu'on a pu observer que dans les modèles à moins de trois trajectoires, et lorsqu'on supposait la dépendance à l'intérieur des groupes, on ajustait mieux les données qu'en supposant l'indépendance, cependant, plus on se rapprochait de quatre trajectoires plus les deux types d'analyses donnaient des ajustements comparables. Le tableau 3.8 donne les paramètres du mélange. On peut voir que le fait d'être un garçon

TABLEAU 3.8. Paramètres du mélange des trajectoires d'agressivité physique

| Paramètres | Valeurs Estimés | é. t. | Z | P(\|Z\| > 0) | exp( V. E. ) |
|---|---|---|---|---|---|
| $\lambda_{01}$ | 0 | - | - | - | 1 |
| $\lambda_{11}$ | 0 | - | - | - | 1 |
| $\lambda_{02}$ | 0,463 | 0,222 | 2,089 | 0,037 | 8,08 |
| $\lambda_{12}$ | 1,265 | 0,403 | 3,142 | 0,002 | 23,15 |
| $\lambda_{03}$ | -0,725 | 0,353 | -2,056 | 0,040 | 0,13 |
| $\lambda_{13}$ | 1,543 | 0,493 | 3,128 | 0,002 | 22,83 |
| $\lambda_{04}$ | -0,611 | 0,392 | -1,560 | 0,119 | 0,21 |
| $\lambda_{14}$ | 1,307 | 0,519 | 2,520 | 0,012 | 12,43 |
| $\lambda_{05}$ | -3,588 | 1,018 | -3,524 | 0,000 | 0,03 |
| $\lambda_{15}$ | 2,750 | 1,155 | 2,382 | 0,017 | 10,83 |

TABLEAU 3.9. Probabilités conditionnelles d'appartenir aux trajectoires d'agressivité physique selon le sexe

| k | $P(g = k\|garçon)$ | $P(g = k\|fille)$ |
|---|---|---|
| 1 | 0,088 | 0,274 |
| 2 | 0,497 | 0,436 |
| 3 | 0,200 | 0,133 |
| 4 | 0,177 | 0,149 |
| 5 | 0,038 | 0,008 |

augmentait de beaucoup la probabilité d'être dans les groupes 2 à 5 lorsqu'on la comparait au groupe 1 qui était le groupe de référence. Pour voir la répartition des probabilités d'appartenance à chacun des groupes conditionnellement au fait d'être un garçon ou une fille il s'agit de revenir à la définition de la probabilité définie à l'équation (3.2.3). Le tableau 3.9 donne les probabilités d'appartenance aux trajectoires conditionnellement au sexe du participant. L'inférence faite sur le groupe 5 bien qu'ayant du sens n'est en fait basée que

sur un effectif d'environ six personnes, on doit donc être prudent dans l'interprétation de ce résultat. En changeant la catégorie de référence, on confirmait que les seules différences significatives étaient entre le groupe 1 et les groupes 2 à 5. Ces résultats ne sont pas tous affichés par souci d'économie puisqu'ils ne modifiaient pas le contenu du tableau 3.9.

### 3.2.2. Hyperactivité de 1,5 à 9 ans

#### 3.2.2.1. *Modèle statistique*

L'hyperactivité a été modélisée avec des courbes de croissance polynomiales cubiques. Un modèle cubique est une réduction sensée par rapport au modèle de moyennes qui requérait le calcul de sept moyennes pour chaque trajectoire. Comme dans le cas de l'agressivité physique, des matrices de covariances à symétrie composée ont été utilisées. Lorsqu'on a utilisé cette structure dans un modèle et qu'on l'a comparé à un modèle avec une matrice de covariance générale on a pu noter que les ajustements étaient comparables tout en passant de 28 paramètres à 2 paramètres. Voici la description plus formelle du modèle utilisé pour ajuster les données d'hyperactivité :

$$E(y_{ij}|g=k) = \beta_{0k} + \beta_{1k}t_j + \beta_{2k}t_j^2 + \beta_{3k}t_j^3,$$

$$\mathrm{Cov}\left[\left.\begin{pmatrix} y_{i1} \\ \vdots \\ y_{i7} \end{pmatrix}\right| g=k\right] = \begin{bmatrix} \sigma_k^2 & \sigma_k^2\rho_k & \dots & \sigma_k^2\rho_k \\ \sigma_k^2\rho_k & \sigma_k^2 & \dots & \sigma_k^2\rho_k \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_k^2\rho_k & \sigma_k^2\rho_k & \dots & \sigma_k^2 \end{bmatrix}, \tag{3.2.4}$$

où $y_{ij}$ représente la mesure d'hyperactivité du participant $i$ au temps $t_j$, $j = 1, 2, \dots, 7$ et $g$ indique les $G$ différentes trajectoires.

Concernant l'estimation des $\pi_{ij}$, c'est-à-dire la probabilité pour le participant $i$ d'appartenir au groupe $j$, la même procédure que pour l'agressivité physique a été utilisée.

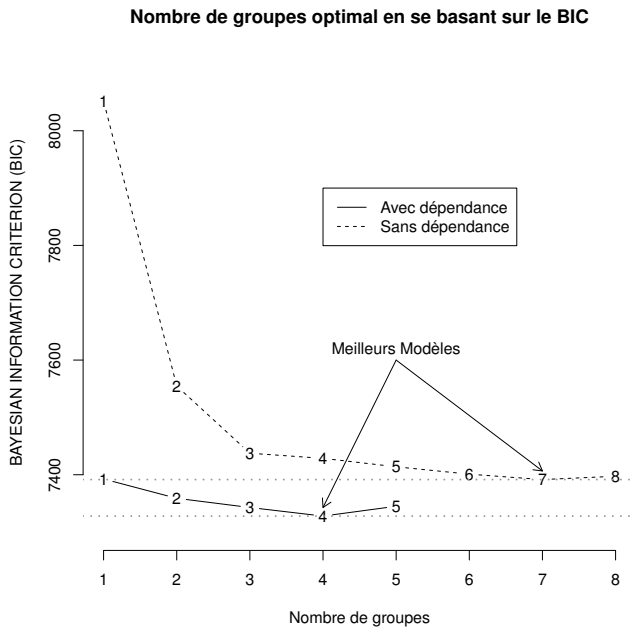



FIGURE 3.7. Sélection du meilleur modèle (selon le BIC)

3.2.2.2. *Résultats*

La figure 3.7 montre le processus de sélection du meilleur modèle. Le meilleur modèle en supposant la dépendance était le modèle à quatre trajectoires. Par contre, en supposant l'indépendance, le meilleur modèle était le modèle à sept groupes. La figure 3.8 montre les moyennes prédites d'hyperactivité pour les quatre trajectoires de 1,5 à 9 ans supposant la dépendance *vs* les sept trajectoires en supposant l'indépendance. À la vue de cette figure, on peut voir que la supposition de dépendance à l'intérieur des trajectoires donne des résultats très différents. Le modèle supposant la dépendance avait un meilleur ajustement en se basant sur le BIC (7327,89 *vs* 7391,58), nous avons donc retenu ce modèle. De plus ce modèle nécessitait 14 paramètres de moins : 27 paramètres en supposant la dépendance *vs* 41 paramètres en supposant l'indépendance. Le modèle a été finalement réévalué en incluant la variable de sexe du participant, tel que décrit dans la méthode. Nous avons donc obtenu le modèle final, incluant le risque associé de se trouver dans l'une ou l'autre des trajectoires selon que l'on soit un garçon ou une fille. Le tableau 3.10 donne les valeurs des

TABLEAU 3.10. Paramètres de localisation et de dispersion des trajectoires d'hyperactivité

| Paramètres | Valeurs Estimés | é. t. | Z | $P(\|Z\| > 0)$ |
|---|---|---|---|---|
| Groupe 1 (12,8%) : Faible hyperactivité | | | | |
| $\beta_{01}$ | 1,727 | 0,227 | 7,618 | 0,000 |
| $\beta_{11}$ | 0,226 | 0,284 | 0,796 | 0,426 |
| $\beta_{21}$ | -0,070 | 0,109 | -0,640 | 0,522 |
| $\beta_{31}$ | 0,003 | 0,010 | 0,316 | 0,752 |
| $\sigma_1^2$ | 1,237 | 0,180 | 6,874 | 0,000 |
| $\rho_1$ | 0,296 | 0,085 | 3,469 | 0,001 |
| Groupe 2 (29,2%) : Hyperactivité descendante | | | | |
| $\beta_{02}$ | 5,744 | 0,288 | 19,964 | 0,000 |
| $\beta_{12}$ | -0,856 | 0,267 | -3,211 | 0,001 |
| $\beta_{22}$ | 0,038 | 0,090 | 0,425 | 0,670 |
| $\beta_{32}$ | 0,003 | 0,008 | 0,358 | 0,721 |
| $\sigma_2^2$ | 3,920 | 0,530 | 7,401 | 0,000 |
| $\rho_2$ | 0,555 | 0,064 | 8,734 | 0,000 |
| Groupe 3 (14,7%) : Hyperactivité courbe concave | | | | |
| $\beta_{03}$ | 2,858 | 0,398 | 7,184 | 0,000 |
| $\beta_{13}$ | 3,713 | 0,353 | 10,513 | 0,000 |
| $\beta_{23}$ | -1,318 | 0,123 | -10,726 | 0,000 |
| $\beta_{33}$ | 0,112 | 0,011 | 10,015 | 0,000 |
| $\sigma_3^2$ | 3,865 | 0,746 | 5,180 | 0,000 |
| $\rho_3$ | 0,678 | 0,067 | 10,173 | 0,000 |
| Groupe 4 (43,2%) : Hyperactivité ascendante | | | | |
| $\beta_{04}$ | 3,894 | 0,240 | 16,205 | 0,000 |
| $\beta_{14}$ | 0,391 | 0,249 | 1,569 | 0,117 |
| $\beta_{24}$ | -0,009 | 0,089 | -0,105 | 0,916 |
| $\beta_{34}$ | -0,005 | 0,008 | -0,587 | 0,557 |
| $\sigma_4^2$ | 3,994 | 0,336 | 11,882 | 0,000 |
| $\rho_4$ | 0,369 | 0,053 | 6,982 | 0,000 |

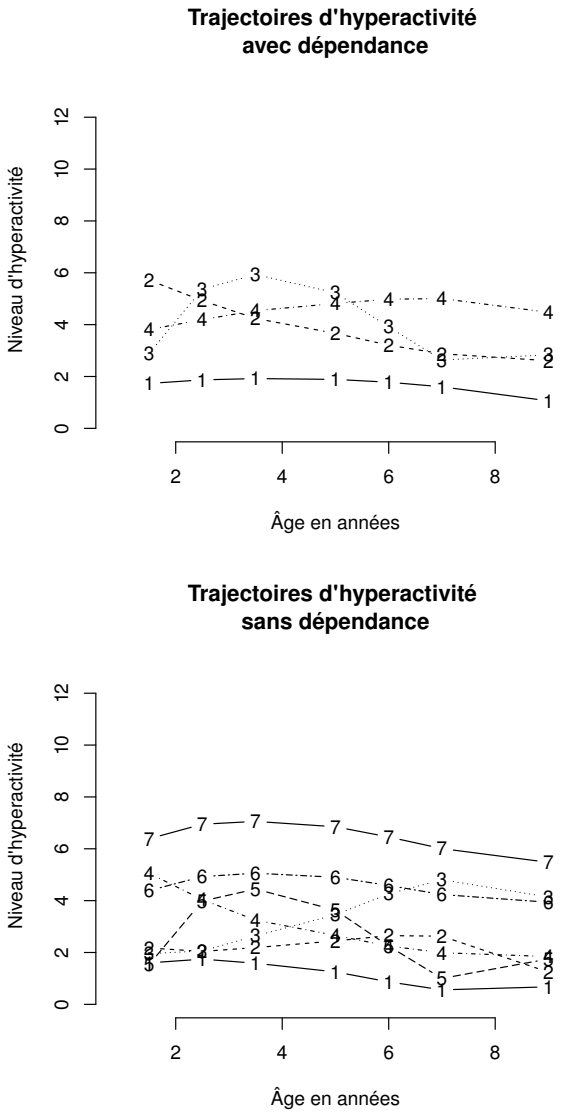



FIGURE 3.8. Moyennes prédites d'hyperactivité de 1,5 à 9 ans

paramètres estimés de localisation et de dispersion correspondants à la série d'équations (3.2.4) décrivant le modèle dans la section méthode. Le premier groupe, composé de 12,8% des participants, avait une variance de 1,24, ce qui était beaucoup plus petit lorsqu'on la comparait aux autres groupes. La trajectoire de ce groupe était presque constante avec un niveau relativement faible ($\approx$ 1,5). Le groupe 2, composé de 29,2% des participants, avait une variance de 3,92. La dispersion était donc beaucoup plus grande à l'intérieur de ce groupe. La trajectoire de ce groupe démarrait relativement haute à 1,5 ans ($\approx$ 5,75) puis

redescendait graduellement pour atteindre un niveau beaucoup plus bas à 9 ans ($\approx 3$). Le groupe 3, composé de 14,7% des participants, avait une variance de 3,87 ce qui était comparable au groupe précédent. Cette trajectoire suivait une courbe concave qui commençait à un niveau légèrement inférieur à 3, à 1,5 ans, puis augmentait pour atteindre un niveau de 6 à 3,5 ans et finalement redescendre pour atteindre un niveau comparable au groupe 2 ($\approx 3$). Le coefficient de corrélation dans les groupes 2 et 3 était respectivement de 0,56 et 0,68. Ces deux groupes démontraient une stabilité assez grande à travers le temps. Au contraire, les groupes 1 et 2 avec des coefficients de corrélation respectivement de 0,30 et 0,37 démontraient une stabilité relativement basse. Les corrélations étaient toutefois significativement différentes de 0 dans tous les groupes.

TABLEAU 3.11. Paramètres du mélange des trajectoires d'hyperactivité

| Paramètres | Valeurs Estimés | é. t. | Z | $P(\|Z\| > 0)$ | exp( V. E. ) |
|---|---|---|---|---|---|
| $\lambda_{01}$ | 0 | - | - | - | 1 |
| $\lambda_{11}$ | 0 | - | - | - | 1 |
| $\lambda_{02}$ | 0,401 | 0,532 | 0,754 | 0,451 | 1,493 |
| $\lambda_{12}$ | 1,394 | 0,095 | 14,685 | 0,000 | 4,031 |
| $\lambda_{03}$ | -0,045 | 0,394 | -0,115 | 0,909 | 0,956 |
| $\lambda_{13}$ | 0,775 | 0,092 | 8,460 | 0,000 | 2,171 |
| $\lambda_{04}$ | 0,345 | 0,385 | 0,894 | 0,371 | 1,412 |
| $\lambda_{14}$ | 2,200 | 0,064 | 34,590 | 0,000 | 9,025 |

TABLEAU 3.12. Probabilités conditionnelles d'appartenir aux trajectoires d'hyperactivité selon le sexe

| k | $P(g = k\|garçon)$ | $P(g = k\|fille)$ |
|---|---|---|
| 1 | 0,046 | 0,206 |
| 2 | 0,276 | 0,307 |
| 3 | 0,095 | 0,197 |
| 4 | 0,583 | 0,290 |

Le tableau 3.11 donne les paramètres du mélange. Le fait d'être un garçon augmentait de beaucoup la probabilité d'être dans les groupes 2 à 4 lorsqu'on la comparait au groupe 1, le groupe de référence. En utilisant le groupe 2 comme référence, les garçons avaient moins de chance d'être dans le groupe 3 mais plus d'être dans le groupe 4. Finalement, en utilisant le groupe 3 comme référence, les garçons avaient plus de chance de se trouver dans le groupe 4. Ces différentes paramétrisations ne sont pas rapportées puisqu'elles sont redondantes avec la paramétrisation présentée au tableau 3.11. Pour voir la répartition des probabilités d'appartenance à chacun des groupes conditionnellement au fait d'être un garçon ou une fille il s'agit de revenir à la définition de la probabilité définie à l'équation (3.2.3). Le tableau 3.12 donne les probabilités d'appartenance aux trajectoires conditionnellement au sexe du participant. On a pu constater que plus de la moitié des garçons se retrouvaient dans le groupe 4, un peu moins du tiers dans le groupe 2 et les autres dans les groupes 1 et 3. Les filles se répartissaient presque également dans les quatre groupes avec une probabilité légèrement plus grande de se retrouver dans les groupes 2 et 4.

### 3.2.3. Mémoire à court terme

#### 3.2.3.1. *Méthode*

La mémoire à court terme a été modélisée avec des courbes de croissance polynomiales cubiques. Pour modéliser la variance, une matrice de covariance avec corrélation à symétrie composée et variance hétérogène à travers le temps a été utilisée. La variance hétérogène a été introduite car plus les participants vieillissaient plus la variance tendait à diminuer. Voici la description plus formelle du modèle utilisé pour ajuster les données de mémoire à court terme :

$$E(y_{ij}|g=k) = \beta_{0k} + \beta_{1k}t_j + \beta_{2k}t_j^2 + \beta_{3k}t_j^3,$$

$$\mathrm{Cov}\left[\left.\begin{pmatrix} y_{i1} \\ \vdots \\ y_{i5} \end{pmatrix}\right| g=k\right] = \begin{bmatrix} \sigma_{1k}^2 & \sigma_{1k}\sigma_{2k}\rho_k & \dots & \sigma_{1k}\sigma_{5k}\rho_k \\ \sigma_{1k}\sigma_{2k}\rho_k & \sigma_{2k}^2 & \dots & \sigma_{2k}\sigma_{5k}\rho_k \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1k}\sigma_{5k}\rho_k & \sigma_{2k}\sigma_{5k}\rho_k & \dots & \sigma_{5k}^2 \end{bmatrix}, \quad (3.2.5)$$

où $y_{ij}$ représente la mesure de mémoire à court terme du participant $i$ au temps $t_j$, $j = 1, 2, \ldots, 5$ et $g = 1, \ldots, G$ indique les G différentes trajectoires.

Concernant l'estimation des $\pi_{ij}$, c'est-à-dire la probabilité pour le participant $i$ d'appartenir au groupe $j$, nous supposons qu'*a priori* tous les participants ont la même chance d'être dans un groupe ou l'autre. On parle donc des probabilités $\pi_j$. Voici la description du modèle :

$$\pi_j = \frac{e^{\lambda_j}}{\sum_{j=1}^{G} e^{\lambda_j}}, \text{ où } \lambda_1 = 0. \tag{3.2.6}$$

3.2.3.2. *Résultats*

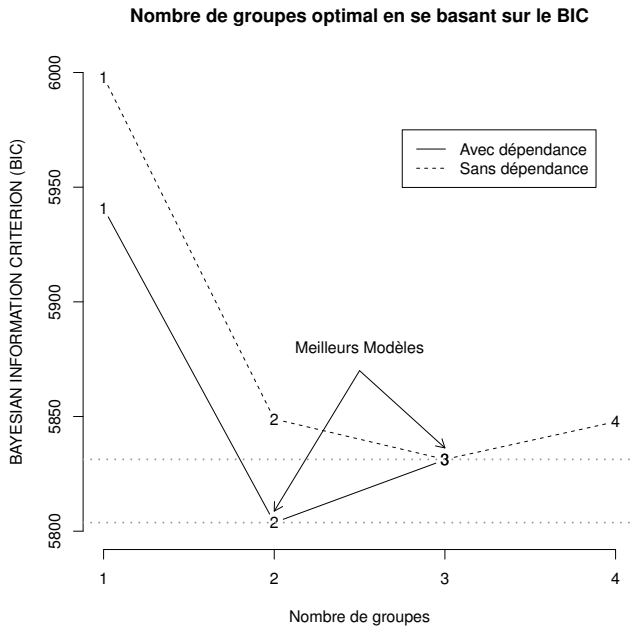



FIGURE 3.9. Sélection du meilleur modèle (selon le BIC)

La figure 3.9 montre le processus de sélection du meilleur modèle. Le meilleur modèle en supposant la dépendance était le modèle à deux trajectoires. En supposant l'indépendance, le meilleur modèle était le modèle à trois trajectoires. La figure 3.10 montre les moyennes prédites de la mémoire à court terme pour les deux trajectoires de 3,5 à 9 ans supposant la dépendance *vs* les trois trajectoires supposant l'indépendance. Le modèle à deux groupes démontrait un meilleur ajustement en se basant sur le BIC (5803,749 *vs* 5831,305), nous avons

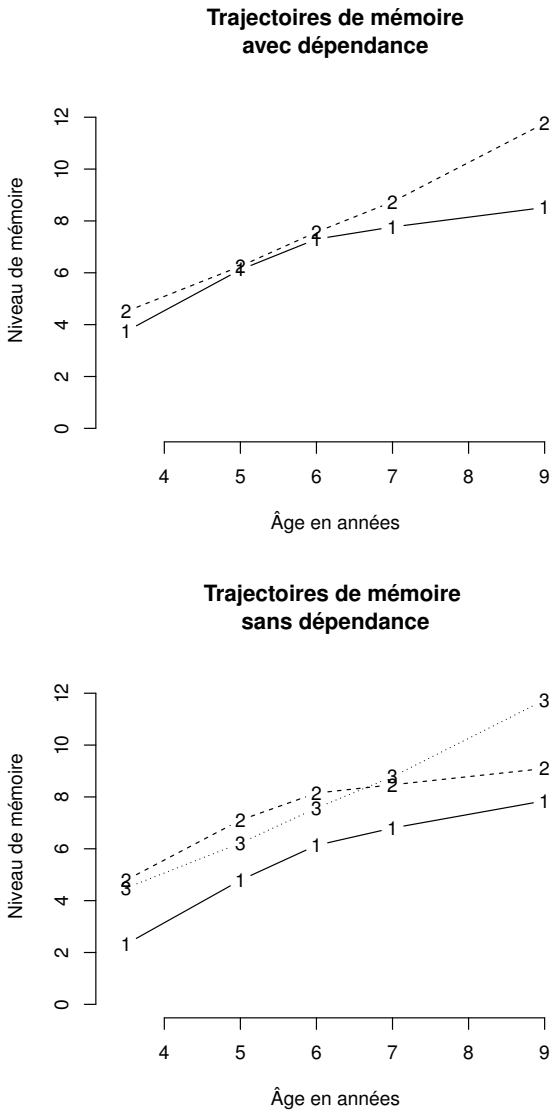



FIGURE 3.10. Moyennes prédites de mémoire à court terme de 3,5 à 9 ans

donc retenu ce modèle. De plus ce modèle nécessitait 8 paramètres de moins : 21 paramètres en supposant la dépendance *vs* 29 paramètres en supposant l'indépendance. Le tableau 3.13 donne les valeurs des paramètres estimés de localisation et de dispersion correspondant à la série d'équations (3.2.5) décrivant le modèle dans la section méthode. Le premier groupe, composé de 33,7% des participants contenait une trajectoire curviligne qui démarrait à un niveau autour de 4 lorsque le participant avait 3,5 ans puis suivait une courbe concave

TABLEAU 3.13. Paramètres de localisation et de dispersion des trajectoires de mémoire à court terme

| Paramètres | Valeurs Estimés | é. t. | Z | P(\|Z\| > 0) |
|---|---|---|---|---|
| Groupe 1 (66,3%) : Trajectoire curviligne | | | | |
| $\beta_{01}$ | 3,745 | 0,182 | 20,567 | 0,000 |
| $\beta_{11}$ | 3,135 | 0,353 | 8,890 | 0,000 |
| $\beta_{21}$ | -0,841 | 0,190 | -4,429 | 0,000 |
| $\beta_{31}$ | 0,081 | 0,025 | 3,215 | 0,001 |
| $\sigma^2_{11}$ | 5,048 | 0,589 | 8,572 | 0,000 |
| $\sigma^2_{21}$ | 5,216 | 0,544 | 9,584 | 0,000 |
| $\sigma^2_{31}$ | 5,319 | 0,550 | 9,679 | 0,000 |
| $\sigma^2_{41}$ | 4,535 | 0,511 | 8,874 | 0,000 |
| $\sigma^2_{51}$ | 2,217 | 0,281 | 7,899 | 0,000 |
| $\rho_1$ | 0,215 | 0,039 | 5,551 | 0,000 |
| Groupe 2 (33,7%) : Trajectoire linéaire | | | | |
| $\beta_{02}$ | 4,509 | 0,318 | 14,197 | 0,000 |
| $\beta_{12}$ | 2,052 | 0,586 | 3,504 | 0,000 |
| $\beta_{22}$ | -0,355 | 0,304 | -1,169 | 0,243 |
| $\beta_{32}$ | 0,047 | 0,040 | 1,169 | 0,242 |
| $\sigma^2_{12}$ | 7,192 | 1,111 | 6,471 | 0,000 |
| $\sigma^2_{22}$ | 4,224 | 0,669 | 6,317 | 0,000 |
| $\sigma^2_{32}$ | 5,288 | 0,824 | 6,420 | 0,000 |
| $\sigma^2_{42}$ | 4,620 | 0,740 | 6,242 | 0,000 |
| $\sigma^2_{52}$ | 0,027 | 0,006 | 4,628 | 0,000 |
| $\rho_2$ | 0,117 | 0,051 | 2,293 | 0,022 |

jusqu'à atteindre un niveau de 8 à l'âge de 9 ans. Le deuxième groupe, suivait une droite qui démarrait à un niveau autour de 4,5 et se terminait à un niveau autour de 11,5 à 9 ans. La variance décroissait à travers le temps pour les deux groupes. La variance du groupe 2 à 9 ans était très faible avec une valeur de 0,027. Cette faible variance indique probablement un effet plafond

pour les participants qui étaient dans le groupe 2. Un effet plafond survient lorsque le niveau maximum possible à une tâche est atteint par un nombre élevé de participants, c'est-à-dire que le vrai potentiel de ces participants n'a pas été mesuré.

### 3.2.4. Vocabulaire réceptif de 3,5 à 8 ans

#### 3.2.4.1. *Méthode*

Le vocabulaire réceptif a été modélisé avec des courbes de croissance polynomiale cubique. Pour modéliser la variance, une matrice de covariance avec corrélation à symétrie composée et variance hétérogène à travers le temps a été utilisée. Voici la description plus formelle du modèle utilisé pour ajuster les données de vocabulaire réceptif :

$$E(y_{ij}|g=k) = \beta_{0k} + \beta_{1k}t_j + \beta_{2k}t_j^2 + \beta_{3k}t_j^3,$$

$$\text{Cov}\left[\left.\begin{pmatrix} y_{i1} \\ \vdots \\ y_{i5} \end{pmatrix}\right| g=k\right] = \begin{bmatrix} \sigma_{1k}^2 & \sigma_{1k}\sigma_{2k}\rho_k & \dots & \sigma_{1k}\sigma_{5k}\rho_k \\ \sigma_{1k}\sigma_{2k}\rho_k & \sigma_{2k}^2 & \dots & \sigma_{2k}\sigma_{5k}\rho_k \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1k}\sigma_{5k}\rho_k & \sigma_{2k}\sigma_{5k}\rho_k & \dots & \sigma_{5k}^2 \end{bmatrix}, \quad (3.2.7)$$

où $y_{ij}$ représente la mesure de vocabulaire réceptif du participant $i$ au temps $t_j$, $j = 1, 2, \dots, 5$ et $g$ indique les $G$ différentes trajectoires.

Concernant l'estimation des $\pi_{ij}$, c'est-à-dire la probabilité pour le participant $i$ d'appartenir au groupe $j$, nous supposons que tous les participants ont la même chance d'appartenir à un des groupes. On parle donc des probabilités $\pi_j$. Voici la description du modèle :

$$\pi_j = \frac{e^{\lambda_j}}{\sum_{j=1}^{G} e^{\lambda_j}}, \text{ où } \lambda_1 = 0. \quad (3.2.8)$$

#### 3.2.4.2. *Résultats*

La figure 3.11 montre le processus de sélection du meilleur modèle. Le meilleur modèle en supposant la dépendance était le modèle à une trajectoire. Par contre, en supposant l'indépendance, le meilleur modèle était le modèle à

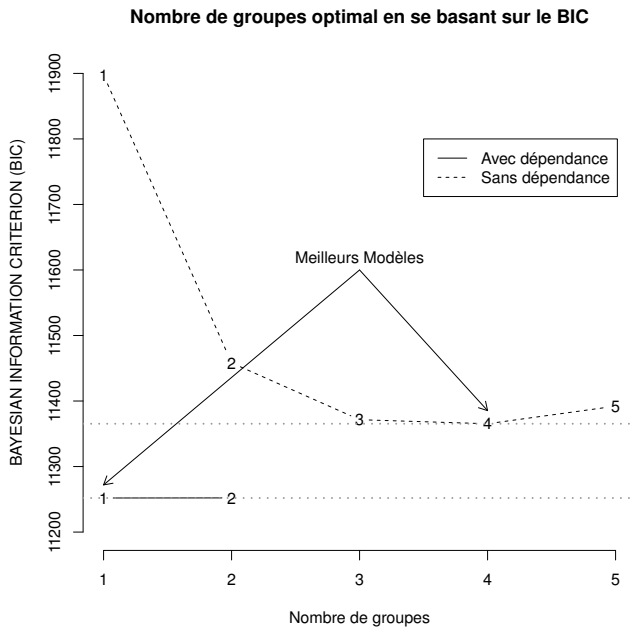



FIGURE 3.11. Sélection du meilleur modèle (selon le BIC)

quatre groupes. La figure 3.12 montre les moyennes prédites du vocabulaire réceptif pour la trajectoire de 3,5 à 8 ans en supposant la dépendance *vs* les quatre trajectoires supposant l'indépendance. Le modèle à un groupe démontrait un meilleur ajustement en se basant sur le BIC (10101,13 *vs* 10215,03), nous avons donc retenu ce modèle. De plus ce modèle nécessitait 29 paramètres de moins : 10 paramètres en supposant la dépendance *vs* 39 paramètres en supposant l'indépendance. Le tableau 3.14 montre les valeurs des paramètres estimés de localisation et de dispersion correspondants à la série d'équations (3.2.7) décrivant le modèle. La seule trajectoire identifiée est presque linéaire avec une petite incurvation vers le bas. Le niveau de vocabulaire réceptif à 3,5 ans est autour de 30 et augmente jusqu'à un niveau avoisinant 110 à 8 ans. Les estimations de variances fluctuent beaucoup à travers le temps. Le coefficient de corrélation $\rho_1$ est relativement élevé et constant. Ceci démontre une bonne stabilité de cette mesure. Une implication pratique de cette analyse est que l'on pourrait mesurer les habiletés de vocabulaire réceptif seulement une année sur deux afin de limiter les coûts des collectes de données.

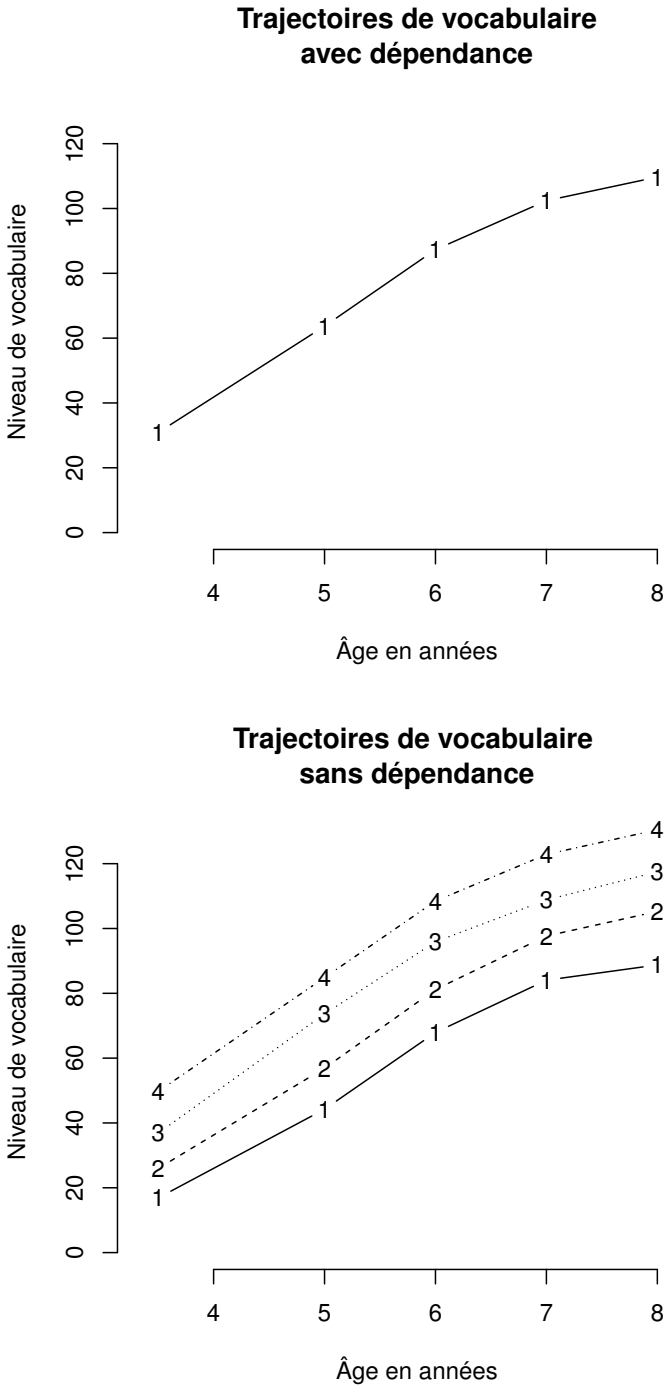



FIGURE 3.12. Moyennes prédites du vocabulaire réceptif de 3,5 à 8 ans

TABLEAU 3.14. Paramètres de localisation et de dispersion de la trajectoire du vocabulaire réceptif

| Paramètres | Valeurs Estimés | é. t. | Z | $P(\|Z\| > 0)$ |
|---|---|---|---|---|
| Groupe 1 (66,3%) : Trajectoire curviligne | | | | |
| $\beta_{01}$ | 30,866 | 0,866 | 35,636 | 0,000 |
| $\beta_{11}$ | 37,657 | 1,604 | 23,474 | 0,000 |
| $\beta_{21}$ | -4,951 | 1,034 | -4,789 | 0,000 |
| $\beta_{31}$ | 0,116 | 0,169 | 0,686 | 0,492 |
| $\sigma^2_{11}$ | 218,726 | 18,512 | 11,815 | 0,000 |
| $\sigma^2_{21}$ | 342,156 | 29,003 | 11,797 | 0,000 |
| $\sigma^2_{31}$ | 201,790 | 16,803 | 12,009 | 0,000 |
| $\sigma^2_{41}$ | 177,821 | 15,077 | 11,794 | 0,000 |
| $\sigma^2_{51}$ | 214,577 | 19,020 | 11,281 | 0,000 |
| $\rho_1$ | 0,596 | 0,026 | 22,836 | 0,000 |

# Chapitre 4

## DISCUSSION

Le but principal de cette étude était de tester la robustesse de l'estimation du nombre de groupes dans le cadre des modèles de mélanges de loi lorsque la structure de covariance était mal spécifiée. La conclusion principale est qu'une mauvaise spécification de la matrice de covariance induit un biais positif dans l'estimation du nombre de groupe, c'est-à-dire on estime plus de groupes qu'il n'en existe réellement. Ceci était surtout vrai lorsqu'on utilisait une structure de covariance simplifiée pour modéliser des données ayant une structure de covariance plus complexe, par exemple supposer que la variance est la même pour tous les temps de mesure, lorsqu'en fait la variance est hétérogène.

Nous nous somme intéressés plus précisément à l'hypothèse d'indépendance des observations d'un individu à travers le temps lorsque les observations sont en fait corrélées. Nous avons constaté, autant dans les simulations du chapitre 2 que dans les analyses sur des données réelles du chapitre 3, que cette supposition peut conduire à estimer beaucoup plus de groupes qu'il n'en existe réellement. Cette supposition d'indépendance a été distinguée des autres cas car elle est commune dans les articles qui décrivent ce type d'analyse.

Bauer (2007) décrivait déjà dans son article des problèmes de surestimation du nombre de groupes en présence d'une matrice de covariance mal spécifiée toutefois cette étude illustre plus en détail, l'ampleur du problème, à l'aide de simulations et d'exemples pratiques. Cette étude indique aussi les cas où la

surestimation du nombre de groupes était plus grave et les cas où elle était moins grave.

La première implication pratique de cette étude est qu'il est préférable de ne pas supposer l'indépendance lorsqu'on fait des analyses de mélange. Cette recommandation prend tout son sens lorsqu'on sait que dans un devis longitudinal les observations sont presque toujours corrélées à travers le temps. Cette modélisation de la structure de covariance peut se faire en utilisant des coefficients aléatoires, par exemple les modèles mixtes, ou en modélisant directement la covariance des observations, telle qu'utilisée dans cette étude. Il est d'ailleurs surprenant de constater que cette supposition d'indépendance soit si commune dans les analyses de mélanges. En effet, il serait presque impensable de faire une analyse de courbes de croissance en supposant l'indépendance des observations à travers le temps. Pourtant ce type d'analyse est un cas particulier d'un mélange de lois dans lequel nous n'avons qu'un seul groupe. Ce cas de figure s'est d'ailleurs produit lors de l'analyse des données du vocabulaire réceptif à la section 3.2.4.2. En supposant la dépendance des observations à travers le temps nous obtenions un mélange à un seul groupe alors qu'en supposant l'indépendance nous obtenions quatre groupes.

La deuxième implication pratique de cette étude est qu'en modélisant la structure de covariance à l'intérieur des groupes nous obtenons plus d'information sur ces groupes. En effet, le niveau de corrélation peut être interprété comme un indicateur de stabilité dans le temps. La variance est aussi un facteur intéressant à interpréter dans le cadre des modèles de mélanges. En effet, un groupe avec une variance très faible peut être interprété comme un groupe ayant des observations très similaires, donc plus homogènes, alors qu'au contraire un groupe ayant des variances plus larges pourrait être vu comme un groupe contenant des caractéristiques plus variées.

Cette étude comporte certaines limites. Dans les simulations toutes les données ont été générées à partir d'un seul groupe afin de montrer qu'on estimait plus de trajectoires lorsque la covariance était mal spécifiée. Nous supposons que de mal spécifier la covariance des données provenant d'un mélange à plus d'un groupe aurait des implications similaires à ce que nous avons trouvé, toutefois ce type de cas doit être vérifié dans des simulations. De plus, nous n'avons effectué les simulations que sur des structures de covariance assez simples que l'on rencontre souvent dans la littérature. Finalement, toutes les analyses se sont basées uniquement sur le critère du BIC pour établir le meilleur modèle. Il existe d'autres critères pour établir le nombre de groupes toutefois ce critère est nettement le plus utilisé dans la littérature.

Les futures recherches devraient tenter de mettre au point des outils plus efficaces afin de mieux estimer le nombre de groupes d'un mélange. De plus, les utilisateurs des modèles de mélange devraient prendre conscience que ceux-ci sont très sensibles à la spécification du modèle. Finalement, il faudrait pouvoir généraliser les mélanges à d'autres types de lois multivariées qui ne supposent pas nécessairement l'indépendance des observations à travers le temps.

# Bibliographie


Bauer, D. J. (2007). Observations on the use of growth mixture models in psychological research. *Multivariate Behavioral Research*, 42(4) :757–786.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3) :297–334.

Cronbach, L. J. et Richard, J. S. (2004). My current thoughts on coefficient alpha and successor procedures. *Educational and Psychological Measurement*, 64(3) :391–418.

Dempster, A. P., Laird, N. M., et Rubin, D. B. (1977). Maximum likelihood from incomplande data via the em algorithm. *Journal of the Royal Statistical Sociandy. Series B (Methodological)*, 39(1) :1–38.

Dobson, A. J. (1990). *An introduction to generalized linear models*. Texts in Statistical Science. Chapman and Hall, New York : 1990., London, 1st edition edition.

Dunn, L. et Dunn, L. (1981). PPVT : Peabody picture vocabulary test-revised : Manual for forms L and M.

Dunn, L., Thériault-Whalen, C., et Dunn, L. (1993). Échelle de Vocabulaire en Images Peabody. adaptation française du Peabody Picture Vocabulary Test-Revised. manuel pour les formes A and B.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(Part II) :179–188.

Giguère, C.-E. (2011). *MMELN : Estimation of multinormal mixture distribution* R package version 1.0.

Jetté, M., Desrosiers, H., et Tremblay, R. E. (1997). "En 2001...j'aurai 5 ans !", enquête auprès des bébés de 5 mois, rapport préliminaire de l'Étude longitudinale du développement des enfants du québec. Technical report, Gouvernement du Québec.

Jones, B. L., Nagin, D. S., et Roeder, K. (2001). A sas procedure based on mixture models for estimating developmental trajectories. *Sociological Methods and Research*, 29(3) :374–393.

Lindstrom, M. J. et Bates, D. M. (1988). Newton-raphson and em algorithms for linear mixed-effects models for repeated-measures data. *Journal of the American Statistical Association*, 83(404) :1014–1022.

McLachlan, G. J. et Peel, D. (1998). Mixfit : An algorithm for the automatic fitting and testing of normal mixture models. In Jain, A. K., Venkatesh, S., and Lovell, B. C., editors, *Fourteenth International Conference on Pattern Recognition, Vols 1 and 2*, International Conference on Pattern Recognition, pages 553–557. IEEE Computer Soc, Los Alamitos.

McLachlan, G. J. et Peel, D. (2000). *Finite mixture models*. Wiley series in probability and statistics. Wiley, New York [u.a.].

Muthén, B., Brown, C. H., Masyn, K., Jo, B., Khoo, S. T., Yang, C. C., Wang, C. P., Kellam, S. G., Carlin, J. B., et Liao, J. (2002). General growth mixture modeling for randomized preventive interventions. *Biostatistics*, 3(4) :459–475.

Muthén, B. et Muthén, L. K. (2000). Integrating person-centered and variable-centered analyses : Growth mixture modeling with latent trajectory classes. *Alcoholism-Clinical and Experimental Research*, 24(6) :882–891.

Muthén, B. et Shedden, K. (1999). Finite mixture modeling with mixture outcomes using the em algorithm. *Biometrics*, 55(2) :463–469.

Nagin, D. S. (1999). Analysing developmental trajectories : A semiparametric, group-based approach. *Psychological Methods*, 4(2) :139–157.

Pearson, K. (1894). Contributions to the theory of mathematical evolution. *Philosophical Transactions of the Royal Society of London A*, 185 :71–110.

Pearson, K. (1895). Contributions to the theory of mathematical evolution, ii : skew variation. *Philosophical Transactions of the Royal Society of London A*, 186 :343–414.

Pinheiro, J. C. et Bates, D. M. (1996). Unconstrained paramandrizations for variance-covariance matrices. *Statistics and Computing*, 6(3) :289–296.

Tremblay, R. E., Vitaro, F., Gagnon, C., Piché, C., et Royer, N. (1992). A prosocial scale for the preschool behavior questionnaire - concurrent and predictive correlates. *International Journal of Behavioral Development*, 15(2) :227–245.

Zelazo, P. D., Jacques, S., Burack, J. A., et Frye, D. (2002). The relation bandween theory of mind and rule use : Evidence from persons with autism-spectrum disorders. *Infant and Child Development*, 11(2) :171–195.

# Annexe A

## DÉTAILS DE LA SIMULATION

Les tableaux A.1-A.8 montrent les résultats complets des simulations. Dans tous les tableaux, p réfère au nombre de mesures répétées, c'est-à-dire le nombre de colonnes de la matrice des observations, et N à la taille échantillonnale, c'est-à-dire le nombre de lignes de la matrice des observations. G réfère au nombre de trajectoires trouvées lors de l'analyse des données simulées. Nous donnons la valeur de (G-1) afin d'illustrer le nombre de trajectoires supplémentaires qui ont été trouvées. Le paramètre ρ représente le niveau de corrélation utilisé pour simuler les données. Finalement, Cov représente la structure de covariance supposée dans les analyses.

TABLEAU A.1. Résultats pour des données générées avec une covariance à symétrie composée (CS) et p=3.

| | | N=200 | | | N=400 | | | N=600 | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | ρ | | | ρ | | | ρ | | |
| Cov | (G-1) | 0 | 0,3 | 0,6 | 0 | 0,3 | 0,6 | 0 | 0,3 | 0,6 |
| CS | 0 | 20 | 20 | 19 | 20 | 20 | 20 | 20 | 20 | 20 |
| | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| UCS | 0 | 20 | 19 | 16 | 20 | 19 | 19 | 20 | 17 | 20 |
| | 1 | 0 | 1 | 4 | 0 | 1 | 1 | 0 | 3 | 0 |
| AR1 | 0 | 20 | 20 | 20 | 20 | 18 | 14 | 20 | 14 | 4 |
| | 1 | 0 | 0 | 0 | 0 | 2 | 6 | 0 | 6 | 10 |
| | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 6 |
| UAR1 | 0 | 19 | 17 | 15 | 20 | 18 | 16 | 20 | 19 | 9 |
| | 1 | 1 | 3 | 5 | 0 | 2 | 4 | 0 | 1 | 10 |
| | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| IND | 0 | 18 | 0 | 0 | 19 | 0 | 0 | 20 | 0 | 0 |
| | 1 | 2 | 17 | 0 | 1 | 16 | 0 | 0 | 11 | 0 |
| | 2 | 0 | 3 | 16 | 0 | 4 | 3 | 0 | 9 | 0 |
| | 3 | 0 | 0 | 4 | 0 | 0 | 14 | 0 | 0 | 13 |
| | 4 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 7 |
| | 5 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| UIND | 0 | 18 | 1 | 0 | 20 | 0 | 0 | 20 | 0 | 0 |
| | 1 | 2 | 17 | 1 | 0 | 20 | 0 | 0 | 18 | 0 |
| | 2 | 0 | 2 | 12 | 0 | 0 | 11 | 0 | 2 | 6 |
| | 3 | 0 | 0 | 7 | 0 | 0 | 9 | 0 | 0 | 13 |
| | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |

Note : Cov réfère au type de covariance supposée. G réfère au nombre de trajectoires estimées.

TABLEAU A.2. Résultats pour des données générées avec une covariance à symétrie composée avec variance hétérogène (UCS) et p=3.

| | | N=200 | | | N=400 | | | N=600 | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | ρ | | | ρ | | | ρ | | |
| Cov | (G-1) | 0 | 0,3 | 0,6 | 0 | 0,3 | 0,6 | 0 | 0,3 | 0,6 |
| CS | 0 | 11 | 4 | 0 | 2 | 0 | 0 | 0 | 0 | 0 |
| | 1 | 9 | 14 | 18 | 18 | 17 | 12 | 20 | 15 | 4 |
| | 2 | 0 | 2 | 2 | 0 | 3 | 8 | 0 | 5 | 16 |
| UCS | 0 | 20 | 18 | 17 | 20 | 19 | 20 | 20 | 20 | 19 |
| | 1 | 0 | 2 | 3 | 0 | 1 | 0 | 0 | 0 | 1 |
| AR1 | 0 | 12 | 4 | 1 | 2 | 0 | 0 | 0 | 0 | 0 |
| | 1 | 8 | 16 | 12 | 18 | 19 | 2 | 20 | 15 | 2 |
| | 2 | 0 | 0 | 7 | 0 | 1 | 12 | 0 | 5 | 3 |
| | 3 | 0 | 0 | 0 | 0 | 0 | 6 | 0 | 0 | 15 |
| UAR1 | 0 | 19 | 17 | 15 | 20 | 19 | 15 | 19 | 19 | 9 |
| | 1 | 1 | 3 | 5 | 0 | 1 | 5 | 1 | 1 | 4 |
| | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 7 |
| IND | 0 | 7 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| | 1 | 12 | 14 | 0 | 18 | 4 | 0 | 20 | 1 | 0 |
| | 2 | 1 | 4 | 9 | 1 | 15 | 2 | 0 | 18 | 0 |
| | 3 | 0 | 2 | 10 | 0 | 1 | 13 | 0 | 1 | 14 |
| | 4 | 0 | 0 | 1 | 0 | 0 | 5 | 0 | 0 | 6 |
| UIND | 0 | 20 | 2 | 0 | 20 | 0 | 0 | 20 | 0 | 0 |
| | 1 | 0 | 18 | 0 | 0 | 20 | 0 | 0 | 18 | 0 |
| | 2 | 0 | 0 | 17 | 0 | 0 | 8 | 0 | 2 | 1 |
| | 3 | 0 | 0 | 3 | 0 | 0 | 11 | 0 | 0 | 19 |
| | 4 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |

Note : Cov réfère au type de covariance supposée. G réfère au nombre de trajectoires estimées.

TABLEAU A.3. Résultats pour des données générées avec une covariance auto-régressive d'ordre 1 (AR1) et p=3.

| | | N=200 | | | N=400 | | | N=600 | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | ρ | | | ρ | | | ρ | | |
| Cov | (G-1) | 0 | 0,3 | 0,6 | 0 | 0,3 | 0,6 | 0 | 0,3 | 0,6 |
| CS | 0 | 20 | 19 | 7 | 20 | 18 | 0 | 20 | 15 | 2 |
| | 1 | 0 | 1 | 12 | 0 | 2 | 20 | 0 | 5 | 18 |
| | 2 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| UCS | 0 | 20 | 17 | 11 | 20 | 19 | 2 | 20 | 18 | 0 |
| | 1 | 0 | 3 | 9 | 0 | 1 | 16 | 0 | 2 | 20 |
| | 2 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 |
| AR1 | 0 | 20 | 20 | 20 | 20 | 19 | 20 | 20 | 20 | 20 |
| | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| UAR1 | 0 | 20 | 15 | 14 | 19 | 15 | 18 | 20 | 19 | 20 |
| | 1 | 0 | 5 | 6 | 1 | 5 | 2 | 0 | 1 | 0 |
| IND | 0 | 20 | 4 | 0 | 20 | 0 | 0 | 20 | 0 | 0 |
| | 1 | 0 | 13 | 3 | 0 | 20 | 0 | 0 | 14 | 0 |
| | 2 | 0 | 3 | 10 | 0 | 0 | 8 | 0 | 6 | 5 |
| | 3 | 0 | 0 | 6 | 0 | 0 | 8 | 0 | 0 | 6 |
| | 4 | 0 | 0 | 1 | 0 | 0 | 3 | 0 | 0 | 9 |
| | 5 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| UIND | 0 | 20 | 7 | 0 | 20 | 0 | 0 | 20 | 0 | 0 |
| | 1 | 0 | 11 | 2 | 0 | 20 | 0 | 0 | 18 | 0 |
| | 2 | 0 | 2 | 16 | 0 | 0 | 17 | 0 | 2 | 12 |
| | 3 | 0 | 0 | 2 | 0 | 0 | 3 | 0 | 0 | 8 |

Note : Cov réfère au type de covariance supposée. G réfère au nombre de trajectoires estimées.

TABLEAU A.4. Résultats pour des données générées avec une covariance auto-régressive d'ordre 1 avec une variance hétérogène (UAR1) et p=3.

| | | N=200 | | | N=400 | | | N=600 | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | ρ | | | ρ | | | ρ | | |
| Cov | (G-1) | 0 | 0,3 | 0,6 | 0 | 0,3 | 0,6 | 0 | 0,3 | 0,6 |
| CS | 0 | 15 | 4 | 0 | 2 | 0 | 0 | 0 | 0 | 0 |
| | 1 | 5 | 14 | 14 | 18 | 20 | 11 | 20 | 16 | 9 |
| | 2 | 0 | 2 | 6 | 0 | 0 | 9 | 0 | 4 | 11 |
| UCS | 0 | 19 | 16 | 18 | 20 | 20 | 7 | 20 | 19 | 1 |
| | 1 | 1 | 4 | 2 | 0 | 0 | 12 | 0 | 1 | 19 |
| | 2 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| AR1 | 0 | 11 | 5 | 0 | 3 | 0 | 0 | 0 | 0 | 0 |
| | 1 | 9 | 14 | 20 | 17 | 20 | 18 | 20 | 17 | 15 |
| | 2 | 0 | 1 | 0 | 0 | 0 | 2 | 0 | 3 | 5 |
| UAR1 | 0 | 19 | 17 | 17 | 19 | 18 | 18 | 20 | 18 | 18 |
| | 1 | 1 | 2 | 3 | 1 | 2 | 2 | 0 | 2 | 2 |
| | 2 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| IND | 0 | 5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | 1 | 15 | 15 | 0 | 19 | 9 | 0 | 19 | 3 | 0 |
| | 2 | 0 | 5 | 13 | 1 | 9 | 7 | 1 | 15 | 2 |
| | 3 | 0 | 0 | 6 | 0 | 2 | 11 | 0 | 2 | 15 |
| | 4 | 0 | 0 | 1 | 0 | 0 | 2 | 0 | 0 | 3 |
| UIND | 0 | 20 | 7 | 0 | 20 | 1 | 0 | 20 | 0 | 0 |
| | 1 | 0 | 11 | 5 | 0 | 19 | 0 | 0 | 20 | 0 |
| | 2 | 0 | 2 | 11 | 0 | 0 | 12 | 0 | 0 | 15 |
| | 3 | 0 | 0 | 4 | 0 | 0 | 8 | 0 | 0 | 5 |

Note : Cov réfère au type de covariance supposée. G réfère au nombre de trajectoires estimées.

TABLEAU A.5. Résultats pour des données générées avec une covariance à symétrie composée (CS) et p=5.

| | | N=200 | | | N=400 | | | N=600 | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | ρ | | | ρ | | | ρ | | |
| Cov | (G-1) | 0 | 0,3 | 0,6 | 0 | 0,3 | 0,6 | 0 | 0,3 | 0,6 |
| CS | 0 | 20 | 20 | 20 | 20 | 20 | 20 | 20 | 20 | 20 |
| UCS | 0 | 20 | 18 | 17 | 20 | 20 | 20 | 20 | 20 | 20 |
| | 1 | 0 | 2 | 3 | 0 | 0 | 0 | 0 | 0 | 0 |
| AR1 | 0 | 20 | 1 | 0 | 20 | 0 | 0 | 20 | 0 | 0 |
| | 1 | 0 | 19 | 5 | 0 | 19 | 0 | 0 | 13 | 0 |
| | 2 | 0 | 0 | 15 | 0 | 1 | 20 | 0 | 7 | 20 |
| UAR1 | 0 | 20 | 10 | 2 | 20 | 1 | 0 | 20 | 0 | 0 |
| | 1 | 0 | 10 | 7 | 0 | 18 | 0 | 0 | 12 | 0 |
| | 2 | 0 | 0 | 11 | 0 | 1 | 16 | 0 | 8 | 13 |
| | 3 | 0 | 0 | 0 | 0 | 0 | 4 | 0 | 0 | 7 |
| IND | 0 | 18 | 0 | 0 | 20 | 0 | 0 | 20 | 0 | 0 |
| | 1 | 2 | 7 | 0 | 0 | 1 | 0 | 0 | 1 | 0 |
| | 2 | 0 | 12 | 0 | 0 | 19 | 0 | 0 | 17 | 0 |
| | 3 | 0 | 1 | 11 | 0 | 0 | 4 | 0 | 2 | 1 |
| | 4 | 0 | 0 | 9 | 0 | 0 | 14 | 0 | 0 | 14 |
| | 5 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 5 |
| UIND | 0 | 20 | 0 | 0 | 20 | 0 | 0 | 20 | 0 | 0 |
| | 1 | 0 | 18 | 0 | 0 | 9 | 0 | 0 | 1 | 0 |
| | 2 | 0 | 2 | 7 | 0 | 11 | 1 | 0 | 19 | 0 |
| | 3 | 0 | 0 | 13 | 0 | 0 | 16 | 0 | 0 | 8 |
| | 4 | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 0 | 12 |

Note : Cov réfère au type de covariance supposée. G réfère au nombre de trajectoires estimées.

TABLEAU A.6. Résultats pour des données générées avec une covariance à symétrie composée avec variance hétérogène (UCS) et p=5.

| | | N=200 | | | N=400 | | | N=600 | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | ρ | | | ρ | | | ρ | | |
| Cov | (G-1) | 0 | 0,3 | 0,6 | 0 | 0,3 | 0,6 | 0 | 0,3 | 0,6 |
| CS | 0 | 12 | 2 | 0 | 2 | 0 | 0 | 0 | 0 | 0 |
| | 1 | 8 | 16 | 11 | 18 | 17 | 0 | 19 | 5 | 0 |
| | 2 | 0 | 2 | 9 | 0 | 3 | 20 | 1 | 15 | 12 |
| | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 8 |
| UCS | 0 | 19 | 13 | 14 | 20 | 18 | 17 | 19 | 20 | 18 |
| | 1 | 1 | 7 | 6 | 0 | 2 | 3 | 1 | 0 | 2 |
| AR1 | 0 | 11 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 |
| | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | 2 | 9 | 20 | 20 | 17 | 20 | 20 | 20 | 20 | 20 |
| UAR1 | 0 | 20 | 5 | 0 | 20 | 0 | 0 | 20 | 0 | 0 |
| | 1 | 0 | 15 | 7 | 0 | 18 | 0 | 0 | 14 | 0 |
| | 2 | 0 | 0 | 12 | 0 | 2 | 14 | 0 | 6 | 4 |
| | 3 | 0 | 0 | 1 | 0 | 0 | 6 | 0 | 0 | 16 |
| IND | 0 | 6 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| | 1 | 14 | 3 | 0 | 19 | 0 | 0 | 18 | 0 | 0 |
| | 2 | 0 | 16 | 1 | 0 | 18 | 0 | 2 | 16 | 0 |
| | 3 | 0 | 1 | 11 | 0 | 2 | 3 | 0 | 4 | 1 |
| | 4 | 0 | 0 | 8 | 0 | 0 | 14 | 0 | 0 | 14 |
| | 5 | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 0 | 5 |
| UIND | 0 | 20 | 0 | 0 | 20 | 0 | 0 | 20 | 0 | 0 |
| | 1 | 0 | 18 | 0 | 0 | 13 | 0 | 0 | 4 | 0 |
| | 2 | 0 | 2 | 6 | 0 | 7 | 0 | 0 | 16 | 0 |
| | 3 | 0 | 0 | 12 | 0 | 0 | 16 | 0 | 0 | 12 |
| | 4 | 0 | 0 | 2 | 0 | 0 | 4 | 0 | 0 | 8 |

Note : Cov réfère au type de covariance supposée. G réfère au nombre de trajectoires estimées.

TABLEAU A.7. Résultats pour des données générées avec une covariance auto-régressive d'ordre 1 (AR1) et p=5.

| | | N=200 | | | N=400 | | | N=600 | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | ρ | | | ρ | | | ρ | | |
| Cov | (G-1) | 0 | 0,3 | 0,6 | 0 | 0,3 | 0,6 | 0 | 0,3 | 0,6 |
| CS | 0 | 20 | 16 | 0 | 20 | 1 | 0 | 20 | 1 | 0 |
| | 1 | 0 | 4 | 18 | 0 | 19 | 17 | 0 | 19 | 12 |
| | 2 | 0 | 0 | 2 | 0 | 0 | 3 | 0 | 0 | 8 |
| UCS | 0 | 20 | 18 | 0 | 20 | 13 | 0 | 19 | 6 | 0 |
| | 1 | 0 | 2 | 20 | 0 | 7 | 18 | 1 | 14 | 15 |
| | 2 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 5 |
| AR1 | 0 | 20 | 20 | 20 | 20 | 20 | 20 | 20 | 20 | 20 |
| UAR1 | 0 | 19 | 20 | 19 | 20 | 20 | 19 | 20 | 20 | 20 |
| | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 |
| IND | 0 | 20 | 2 | 0 | 20 | 0 | 0 | 20 | 0 | 0 |
| | 1 | 0 | 9 | 0 | 0 | 5 | 0 | 0 | 2 | 0 |
| | 2 | 0 | 9 | 0 | 0 | 12 | 0 | 0 | 11 | 0 |
| | 3 | 0 | 0 | 9 | 0 | 3 | 0 | 0 | 7 | 0 |
| | 4 | 0 | 0 | 9 | 0 | 0 | 7 | 0 | 0 | 5 |
| | 5 | 0 | 0 | 2 | 0 | 0 | 9 | 0 | 0 | 6 |
| | 6 | 0 | 0 | 0 | 0 | 0 | 4 | 0 | 0 | 8 |
| | 7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| UIND | 0 | 20 | 7 | 0 | 20 | 0 | 0 | 20 | 0 | 0 |
| | 1 | 0 | 13 | 0 | 0 | 19 | 0 | 0 | 10 | 0 |
| | 2 | 0 | 0 | 9 | 0 | 1 | 1 | 0 | 10 | 0 |
| | 3 | 0 | 0 | 10 | 0 | 0 | 11 | 0 | 0 | 2 |
| | 4 | 0 | 0 | 1 | 0 | 0 | 8 | 0 | 0 | 18 |

Note : Cov réfère au type de covariance supposée. G réfère au nombre de trajectoires estimées.

TABLEAU A.8. Résultats pour des données générées avec une covariance auto-régressive d'ordre 1 avec variance hétérogène(UAR1) et p=5.

| | | N=200 | | | N=400 | | | N=600 | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | ρ | | | ρ | | | ρ | | |
| Cov | (G-1) | 0 | 0,3 | 0,6 | 0 | 0,3 | 0,6 | 0 | 0,3 | 0,6 |
| CS | 0 | 7 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 |
| | 1 | 13 | 18 | 7 | 18 | 14 | 0 | 20 | 10 | 0 |
| | 2 | 0 | 2 | 12 | 0 | 6 | 18 | 0 | 10 | 18 |
| | 3 | 0 | 0 | 1 | 0 | 0 | 2 | 0 | 0 | 2 |
| UCS | 0 | 20 | 17 | 0 | 20 | 16 | 0 | 20 | 8 | 0 |
| | 1 | 0 | 3 | 18 | 0 | 4 | 18 | 0 | 12 | 10 |
| | 2 | 0 | 0 | 2 | 0 | 0 | 2 | 0 | 0 | 9 |
| | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| AR1 | 0 | 16 | 1 | 0 | 2 | 0 | 0 | 0 | 0 | 0 |
| | 1 | 0 | 1 | 5 | 0 | 0 | 0 | 0 | 0 | 0 |
| | 2 | 4 | 18 | 15 | 18 | 20 | 20 | 20 | 20 | 20 |
| UAR1 | 0 | 19 | 15 | 15 | 20 | 17 | 19 | 20 | 20 | 19 |
| | 1 | 1 | 5 | 5 | 0 | 3 | 1 | 0 | 0 | 1 |
| IND | 0 | 8 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 |
| | 1 | 12 | 12 | 0 | 18 | 6 | 0 | 20 | 1 | 0 |
| | 2 | 0 | 8 | 3 | 0 | 14 | 0 | 0 | 19 | 0 |
| | 3 | 0 | 0 | 9 | 0 | 0 | 2 | 0 | 0 | 0 |
| | 4 | 0 | 0 | 6 | 0 | 0 | 8 | 0 | 0 | 7 |
| | 5 | 0 | 0 | 1 | 0 | 0 | 10 | 0 | 0 | 8 |
| | 6 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 5 |
| UIND | 0 | 20 | 8 | 0 | 20 | 0 | 0 | 20 | 0 | 0 |
| | 1 | 0 | 12 | 2 | 0 | 19 | 0 | 0 | 16 | 0 |
| | 2 | 0 | 0 | 12 | 0 | 1 | 3 | 0 | 4 | 0 |
| | 3 | 0 | 0 | 6 | 0 | 0 | 12 | 0 | 0 | 6 |
| | 4 | 0 | 0 | 0 | 0 | 0 | 4 | 0 | 0 | 13 |
| | 5 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 |

Note : Cov réfère au type de covariance supposée. G réfère au nombre de trajectoires estimées.

# Annexe B

## SYNTAXE DES ANALYSES

Voici les syntaxes utilisées pour les modèles finaux présentés au chapitre 3. Les analyses sont effectuées avec la version 6.11 du logiciel Mplus.

### B.1. AGRESSIVITÉ PHYSIQUE, MODÈLE À 5 GROUPES, AVEC RÉGRESSION CUBIQUE, SUPPOSANT L'INDÉPENDANCE À TRAVERS LE TEMPS

```
TITLE: Syntaxe du modèle d'agressivité à 5 groupes
       supposant l'indépendance;
DATA: FILE IS agp17_108.dat;
VARIABLE: NAMES ARE ID SEXE AGP17 AGP30 AGP42 AGP60
                                  AGP72 AGP84 AGP108;
USEVARIABLES ARE SEXE AGP17 AGP30 AGP42 AGP60
                     AGP72 AGP84 AGP108;
CLASSES = Grpagp(5);
MISSING AGP17 AGP30 AGP42 AGP60 AGP72 AGP84 AGP108 (999);
ANALYSIS: TYPE=MIXTURE;
          ESTIMATOR=ML;
          start=1000 10;
MODEL:
%OVERALL%
I S Q C | AGP17@0 AGP30@1 AGP42@2 AGP60@3
```

```
          AGP72@4 AGP84@5 AGP108@7;
I-C@0;
I with S-C@0;
S with Q-C@0;
Q with C@0;
GRPAGP on sexe;
%grpagp#1%
[i*0.066 s@0 q@0 C@0];
AGP17-AGP108 (v1);
%grpagp#2%
[I*0.524 S*0.521 Q*-0.182 C*0.015];
AGP17-AGP108 (v2);
%grpagp#3%
[I*0.706 S*1.415 Q*-0.380 C*0.031];
AGP17-AGP108 (v2);
%grpagp#4%
[I*1.166 S*1.874 Q*-0.699 C*0.060];
AGP17-AGP108 (v2);
%grpagp#5%
[I*1.148 S*1.828 Q*-0.230 C*0.001];
AGP17-AGP108 (v2);
PLOT:
  TYPE=PLOT3;
  series =    AGP17 (0) AGP30 (1) AGP42 (2) AGP60 (3)
              AGP72 (4) AGP84 (5) AGP108 (7);
OUTPUT: TECH1 TECH4 TECH8;
```

## B.2. HYPERACTIVITÉ, MODÈLE À 4 GROUPES, AVEC RÉGRESSION CUBIQUE ET EN SUPPOSANT UNE COVARIANCE À SYMÉTRIE COMPOSÉE

```
TITLE: Hyperactivité à 2 groupes avec covariance
```

```
       à symétrie composée;
DATA: FILE IS HYP17_108.dat;
VARIABLE: NAMES ARE ID SEXE HYP17 HYP30 HYP42 HYP60
                                  HYP72 HYP84 HYP108;
USEVARIABLES ARE SEXE HYP17 HYP30 HYP42 HYP60
                            HYP72 HYP84 HYP108;
CLASSES = GrpHYP(4);
MISSING HYP17 HYP30 HYP42 HYP60 HYP72 HYP84 HYP108 (999);
ANALYSIS: TYPE=MIXTURE;
          ESTIMATOR=ML;
          start=1000 10;
MODEL:
%OVERALL%
i s q c| HYP17@0 HYP30@1 HYP42@2 HYP60@3
                 HYP72@4 HYP84@5 HYP108@7;
i-c@0;
i with s-c@0;
s with q-c@0;
q with c@0;
GRPHYP on sexe;
%GrpHYP#1%
[I*1.732 S*0.184 Q*-0.047 C*0.001];
HYP17-HYP108*1.226 (VAR1);
HYP17 with HYP30-HYP108 (COV1);
HYP30 with HYP42-HYP108 (COV1);
HYP42 with HYP60-HYP108 (COV1);
HYP60 with HYP72-HYP108 (COV1);
HYP72 with HYP84-HYP108 (COV1);
HYP84 with HYP108 (COV1);
%GrpHYP#2%
[I*5.717 S*-0.843 Q*0.050 C*0.001];
```

```
HYP17–HYP108*4 (VAR2);
HYP17 with HYP30–HYP108 (COV2);
HYP30 with HYP42–HYP108 (COV2);
HYP42 with HYP60–HYP108 (COV2);
HYP60 with HYP72–HYP108 (COV2);
HYP72 with HYP84–HYP108 (COV2);
HYP84 with HYP108 (COV2);
%GrpHYP#3%
[I*2.889 S*3.660 Q*–1.287 C*0.109];
HYP17–HYP108*3.983 (VAR3);
HYP17 with HYP30–HYP108 (COV3);
HYP30 with HYP42–HYP108 (COV3);
HYP42 with HYP60–HYP108 (COV3);
HYP60 with HYP72–HYP108 (COV3);
HYP72 with HYP84–HYP108 (COV3);
HYP84 with HYP108 (COV3);
%GrpHYP#4%
[I*3.810 S*0.389 Q*0.000 C*–0.006];
HYP17–HYP108*3.884 (VAR4);
HYP17 with HYP30–HYP108 (COV4);
HYP30 with HYP42–HYP108 (COV4);
HYP42 with HYP60–HYP108 (COV4);
HYP60 with HYP72–HYP108 (COV4);
HYP72 with HYP84–HYP108 (COV4);
HYP84 with HYP108 (COV4);
PLOT:
  TYPE=PLOT3;
  series =    HYP17 (0) HYP30 (1) HYP42  (2) HYP60 (3)
              HYP72 (4) HYP84 (5) HYP108 (7);
OUTPUT: TECH1 TECH4;
Model constraint:
```

```
new (rho1);
new (rho2);
new (rho3);
new (rho4);
cov1=var1*rho1;
cov2=var2*rho2;
cov3=var3*rho3;
cov4=var4*rho4;
```

## B.3. Mémoire à court terme, modèle à 2 groupes, avec régression cubique et en supposant une covariance à symétrie composée avec variance hétérogène

```
TITLE: Mémoire à 2 groupes avec régression cubique et
       covariance à symétrie composée;
DATA: FILE IS vcr42_108.dat;
VARIABLE: NAMES ARE IDME SEXE VCR42 VCR60
                        VCR72 VCR84 VCR108;
USEVARIABLES ARE  VCR42 VCR60 VCR72 VCR84 VCR108;
CLASSES = GrpVCR(2);
MISSING VCR42 VCR60 VCR72 VCR84 VCR108 (999);
ANALYSIS: TYPE=MIXTURE;
          ESTIMATOR=ML;
          start=500 5;
MODEL:
%OVERALL%
i s q c| VCR42@0 VCR60@1 VCR72@2 VCR84@3 VCR108@5;
i-c@0;
i with s-c@0;
s with q-c@0;
q with c@0;
%GrpVCR#1%
```

```
VCR42-VCR108 (V11-V15);
VCR42 with VCR60-VCR108 (COV112-COV115);
VCR60 with VCR72-VCR108 (COV123-COV125);
VCR72 with VCR84-VCR108 (COV134-COV135);
VCR84 with VCR108 (COV145);
%GrpVCR#2%
VCR42-VCR108 (V21-V25);
VCR42 with VCR60-VCR108 (COV212-COV215);
VCR60 with VCR72-VCR108 (COV223-COV225);
VCR72 with VCR84-VCR108 (COV234-COV235);
VCR84 with VCR108 (COV245);
PLOT:
  TYPE=PLOT3;
  series = VCR42 (0) VCR60 (1) VCR72 (2) VCR84  (3)
                                         VCR108 (5);
OUTPUT: TECH1 TECH4;
MODEL CONSTRAINT :
new (RHO1*.10);
new (RHO2*.10);
RHO1>0;
RHO2>0;
RHO1<1;
RHO2<1;
COV112=RHO1*sqrt(V11)*sqrt(V12);
COV113=RHO1*sqrt(V11)*sqrt(V13);
COV114=RHO1*sqrt(V11)*sqrt(V14);
COV115=RHO1*sqrt(V11)*sqrt(V15);
COV123=RHO1*sqrt(V12)*sqrt(V13);
COV124=RHO1*sqrt(V12)*sqrt(V14);
COV125=RHO1*sqrt(V12)*sqrt(V15);
COV134=RHO1*sqrt(V13)*sqrt(V14);
```

```
COV135=RHO1*sqrt(V13)*sqrt(V15);
COV145=RHO1*sqrt(V14)*sqrt(V15);
COV212=RHO2*sqrt(V21)*sqrt(V22);
COV213=RHO2*sqrt(V21)*sqrt(V23);
COV214=RHO2*sqrt(V21)*sqrt(V24);
COV215=RHO2*sqrt(V21)*sqrt(V25);
COV223=RHO2*sqrt(V22)*sqrt(V23);
COV224=RHO2*sqrt(V22)*sqrt(V24);
COV225=RHO2*sqrt(V22)*sqrt(V25);
COV234=RHO2*sqrt(V23)*sqrt(V24);
COV235=RHO2*sqrt(V23)*sqrt(V25);
COV245=RHO2*sqrt(V24)*sqrt(V25);
```

## B.4. Vocabulaire réceptif, modèle à 1 groupe, avec régression cubique et en supposant une covariance à symétrie composée avec variance hétérogène

```
TITLE: Vocabulaire à 1 groupe avec régression cubique et covariance
       à symétrie composée avec variance hétérogène;
DATA: FILE IS evip42_96.dat;
VARIABLE: NAMES ARE IDME SEXE   EVIP42 EVIP60
                        EVIP72 EVIP84 EVIP96;
USEVARIABLES ARE  EVIP42 EVIP60 EVIP72 EVIP84 EVIP96;
CLASSES = GrpEVIP(1);
MISSING EVIP42 EVIP60 EVIP72 EVIP84 EVIP96 (999);
ANALYSIS: TYPE=MIXTURE;
          ESTIMATOR=ML;
          start=200 5;
MODEL:
%OVERALL%
i s q c| EVIP42@0 EVIP60@1 EVIP72@2 EVIP84@3 EVIP96@4;
i-c@0;
```

```
i with s-c@0;
s with q-c@0;
q with c@0;
%GrpEVIP#1%
EVIP42-EVIP96 (V11-V15);
EVIP42 with EVIP60-EVIP96(COV112-COV115);
EVIP60 with EVIP72-EVIP96(COV123-COV125);
EVIP72 with EVIP84-EVIP96(COV134-COV135);
EVIP84 with EVIP96(COV145);
PLOT:
  TYPE=PLOT3;
  series = EVIP42 (0) EVIP60 (1) EVIP72 (2)
                      EVIP84 (3) EVIP96 (5);
OUTPUT: TECH1 TECH4;
MODEL CONSTRAINT:
NEW (RHO1*0);
NEW (SD1);
NEW (SD2);
NEW (SD3);
NEW (SD4);
NEW (SD5);
V11=SD1*SD1;
V12=SD2*SD2;
V13=SD3*SD3;
V14=SD4*SD4;
V15=SD5*SD5;
COV112=SD1*SD2*RHO1;
COV113=SD1*SD3*RHO1;
COV114=SD1*SD4*RHO1;
COV115=SD1*SD5*RHO1;
COV123=SD2*SD3*RHO1;
```

```
COV124=SD2*SD4*RHO1;
COV125=SD2*SD5*RHO1;
COV134=SD3*SD4*RHO1;
COV135=SD3*SD5*RHO1;
COV145=SD4*SD5*RHO1;
```