

Faculté des sciences

Analyse longitudinale de données à haute dimension

Application aux données -omiques

Auteure : Mathilde Tournay

Promotrice : Bernadette Govaerts

Lecteurs : Catherine Legrand et Michel Thiel

Année académique 2019-2020

Mémoire de fin d'études présenté en vue de l'obtention
du master en statistiques : **orientation biostatistiques**

Remerciements

En préambule à ce mémoire, je souhaite adresser mes remerciements à toutes les personnes qui ont contribué de près ou de loin à la réalisation de ce travail.

Je tiens tout d'abord à remercier ma promotrice, le Professeur Bernadette Govaerts, pour m'avoir permis de réaliser ce projet ainsi que pour son soutien et ses conseils prodigués tout au long de l'année.

Je tiens également à remercier Michel Thiel et le Professeur Catherine Legrand pour avoir accepté d'être lecteurs et membres du jury de ce mémoire.

Enfin, merci à ma famille et à mes amis pour leur soutien lors de cette période un peu particulière et pour avoir pris le temps de relire ce travail.

Mathilde Tournay

Liste des abréviations

ACP	Analyse en composantes principales
APCA	ANOVA-principal component analysis
ASCA	ANOVA-simultaneous component analysis
CP	Composante principale
CS	Composante symétrique
dl	Degré de liberté
EA	Effet aléatoire
GLM	Modèle linéaire général
ML	Maximum de vraisemblance
LiMM-PCA	Linear Mixed Models-PCA
LMM	Modèle linéaire mixte
REML	Maximum de vraisemblance restreinte
RMN	Résonance magnétique nucléaire
ppm	Parties par million

Table des matières

Chapitre I : Introduction	1
1 Contexte	1
2 Objectifs	4
3 Contenu du mémoire	5
Chapitre II : Bases de données et études longitudinales	6
1 La métabolomique	6
2 Bases de données	7
2.1 Données multivariées : Choo	7
2.2 Données univariées : Orthodont	9
3 Données longitudinales	10
3.1 ANOVA à mesures répétées	11
3.2 MANOVA	13
3.3 Les modèles linéaires mixtes	14
Chapitre III : Modèles linéaires mixtes	15
1 Définition et notations	16
1.1 Modèles à effets aléatoires	17
1.2 Modèles à covariance pattern	19
1.3 Modèles à coefficients aléatoires	21
2 Estimation	23
3 Tests d'hypothèses	24
Chapitre IV : Extensions des LMMs aux réponses de grande dimension	25
1 ASCA ⁺ et APCA ⁺	25
1.1 Étape 1 : Modèle linéaire général et matrice du modèle	27
1.2 Étape 2 : Estimation des paramètres et décomposition en matrices d'effet	28

1.3	Étape 3 : Analyse en composantes principales sur les matrices d'effets	29
1.4	Étape 4 : Pourcentage de variation et tests de significativité	31
2	LiMM-PCA	32
2.1	Étape 1 : Orthogonalisation et réduction de dimension de la matrice de réponses par ACP	33
2.2	Étape 2 : Ajustement en parallèle de LMM	33
2.3	Étape 3 : Décomposition en matrices d'effets	34
2.4	Étape 4a : Quantification du pourcentage de variance expliqué par chaque facteur	34
2.5	Étape 4b : Tests de significativité des facteurs	35
2.6	Étape 5 : ACP et représentation visuelle des matrices d'effets	37
 Chapitre V : Analyse de données longitudinales univariées : <i>Implémentation et application</i>		39
1	Outils disponibles en R	39
1.1	Packages nlme et lme4	39
1.2	Fonctions R	40
2	Application aux données Orthodont	43
2.1	lme vs lmer	43
2.2	lme/lmer vs gls	44
3	Conclusions	48
 Chapitre VI : Analyses de données longitudinales à réponses multivariées : <i>Application aux données Choo</i>		51
1	Pré-traitement des données	51
2	Analyse en composantes principales	52
3	ASCA+ et APCA+	54
3.1	Étape 1 : Modèle GLM	54
3.2	Étape 2 : Estimation des paramètres et décomposition en matrices d'effets	55
3.3	Étape 3 : ACP sur les matrices d'effets et visualisation	55
3.4	Étape 4 : Pourcentage de variation et significativité	58
4	LiMM-PCA	59
4.1	Étape 1 : Orthogonalisation et réduction de dimension de la matrice réponse par ACP	59
4.2	Étapes 2 et 3 : Ajustement en parallèle de LMM et décomposition en matrices d'effets	59

TABLE DES MATIÈRES

4.3	Étape 4 : Quantification de l'importance de chaque effet et significativité	60
4.4	Étape 5 : Visualisation des matrices d'effets par ACP	63
5	Implémentation et analyse d'un modèle multivarié à structure de covariance CS	68
Chapitre VII : Conclusions et perspectives		72
Bibliographie		77
Annexe A : Normalité des résidus		82
Annexe B : Distributions des GLLR - procédure bootstrap		83

Chapitre I

Introduction

1 Contexte

Les technologies modernes à haut débit permettent l'acquisition simultanée d'un nombre important de données sur un grand nombre de variables au cours d'une seule expérience. Ces technologies, dites "omiques", incluent diverses disciplines telles que la génomique, la transcriptomique, la protéomique et la métabolomique et visent à étudier la réponse globale d'un système biologique à des niveaux multiples ; du séquençage des gènes à l'expression des métabolites (Liquet *et al.*, 2012).

Les récents progrès et la diminution subséquente des coûts liés à ces technologies offrent aujourd'hui la possibilité d'effectuer des mesures répétées plusieurs fois dans le temps sur les mêmes individus ou échantillons biologiques (Liquet *et al.*, 2012). Ce type d'expérience a pour avantage d'offrir une capture dynamique des processus métaboliques engendrés par un facteur biologique d'intérêt. En outre, elles permettent d'économiser en sujets et de réduire le bruit lié à la variation inter-individus (Li *et al.*, 2004). Cependant, le design expérimental sous-jacent à ces études engendre une structure multiniveaux dans les données qui contiennent alors plusieurs sources de variabilité (eg. inter- et intra- individus) (Jansen *et al.*, 2005a). Cette structure complexe doit être prise en compte dans l'analyse au même titre que le nombre important de variables corrélées. Ces deux composantes sont résumées à la Fig.I.1.

La littérature regorge de nombreux articles concernant l'analyse longitudinale de données -omiques (Smilde *et al.*, 2005; Keun *et al.*, 2004; Liquet *et al.*, 2012). L'approche la plus courante consiste à appliquer une analyse de la variance (ANOVA) ou un modèle linéaire mixte (LMM) sur chaque variable individuellement (Straube *et al.*, 2015; Li *et al.*, 2004; Karlovich *et al.*, 2009). Cependant, ces approches ignorent la nature multivariée des données et la dépendance entre les différentes variables

(Zwanenburg *et al.*, 2011). Des extensions multivariées de ces modèles existent (eg. MANOVA) mais se montrent inefficaces lorsque le nombre de variables est supérieur au nombre d'observations ce qui est souvent le cas des données -omiques (Jansen *et al.*, 2005b).

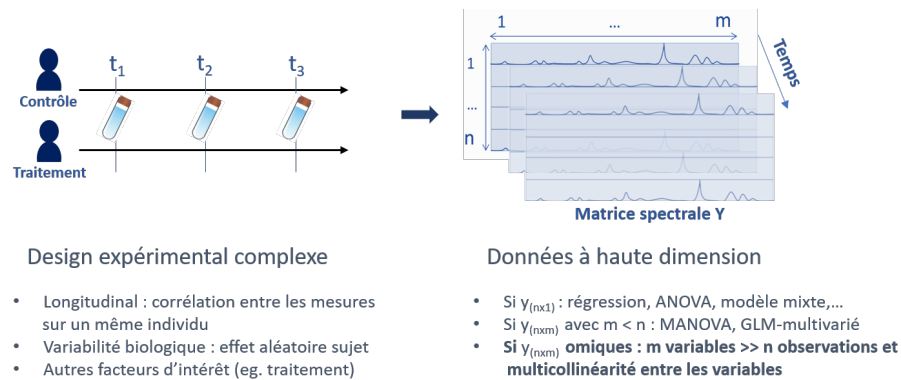


Figure I.1 – Planification et contraintes d’une expérience longitudinale type en sciences -omiques : des sujets sont assignés aléatoirement dans un groupe de traitement et des mesures sont effectuées à plusieurs temps. Typiquement, une telle expérience génère une quantité importante de données caractérisées par un nombre de variables corrélées supérieur au nombre d’observations.

L’analyse en composantes principales (ACP) est une technique couramment utilisée pour analyser des données -omiques (corrélées et à haute dimension). Elle permet de représenter la variabilité présente dans un jeu de données dans un espace à dimensions réduites. Cependant, cette méthode n’a pas la capacité de tenir compte de la structure du plan d’expérience et donc de différencier les différentes sources de variation qui sont alors confondues.(Jansen *et al.*, 2005a).

Face à cette contrainte, le monde statistique a vu l’émergence de plusieurs méthodes qui visent à combiner aussi bien la modélisation statistique de designs complexes que l’utilisation d’outils d’analyse multivariée. Parmi ces méthodes, l’ASCA (ANOVA-Simultaneous Component Analysis) et l’APCA (ANOVA-Principal Component Analysis) sont couramment utilisées pour analyser des données -omiques (Smilde *et al.*, 2005; Jansen *et al.*, 2005b). Elles consistent, dans un premier temps, à décomposer la matrice de réponses en une somme de matrices d’effets correspondant aux différents termes d’un modèle ANOVA. Dans un deuxième temps, une ACP est réalisée sur chacune de ces matrices afin de représenter l’information dans un espace à dimensions réduites. L’APCA se différencie de l’ASCA à cette étape en incorporant la matrice des résidus aux matrices d’effets avant d’appliquer l’ACP (Harrington *et al.*, 2005; Zwanenburg *et al.*, 2011).

La méthodologie ASCA/APCA a servi de base pour l'élaboration de méthodes plus récentes et également basées sur la décomposition ANOVA mais combinée à d'autres outils multivariés. C'est le cas notamment des approches AoV-PLS (El Ghaziri *et al.*, 2015) et ANOVA-TP (Marini *et al.*, 2015) qui appliquent respectivement sur les matrices d'effets une régression des moindres carrés partiels (PLS) et une PLS avec projection ciblée. Les méthodes AComDim (Bouveresse *et al.*, 2011), PARAFASCA (Jansen *et al.*, 2008) et AMOPLS (Boccard et Rudaz, 2016) proposent quant à elles d'explorer les sources de variation d'intérêt simultanément à l'aide d'un modèle unique au lieu d'effectuer une analyse séparée sur chaque matrice d'effet.

Ces méthodes ont cependant le désavantage de ne pas être adaptées aux cas où les designs sont non-balancés. En effet, elles font intervenir des approches standards qui fournissent des estimateurs valables uniquement lorsque les données sont balancées. Thiel *et al.* (2017) ont alors proposé d'estimer les matrices d'effets sur base d'un modèle linéaire général (GLM) à défaut d'un modèle ANOVA dans le but d'étendre l'utilisation des méthodes ASCA/APCA à tout type de designs expérimentaux comportant des facteurs fixes catégoriels. Guisset *et al.* (2019) a par après montré que cette nouvelle approche, appelée ASCA/APCA+, pouvait profiter à d'autres méthodes, citées précédemment, dont l'étape de modélisation est basée initialement sur l'ANOVA.

Les méthodes présentées jusqu'ici sont basées sur l'hypothèse d'indépendance entre les observations et sont destinées aux plans d'expériences incluant exclusivement des facteurs fixes. Or, les designs longitudinaux (tout comme les plans à mesures répétées, hiérarchisés ou multiniveaux) induisent une structure dans les données et nécessitent donc l'utilisation de modèles statistiques plus poussés. Des méthodes, dérivées de l'ASCA, ont été développées afin d'analyser des données multiniveaux à réponses multiples. Timmerman (2006) a proposé une méthode, la Multilevel-SCA (MSCA), capable de modéliser séparément les variations inter- et intra- individus. La MSCA a tout d'abord été utilisée dans le domaine de la psychologie pour analyser l'évolution temporelle de réponses d'individus à un questionnaire Timmerman (2006) puis en métabolomique pour l'analyse de spectres RMN d'échantillons d'urine de singes collectés plusieurs fois dans le temps (Jansen *et al.*, 2005a). En parallèle, de Noord et Theobald (2005) ont développé la méthode Multilevel-PLS (MLPLS) dans le but de suivre la dynamique d'un procédé de production.

Récemment, Martin et Govaerts (2020) ont mis au point une méthode, appelée Linear-Mixed Models-PCA (LiMM-PCA), qui vise à étendre l'ASCA/APCA+ à des designs plus complexes. La contribution majeure de cette méthode réside dans la substitution du modèle GLM par un LMM. Via l'introduction d'effets aléatoires, cette classe de modèles permet de prendre en compte des structures de corrélation complexes et conduit à une modélisation plus précise des données. En outre, les conclusions tirées du modèle seront généralisables à d'autres observations (Brown et Prescott, 2014). La méthode LiMM-PCA s'est déjà montrée efficace pour mettre en évidence et comparer les différentes sources de variabilité dans deux études de répétabilité mais pourrait également être utilisée pour analyser des données longitudinales (Martin et Govaerts, 2020).

2 Objectifs

Les LMMs permettent de tenir compte de la corrélation présente dans des données longitudinales par l'introduction d'effets aléatoires. Ces outils statistiques comprennent également une autre classe de modèles, dits à structures de covariance, couramment utilisés dans l'analyse longitudinale et qui consistent à spécifier directement la corrélation dans la matrice des résidus. A l'heure actuelle, la méthode LiMM-PCA n'implémente pas de fonctions permettant d'analyser ce deuxième type de modèle. Ce mémoire aura pour objectifs principaux d'appliquer la LiMM-PCA à un nouveau jeu de données et d'étendre les concepts théoriques sous-jacents afin de généraliser son utilisation à tous types de modèles longitudinaux.

Par ailleurs, la méthodologie ASCA/APCA+ a récemment été implémentée par Franceschini (2019) au sein d'un package R nommé LMWiRE. Ce package n'est actuellement pas disponible sur le R CRAN mais une version en développement est accessible sur GitHub¹. Ce mémoire aura pour objectif secondaire d'appliquer différentes fonctions incluses dans LMWiRE afin de rapporter d'éventuels bugs et de proposer des améliorations.

1. <https://github.com/FranceschiniS/LMWiRe>

3 Contenu du mémoire

Le second chapitre proposera une description succincte de la métabolomique et détaillera les bases de données *Orthodont* et *Choo* utilisées dans le cadre dans ce mémoire. Les méthodes statistiques univariées classiques utilisées dans l'analyse de données longitudinales seront ensuite introduites. Au chapitre III, une attention particulière sera portée aux modèles linéaires mixtes. Les concepts théoriques pertinents de ce modèle général et flexible seront présentés et nous verrons dans quelles mesures ils peuvent être utilisés pour analyser des données longitudinales. En effet, la compréhension des LMMs dans le cas univarié est primordiale afin d'étendre ces modèles aux cas de réponses multivariées. Le chapitre IV introduira dans un premier temps la méthodologie des méthodes d'analyses multivariées ASCA/APCA+ développées par Thiel *et al.* (2017). La méthode LiMMPCA sera ensuite présentée en détaillant plus en profondeur les nouvelles contributions apportées par rapport à l'ASCA/APCA+.

Les fonctions disponibles dans le logiciel R afin d'analyser des données longitudinales univariées seront décrites et comparées dans le chapitre V. Au chapitre VI, les différentes méthodes multivariées seront ensuite appliquées sur le jeu de données *Choo*. Nous présenterons par ailleurs les adaptations effectuées sur la méthode LiMM-PCA pour analyser les modèles à structures de covariance.

Chapitre II

Bases de données et études longitudinales

1 La métabolomique

La métabolomique est un domaine de recherche qui consiste à identifier et quantifier de manière exhaustive et non sélective l'ensemble des métabolites¹ présents dans un système biologique (biofluides, tissus, cellules, organes, etc.) (Puchades-Carrasco et Pineda-Lucena, 2015). Ces molécules sont les produits finaux de processus complexes liés à la réponse cellulaire. En fournissant un aperçu du profil métabolique à un instant donné, cette approche permet dès lors de caractériser de manière globale les perturbations du métabolisme en réponse à des stimuli tels qu'une modification génétique, une pathologie ou une exposition à un composé xénobiotique (Canlet *et al.*, 2012; Bonvallot, 2014).

Les techniques analytiques les plus couramment utilisées pour générer des données métabolomiques sont la spectrométrie de masse (MS) et la spectrométrie par résonance magnétique nucléaire du proton (¹H-RMN). Ces méthodes ont leurs avantages et inconvénients et le choix de l'une ou l'autre sera dicté par un compromis entre les objectifs à atteindre et la faisabilité technique. En effet, bien que la MS soit plus sensible, la RMN a pour avantages d'être non-destructive, peu coûteuse, non-sélective et de ne demander que peu de temps de préparation (Rousseau, 2011; Beyoğlu et Idle, 2013).

1. molécules inférieures à 1 kilodalton et présentes dans les fluides biologiques

Une étude métabolomique va générer pour chaque échantillon un spectre complexe qui peut ensuite être transformé en un ensemble de variables corrélées dont le nombre dépasse largement le nombre d'unités expérimentales (Rousseau, 2011). Les valeurs observées correspondent à un ou plusieurs métabolites présents dans l'échantillon. Toute variation du profil spectral sera dès lors le reflet de modifications du métabolisme en réponse à un facteur biologique. L'objectif va être de comparer les différents profils et de détecter les métabolites, appelés biomarqueurs, à l'origine de ces différences. Cette étape nécessitera l'utilisation d'outils statistiques multivariés.

2 Bases de données

2.1 Données multivariées : Choo

Les données multivariées analysées dans le cadre de ce mémoire proviennent de l'article de Choo *et al.* (2017). Cette étude métabolomique consistait à caractériser l'effet de la prise d'un antibiotique important du point de vue clinique, la vancomycine-imipenem, sur le microbiote de souris. En effet, les perturbations induites sur le microbiote intestinal par les antibiotiques impactent indirectement la composition du microbiome fécal dont la composition peut être déterminée par spectrométrie RMN. L'analyse a, par ailleurs, été effectuée longitudinalement afin d'avoir une image dynamique et progressive des changements métabolomiques engendrés.

Pour cette étude, 16 souris génétiquement identiques ont été divisées en deux groupes égaux : traitement et contrôle. Pour chaque souris, des échantillons de matières fécales ont été collectés à trois temps (Fig. II.1) :

- Avant le traitement (T1)
- A la fin de l'antibiothérapie, soit après 14 jours (T2)
- 9 jours après la fin du traitement (T3)

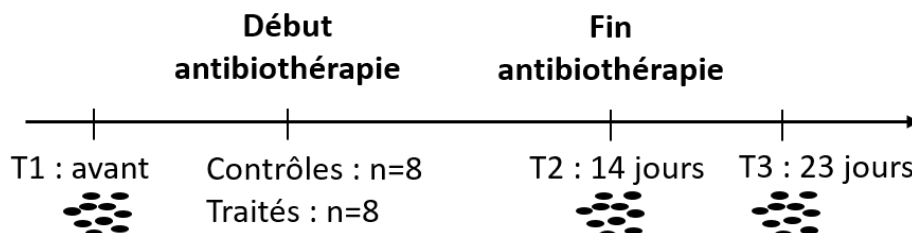


Figure II.1 – Design d'expérience sous-jacent aux données Choo.

Chaque échantillon a ensuite été analysé par spectroscopie RMN. Les signaux bruts obtenus ont alors subi une série de transformations et de prétraitements afin d'obtenir des données directement exploitables. Il devrait donc y avoir $n = 2 \times 8 \times 3 = 48$ observations. Cependant, deux d'entre elles comportaient des valeurs aberrantes et ont été retirées de l'analyse ce qui porte le total à 46 observations. Pour plus d'informations concernant la préparation des données, voir l'article de Choo *et al.* (2017).

Deux matrices sont incluses dans la base de données :

- la matrice **outcomes** ($n \times m$) ou matrice spectrale **Y** : les signaux bruts obtenus par RMN sont convertis en spectres et décomposés en m régions de largeur fixe, appelées descripteurs et dont les noms sont exprimés en parties par million (ppm). Pour chaque échantillon, les valeurs de la matrice représentent les intensités normalisées du signal au niveau de chaque descripteur. Cette matrice est donc caractérisée par un nombre m de variables supérieur au nombre d'observations n et elle synthétise l'ensemble des données spectrales. Dans notre cas, $m = 1452$ (entre 8,5 et 0,5 ppm). La figure II.2 montre les spectres obtenus aux deux premiers temps pour un individu du groupe "traitement".
- la matrice **design** ($n \times l$) qui, pour chaque observation, comprend les niveaux des l facteurs pris en compte dans le plan d'expérience. Dans notre cas, trois facteurs sont considérés : le *temps* (T1, T2 ou T3), le *traitement* (contrôle ou non) ainsi que le *sujet* (1 à 16). Notons que suite au retrait de deux observations, le design est non-balancé.

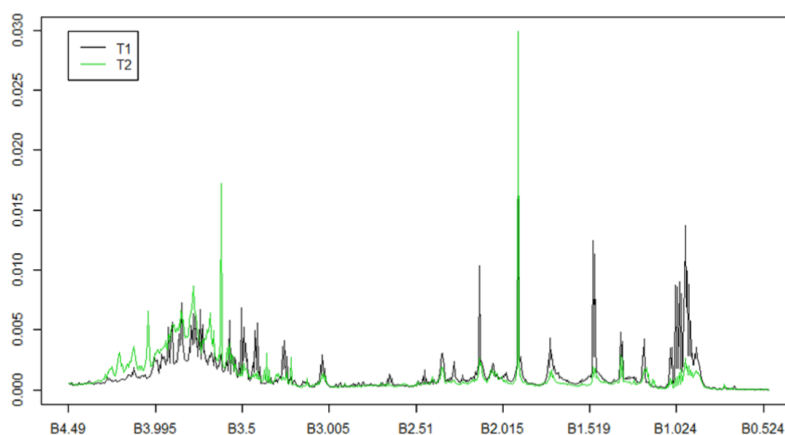


Figure II.2 – Profils spectraux ^1H -RMN d'un même individu aux temps T1 et T2. Notons que la fenêtre spectrale est comprise ici entre 4,5 et 0,5 ppm.

2.2 Données univariées : Orthodont

Ce jeu de données, introduit par Potthoff et Roy (1964), constitue un exemple classique et couramment repris dans la littérature traitant de l'analyse longitudinale. Elles sont accessibles en R sous le nom `Orthodont` dans le package `nlme`.

Ces données consistent en des mesures dentaires effectuées sur 27 enfants (16 garçons et 11 filles) au cours de leur croissance. Pour chaque sujet, la distance entre la glande pituitaire et la fosse ptérygo-maxillaire a été mesurée sur base de clichés de radiologie tous les deux ans entre 8 et 14 ans. Il y a donc un total de $4 \times 27 = 108$ observations.

Le jeu de données contient les variables suivantes :

- *Sex* (masculin ou féminin)
- *Subject* (eg. F01 ou M02)
- *Age* (8, 10, 12 et 14)
- *Distance* (en millimètres)

La figure II.3 montre l'évolution au fil du temps de la distance pour les filles et les garçons.

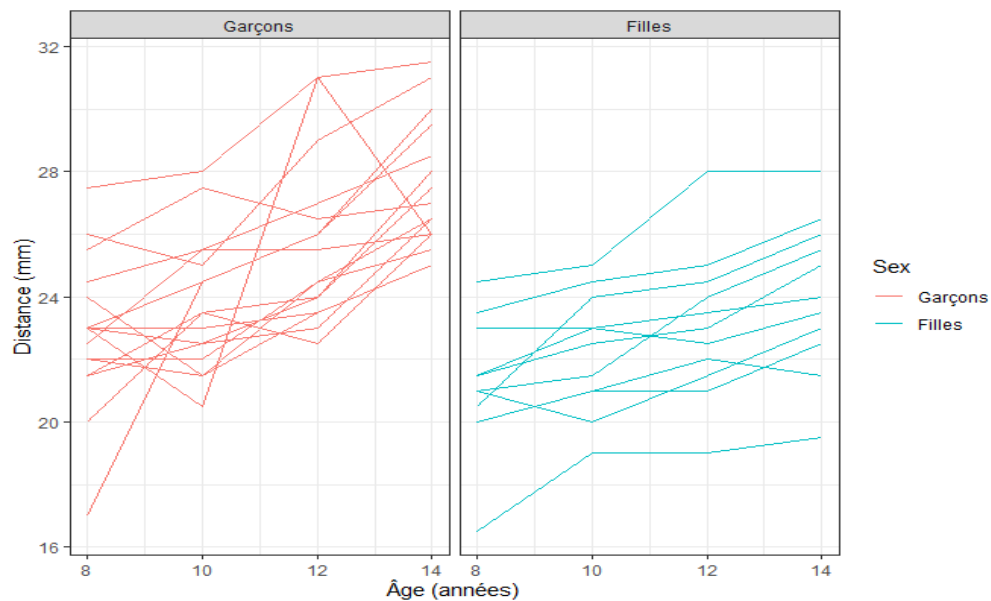


Figure II.3 – Croissance de la distance entre la glande pituitaire et la fosse ptérygo-maxillaire chez les enfants.

3 Données longitudinales

Données longitudinales est la terminologie utilisée pour désigner des observations mesurées de manière répétée au fil du temps sur un échantillon d'unités expérimentales telles que des individus, des sujets, des parcelles agricoles, etc. (Diggle *et al.*, 2002)

Depuis plusieurs décennies, il y a un intérêt croissant pour les études utilisant des plans longitudinaux dans de nombreux domaines tels que la médecine ou l'agriculture. En effet, les études longitudinales offrent de nombreux avantages (Hedeker et Gibbons, 2006) :

- elles permettent d'acquérir des informations sur l'évolution temporelle d'une réponse au sein d'un individu et sur les facteurs qui l'influencent.
- elles permettent d'économiser sur le nombre de sujets à étudier par rapport à une étude transversale (pour un même niveau de puissance statistique). En effet, les mesures répétées sur un même sujet fournissent plus d'informations qu'une seule mesure obtenue sur un seul sujet.
- les comparaisons peuvent être faites au sein d'un même sujet plutôt qu'entre plusieurs sujets. En effet, chaque sujet peut servir de son propre contrôle. Cela permet également d'éliminer les sources de variabilité inter-sujets de l'erreur expérimentale.

Cependant, l'analyse de données longitudinales doit également faire face à plusieurs challenges. Tout d'abord, les mesures répétées au sein d'un même groupe présentent une structure de covariance complexe qui doit être prise en compte de manière appropriée dans l'analyse. L'hypothèse d'indépendance des modèles classiques ne tient donc plus. Par ailleurs, ce type d'études est particulièrement sujet aux données manquantes (eg. perte de suivi ou décès du sujet) et sont donc souvent non-balancées (Verbeke, 1997; Fitzmaurice *et al.*, 2008).

Ces dernières années ont vu de nombreux progrès en matière de méthodologie statistique pour les données longitudinales. La suite de ce chapitre propose une description de ces méthodes en se focalisant sur l'analyse de réponses longitudinales continues et normalement distribuées.

3.1 ANOVA à mesures répétées

Une des premières méthodes proposées pour analyser des données longitudinales est l'ANOVA à *mesures répétées*. Une particularité du RM-ANOVA est qu'il prend en compte la variabilité inter-individus et l'exclut de la variabilité résiduelle. Il suppose que chaque sujet a un niveau de réponse sous-jacent qui persiste au cours du temps (Kutner *et al.*, 2005). Le design expérimental le plus couramment rencontré est le suivant : chaque sujet est assigné à un groupe défini par un traitement et suivi au cours du temps. Ici, les sujets sont emboîtés dans des groupes et croisés avec le facteur temps. De ce fait, chaque sujet n'est observé qu'à l'intérieur d'un seul groupe, et chaque sujet n'est observé qu'une seule fois par temps (Hedeker et Gibbons, 2006).

Un modèle ANOVA à trois facteurs est alors considéré avec les effets temps et traitements considérés comme fixe et l'effet sujet comme aléatoire. Ce modèle mixte est décrit comme suit :

$$y_{ijk} = \mu + \rho_{i(j)} + \alpha_j + \beta_k + (\alpha\beta)_{jk} + \epsilon_{ijk} \text{ avec } i = 1, \dots, s; j = 1, \dots, a; k = 1, \dots, b$$

où y_{ijk} est la réponse observée pour le $i^{\text{ème}}$ sujet, pour le traitement j et au temps k . μ est la moyenne générale, $\rho_{i(j)}$ est l'effet aléatoire lié au sujet i (emboité dans le groupe j), α_j l'effet du groupe traitement j , β_k l'effet du temps k et $(\alpha\beta)_{jk}$ l'interaction entre le groupe j et le temps k . Les $\rho_{i(j)}$ sont supposés être des variables aléatoires indépendantes $\sim iN(0, \sigma_p^2)$ et les erreurs $\epsilon_{ijk} \sim iN(0, \sigma_e^2)$. Par ailleurs, pour éviter une surparamétrisation du modèle, des contraintes sont imposées sur les paramètres : $\sum_j \alpha_j = 0$, $\sum_k \beta_k = 0$, $\sum_k (\alpha\beta)_{jk} = 0 \forall j$ et $\sum_j (\alpha\beta)_{jk} = 0 \forall k$. Ce modèle considère également qu'il n'y a pas d'interaction entre le sujet et le traitement et que le nombre de sujets est identique dans chaque groupe.

Dans le cas d'un design équilibré, la somme des carrés totale, soit la somme des déviations au carré de chaque observation par rapport à la moyenne globale, est décomposée comme suit :

$$\begin{aligned} SS_T &= \sum_{ijk} (y_{ijk} - \bar{y}_{...})^2 \\ &= bs \sum_j (\bar{y}_{.j} - \bar{y}_{...})^2 + as \sum_k (\bar{y}_{..k} - \bar{y}_{...})^2 + s \sum_{jk} (\bar{y}_{.jk} - \bar{y}_{.j} - \bar{y}_{..k} + \bar{y}_{...})^2 + \\ &\quad b \sum_{ij} (\bar{y}_{ij.} - \bar{y}_{.j.}) + \sum_{ijk} (y_{ijk} - \bar{y}_{ij.} - \bar{y}_{.jk} + \bar{y}_{.j.})^2 \\ &= SS_A + SS_B + SS_{AB} + SS_{S(A)} + SS_E \end{aligned} \tag{II.1}$$

où

- $\bar{y}_{...} = \sum_{i=1}^s \sum_{j=1}^a \sum_{k=1}^b \frac{y_{ijk}}{abs}$ est la moyenne estimée de l'ensemble des observations.
- $\bar{y}_{.j} = \sum_{i=1}^s \sum_{k=1}^b \frac{y_{ijk}}{bs}$ est la moyenne estimée au sein du groupe de traitement j.
- $\bar{y}_{..k} = \sum_{i=1}^s \sum_{j=1}^a \frac{y_{ijk}}{as}$ est la moyenne estimée des observations au temps k.
- $\bar{y}_{.jk} = \sum_{i=1}^s \frac{y_{ijk}}{s}$ est la moyenne estimée des s sujets du groupe j au temps k.
- $\bar{y}_{ij.} = \sum_{k=1}^b \frac{y_{ijk}}{b}$ est la moyenne estimée des mesures du sujet i du groupe j.

Le tableau II.1 reprend les degrés de libertés (dl) associés à chaque effet, les carrés moyens (MS) ainsi que leurs espérances ($E(MS)$). Les MS sont obtenus en divisant les sommes des carrés par les dl.

Effet	SS	dl	MS	$E(MS)$	F-test
A	SS_A	$a - 1$	$MS_A = SS_A/dl_A$	$\sigma_\epsilon^2 + b\sigma_p^2 + \frac{bs \sum \alpha_i^2}{(a-1)}$	$MS_A/MS_{S(A)}$
B	SS_B	$b - 1$	$MS_B = SS_B/dl_B$	$\sigma_\epsilon^2 + \frac{as \sum \beta_j^2}{(b-1)}$	MS_B/MS_E
AB	SS_{AB}	$(a - 1)(b - 1)$	$MS_{AB} = SS_{AB}/dl_{AB}$	$\sigma_\epsilon^2 + s \frac{\sum \sum (\alpha\beta)_{ij}^2}{(a-1)(b-1)}$	MS_{AB}/MS_E
$S(A)$	$SS_{S(A)}$	$a(s - 1)$	$MS_{S(A)} = SS_{S(A)}/dl_{S(A)}$	$\sigma_\epsilon^2 + b\sigma_p^2$	$MS_{S(A)}/MS_E$
Error	SS_E	$a(s - 1)(b - 1)$	$MS_E = SS_E/dl_E$	σ_ϵ^2	
Total	SS_T	$asb - 1$			

Tableau II.1 – Table d'ANOVA et $E(MS)$ pour un modèle ANOVA à mesures répétées dans le cas balancé.

En ANOVA, la significativité des effets peut être attestée au moyen de tests F présentés ci-dessous. Dans le cas où il n'y a que des facteurs fixes, ce test compare l'importance d'un effet par rapport aux erreurs résiduelles sur base de la statistique de test $F_{obs} = \frac{MS_f}{MS_E}$. Cependant, dans le cas d'une ANOVA mixte, plusieurs sources aléatoires sont présentes dans le modèle. Il faut dès lors se référer à l' $E(MS)$ pour décider quelle MS doit être assignée au dénominateur de la statistique de test.

— **Facteur interaction AB :**

H_0 : tous les $(\alpha\beta)_{jk} = 0 \forall j, k$

H_1 : au moins un des $(\alpha\beta)_{jk}$ est différent de 0

La statistique de test suivante est utilisée :

$$F^* = \frac{MS_{AB}}{MS_E} \sim_{H_0} F_{(dl_A, dl_E)}$$

— **Facteur temps B :**

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_b = 0$$

H_1 : au moins un des β_k est différent de 0

$$F^* = \frac{MS_B}{MS_E} \sim_{H_0} F_{(dl_B, dl_E)}$$

— **Facteur traitement A :**

$$H_0 : \alpha_1 = \alpha_2 = \dots = \alpha_a = 0$$

H_1 : au moins un des α_j est différent de 0

$$F^* = \frac{MS_A}{MS_{S(A)}} \sim_{H_0} F_{(dl_A, dl_{S(A)})}$$

Pour cet effet, nous pouvons constater sur base des E(MS) que le carré moyen lié à l'effet sujet doit être utilisé pour construire la statistique de test.

— **Facteur sujet S :**

$$H_0 : \sigma_p^2 = 0$$

$$H_1 : \sigma_p^2 > 0$$

$$F^* = \frac{MS_{S(A)}}{MS_E} \sim_{H_0} F_{(dl_{S(A)}, dl_E)}$$

Ces informations s'avéreront utiles ultérieurement dans ce mémoire lors de la dernière étape de la méthode LiMM-PCA.

3.2 MANOVA

L'ANOVA multivariée (MANOVA) a également été proposée et utilisée par le passé pour analyser les données longitudinales. Le principe général de cette méthode, qui ne sera pas détaillée dans ce mémoire, est qu'elle considère chaque mesure temporelle comme une variable distincte. Cette caractéristique ne lui permet cependant pas de décrire des tendances car elle ne reconnaît pas l'ordre temporel des variables (Hedeker et Gibbons, 2006).

Les deux méthodes classiques présentées montrent des caractéristiques qui limitent leur usage pour l'analyse de données longitudinales. D'une part, elles nécessitent un plan parfaitement équilibré pour les données. En effet, elles ne sont pas en mesure de gérer les données manquantes et ne tolèrent ni des temps d'observation différents pour

chaque sujet, ni des covariables qui varient dans le temps. D'autres part, ces modèles suggèrent des hypothèses très restrictives vis à vis de la matrice de covariance et ne reconnaissent pas l'ordre temporel des observations. Ils ne sont donc pas extensibles à des designs plus complexes (Fitzmaurice *et al.*, 2008).

3.3 Les modèles linéaires mixtes

Les méthodes ANOVA et MANOVA sont en fait des cas spéciaux de modèles plus généraux, qui permettent de gagner en flexibilité et de palier la plupart de ces contraintes : les modèles linéaires mixtes. Cette classe de modèles a gagné en intérêt ces dernières années et constitue aujourd'hui la référence en ce qui concerne l'analyse de données longitudinales. Les LMMs seront présentés dans le chapitre suivant.

Chapitre III

Modèles linéaires mixtes

Les modèles linéaires classiques sont principalement basés sur l'hypothèse d'indépendance entre les observations. Cette hypothèse ne peut cependant pas être vérifiée pour certains designs expérimentaux qui induisent une structure dans les données. C'est le cas notamment pour les données groupées telles que les designs en blocks, les données hiérarchisées, répétées ou encore longitudinales.

Les modèles linéaires mixtes offrent un cadre général qui permet de prendre en compte ces structures de corrélation complexes en incluant aussi bien des facteurs fixes qu'aléatoires. Un facteur est *aléatoire* s'il est approprié de considérer les niveaux de ce facteur comme issus d'une population sous-jacente (eg. sujet). Les effets sont dès lors supposés être des réalisations d'une distribution aléatoire. Cependant, si les niveaux spécifiques d'un facteur sont d'intérêt alors celui-ci est considéré comme *fixe* (eg. traitement) (Martin et Govaerts, 2020).

Cette section a pour but de présenter de manière générale les LMMs et de décrire dans quelles mesures ils sont utilisés pour analyser des données longitudinales. Sauf contre indication, les explications proviennent principalement de l'article de Govaerts *et al.* (2019), ainsi que des ouvrages des références suivant : "Applied mixed models in medicine" (Brown et Prescott, 2014), "Longitudinal data analysis" (Hedeker et Gibbons, 2006) et "Linear mixed models for longitudinal data" (Verbeke, 1997)

1 Définition et notations

Sous sa forme la plus générale, le modèle linéaire mixte s'écrit :

$$\mathbf{y} = \mathbf{X}\boldsymbol{\alpha} + \mathbf{Z}\boldsymbol{\beta} + \boldsymbol{\epsilon} \quad (\text{III.1})$$

où $\mathbf{X}$ la matrice de modèle des effets fixes. Elle est de dimension $n \times p$ où p est le nombre de paramètres fixes du modèle (terme constant inclus). Cette matrice est organisée en blocs et les variables catégorielles peuvent être codées de différentes manières. La méthode *sum-to-zero coding* sera néanmoins utilisée. Nous verrons plus en détails comment cette matrice est construite dans le cadre de la méthode ASCA au chapitre suivant. Le deuxième bloc est lié aux effets aléatoires. La matrice de design $\mathbf{Z}$ est de dimension $n \times q$ où q correspond à la somme du nombre de niveaux des effets aléatoires. Elle est construite à l'aide d'un codage fictif (avec des 0 et des 1) pour les variables catégorielles (aléatoires) du modèle et aucun intercept n'est inclus. $\boldsymbol{\beta}$ est un vecteur de variables aléatoires associées à chaque effet aléatoire. $\boldsymbol{\epsilon}$ représente la variation résiduelle, c'est à dire celle qui n'est incluse dans aucune autre composante du modèle.

Les hypothèses principales pour les effets aléatoires sont $\boldsymbol{\beta} \sim N(0, \mathbf{G})$ et $\boldsymbol{\epsilon} \sim N(0, \mathbf{R})$ avec $\mathbf{G}$ et $\mathbf{R}$ les matrices de covariance de tailles respectives $q \times q$ et $n \times n$. Par ailleurs, les effets aléatoires et les résidus sont supposés indépendants.

Le modèle linéaire mixte est également caractérisé par son espérance et sa variance qui s'écrivent :

$$\mathbb{E}[\mathbf{y}] = \mathbf{X}\boldsymbol{\alpha}$$

et

$$\mathbf{V} = \text{var}(\mathbf{y}) = \text{var}(\mathbf{X}\boldsymbol{\alpha} + \mathbf{Z}\boldsymbol{\beta} + \boldsymbol{\epsilon}) = \mathbf{Z}\text{var}(\boldsymbol{\beta})\mathbf{Z}' + \text{var}(\boldsymbol{\epsilon}) = \mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R}$$

On en déduit que la distribution marginale de $\mathbf{y}$ suit une loi normale multivariée : $\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\alpha}, \mathbf{V})$. Contrairement aux GLM, la matrice $\mathbf{V}$ n'est plus diagonale mais inclut des termes de covariance entre les observations liés aux termes aléatoires. En effet, deux mesures effectuées sur un même sujet seront corrélées là où, pour des sujets différents, elles seront indépendantes.

Le modèle linéaire mixte peut être utilisé de différentes manières afin d'analyser des données longitudinales. La première approche, certainement la plus simpliste, consiste à considérer un modèle à *effets aléatoires* avec l'effet sujet ajusté comme tel (aléatoire). Les observations faites sur un même sujet partageront dès lors la même covariance (seront corrélées). Un deuxième modèle, dit à *covariance pattern*, permet de modéliser cette corrélation en imposant différentes structures de covariance à la matrice des résidus $\mathbf{R}$. Enfin, lorsque l'intérêt de l'étude se porte sur la relation entre la réponse et le temps, un modèle à *coefficients aléatoires* peut être utilisé (Brown et Prescott, 2014). Nous verrons que ces trois approches se différencient principalement par la structure des matrices de covariance $\mathbf{G}$ et $\mathbf{R}$ qu'elles engendrent.

1.1 Modèles à effets aléatoires

Ces modèles visent à prendre en compte la structure groupée de certaines données (multiniveaux) en introduisant des effets aléatoires. De cette manière, la variabilité totale va être décomposée en différentes composantes de variances qui pourront être analysées indépendamment.

Typiquement, dans le cas de données longitudinales, les mesures sont répétées au sein d'un même bloc (eg. sujet) et les effets des différents blocs vont être considérés comme aléatoires. Cela va permettre de prendre en compte la variabilité biologique propre à chaque individu et de généraliser les résultats à une population plus large.

Le modèle décrit ci dessus avec l'effet sujet comme aléatoire est équivalent au modèle ANOVA à mesures répétées dans le cas où les données sont balancées (en considérant toujours les effets temps, traitement et leur interaction comme fixes). En effet, le modèle présenté à la section 3.1 du chapitre II

$$y_{ijk} = \mu + \rho_{i(j)} + \alpha_j + \beta_k + (\alpha\beta)_{jk} + \epsilon_{ijk}$$

peut être généralisé et réécrit sous la forme d'un modèle linéaire mixte pour deux traitements, trois temps et deux sujets par traitement :

$$\mathbf{y} = \mathbf{X}\boldsymbol{\alpha} + \mathbf{Z}\boldsymbol{\beta} + \boldsymbol{\epsilon} = \begin{pmatrix} 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & -1 & 1 & 0 & -1 & 0 \\ 1 & -1 & 0 & 1 & 0 & -1 \\ 1 & -1 & -1 & -1 & 1 & 1 \\ 1 & -1 & 1 & 0 & -1 & 0 \\ 1 & -1 & 0 & 1 & 0 & -1 \\ 1 & -1 & -1 & -1 & 1 & 1 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \beta_1 \\ \beta_2 \\ \alpha\beta_{11} \\ \alpha\beta_{12} \end{pmatrix} + \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \rho_{1(1)} \\ \rho_{2(1)} \\ \rho_{1(2)} \\ \rho_{2(2)} \end{pmatrix} + \boldsymbol{\epsilon}$$

$\boldsymbol{\beta}$ suit une loi normale multivariée $N(0, \mathbf{G})$ avec

$$\mathbf{G} = \begin{pmatrix} \sigma_p^2 & 0 & 0 & 0 \\ 0 & \sigma_p^2 & 0 & 0 \\ 0 & 0 & \sigma_p^2 & 0 \\ 0 & 0 & 0 & \sigma_p^2 \end{pmatrix}$$

où σ_p^2 est la composante de variance liée à l'effet sujet. Les résidus sont quant à eux supposés indépendants et la matrice $\mathbf{R}$ est diagonale ($\mathbf{R} = \sigma^2 I_n$). La matrice $\mathbf{V}$ du modèle marginal est donc bloc-diagonale et s'écrit comme suit :

$$\mathbf{V} = \begin{pmatrix} \mathbf{V}_1 & 0 & 0 & 0 \\ 0 & \mathbf{V}_2 & 0 & 0 \\ 0 & 0 & \mathbf{V}_3 & 0 \\ 0 & 0 & 0 & \mathbf{V}_4 \end{pmatrix}$$

où

$$\mathbf{V}_i = \begin{pmatrix} \sigma^2 + \sigma_p^2 & \sigma_p^2 & \sigma_p^2 \\ \sigma_p^2 & \sigma^2 + \sigma_p^2 & \sigma_p^2 \\ \sigma_p^2 & \sigma_p^2 & \sigma^2 + \sigma_p^2 \end{pmatrix} \text{ pour } i = 1, \dots, 4$$

Chaque bloc $\mathbf{V}_i$ représente un sujet et est de taille équivalente au nombre de mesures effectuées sur ce sujet. Les mesures faites sur un même sujet partagent le même effet aléatoire, elles sont donc corrélées :

$$\text{cor}(y_{ijk}, y_{ijk'}) = \frac{\text{cov}(y_{ijk}, y_{ijk'})}{\sqrt{\text{var}(y_{ijk})}\sqrt{\text{var}(y_{ijk'})}} = \frac{\sigma_p^2}{\sigma^2 + \sigma_p^2}$$

La variance d'une seule observation dans un modèle à effet aléatoire est égale à la somme des composantes de variance : $\text{var}(y_{ijk}) = \sigma^2 + \sigma_p^2$.

Ce modèle est assez restrictif dans le sens où il suppose une covariance constante entre les différentes mesures faites sur un même sujet. Or, souvent, cette corrélation n'est pas constante mais tend, par exemple, à diminuer au plus les observations sont éloignées dans le temps. Le modèle suivant va permettre de rendre compte de ces structures plus complexes de corrélation (Brown et Prescott, 2014).

1.2 Modèles à covariance pattern

Les modèles à covariance pattern visent à induire une structure de corrélation, non pas par l'introduction d'effets aléatoires mais en la spécifiant explicitement dans la matrice de corrélation des termes d'erreurs $\mathbf{R}$. Le modèle suivant à effets fixes, et sans l'effet sujet aléatoire, est alors considéré :

$$y_{ijk} = \mu + \alpha_j + \beta_k + (\alpha\beta)_{jk} + \epsilon_{ijk}$$

Compte tenu que le modèle n'inclut pas d'effets aléatoires, la matrice de covariance $\mathbf{V}$ est égale à la matrice des résidus $\mathbf{R}$. En effet, $\mathbf{V} = \mathbf{ZGZ}' + \mathbf{R} = \mathbf{R}$. Cependant, la différence avec un modèle linéaire simple est que, dans un modèle à structure de covariance, la matrice $\mathbf{R}$ n'est plus diagonale mais bloc-diagonale afin de tenir compte de la corrélation qui existe entre les mesures répétées dans un même bloc. Pour quatre sujets, $\mathbf{R}$ est obtenue comme suit :

$$\mathbf{R} = \begin{pmatrix} \mathbf{R}_1 & 0 & 0 & 0 \\ 0 & \mathbf{R}_2 & 0 & 0 \\ 0 & 0 & \mathbf{R}_3 & 0 \\ 0 & 0 & 0 & \mathbf{R}_4 \end{pmatrix}$$

Chaque bloc $\mathbf{R}_i$ ($i = 1, \dots, 4$) correspond à un sujet et est de dimension égale au nombre d'observations répétées sur ce sujet. Les covariances entre des observations provenant de sujets différents sont nulles.

Plusieurs structures de dépendance pour les observations au sein d'un bloc sont disponibles. Les plus courantes sont présentées ci-dessous en considérant toujours une étude à trois points dans le temps.

- *Corrélation de type "non structurée"* : les variances des réponses à chaque temps et les covariances pour chaque paire de temps peuvent être différentes.

$$\mathbf{R}_i = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_2^2 & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 \end{pmatrix}$$

$$\text{cov}(\epsilon_{ij}, \epsilon_{ij'}) = \sigma_{jj'}$$

Nombre de paramètres : $b(b+1)/2$

- *Composante symétrique* : toutes les covariances sont égales.

$$\mathbf{R}_i = \sigma^2 \begin{pmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{pmatrix}$$

$$\text{cov}(\epsilon_{ij}, \epsilon_{ij'}) = \rho\sigma^2$$

Nombre de paramètres : 2

- *Autorégressif d'ordre 1* : la corrélation décroît exponentiellement lorsque les observations sont plus éloignées dans le temps.

$$\mathbf{R}_i = \sigma^2 \begin{pmatrix} 1 & \rho & \rho^2 \\ \rho & 1 & \rho \\ \rho^2 & \rho & 1 \end{pmatrix}$$

$$\text{cov}(\epsilon_{ij}, \epsilon_{ij'}) = \sigma^2 \rho^{|j-j'|}$$

Nombre de paramètres : 2

- *Toeplitz* : les covariances sont différentes pour chaque niveau de séparation entre les temps.

$$\mathbf{R}_i = \sigma^2 \begin{pmatrix} 1 & \rho_1 & \rho_2 \\ \rho_1 & 1 & \rho_1 \\ \rho_2 & \rho_1 & 1 \end{pmatrix}$$

$$\text{cov}(\epsilon_{ij}, \epsilon_{ij'}) = \alpha^{|j-j'|}$$

Nombre de paramètres : b

Il est à noter qu'un modèle à intercept (eg. effet sujet) aléatoire est similaire à un modèle à composante symétrique en termes de résultats (eg. estimations des effets fixes, matrice $\mathbf{V}$, log-vraisemblance). Ces modèles seront comparés ultérieurement au chapitre V. Notons également que de nombreuses autres structures existent et sont

décrites dans la littérature (Brown et Prescott, 2014). Il est également envisageable d'ajuster des paramètres de covariance différents pour chaque groupe de traitement (hétéroscédasticité). C'est le cas notamment lorsque les mesures sont plus variables pour le traitement que pour le contrôle.

Choisir la structure la plus appropriée n'est pas chose aisée. Un compromis doit être trouvé quant au nombre de paramètres à inclure. Il faut sélectionner la structure qui ajuste au mieux la covariance des données observées tout en étant parcimonieux pour éviter une surparamétrisation du modèle (Wolfinger, 1997).

Pour des modèles emboîtés, le test de maximum de vraisemblance permettra de vérifier si l'ajout de paramètres améliore significativement le modèle. Sinon, des critères d'ajustement du modèle tels que l'AIC ou le BIC peuvent servir de base de comparaison.

1.3 Modèles à coefficients aléatoires

La dernière approche consiste à concevoir un modèle qui décrit linéairement la relation entre la réponse et la variable temps en considérant cette dernière comme continue et non plus catégorielle. Une structure de covariance entre les observations pour chaque sujet est alors obtenue en permettant à l'ordonnée à l'origine et à la pente-temps de varier entre les différents sujets.

Ce modèle, aussi appelé *random slope and intercept*, est obtenu en considérant les effets sujet et l'interaction pente-sujet comme aléatoires :

$$y_{ijk} = \mu + \tau_j + p_{i(j)} + \lambda t_{ijk} + (p\lambda)_{i(j)} t_{ijk} + \epsilon_{ijk}$$

où μ représente l'intercept moyen, λ la pente moyenne liée au temps et τ_j l'effet du traitement j (afin d'alléger les notations l'interaction temps-traitement est négligée).

Les composantes aléatoires s'interprètent comme ceci :

- $p_{i(j)}$ représente l'effet aléatoire sujet, soit la déviation de l'individu par rapport à l'intercept moyen ;
- $(p\lambda)_{i(j)}$ est la correction pour la pente de chaque sujet i par rapport à la pente moyenne ;
- ϵ_{ijk} , est la variabilité des différentes mesures sur un même individu par rapport à sa pente moyenne, $\epsilon_{ijk} \sim N(0, \sigma^2)$

Pour chaque sujet, les deux effets aléatoires sont la réalisation de variables qui suivent une loi normale multivariée :

$$\begin{pmatrix} p_{i(j)} \\ (p\lambda)_{i(j)} \end{pmatrix} \sim \mathcal{N}(0, \mathbf{G}_i) \text{ avec } \mathbf{G}_i = \begin{pmatrix} \sigma_p^2 & \sigma_{p,p\lambda} \\ \sigma_{p,p\lambda} & \sigma_{p\lambda}^2 \end{pmatrix}$$

Considérons toujours un modèle à quatre individus pour lesquels il y a trois observations à des temps différents. Les matrices $\mathbf{Z}$ et $\mathbf{G}$ s'écrivent

$$\mathbf{Z} = \begin{pmatrix} p_1 & p\lambda_1 & p_2 & p\lambda_2 & p_3 & p\lambda_3 & p_4 & p\lambda_4 \\ 1 & t_{11} & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & t_{12} & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & t_{13} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & t_{21} & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & t_{22} & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & t_{23} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & t_{31} & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & t_{32} & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & t_{33} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & t_{41} \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & t_{42} \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & t_{43} \end{pmatrix}$$

$$\mathbf{G} = \begin{pmatrix} \mathbf{G}_1 & 0 & 0 & 0 \\ 0 & \mathbf{G}_2 & 0 & 0 \\ 0 & 0 & \mathbf{G}_3 & 0 \\ 0 & 0 & 0 & \mathbf{G}_4 \end{pmatrix}$$

Par ailleurs, étant donné que la matrice des résidus $\mathbf{R}$ est diagonale, la covariance du modèle $\mathbf{V} = \mathbf{ZGZ}' + \mathbf{R}$ est bloc-diagonale et égale à

$$\mathbf{V} = \begin{pmatrix} \mathbf{V}_1 & 0 & 0 & 0 \\ 0 & \mathbf{V}_2 & 0 & 0 \\ 0 & 0 & \mathbf{V}_3 & 0 \\ 0 & 0 & 0 & \mathbf{V}_4 \end{pmatrix} \text{ avec } \mathbf{V}_i = \begin{pmatrix} \sigma^2 + c_{i,11} & c_{i,12} & c_{i,13} \\ c_{i,12} & \sigma^2 + c_{i,22} & c_{i,23} \\ c_{i,13} & c_{i,23} & \sigma^2 + c_{i,33} \end{pmatrix}$$

où $i = 1, \dots, 4$ et $c_{i,kk'} = \sigma_p^2 + (t_{ik} + t_{ik'})\sigma_{p,p\lambda} + t_{ik}t_{ik'}\sigma_{p\lambda}^2$. La variance augmente donc avec le temps.

Dans le cas où seul l'intercept varie (*random intercept*), toutes les trajectoires individuelles sont parallèles à la trajectoire globale et la structure de covariance est similaire à celle du modèle à composante symétrique.

$$y_{ijk} = (\mu + p_{ij}) + \lambda t_{ijk} + \epsilon_{ijk}$$

Le modèle à coefficients aléatoires est le plus adapté lorsque les données sont mesurées à des temps variables pour les sujets ou lorsque la relation entre le temps et la réponse est supposée linéaire. Cependant, il nécessite de considérer le temps comme une variable continue. Or les méthodes multivariées développées jusqu'ici ne permettent que des designs avec exclusivement des variables catégorielles et ce modèle ne sera donc pas étendu dans le cadre de ce mémoire au cas de réponses multivariées.

2 Estimation

La procédure d'ajustement se fait en trois grandes étapes : estimation des effets fixes, prédictions des coefficients pour les effets aléatoires et estimation des composantes de variances (σ^2 inclus). Dans des cas très simples, l'estimation des effets aléatoires peut se faire en utilisant les tables d'ANOVA. Cependant, en général, la méthode du maximum de vraisemblance (ML) est utilisée pour ajuster les modèles linéaires mixtes.

Étant donné que les observations ne sont pas indépendantes, la fonction de vraisemblance se construit sur la densité multivariée des $\mathbf{y}_i$: $\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\alpha}, \mathbf{V})$. Le logarithme de la fonction de vraisemblance est donné par :

$$l = \log(L) = k - \frac{1}{2}[\log|\mathbf{V}| + (\mathbf{y} - \mathbf{X}\boldsymbol{\alpha})'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\alpha})], \quad (\text{III.2})$$

où k est une constante. Les valeurs des paramètres qui maximisent cette vraisemblance peuvent maintenant être déterminées. Cependant, cette méthode suppose que $\mathbf{V}$ est connue et donnera des estimateurs biaisés dans le cas où elle est estimée. Il est alors recommandé, d'estimer dans un premier temps les paramètres de variance inclus dans $\mathbf{V}$ par maximisation de la vraisemblance restreinte (REML) en considérant les effets fixes comme constants. Ces estimateurs sont obtenus à l'aide d'algorithmes d'optimisation itératifs étant donné qu'aucune solution analytique n'est dérivable.

Dans un deuxième temps, les paramètres des effets fixes sont estimés en maximisant l'équation III.2 avec $\mathbf{V}$ égal à l'estimateur REML $\hat{\mathbf{V}}$:

$$\hat{\boldsymbol{\alpha}} = (\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{y}. \quad (\text{III.3})$$

Enfin, les prédictions pour les effets aléatoires sont obtenues comme suit :

$$\hat{\boldsymbol{\beta}} = \hat{\mathbf{G}}\mathbf{Z}'\hat{\mathbf{V}}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\alpha}}). \quad (\text{III.4})$$

Pour plus d'informations concernant les méthodes d'estimations en LLM, voir Brown et Prescott (2014).

3 Tests d'hypothèses

La significativité des effets fixes peut être déterminée à l'aide des tests F et t de Wald moyennant une approximation des ddl par la méthode de Satterthwaite ou de Kenward et Roger (Verbeke, 1997). Le test de rapport de vraisemblance (LRR) peut également être utilisé afin de tester la significativité des paramètres de variances ou des effets fixes en comparant deux modèles emboîtés. La statistique de test suivante est alors calculée :

$$\lambda = 2 [\log(M_1) - \log(M_0)]$$

où M_1 et M_0 sont les vraisemblances des modèles qui incluent ou excluent respectivement l'effet testé. Sous l'hypothèse nulle, cette statistique suit une loi chi-carré χ_d^2 où d , le nombre de degrés de liberté, est égal à la différence entre le nombre de paramètres libres dans les modèles réduit et complet. Néanmoins, pour les paramètres de variance, l'hypothèse nulle $H_0 : \sigma^2 = 0$ fait intervenir des valeurs sur la frontière. La théorie asymptotique classique n'est donc plus valable mais il a été démontré que, sous certaines conditions, la statistique de test peut être approximée par un mélange de χ^2 : $H_0 : 0.5\chi_0^2 + 0.5\chi_1^2$ (Pinheiro et Bates, 2000).

Une méthode alternative, et qui sera utilisée dans le cadre de ce mémoire, consiste à déterminer la distribution de la statistique de test LRR sous H_0 par une approche bootstrap (Davison et Hinkley, 1997).

Il est important de noter que pour tester les effets fixes, les ML doivent être utilisées dans le test LRR alors que pour les paramètres de variance les REML sont préférées (Verbeke, 1997).

Chapitre IV

Extensions des LMMs aux réponses de grande dimension

Ce chapitre a pour objectif de présenter les méthodes statistiques multivariées qui combinent à la fois réduction de dimension et modélisation de designs complexes quand un grand nombre de réponses sont observées simultanément pour chaque expérience. Dans un premier temps, les méthodes ASCA⁺ et APCA⁺ seront brièvement décrites. Nous verrons ensuite leurs limites face à la présence de facteurs aléatoires dans le design expérimental. Enfin, la méthode LiMM-PCA, développée par Martin et Govaerts (2020) sera présentée.

1 ASCA⁺ et APCA⁺

Dans le cadre d’une étude métabolomique, chaque analyse réalisée donne lieu à un vecteur de réponses (eg. spectre). Les vecteurs obtenus sur les différents échantillons peuvent être regroupés au sein d’une matrice de réponses $\mathbf{Y}$. Les méthodes ASCA/APCA peuvent ensuite être utilisées pour mettre en évidence les effets des facteurs qui composent le plan d’expérience. Ces méthodes fournissent néanmoins des estimateurs biaisés dans le cas où les données ne sont pas balancées. Dans ce contexte, Thiel *et al.* (2017) a récemment proposé deux nouvelles approches, l’ASCA+ et l’APCA+. Celles-ci sont basées sur l’utilisation d’un modèle linéaire général (GLM) au lieu d’une ANOVA et permettent de généraliser l’utilisation des méthodes classiques à tout type de designs expérimentaux comportant des facteurs fixes catégoriels.

Le cadre général de l'ASCA⁺ est représenté à la figure IV.1. La suite de cette section est consacrée à une présentation succincte des concepts principaux de cette approche décrite plus en détails dans Thiel *et al.* (2017).

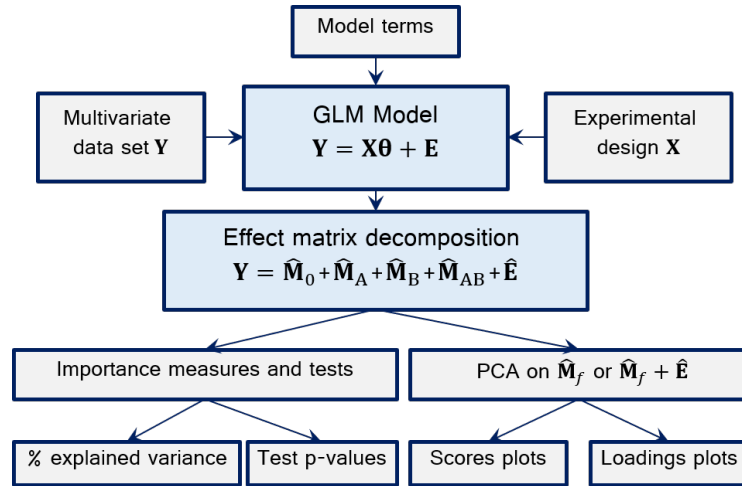


Figure IV.1 – Etapes principales des méthodes ASCA⁺ et APCA⁺ (Guisset *et al.*, 2019).

La méthode s'articule en deux grandes étapes. Dans un premier temps, la matrice de réponses $\mathbf{Y}$ est décomposée en matrices d'effets selon un modèle GLM lié au plan d'expérience. Dans un deuxième temps, une technique multivariée de dimension de réduction (PCA) est appliquée à chaque matrice dans le but de visualiser graphiquement les effets des facteurs.

1.1 Étape 1 : Modèle linéaire général et matrice du modèle

Le modèle GLM multivarié s'écrit comme suit :

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\Theta} + \mathbf{E},$$

où $\mathbf{Y}$ est la matrice $n \times m$ des réponses observées, $\mathbf{X}$ la matrice du modèle de dimension $n \times p$, $\boldsymbol{\Theta}$ la matrice $p \times m$ des paramètres et $\mathbf{E}$ la matrice $n \times m$ des erreurs dont les colonnes sont supposées suivre une loi normale multivariée $N(0, \sigma^2 \mathbf{I}_n)$. Notons que m correspond au nombre de variables réponses.

La clé du modèle GLM réside dans la construction de la matrice modèle $\mathbf{X}$. En effet, cette matrice est encodée de manière bien spécifique selon une méthode appelée *sum-to-zero* ou *deviation coding* et sur base du plan d'expérience utilisé pour collecter les données. Elle est organisée en $F + 1$ blocs correspondants au terme constant et à chacun des effets du modèle : $\mathbf{X} = (\mathbf{X}_0 \mid \mathbf{X}_1 \mid \dots \mid \mathbf{X}_F)$. Pour un facteur A comportant a niveaux, une matrice $\mathbf{X}_a$ avec $a - 1$ colonnes est créée. Les $a - 1$ premiers niveaux sont codés par des 0 et 1 alors que le dernier niveau, qui peut être estimé sur base des autres, est codé par des -1. Les colonnes pour les interactions sont obtenues en multipliant deux à deux les colonnes des effets principaux.

Pour illustrer les propos de cette section, un modèle à deux facteurs A et B avec interaction sera considéré. Si les facteurs ont respectivement 2 et 3 niveaux et que le plan ne comprend pas de répétitions, la matrice du modèle est encodée comme suit :

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_0 & \mathbf{X}_A & \mathbf{X}_B & \mathbf{X}_{AB} \\ \begin{array}{c|c|c|c} 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & 0 \\ 1 & -1 & 0 & 1 \\ 1 & -1 & -1 & -1 \end{array} & \begin{array}{c|c} 1 & 0 \\ 0 & 1 \\ -1 & -1 \\ -1 & 0 \\ 0 & -1 \\ 1 & 1 \end{array} \end{pmatrix}$$

La matrice $\mathbf{X}$ ne dépend pas du nombre de variables réponses et sera donc similaire dans les cas uni- et multi-variés.

1.2 Étape 2 : Estimation des paramètres et décomposition en matrices d'effet

Des estimateurs non biaisés des paramètres du modèle multivarié peuvent être obtenus par la méthode des moindres carrés ordinaires :

$$\hat{\Theta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

La matrice des paramètres estimés, de dimension $p \times m$, sert alors de base pour la décomposition de la matrice de réponses en matrices correspondantes aux différentes sources de variation : effets principaux, interactions et erreurs. En effet, tout comme $\mathbf{X}$, $\hat{\Theta}$ est divisée en blocs par rapport à chaque effet : $\hat{\Theta}' = (\hat{\Theta}'_0 \mid \hat{\Theta}'_1 \mid \dots \mid \hat{\Theta}'_F)$. Pour l'exemple considéré, la matrice des paramètres estimés est la suivante :

$$\hat{\Theta} = \begin{pmatrix} \hat{\mu}_{..1} & \hat{\mu}_{..2} & \dots & \hat{\mu}_{..m} \\ \hat{\alpha}_{11} & \hat{\alpha}_{12} & \dots & \hat{\alpha}_{1m} \\ \hat{\beta}_{11} & \hat{\beta}_{12} & \dots & \hat{\beta}_{1m} \\ \hat{\beta}_{21} & \hat{\beta}_{22} & \dots & \hat{\beta}_{2m} \\ \widehat{(\alpha\beta)}_{111} & \widehat{(\alpha\beta)}_{112} & \dots & \widehat{(\alpha\beta)}_{11m} \\ \widehat{(\alpha\beta)}_{121} & \widehat{(\alpha\beta)}_{122} & \dots & \widehat{(\alpha\beta)}_{12m} \end{pmatrix} = \begin{pmatrix} \hat{\theta}_0 \\ \hat{\theta}_A \\ \hat{\theta}_B \\ \hat{\theta}_{AB} \end{pmatrix}$$

La matrice pour l'effet F est alors obtenue en multipliant les blocs des matrices $\mathbf{X}$ et $\hat{\Theta}$ liés à cet effet :

$$\hat{M}_f = \mathbf{X}_f \hat{\Theta}_f$$

Chaque matrice d'effet est de même dimension que $\mathbf{Y}$ ($n \times m$) et représente la part de $\mathbf{Y}$ expliquée par chaque effet du modèle (terme constant, effets principaux et interactions).

Par exemple, pour l'effet B :

$$\hat{M}_B = \mathbf{X}_B \hat{\Theta}_B = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ -1 & -1 \\ 1 & 0 \\ 0 & 1 \\ -1 & -1 \end{pmatrix} \begin{pmatrix} \hat{\beta}_{11} & \hat{\beta}_{12} & \dots & \hat{\beta}_{1m} \\ \hat{\beta}_{21} & \hat{\beta}_{22} & \dots & \hat{\beta}_{2m} \end{pmatrix}$$

$$\hat{\mathbf{M}}_B = \begin{pmatrix} \hat{\beta}_{11} & \hat{\beta}_{12} & \dots & \hat{\beta}_{1m} \\ \hat{\beta}_{21} & \hat{\beta}_{22} & \dots & \hat{\beta}_{2m} \\ -\hat{\beta}_{11} - \hat{\beta}_{21} & -\hat{\beta}_{12} - \hat{\beta}_{22} & \dots & -\hat{\beta}_{1m} - \hat{\beta}_{2m} \\ \hat{\beta}_{11} & \hat{\beta}_{12} & \dots & \hat{\beta}_{1m} \\ \hat{\beta}_{21} & \hat{\beta}_{22} & \dots & \hat{\beta}_{2m} \\ -\hat{\beta}_{11} - \hat{\beta}_{21} & -\hat{\beta}_{12} - \hat{\beta}_{22} & \dots & -\hat{\beta}_{1m} - \hat{\beta}_{2m} \end{pmatrix}$$

La décomposition ASCA⁺ de la matrice réponse d'un modèle à F effets (hors terme constant) est donnée par :

$$\mathbf{Y} = \hat{\mathbf{M}}_0 + \sum_{f=1}^F \hat{\mathbf{M}}_f + \hat{\mathbf{E}} ; \quad (\text{IV.1})$$

où

- $\hat{\mathbf{E}} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\Theta}}$ est la matrice des résidus estimés
- $\hat{\mathbf{M}}_0$ est la matrice d'effet générale (liée aux termes constants). Elle contient m colonnes comprenant les moyennes globales $\bar{y}_{...l}$ pour chaque variable réponse $l = 1, \dots, m$.

1.3 Étape 3 : Analyse en composantes principales sur les matrices d'effets

Cette étape consiste à appliquer une analyse en composante principale sur les matrices d'effets. Pour rappel, l'ACP est une méthode statistique qui permet de représenter dans un espace, réduit à un certain nombre de composante, des réponses multivariées contenues dans une matrice $\mathbf{Y}$ de dimension $n \times m$ tout en conservant un maximum d'information.

La matrice $\mathbf{Y}$ est alors décomposée en

$$\mathbf{Y} = \mathbf{T}\mathbf{P}'$$

où

- $\mathbf{T}$ ($n \times r$), la matrice des scores, contient les coordonnées des observations initiales dans le nouvel espace.
- $\mathbf{P}$ ($m \times r$), la matrice des loadings, regroupe les poids des m variables initiales sur les différentes composantes principales (CP).
- $r = \min(n, m)$ est le nombre de CP. Celles-ci sont orthogonales et construites de manière à maximiser la variance entre les observations.

Lorsqu'une ACP est appliquée aux matrices d'effets, les graphes des loadings et des scores obtenus constituent une source importante d'informations dans l'interprétation des résultats du modèle GLM multivarié. La représentation des scores sur le plan formé par les deux premières composantes principales permet de visualiser comment les niveaux des effets sont situés les uns par rapport aux autres. Quant aux loadings, ils permettent d'identifier quelles variables $\mathbf{y}_i$ sont liées à chaque effet du modèle. Au plus les descripteurs sont intenses, au plus les variables sont bien représentées sur la CP.

La différence entre l'ASCA⁺ et l'APCA⁺ se marque exclusivement dans cette étape. En effet, la première applique l'ACP sur les matrices d'effets "pures", directement obtenues sur base de la décomposition, alors qu'en l'APCA⁺, la matrice des résidus est ajoutée aux matrices d'effets en amont de l'ACP :

$$\text{ASCA} : \hat{\mathbf{M}}_f = \mathbf{T}_f \mathbf{P}'_f \quad \text{APCA} : \hat{\mathbf{M}}_f + \hat{\mathbf{E}} = \mathbf{T}_f \mathbf{P}'_f$$

où $\mathbf{T}_f$ et $\mathbf{P}'_f$ correspondent respectivement aux matrices de scores et de loadings du facteur f .

Là où les scores ne représentent que les moyennes des niveaux des facteurs en ASCA, les variations individuelles dues aux résidus (et donc non expliquées par le modèle) peuvent être observées sur les graphiques des scores en APCA. De manière générale, les loadings en APCA comporteront plus de bruit lié à l'ajout des résidus et l'on préférera donc les loadings ASCA.

Une troisième méthode, l'ASCA-E, réalise l'ACP sur les matrices d'effets purs pour obtenir les loadings mais calcule une matrice de scores "augmentée" à partir de la formule suivante (Zwanenburg *et al.*, 2011; Thiel *et al.*, 2017) :

$$\mathbf{T}_f^E = (\hat{\mathbf{M}}_f + \hat{\mathbf{E}}) \mathbf{P}'_f$$

Cette méthode permet de visualiser la variabilité au sein de chaque facteur par rapport à la variabilité résiduelle et ce sur base des mêmes loadings que pour l'ASCA. Contrairement à dans l'APCA, la variabilité dans la matrice des scores $\mathbf{T}_f^E$ ne sera donc pas maximisée.

1.4 Étape 4 : Pourcentage de variation et tests de significativité

L'étape 4 consiste à évaluer l'importance relative de chaque effet principal et des éventuelles interactions du modèle. Thiel *et al.* (2017) se base sur la somme des carrés de Type III et la norme de Frobenius $||.||^2$ pour calculer le pourcentage de variance expliqué par un effet f :

$$\%Var f = \frac{||\hat{\mathbf{E}}_{/f}||^2 - ||\hat{\mathbf{E}}_{full}||^2}{||\mathbf{Y} - \hat{\mathbf{M}}_0||^2} \times 100$$

où $\hat{\mathbf{E}}_{full}$ et $\hat{\mathbf{E}}_{/f}$ sont les matrices des résidus estimées respectivement sur base des modèles complet et réduit, c'est à dire dont l'effet f a été retiré. Notons que quand le design n'est pas balancé, la somme des pourcentages de variance ne sera pas exactement égale à 100%.

Il est ensuite possible de déterminer quels effets sont globalement significatifs sur base d'une statistique de test $F_{obs} = MS_f/MS_E$. La méthode initialement utilisée dans l'article de Thiel *et al.* (2017) consiste à calculer des p-valeurs via des tests de permutation. Elle ne sera pas vue en détails ici mais est décrite dans la littérature (Anderson et Braak (2003); Zwanenburg *et al.* (2011)). Une autre méthode, plus générale et qui sera utilisée dans le cadre de ce mémoire, consiste à attester de la significativité des facteurs sur base d'une procédure de bootstrap paramétrique.

Pour tester la significativité d'un effet f , la procédure est la suivante :

1. Définir le modèle sous H_0 , obtenu en excluant l'effet f du modèle complet H_1 .
2. Ajuster les modèles réduit H_0 et complet H_1 à la matrice des réponses $\mathbf{Y}$.
3. Calculer la statistique de test $F_{obs} = \frac{(SSE_{/f} - SSE_{full})/dl_{/f} - dl_{full}}{SSE_{full}/dl_{full}}$.
4. Calculer les valeurs prédites $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\Theta}}$ et les résidus $\hat{\mathbf{E}} = \mathbf{Y} - \hat{\mathbf{Y}}$ du modèle sous H_0 . Ces deux matrices seront utilisées dans la boucle bootstrap.
5. Pour un nombre fixé B d'itérations ($b = 1, \dots, B$), répéter les étapes suivantes :
 - Échantillonner $\hat{\mathbf{E}}_b$ avec remise parmi $\hat{\mathbf{E}}$
 - Générer une matrice de réponses bootstrap $\mathbf{Y}_b = \hat{\mathbf{Y}} + \hat{\mathbf{E}}_b$
 - Estimer les modèles complet et restreint sur base de cette nouvelle matrice $\mathbf{Y}_b$
 - Calculer la statistique de test F_b .
6. Calculer la p-valeur pour la statistique de test : $p_f^{boot} = \frac{\sum_{b=1}^B I(F_b \geq F_{obs}) + 1}{B+1}$.

2 LiMM-PCA

Les méthodes ASCA+ et APCA+ sont construites sur base de l'utilisation de modèles linéaires généraux pour décomposer la matrice de réponse en matrices d'effets. Les GLM considèrent tous les effets du modèle comme fixes. Cependant, des designs expérimentaux plus complexes qui incluent aussi bien des effets fixes qu'aléatoires sont couramment rencontrés. Il y a dès lors un intérêt certain à pouvoir étendre la méthodologie ASCA+ à des modèles linéaires mixtes. C'est ce qu'ont réalisé récemment Martin et Govaerts (2020) au travers d'une nouvelle approche nommée Linear Mixed Models-PCA (LiMM-PCA).

Dans le cas de réponses multivariées, l'utilisation de LMMs va néanmoins imposer plusieurs adaptations dans la méthodologie car ces modèles nécessitent que certaines hypothèses soient vérifiées. La matrice $\mathbf{Y}$ devra alors répondre à deux propriétés statistiques majeures : la normalité de chaque réponse et l'indépendance entre les variables. La première hypothèse pourra être satisfaite au moyen d'une éventuelle transformation mathématique sur les données. La deuxième nécessite qu'une ACP soit appliquée sur la matrice $\mathbf{Y}$ en amont de l'étape de modélisation et implique inévitablement des adaptations subséquentes de la méthodologie. Notons que le fait que les réponses ne soient pas indépendantes n'a pas d'impact en ASCA+ car les estimateurs sont obtenus par la méthode des moindres carrés ordinaires. Cependant, dans le cadre des LMMs, les paramètres sont estimés par maximum de vraisemblance et l'indépendance des réponses aura alors son importance.

Les différentes étapes de la méthode LiMM-PCA sont résumées dans la figure IV.2 et seront détaillées dans la suite de cette section. Un accent tout particulier sera mis sur les nouvelles contributions apportées par rapport à l'ASCA+ et mises en évidence en vert dans la figure.

La description des étapes de la LiMM-PCA se base essentiellement sur l'article de Martin et Govaerts (2020).

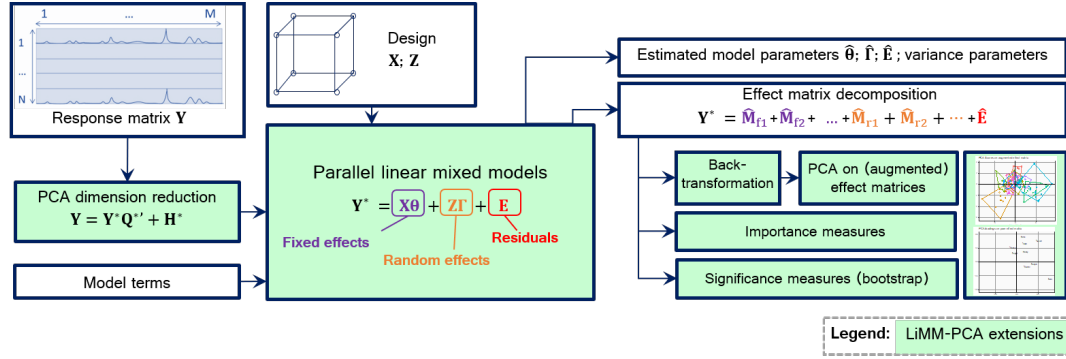


Figure IV.2 – Etapes de la méthode LiMM-PCA. Les encadrés surlignés correspondent aux extensions apportées à la méthode ASCA⁺

2.1 Étape 1 : Orthogonalisation et réduction de dimension de la matrice de réponses par ACP

Comme mentionné précédemment, une des limitations de la méthodologie ASCA⁺ est qu'elle ne prend pas en compte la corrélation entre les variables réponses. Pour pallier ce problème, une analyse en composantes principales est appliquée sur la matrice de réponses avant modélisation afin de la transformer en un nombre réduit de composantes orthogonales sans perdre d'information.

La matrice transformée $\mathbf{Y}^*$ de dimension $(n \times m^*)$ est telle que :

$$\mathbf{Y} = \mathbf{Y}^* \mathbf{Q}^{*'} + \mathbf{H}^* \quad (\text{IV.2})$$

où m^* correspond au nombre de composantes principales (PC) retenues. Ce nombre est déterminé de façon à ce que la somme des pourcentages de variance expliquée par les CP soit supérieure à un certain seuil (typiquement, $\sum_{m^*} \text{var}(CP) \geq 99\%$). La matrice des loadings, $\mathbf{Q}^*$ ($m \times m^*$), sera utilisée ultérieurement à l'étape 5 afin de retransformer les loadings et permettre ainsi leur visualisation dans un espace propre aux variables initiales.

2.2 Étape 2 : Ajustement en parallèle de LMM

Dans cette étape, un modèle linéaire mixte est ajusté à chaque colonne $\mathbf{y}_j^*$ de la matrice $\mathbf{Y}^*$ au même titre qu'un modèle GLM est ajusté à chaque variable réponse dans la méthode ASCA⁺. Les matrices $\mathbf{X}$ et $\mathbf{Z}$ sont construites selon le design expérimental sous-jacent et organisées en blocs : $\mathbf{X} = (\mathbf{X}_0 \mid \mathbf{X}_1 \mid \mathbf{X}_2 \mid \dots \mid \mathbf{X}_F)$ et $\mathbf{Z} = (\mathbf{Z}_1 \mid \mathbf{Z}_2 \mid \dots \mid \mathbf{Z}_R)$. Les codages "sum deviation" et binaires sont utilisés pour coder respectivement les F effets fixes et les R effets aléatoires.

Pour $l = 1, \dots, m^*$, nous avons

$$\mathbf{y}_l^* = \mathbf{X}\boldsymbol{\theta}_l + \mathbf{Z}\boldsymbol{\gamma}_l + \boldsymbol{\epsilon}_l \quad (\text{IV.3})$$

où les paramètres fixes $\boldsymbol{\theta}_l$ et les prédictions aléatoires $\boldsymbol{\gamma}_l$ sont définis et estimés par maximum de vraisemblance en accord avec ce qui a été présenté au chapitre III. Par ailleurs, les composantes de variance, $\hat{\sigma}_{rl}^2$ ($r = 1, \dots, R$) et $\hat{\sigma}_{el}^2$, associées aux R effets aléatoires et aux résidus sont estimés.

Les différents modèles estimés peuvent être écrits sous la forme multivariée suivante :

$$\mathbf{Y}^* = \mathbf{X}\hat{\boldsymbol{\Theta}} + \mathbf{Z}\hat{\boldsymbol{\Gamma}} + \hat{\mathbf{E}} \quad (\text{IV.4})$$

où la matrice des effets fixes estimés $\hat{\boldsymbol{\Theta}}$ est de dimension $(p \times m^*)$ et celle des prédictions aléatoires $\hat{\boldsymbol{\Gamma}}$ de dimension $(q \times m^*)$. A l'instar de $\mathbf{X}$ et $\mathbf{Z}$, les transposées de ces matrices sont organisées en blocs.

2.3 Étape 3 : Décomposition en matrices d'effets

Cette étape se fait de manière analogue à l'ASCA⁺ si ce n'est que pour la méthode LiMM-PCA, la matrice de réponses est décomposée en une somme de matrices d'effets fixes et d'effets aléatoires :

$$\begin{aligned} \mathbf{Y}^* &= \hat{\mathbf{M}}_0 + \sum_{f=1}^F \mathbf{X}_f \hat{\boldsymbol{\Theta}}_f + \sum_{r=1}^R \mathbf{Z}_r \hat{\boldsymbol{\Gamma}}_r + \hat{\mathbf{E}} \\ \mathbf{Y}^* &= \hat{\mathbf{M}}_0 + \sum_{f=1}^F \hat{\mathbf{M}}_f + \sum_{r=1}^R \hat{\mathbf{M}}_r + \hat{\mathbf{E}} \end{aligned}$$

2.4 Étape 4a : Quantification du pourcentage de variance expliqué par chaque facteur

Dans le contexte des modèles mixtes, déterminer le pourcentage de variation expliqué par chaque effet du modèle n'est pas aussi aisé qu'il ne l'est pour l'ASCA⁺. En effet, d'une part, les effets fixes et aléatoires doivent pouvoir être comparés sur une même échelle. Il n'est dès lors plus adéquat de baser la quantification des effets sur base du calcul de la norme de Frobenius. D'autre part, dans le cas de designs non balancés, il faut pouvoir prendre en compte le fait que les matrices ne sont plus orthogonales entre elles.

La méthode suggérée par Manon Martin repose sur le travail de Nakagawa et Schielzeth (2013) qui proposent une solution pour quantifier les effets fixes et aléatoires sur base du calcul de coefficients de détermination R^2 marginaux et conditionnels.

Pour chaque réponse $\mathbf{y}_l^*$, $j = 1, \dots, m^*$:

- Effets aléatoires : conditionnellement à $\mathbf{X}$, la variance est estimée en additionnant les estimations des composantes de variance, $\widehat{var}(\mathbf{y}_l^*|\mathbf{X}) = \sum_{r=1}^R \hat{\sigma}_{rl}^2 + \hat{\sigma}_{el}^2$.
- Effets fixes : la variance pour un facteur est définie comme $\hat{\sigma}_{fj}^2 = var(\mathbf{X}_f \hat{\boldsymbol{\theta}}_{fl})$, ce qui équivaut à la variance de la $l^{ème}$ colonne de la matrice d'effet correspondante, $\hat{\sigma}_{fl}^2 = var(\hat{\mathbf{M}}_{fl})$.

Dans le cas multivarié, pour toutes les réponses :

- Variance globale : étant donné que les réponses sont orthogonales, la variance totale est définie comme la somme des variances individuelles,

$$\widehat{var}(\mathbf{Y}^*) = \sum_{l=1}^{m^*} \widehat{var}(\mathbf{y}_l^*) = \sum_{l=1}^{m^*} \left(\sum_{f=1}^F \hat{\sigma}_{fl}^2 + \sum_{r=1}^R \hat{\sigma}_{rl}^2 + \hat{\sigma}_{el}^2 \right) ;$$

- Variance liée à l'effet g , $g \in \{1 \dots R, 1 \dots F\}$: $\widehat{var} g = \sum_{l=1}^{m^*} \hat{\sigma}_{gl}^2$
- Pourcentage de variance expliqué par l'effet g : $\% \widehat{var} g = \widehat{var}(g) / \widehat{var}(\mathbf{Y}^*)$.

2.5 Étape 4b : Tests de significativité des facteurs

Après avoir quantifié le pourcentage de variance expliqué par chaque effet, il est important de pouvoir déterminer lesquels sont significatifs. La procédure du test de rapport de vraisemblance (LLR) vue pour les modèles univariés dans la section III.3 peut être étendue aux cas multivariés. A noter que la méthode proposée dépend à nouveau de l'hypothèse d'indépendance entre les variables réponses et prend en compte le problème de multiplicité des tests.

Soit, pour une réponse $\mathbf{y}_l^*$, $L_{H_1,l}$ et $L_{H_0,j/g}$ les log-vraisemblances des modèles respectivement avec et sans l'effet $g \in \{1 \dots R, 1 \dots F\}$ testé. Le rapport global de vraisemblance pour cet effet s'obtient comme suit :

$$GLLR_g : \Lambda_g^{obs} = 2[\sum_{l=1}^{m^*} (\log(L_{H_1,l}) - \log(L_{H_0,l/g}))];$$

Pour les effets fixes, sous H_0 , la statistique suit une loi de chi-carré $\chi_{m^* \times k}^2$ ou k correspond à la différence entre le nombre de paramètres libres dans les modèles complet et restreint. En ce qui concerne les effets aléatoires, l'hypothèse nulle ($H_0 : \hat{\sigma}_r^2 = 0$) fait intervenir des valeurs sur la frontière de l'espace de paramètres. Sous certaines conditions, la distribution de la statistique de test est alors inconnue et peut être approximée par un mélange de χ^2 . Cependant, Martin et Govaerts (2020) propose de remplacer cette méthode et d'évaluer la significativité des effets sur base d'une procédure bootstrap.

La procédure de bootstrap utilisée est décrite dans Davison et Hinkley (1997) et Halekoh *et al.* (2014) et a été adaptée pour satisfaire au contexte multivarié.

Pour tester la significativité d'un effet g , la méthode est la suivante :

1. Définir le modèle sous H_0 , obtenu en excluant l'effet g du modèle complet H_1 .
2. Ajuster les modèles réduit H_0 et complet H_1 à la matrice des réponses $\mathbf{Y}^*$.
3. Calculer la statistique GLLR observée Λ_g^{obs} .
4. Conserver la matrice des effets fixes $\hat{\Theta}$ ainsi que les composantes de variance du modèle sous H_0 . Ceux-ci seront utilisés dans la boucle bootstrap.
5. Pour un nombre fixé B d'itérations ($b = 1, \dots, B$), répéter les étapes suivantes :
 - Générer une matrice de réponse bootstrap $\mathbf{Y}_b^*$ avec $\mathbf{Y}_b^* = \mathbf{X}\hat{\Theta} + \mathbf{Z}\mathbf{\Gamma}_b + \mathbf{E}_b$ où les matrices $\mathbf{\Gamma}_b$ et $\mathbf{E}_b$ sont générées aléatoirement à partir de distributions normales de moyenne 0 et de variances issues du modèle estimé sous H_0 . Les matrices $\mathbf{X}$ et $\mathbf{Z}$ proviennent également du modèle nul.
 - Estimer les modèles complet et restreint sur base de cette nouvelle matrice $\mathbf{Y}_b^*$.
 - Calculer la statistique GLLR Λ_g^b . A noter que cette statistique doit être proche de 0 étant donné que les réponses ont été générées sur base du modèle restreint.
6. Calculer la p-valeur pour la statistique de test : $p_g^{\text{boot}} = \frac{\sum_{b=1}^B I(\Lambda_g^b \geq \Lambda_g^{\text{obs}}) + 1}{B+1}$.

2.6 Étape 5 : ACP et représentation visuelle des matrices d'effets

L'étape 5 de la LiMM-PCA permet de visualiser pour chaque effet les résultats de la modélisation multivariée à travers une série de graphiques des scores et des loadings. Elle s'articule en deux axes et est analogue aux dernières étapes des méthodes ASCA et APCA sous réserve de quelques modifications.

Matrices d'effets purs

Tout comme pour l'ASCA, une analyse en composantes principales est appliquée sur chaque matrice d'effets purs $\hat{M}_g$:

$$\hat{M}_g = T_g P_g^{*'}$$

où T_g ($n \times r_g$) est la matrice des scores, $P_g^{*'} (r_g \times m^*)$ la matrice des loadings et $r_g = \min(n, m^*)$. Il est important de rappeler qu'avant la modélisation, la matrice réponse Y a été réduite par ACP en une matrice Y^* dont les colonnes sont orthogonales. Avant d'être mis en graphique, les loadings doivent dès lors être retransformés à l'aide de la matrice Q^* obtenue lors de la première étape :

$$P_g = P_g^* Q^{*'} .$$

La matrice P_g ainsi obtenue est de taille ($c_g \times m$) et comprend les loadings des m variables réponses dans leur espace initial.

Matrices d'effets augmentées

Dans la méthodologie APCA, les matrices de résidus sont ajoutées aux matrices d'effets avant l'ACP afin de mettre en évidence l'effet d'un facteur par rapport à la variabilité résiduelle. Cependant, dans le cadre des modèles linéaires mixtes, plusieurs sources de variation sont présentes dans le modèle. Dès lors, pour savoir quelle matrice ajouter, il faut se référer aux rapports des carrés moyens attendus utilisés pour les tests F en ANOVA.

Au deuxième chapitre, la Table II.1 pour l'ANOVA à mesures répétées nous indique que la statistique de test F pour l'effet traitement A prend au dénominateur les carrés moyens liés à l'effet sujet ($MS_{S(A)}$).

Dans la méthode LiMM-PCA, la matrice d'effet fixe A sera alors "augmentée" par l'ajout de la matrice d'effet $\hat{M}_{S(A)}$:

$$\hat{M}_A + C \times \hat{M}_{S(A)}$$

où C est un facteur de correction qui vise à prendre en compte les degrés de liberté et le quantile de la distribution associé au test F.

Dans notre exemple,

$$C = \sqrt{dl_A / (dl_{S(A)}) \times F_{dl_A, dl_{S(A)}, 1-\alpha}}$$

où $F_{dl_A, dl_{S(A)}, 1-\alpha}$ est le $(1 - \alpha)$ ème percentile de la distribution $F_{dl_A, dl_{S(A)}}$.

Il est important de préciser que les degrés de liberté associés aux effets aléatoires ne correspondent pas à ceux donnés dans la table d'ANOVA. En effet, la matrice d'un effet aléatoire r est construite sur base des prédictions estimées pour cet effet par le LMM. Or, ces estimateurs sont plus proches de 0 par rapport à ce qu'ils auraient été s'ils avaient été ajustés comme des effets fixes. Dans leur article, Martin et Govaerts (2020) détaillent une procédure qui vise à calculer des dl effectifs (ED) pour chaque CP sur base de la variance σ_r^2 estimée. Ceux-ci seront compris en 0 et le nombre de coefficients liés à l'effet aléatoire r .

Chapitre V

Analyse de données longitudinales univariées

Implémentation et application

Comme il a été vu au chapitre précédent, la deuxième étape de la LiMM-PCA vise à ajuster un modèle linéaire mixte à chacune des composantes principales retenues à l'étape 1. La LiMM-PCA a actuellement été implémentée par M. Martin sur base de la fonction `lmer` du package **lme4**. Cependant, cette fonction ne permet pas de spécifier directement une structure particulière pour la matrice de covariance des résidus et ne sera donc pas adaptée à tous les types de modèles longitudinaux présentés.

L'objectif de ce chapitre est de fournir une description des outils disponibles en R pour ajuster des LMMs à des données de type longitudinal et de les comparer. L'analyse sera faite sur le jeu de données univariées `orthodont` en prenant soin de considérer toutes les variables explicatives comme catégorielles. En effet, cette caractéristique est indispensable pour pouvoir étendre les modèles aux réponses multivariées.

1 Outils disponibles en R

1.1 Packages `nlme` et `lme4`

Plusieurs packages R contiennent des fonctions qui permettent d'ajuster et d'analyser des modèles linéaires mixtes. Cependant, deux d'entre eux sont plus couramment utilisés de part l'aptitude de leurs fonctions respectives à couvrir une large gamme de LMMs et à offrir de nombreux outils de diagnostic : **lme4** (Bates, 2010; Bates

et al., 2015) et **nlme** (Pinheiro et Bates, 2000; Gałęcki et Burzykowski, 2013; Pinheiro *et al.*, 2020).

Ces deux packages ont été écrits par Douglas M. Bates. Par conséquent, ils sont fort semblables en terme de syntaxe et incluent un grand nombre de fonctions similaires. Cependant, ils ont tous les deux leurs avantages et leurs inconvénients.

Avantages de **lme4** par rapport à **nlme**

- les méthodes d’algèbre linéaire implémentées sont plus efficaces ce qui permet de résoudre plus rapidement des problèmes sur des jeux de données conséquents. Le package **lme4** est également recommandé lorsqu’il y a plusieurs effets aléatoires.
- la syntaxe est plus simple et plus flexible ce qui permet notamment d’implémenter des modèles avec effets aléatoires croisés, plus difficilement ajustables avec **nlme**
- il offre la possibilité d’ajuster des modèles linéaires mixtes généralisés via la fonction **glmer** (eg. régression logistique, modèle log-linéaire,...)
- il corrige les degrés de liberté pour les tests F et t.

Avantages de **nlme** par rapport à **lme4**

- il implémente des fonctionnalités qui permettent de tenir compte de l’hétéroscédasticité des résidus.
- il permet de modéliser les corrélations entre les résidus ainsi que des structures de variance-covariance complexes.
- à ce jour, le package **nlme** est bien mieux documenté que le package **lme4** qui est plus récent.

1.2 Fonctions R

Dans le cadre des LMMs et plus particulièrement pour l’analyse de modèles longitudinaux, plusieurs fonctions issues de ces deux packages peuvent être utilisées. Celles-ci seront décrites très brièvement, plus d’informations sont disponibles sur le R CRAN. Pour les différents exemples, comme précédemment dans ce travail, un cas classique d’étude longitudinale sera considéré : des sujets sont assignés à un groupe et des mesures sont effectuées à plusieurs temps. Le temps, le groupe et l’interaction entre les deux sont considérés comme des effets fixes.

Fonction `lmer` du package `lme4`

Cette fonction est utilisée dans la méthode LiMM-PCA. Elle permet d'ajuster des LMMs dont la réponse est continue et de distribution gaussienne.

ex : `lmer(y ~ groupe * temps + (1|sujet), REML=T)`

Le code `(1|sujet)` indique que l'intercept (1) varie (|) pour chaque niveau d'un bloc (sujet). Cette formule vise à ajuster un LMM relativement simple (à intercept aléatoire) mais des modèles plus complexes peuvent également être estimés (Bates *et al.*, 2015). Par exemple, si les observations sont groupées par sujet qui sont eux-mêmes groupés en blocs, la formule `(1|bloc/sujet)` peut être utilisée pour indiquer les effets aléatoires.

Fonction `lme` du package `nlme`

A l'instar de `lmer`, cette fonction ajuste des LMMs. Cependant, elle offre également la possibilité de moduler la matrice des résidus R au travers de deux nouveaux arguments : `correlation` et `weight`.

ex : `lme(y ~ groupe * temps, random = ~1|sujet, method="REML", correlation=C, weight=W)`

L'argument `correlation` permet de spécifier une forme de corrélation pour les résidus. Par exemple, pour un modèle à composante symétrique, nous utiliserons `C=corCompSymm(~1|sujet)` où `|sujet` indique au modèle que les observations sont corrélées au sein d'un même individu.

L'argument `weight` permet quant à lui de tenir compte de l'hétéroscédasticité en spécifiant la forme des termes de variance. Par exemple, si on veut estimer un paramètre de variance différent pour chaque combinaison groupe-temps, `W=varIdent(~1|groupe*temps)` peut être utilisé.

C et W peuvent correspondre à différents arguments dont les principaux sont repris au Tableau V.1.

C	Corrélation	W	Variances
<code>corCompSymm</code>	composante symétrique	<code>varFixed</code>	variance fixe
<code>corSymm</code>	générale (non-structurée)	<code>varIdent</code>	différentes variances par groupe
<code>corAR1</code>	autorégressive d'ordre 1	<code>varExp</code>	la variance (eg. au fil du temps)
<code>corExp</code>	exponentielle		peut évoluer exponentiellement

Tableau V.1 – Exemples de structures de corrélations et de variances disponibles.

Fonction `gls` du package `nlme`

Cette fonction s'utilise exactement de la même manière que la fonction `lme`, excepté le fait qu'elle est destinée aux modèles avec structure de covariance sans effets aléatoires.

```
ex : lme(y ~ group * temps,
        method="REML", correlation=C, weight=W)
```

Comparaison

Les fonctions présentées ont chacune leurs avantages et inconvénients. Selon le type de données à analyser et le design d'expérience, nous verrons que l'une ou l'autre sera souvent préférée. Cependant, il est important de noter que toutes utilisent les mêmes méthodes pour estimer les paramètres. En effet, elles peuvent ajuster le modèle aussi bien par maximisation de la log-vraisemblance restreinte (`method = "REML"`) que par maximisation de la log-vraisemblance (`method = "ML"`). Pour des modèles similaires, les estimations des différents paramètres et composantes de variance obtenus via les fonctions `lmer` et `lme` devraient dès lors être sensiblement identiques.

En ce qui concerne la fonction `gls`, elle ne permet pas d'inclure d'effets aléatoires dans le modèle et ne pourra donc pas être comparée directement. Cependant, certaines structures de covariance *types* peuvent être induites par certains effets aléatoires. Des modèles marginaux avec ces structures de covariance pourront dès lors être ajustés avec la fonction `gls` et comparés à des modèles à effets aléatoires obtenus via les fonctions `lme` et `lmer`.

En effet,

1. le modèle à structure de covariance *composante symétrique* est équivalent à un modèle mixte avec intercept (sujet) aléatoire.
2. une structure de covariance *générale (non-structurée)* peut être obtenue par l'introduction d'effets aléatoires *temps* propres à chaque sujet. Notons que cette structure implique un nombre de paramètres qui augmente de manière quadratique avec le nombre de mesures effectuées par individu. Elle sera dès lors rarement adaptée car elle aboutit souvent à une sur-paramétrisation du modèle. Néanmoins, elle sera également utilisée comme base de comparaison.

2 Application aux données Orthodont

Dans cette section, les fonctions `lmer`, `lme` et `gls` vont être appliquées au jeu de données univariées *Orthodont* et comparées entre elles. Pour rappel, ces données consistent en des mesures de distances effectuées à 8, 10, 12 et 14 ans sur des filles et des garçons. Les effets fixes considérés seront les mêmes pour tous les modèles ajustés ; à savoir le temps, le sexe et l'interaction entre les deux. Précisons que l'objectif ici n'est pas d'interpréter les résultats mais de les comparer.

Rappelons également que pour un LMM, la matrice de covariance globale est égale à $\mathbf{V} = \mathbf{ZGZ}' + \mathbf{R}$ où $\mathbf{Z}$ est la matrice de design des effets aléatoires, $\mathbf{G}$ la matrice de covariance de ces effets et $\mathbf{R}$ la matrice des résidus. $\mathbf{V}$ est bloc-diagonale et nous noterons $\mathbf{V}_i$, le bloc lié au sujet i . Si aucun effet aléatoire n'est inclus dans le modèle, la matrice de covariance se résume à $\mathbf{V}_i = \mathbf{R}_i$.

2.1 lme vs lmer

Dans un premier temps, des modèles ont été ajustés à l'aide des fonctions `lme` et `lmer` en considérant l'effet sujet comme aléatoire et en laissant la matrice $\mathbf{R}_i$ diagonale (hypothèse que les résidus sont non corrélés).

```
lmer(distance ~ age * sex + (1|subject), REML=T, orthodont)
```

```
lme(distance ~ age * sex, random = ~1|subject, method = "REML", orthodont)
```

Les estimations pour les effets fixes (Tableau V.2), les prédictions des effets aléatoires, les log-vraisemblances ainsi que les matrices de covariances obtenues sont presque identiques pour les deux fonctions, exceptées de légères différences liées aux arrondis. Ces résultats nous prouvent que les méthodes d'ajustement de `lmer` et `lme` sont identiques. Cette dernière pourra dès lors également être utilisée dans la méthode LiMM-PCA.

	LMER	LME	GLS CS
Intercept	23.81 (0.38) ^{***}	23.81 (0.38) ^{***}	23.81 (0.38) ^{***}
Age8	-1.78 (0.24) ^{***}	-1.78 (0.24) ^{***}	-1.78 (0.24) ^{***}
Age10	-0.79 (0.24) ^{***}	-0.79 (0.24) ^{**}	-0.79 (0.24) ^{**}
Age12	0.60 (0.24) [*]	0.60 (0.24) [*]	0.60 (0.24) [*]
Garçon (G)	1.16 (0.38) ^{**}	1.16 (0.38) ^{**}	1.16 (0.38) ^{**}
Age8 :G	-0.31 (0.24)	-0.31 (0.24)	-0.31 (0.24)
Age10 :G	-0.37 (0.24)	-0.37 (0.24)	-0.37 (0.24)
Age12 :G	0.15 (0.24)	0.15 (0.24)	0.15 (0.24)
AIC	454.50	454.50	454.50
Log Likelihood	-217.25	-217.25	-217.25
Num. obs.	108	108	108
Num. groups : Subject	27	27	
Var : Subject (Intercept)	3.29	3.29	
Var : Residual	1.98	1.98	

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$

Tableau V.2 – Comparaison des estimations des effets fixes pour les modèles lmer et lme (effet sujet aléatoire) et le modèle GLS à composante symétrique. Les coefficients doivent être interprétés sur base du codage "sum coding".

2.2 lme/lmer vs gls

Composante symétrique

Au chapitre II, il a été vu que pour le modèle à intercepts aléatoires ajusté au point précédent, chaque bloc $\mathbf{V}_i$ de la matrice de covariance était obtenu comme suit

$$\mathbf{V}_i = \mathbf{Z}_i \mathbf{G}_i \mathbf{Z}_i' + \mathbf{R}_i = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \sigma_p^2 \begin{pmatrix} 1 & 1 & 1 & 1 \end{pmatrix} + \begin{pmatrix} \sigma^2 & 0 & 0 & 0 \\ 0 & \sigma^2 & 0 & 0 \\ 0 & 0 & \sigma^2 & 0 \\ 0 & 0 & 0 & \sigma^2 \end{pmatrix}$$

$$\mathbf{V}_i = \begin{pmatrix} \sigma^2 + \sigma_p^2 & \sigma_p^2 & \sigma_p^2 & \sigma_p^2 \\ \sigma_p^2 & \sigma^2 + \sigma_p^2 & \sigma_p^2 & \sigma_p^2 \\ \sigma_p^2 & \sigma_p^2 & \sigma^2 + \sigma_p^2 & \sigma_p^2 \\ \sigma_p^2 & \sigma_p^2 & \sigma_p^2 & \sigma^2 + \sigma_p^2 \end{pmatrix}$$

Nous pouvons constater que cette matrice a une structure à composante symétrique (CS) avec une corrélation égale à $\rho = \frac{\sigma_p^2}{\sigma^2 + \sigma_p^2}$. Nous allons la comparer avec la matrice de variance d'un autre modèle ajusté avec la fonction `gls` et qui spécifie une structure de dépendance CS entre les observations faites sur un même sujet.

```
gls(model = distance ~ age*Sex,
correlation = corCompSymm(form =~ 1|Subject),
orthodont, method="REML")
```

Sur base des résultats obtenus pour les deux modèles, les matrices $\mathbf{V}_i$ respectives ont pu être construites (voir Tableau V.3). Nous pouvons constater que celles-ci sont exactement similaires, tout comme les estimations des effets fixes et des log-vraisemblances reprises dans le Tableau V.2.

Non-structurée

Les deux modèles qui suivent ont été ajustés aux données de telle sorte qu'au sein d'un même sujet, les variances des réponses à chaque temps et les covariances pour chaque paire de temps puissent être différentes.

```
lmer(distance ~ age*sex + (age | Subject), orthodont, REML=T)

gls(model = distance ~ age*Sex, orthodont,
correlation = corSymm(form =~ as.numeric(ageC)|Subject),
weights = varIdent(form =~ 1|ageC), method='REML')
```

Les résultats des modèles sont repris et comparés dans le Tableau V.4. Une première constatation est que la vraisemblance obtenue pour ces modèles est identique mais également meilleur que celle rencontrée pour les modèles à composante symétrique (-212.56 contre -217.25). Ce résultat est logique compte tenu que plus de flexibilité est laissée au modèle.

Nous pouvons ensuite constater que les matrices de covariance sont similaires. Pour le modèle à effets aléatoires, $\mathbf{V}_i$ a pu être obtenue sur base des matrices $\mathbf{G}_i$, $\mathbf{Z}_i$ et $\mathbf{R}_i$ (Tableau V.4). La matrice $\mathbf{Z}_i$ indique par des 0 et des 1 pour chaque observation quelles composantes de variance sont prises en compte (le temps 8 est considéré comme l'intercept). Le modèle `gls` fournit quant à lui des valeurs de corrélation entre les différents temps ainsi que des facteurs d'inflation de la variance à partir desquelles la matrice $\mathbf{V}_i = \mathbf{R}_i$ peut être déduite.

Notons que pour le modèle à effet temps aléatoire, bien que des log-vraisemblances égales soient obtenues, un message d'erreur nous indique que la procédure d'estimation ne converge pas correctement via la fonction `lmer`. Les paramètres estimés et les matrices de covariance sont dès lors légèrement différentes. Cependant, ces différences sont minimales par rapport à l'ordre de grandeur des estimations.

Modèle à effet aléatoire	Modèle à covariance pattern
Fonction	
lme, lmer Effet <i>sujet</i> aléatoire	glb Structure de corrélation à composante symétrique
Log-vraisemblance	
-217.25	-217.25
Matrice de variance-covariance	
Variance : - sur les résidus $\sigma^2 = 1.975038$ - sur l'effet aléatoire sujet $\sigma_p^2 = 3.285388$ - totale $\sigma^2 + \sigma_p^2 = 5.260426$ Corrélation intra-sujet $\rho = \frac{\sigma_p^2}{\sigma^2 + \sigma_p^2} = 0.6245479$ $\mathbf{V}_i = \mathbf{Z}_i \mathbf{G} \mathbf{Z}_i' + \mathbf{R}_i$ $= \begin{pmatrix} \sigma^2 + \sigma_p^2 & \sigma_p^2 & \sigma_p^2 & \sigma_p^2 \\ \sigma_p^2 & \sigma^2 + \sigma_p^2 & \sigma_p^2 & \sigma_p^2 \\ \sigma_p^2 & \sigma_p^2 & \sigma^2 + \sigma_p^2 & \sigma_p^2 \\ \sigma_p^2 & \sigma_p^2 & \sigma_p^2 & \sigma^2 + \sigma_p^2 \end{pmatrix}$ $\sigma^2 + \sigma_p^2 = 5.260426$ $\sigma_p^2 = 3.285388$ $= \begin{pmatrix} 5.2604 & 3.2854 & 3.2854 & 3.2854 \\ 3.2854 & 5.2604 & 3.2854 & 3.2854 \\ 3.2854 & 3.2854 & 5.2604 & 3.2854 \\ 3.2854 & 3.2854 & 3.2854 & 5.2604 \end{pmatrix}$	Variance : sur les résidus $\sigma^2 = 5.260422$ Corrélation intra-sujet : $\rho = 0.6245472$ $\mathbf{V}_i = \mathbf{Z}_i \mathbf{G} \mathbf{Z}_i' + \mathbf{R}_i = \mathbf{R}_i$ $= \begin{pmatrix} \sigma^2 & \sigma^2 \rho & \sigma^2 \rho & \sigma^2 \rho \\ \sigma^2 \rho & \sigma^2 & \sigma^2 \rho & \sigma^2 \rho \\ \sigma^2 \rho & \sigma^2 \rho & \sigma^2 & \sigma^2 \rho \\ \sigma^2 \rho & \sigma^2 \rho & \sigma^2 \rho & \sigma^2 \end{pmatrix}$ $\sigma^2 = 5.260422$ $\sigma^2 \rho = 3.285382$ $= \begin{pmatrix} 5.2604 & 3.2854 & 3.2854 & 3.2854 \\ 3.2854 & 5.2604 & 3.2854 & 3.2854 \\ 3.2854 & 3.2854 & 5.2604 & 3.2854 \\ 3.2854 & 3.2854 & 3.2854 & 5.2604 \end{pmatrix}$

Tableau V.3 – Comparaison des résultats du modèle à effet sujet aléatoire et du modèle à composante symétrique.

Modèle à effets aléatoires	Modèle à covariance pattern
Fonction	
lme, lmer Effet <i>temps</i> aléatoire au sein de chaque sujet	glS Structure de corrélation générale (non structurée) Paramètres de variance différents à chaque temps
Log-vraisemblance	
-212.56	-212.56
Matrice de variance-covariance	
$V_i = Z_i G Z_i' + R_i$ $G_i = \begin{pmatrix} \text{Int} & \text{Age}_{10} & \text{Age}_{12} & \text{Age}_{14} \\ 4.7693 & -1.9806 & -0.9251 & -2.1451 \\ -1.9806 & 2.7997 & 1.0195 & 2.5281 \\ -0.9251 & 1.0195 & 2.9044 & 2.6032 \\ -2.1451 & 2.5281 & 2.6032 & 4.2912 \end{pmatrix}$ $Z_i = \begin{pmatrix} \text{Int} & \text{Age}_{10} & \text{Age}_{12} & \text{Age}_{14} \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix}$ R_i est diagonale et $\sigma^2 = 0.4820$ $V_i = \begin{pmatrix} 5.4154 & 2.7168 & 3.9102 & 2.7102 \\ 2.7168 & 4.1848 & 2.9271 & 3.3172 \\ 3.9102 & 2.9271 & 6.4557 & 4.1307 \\ 2.7102 & 3.3172 & 4.1307 & 4.9857 \end{pmatrix}$	$V_i = R_i$ $= \sigma^2 \begin{pmatrix} 1 & . & . & . \\ \rho_{12}\kappa_2 & \kappa_2^2 & . & . \\ \rho_{13}\kappa_3 & \rho_{23}\kappa_2\kappa_3 & \kappa_3^2 & . \\ \rho_{14}\kappa_4 & \rho_{24}\kappa_2\kappa_4 & \rho_{34}\kappa_3\kappa_4 & \kappa_4^2 \end{pmatrix}$ $\rho_{ij} : \begin{matrix} & 1 & 2 & 3 \\ \begin{matrix} 2 \\ 3 \\ 4 \end{matrix} & \begin{pmatrix} 0.586 & & \\ 0.661 & 0.563 & \\ 0.522 & 0.726 & 0.728 \end{pmatrix} \end{matrix}$ $\sigma^2 = 5.4154$ Facteurs d'inflation ($\kappa_1, \dots, \kappa_4$) : 1, 0.8791, 1.0918, 0.9595 $= \begin{pmatrix} 5.4154 & 2.7168 & 3.9102 & 2.7102 \\ 2.7168 & 4.1848 & 2.9271 & 3.3172 \\ 3.9102 & 2.9271 & 6.4557 & 4.1307 \\ 2.7102 & 3.3172 & 4.1307 & 4.9857 \end{pmatrix}$

Tableau V.4 – Comparaison des fonctions pour obtenir un modèle à matrice de covariance des résidus non structurée.

lme vs gls

Nous avons vu que la fonction `lme` permettait également de spécifier une structure de corrélation entre les résidus. Cependant, pour fonctionner, `lme` nécessite qu'un effet aléatoire soit inclus dans le modèle. La question suivante s'est alors posée : "Qu'advient-il si une structure à composante symétrique est indiquée en plus d'un effet aléatoire sujet" ? Les modèles suivants ont été comparés :

```
lme(distance ~ age*Sex, random = ~1 | Subject ,
     orthodont , method = "REML")
```

```
lme(distance ~ age*Sex, random = ~1 | Subject ,
     correlation = corCompSymm(form =~ 1|Subject) ,
     orthodont , method = "REML")
```

```
gls(model = distance ~ age*Sex,
     correlation = corCompSymm(form =~ 1|Subject) ,
     orthodont , method="REML")
```

Ces trois modèles ont donné des résultats similaires. En effet, lorsque la corrélation est spécifiée pour les observations faites au sein d'un sujet dont l'effet est lui-même inclus comme aléatoire dans le modèle, il y a une redondance dans les arguments et les paramètres spécifiés par `correlation` ne seront pas utilisés.

Cependant, ce n'est pas le cas pour d'autres corrélations telles que l'autorégressive d'ordre 1. En effet, il semblerait que la fonction `lme` considère aussi bien l'effet aléatoire "sujet" que la structure d'autocorrélation au niveau des résidus.

3 Conclusions

La fonction `lmer` du package **lme4** utilisée dans la procédure LiMM-PCA permet de modéliser des modèles linéaires mixtes. Cependant, cette fonction ne permet pas de spécifier une structure particulière pour la matrice des résidus. La corrélation entre les observations faites sur un même individu est dès lors modélisée exclusivement de manière implicite via l'introduction d'effets aléatoires. Cette fonction, plus récente et plus rapide, est souvent préférée à l'heure actuelle car elle se montre efficace lorsque plusieurs composantes de variances sont présentes dans le modèle et dans le cas d'effets croisés. Par ailleurs, `lmer` est couramment utilisée dans l'étude de données longitudinales pour ajuster des modèles à *intercepts et pentes aléatoires*. Cependant, comme mentionné précédemment, ces modèles nécessitent de considérer le temps comme variable continue et ne peuvent donc être étendus aux cas de réponses multivariées.

La fonction `lme` du package **nlme** quant à elle offre la possibilité de modéliser la dépendance entre les observations via l'option `correlation` (forme générale, composante symétrique, autorégressive d'ordre 1, etc.). Elle peut également permettre à la variance des résidus d'être différente pour différents groupes d'individus (hétéroscédasticité : eg. traitement vs contrôle). Elle sera préférée dans le cadre d'analyse de données répétées ou longitudinales.

Il est important de noter que les fonctions `lmer` et `lme` nécessitent la présence d'au moins un effet aléatoire dans le modèle. Cette caractéristique peut être vue comme une contrainte dans plusieurs situations :

1. Dans le cadre de la méthode LiMM-PCA, afin d'attester la significativité d'un effet au moyen d'un test LRR, il est nécessaire d'être en mesure d'obtenir une valeur de vraisemblance pour un modèle ne comprenant pas cet effet. Dans le cas où le modèle ne comprend qu'un seul effet aléatoire, les fonctions ne pourront dès lors plus être utilisées afin d'obtenir cette information.
2. Dans le cas des modèles à *covariance pattern*, la corrélation entre les observations faites au sein d'un même groupe est spécifiée explicitement. Si un seul niveau de regroupement est présent dans les données, ce qui est le cas dans notre exemple où les données sont groupées par sujet, il se peut qu'aucun effet aléatoire ne soit dès lors inclus dans le modèle. La fonction `lme` sera donc inutilisable, excepté pour un modèle à composante symétrique comme mentionné ci-dessus (i.e redondance).

La fonction `lme` pourra cependant être utilisée pour modéliser des données longitudinales multi-niveaux. Prenons l'exemple d'une étude menée dans différents centres, au sein de chaque centre, des mesures sont effectuées plusieurs fois dans le temps sur chaque patient. Un modèle peut être ajusté avec `lme` en incluant l'effet centre comme aléatoire et en utilisant une structure de corrélation pour tenir compte de la dépendance entre les observations faites sur un même patient.

La fonction `gls` du package **nlme** offre les mêmes avantages que `lme` au niveau de la modélisation de la structure de corrélation et de l'hétéroscédasticité. Elle permet en outre de pallier la première contrainte rencontrée pour les deux autres fonctions car elle tolère exclusivement des effets fixes tout en utilisant des méthodes d'estimations similaires. La fonction `glms` sera préférée pour ajuster un modèle à covariance pattern à des données longitudinales à un seul niveau de regroupement.

Notons que les mêmes analyses ont été effectuées sur un jeu de données non balancées (retrait de 2 observations). Les différentes méthodes donnent des résultats similaires et les mêmes conclusions ont pu être tirées que dans le cadre de données balancées.

Chapitre VI

Analyses de données longitudinales à réponses multivariées

Application aux données Choo

1 Pré-traitement des données

Les données Choo ont été fournies dans un format directement utilisable. Cependant, en visualisant les spectres, il a été constaté que ceux-ci n'avaient pas subi de correction au niveau de la ligne de base lors du pré-traitement. Il a dès lors été décidé de repartir des données brutes, disponibles sous forme de signaux FID (Free Induction Decay) et d'y appliquer successivement une série de pré-traitements afin d'obtenir des spectres de qualité, de variance moindre, et de faciliter leur analyse subséquente par des outils statistiques.

Le pré-traitement a été réalisé en R à l'aide du package **PepsNMR** développé à l'UCL. Les paramètres ont été choisis de manière à obtenir un jeu de données fidèle à celui qui a été fourni initialement. Les différentes étapes sont décrites très brièvement ci-dessous. Pour plus d'informations, voir la thèse de Martin (2020).

1. **Correction du délai** : correction du retard infligé par le filtre de l'appareil de mesure.
2. **Suppression du solvant** : une technique de lissage est utilisée pour enlever le signal lié au solvant (souvent de l'eau) qui n'apporte aucune information.
3. **Apodisation** : amélioration de la résolution du signal FID en mettant l'accent sur les zones spectrales présentant un rapport signal/bruit élevé.

4. **Transformée de Fourier** : la FT transforme le signal FID en un spectre complexe. L'échelle de fréquences est ensuite calibrée et convertie en déplacements chimiques (ppm).
5. **Correction de phase d'ordre 0** : une rotation dans le domaine complexe est opérée dans le but de maximiser la positivité de la partie réelle du spectre.
6. **Étalonnage** : le spectre est calibré sur base d'un composé standard dont les pics servent de référence (eg. TMS).
7. **Correction de la ligne de base** : la ligne de base est estimée et soustraite du spectre dans le but d'obtenir une ligne la plus plane possible.
8. **Suppression des valeurs négatives** : la suppression de la ligne de base peut entraîner des valeurs négatives qui sont alors fixées à 0.
9. **Warping** : déformation de l'axe des ppm dans le but d'aligner les pics des spectres sur un spectre de référence.
10. **Sélection de la fenêtre et "bucketing"** : la région informative du spectre qui sera analysée est sélectionnée. Au sein de cette région, les intensités sont ensuite agrégées afin de réduire le nombre de points. Afin d'obtenir des données semblables à celles analysées dans l'article de Choo *et al.* (2017), un bucketing à pas fixe de 0.005 ppm a été réalisé entre 8.5 et 0.5 ppm.
11. **Suppression de régions** : certaines régions non-informatives ou instables sont fixées à 0.
12. **Normalisation** : les spectres sont normalisés pour tenir compte des variations de la concentration globale des échantillons causées par les différentes dilutions. Dans ce cas-ci, la normalisation est opérée à somme constante.

2 Analyse en composantes principales

Dans un premier temps, une analyse en composantes principales a été effectuée sur le jeu de données afin de représenter les observations dans un espace à dimensions réduites. La Fig. VI.1 présente les graphiques des scores sur les quatre premières composantes principales (1-2 à gauche et 3-4 à droite). Les niveaux des facteurs *temps* et *traitement* sont mis en évidence à l'aide de couleurs et de symboles.

Le graphe des deux premières composantes montre un regroupement des observations effectuées au T1 (en rouge). Cela est conforme à ce qui était attendu étant donné que le traitement n'a pas encore été administré à ce stade de l'étude. Au T2, nous pouvons constater une séparation des individus traités à la vancomycine (triangles verts) par rapport aux autres observations le long de la deuxième composante. Les

individus du groupe contrôle quant à eux montrent plus de disparité le long de la CP1 mais aucune séparation nette n'est constatée avec le groupe initial. Enfin, sept jours après la fin du traitement (T3), les scores des individus traités affichent des valeurs proches de celles observées au T1. Cette dynamique pourrait s'expliquer par une certaine capacité de la flore intestinale à se restaurer suite à l'arrêt de la prise d'antibiotique et par conséquent un retour à l'équilibre du microbiote est observé.

Nous pouvons constater que la première composante explique 73,13% de l'information contenue dans les données. Cependant, les effets des facteurs sont difficilement observables sur cette CP qui s'avère dès lors peu informative. Il faut alors s'intéresser aux composantes suivantes. En effet, l'effet *traitement* est visible sur la CP 2 (séparation des T2-traités). Par ailleurs, une distinction entre les temps semble se dessiner le long de la CP 3.

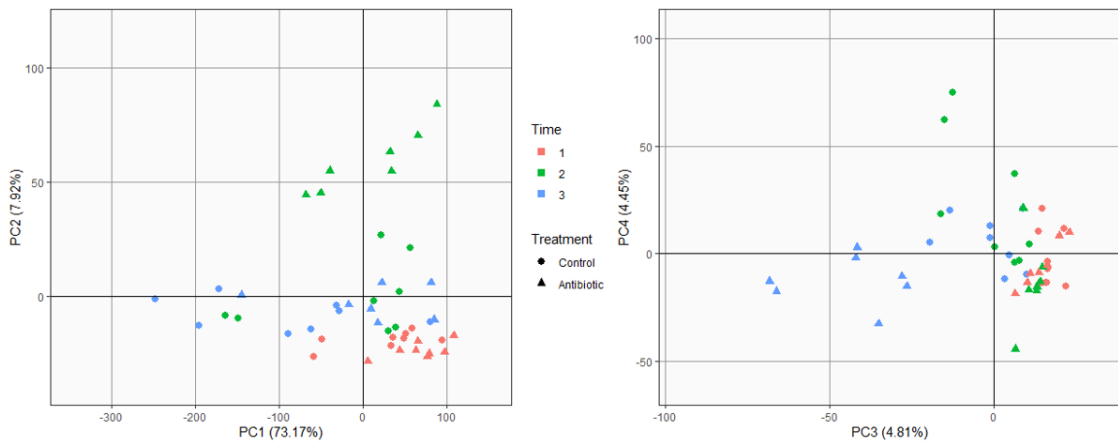


Figure VI.1 – Graphiques des scores sur les CPs 1-2 (à gauche) et 3-4 (à droite).

La Fig. VI.2 présente le graphique des loadings pour les deux premières composantes. Les pics correspondent aux variables les mieux représentées sur chacune d'elles. L'axe des abscisses est donné en ppm, à l'instar des spectres RMN, dans le but de pouvoir caractériser à quel métabolite correspondent les pics observés.

L'ACP est une méthode couramment utilisée dans le cadre des données à haute dimension. Cependant, elle ne permet pas de prendre en compte l'information liée au plan d'expérience au contraire des méthodes ASCA et APCA appliquées à la section suivante.

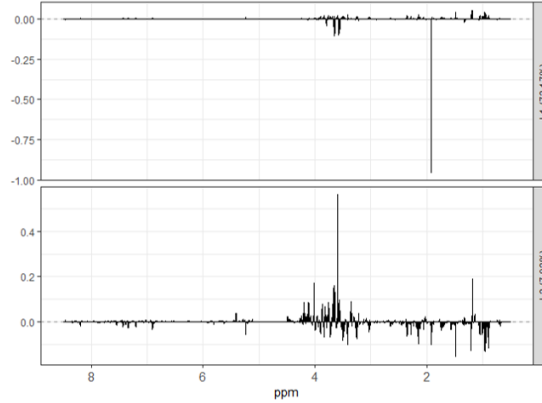


Figure VI.2 – Graphiques des loadings sur les deux premières composantes principales.

3 ASCA+ et APCA+

Dans cette section, les différentes méthodes multivariées vont être appliquées à la matrice $\mathbf{Y}$ en considérant un design expérimental à 2 facteurs croisés fixes, le temps et le groupe de traitement. Cette application a pour but également de déboguer et d'améliorer le package *LMWiRe* développé par Franceschini (2019).

Dans ce modèle, l'effet *sujet* a été omis. D'une part, parce qu'un modèle comprenant ce facteur comme fixe n'a pas pu être ajusté par manque de degrés de liberté. D'autre part, il est peu conseillé de considérer cet effet de la sorte car les résultats seront alors spécifiques aux sujets de l'étude et non généralisables à la population entière de souris.

3.1 Étape 1 : Modèle GLM

Pour chaque colonne de la matrice, le modèle ANOVA est le suivant :

$$y_{ijk} = \mu + \alpha_j + \beta_k + (\alpha\beta)_{jk} + \epsilon_{ijk} \text{ avec } i = 1, \dots, a; j = 1, \dots, b;$$

où y_{ijk} est la réponse observée pour le sujet i du traitement j et au temps k ; a est égal à 2 et b est égal à 3. μ est la moyenne générale, α_j l'effet du traitement j et β_k l'effet du temps k et $(\alpha\beta)_{jk}$ l'interaction entre le groupe i et le temps j . Les erreurs ϵ_{ijk} sont supposées $\sim iN(0, \sigma_\epsilon^2)$.

L'ANOVA 2 croisée peut être généralisée sous la forme d'un modèle GLM qui, sous sa forme multivariée, est défini comme suit :

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\Theta} + \mathbf{E} = \mathbf{X}_0\boldsymbol{\Theta}_0 + \mathbf{X}_T\boldsymbol{\Theta}_T + \mathbf{X}_G\boldsymbol{\Theta}_G + \mathbf{X}_{GT}\boldsymbol{\Theta}_{GT} + \mathbf{E}$$

où 0, T , G et GT correspondent respectivement à l'intercept, le temps, le groupe de traitement et à l'interaction. La matrice du modèle $\mathbf{X}$ est de taille 46×6 et $\boldsymbol{\Theta}$ de taille 6×1452 .

3.2 Étape 2 : Estimation des paramètres et décomposition en matrices d'effets

Les paramètres du modèle de régression sont ensuite estimés par la méthode des moindres carrés ordinaires et la matrice $\mathbf{Y}$ peut être décomposée en matrices d'effets qui contiennent les moyennes des niveaux des facteurs :

$$\mathbf{Y} = \hat{\mathbf{M}}_0 + \hat{\mathbf{M}}_T + \hat{\mathbf{M}}_G + \hat{\mathbf{M}}_{GT} + \hat{\mathbf{E}}$$

Notons qu'étant donné que les données ne sont pas balancées, les matrices ne sont pas tout à fait orthogonales.

3.3 Étape 3 : ACP sur les matrices d'effets et visualisation

Les méthodes ASCA, APCA et ASCA-E, décrites précédemment, ont été appliquées à chacune des trois matrices d'effets $\hat{\mathbf{M}}_T$, $\hat{\mathbf{M}}_G$ et $\hat{\mathbf{M}}_{GT}$ (liées respectivement au *temps*, au *traitement* et à l'*interaction*).

La Fig. VI.3 présente les graphiques des scores des deux premières composantes pour chacune des trois méthodes. En ASCA, seuls les niveaux moyens des niveaux des facteurs sont représentés. Pour le traitement, la première composante explique 100% de l'information car le facteur ne comporte que deux niveaux ce qui correspond à $2 - 1 = 1$ degré de liberté. Les variances liées à l'effet temps (trois niveaux) et à l'interaction $((3 - 1) \times (2 - 1) = 2$ ddl) sont expliquées entièrement par les deux premières CPs. L'effet d'interaction est mis en évidence par des flèches qui montrent une dynamique différente pour les deux groupes de traitement. Cependant, en ASCA, la variation naturelle liée aux résidus n'est pas observable et cette méthode ne permet pas de tirer des conclusions sur l'effet réel des facteurs par rapport à cette variabilité.

VI Analyses de données longitudinales à réponses multivariées: *Application aux données Choo*

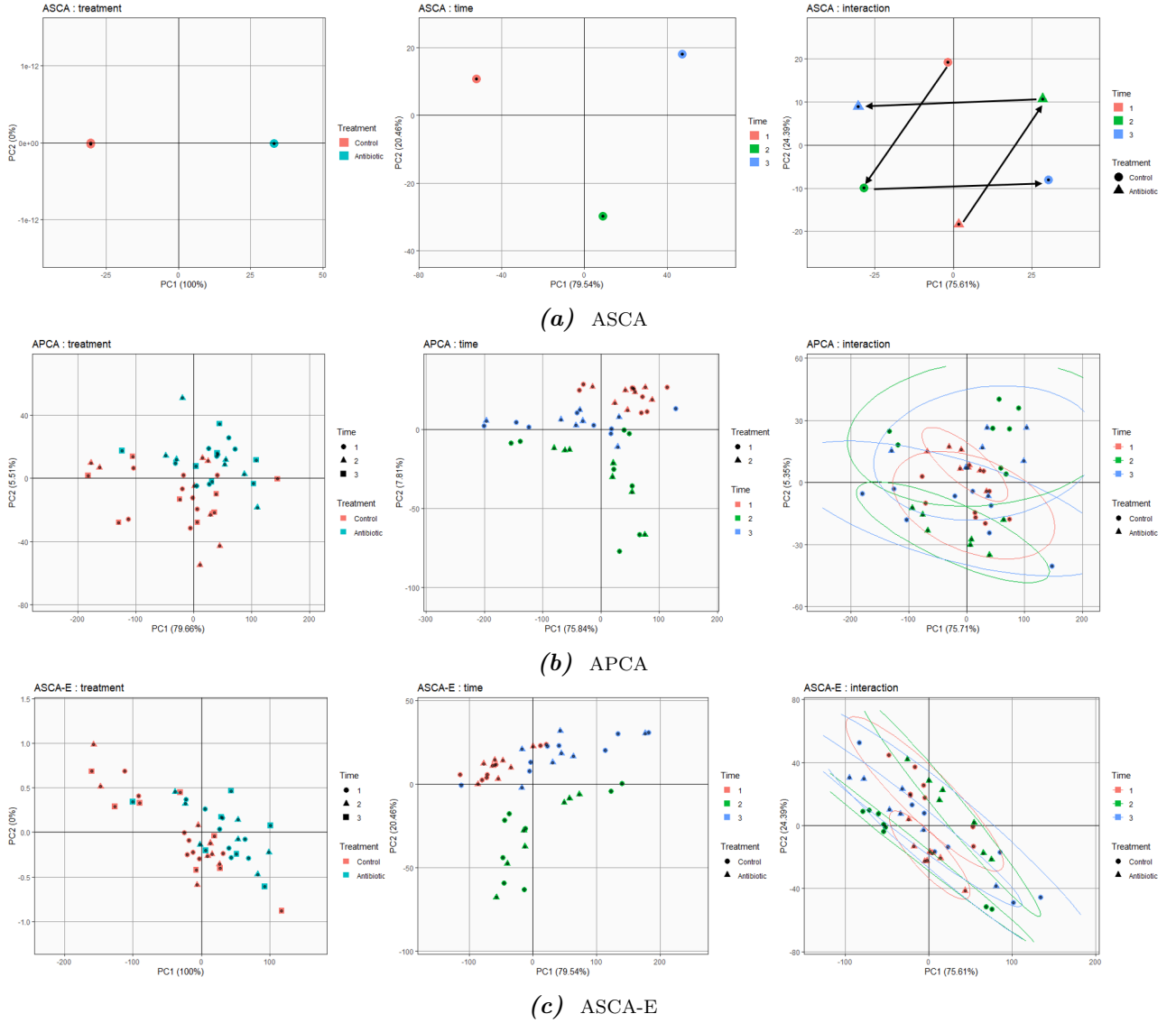


Figure VI.3 – Graphique des scores pour l'ANOVA-principal component analysis (APCA), l'ANOVA-simultaneous component analysis (ASCA) et l'ASCA-E pour les effets *temps*, *traitement* et l'interaction.

Les méthodes APCA et ASCA-E prennent en compte la matrice des résidus estimés et chaque observation est dès lors représentée sur les graphiques des scores (Fig. VI.3b et VI.3c). En APCA, nous pouvons constater que les deux premières CPs expliquent entre 80 et 85% de la variabilité pour les différents facteurs. Néanmoins, la grande disparité observée suite à l'ajout des résidus porte à penser que la variabilité résiduelle est très importante et qu'ils sont affectés par plusieurs sources de variations aléatoires. En effet, pour les deux méthodes, les niveaux des facteurs sont difficilement discernables si ce n'est que pour l'effet *temps* où trois groupes semblent se dessiner (ce qui suggère en outre un effet plus important de ce facteur). Les graphes des scores ASCA-E révèlent néanmoins qu'un effet d'interaction est présent dans le modèle. Nous verrons plus tard que l'interprétation de cette effet seul n'est pas évidente et qu'un nouvel angle d'approche devra être envisagé.

Les graphiques des loadings sont représentés à la Fig. VI.4a. Leur interprétation en lien avec les graphes des scores sera faite plus en profondeur à la section suivante dans la méthode LiMM-PCA. Néanmoins, il est important de constater que l'ajout des résidus en APCA influence considérablement les loadings qui ont alors un profil similaire pour les trois effets. Cela a pour conséquence de masquer certains descripteurs pourtant visibles sur les loadings ASCA.

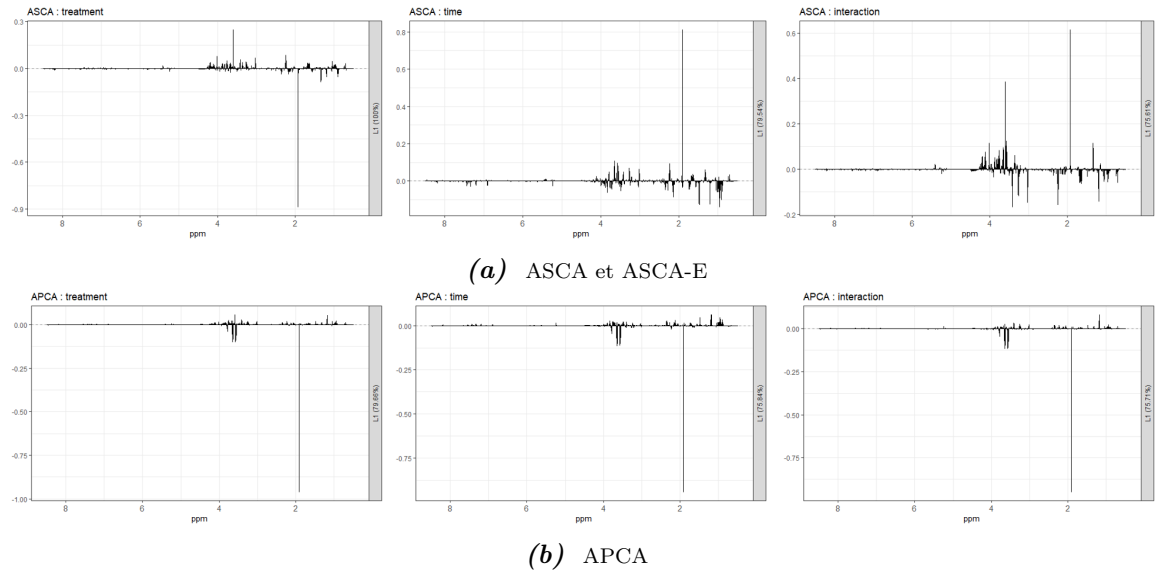


Figure VI.4 – Graphiques des loadings pour la première composante principale en ASCA/ASCA-E et APCA pour les effets *temps*, *traitement* et l'interaction.

3.4 Étape 4 : Pourcentage de variation et significativité

Les pourcentages de variation des facteurs sont repris à la Fig. VI.5. Les p-valeurs associées ont été obtenues à l'aide d'une procédure de bootstrap (1000 rééchantillonnages). Nous pouvons constater que les trois effets sont significatifs pour un niveau de confiance de 5%. Le facteur *temps* est la première source de variation (après les résidus) avec 21.37%. Les effets du traitement et de l'interaction sont moindres avec respectivement 10.02 et 7.42 %.

Comme attendu, une grande part de la variabilité est capturée par les résidus. Rappelons en effet que l'étude a été menée sur des sujets vivants. Or, la première source de variabilité est la variabilité biologique, inhérente à chaque individu. En effet, face à un même traitement, la réponse de sujets différents est très variable. Évaluer l'effet réel de ce traitement peut donc s'avérer difficile dès lors que les individus n'y répondent pas de la même manière.

Afin de prendre en compte la variabilité inter-individus, il est nécessaire d'ajouter un effet aléatoire dans notre modèle. Cependant, les méthodes ASCA et APCA ne tolèrent que des effets fixes. La procédure LiMM-PCA permet quant à elle de prendre en compte des sources de variations aléatoires et sera appliquée au jeu de données *Choo* à la section suivante.

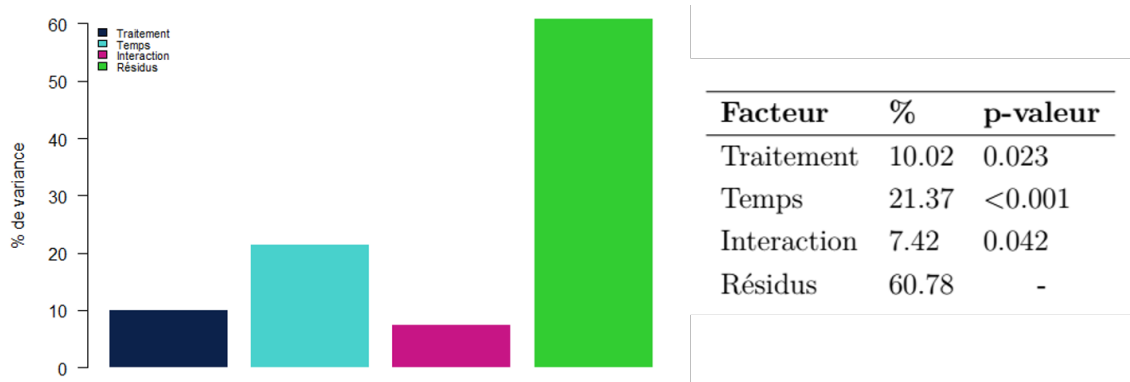


Figure VI.5 – Pourcentage de variance et significativité de chaque facteur.

4 LiMM-PCA

4.1 Étape 1 : Orthogonalisation et réduction de dimension de la matrice réponse par ACP

La première étape de la LiMM-PCA vise à appliquer une PCA sur la matrice réponse $\mathbf{Y}$ afin d'obtenir une matrice avec un nombre réduit de colonnes orthogonales. Les graphiques des scores et des loadings pour les deux premières CPs ont déjà été présentés à la section 2. L'ACP montre que 14 CPs des 46 disponibles cumulent plus de 99% de la variance totale présente dans les profils spectraux. Elles seront donc retenues pour la suite des analyses.

4.2 Étapes 2 et 3 : Ajustement en parallèle de LMM et décomposition en matrices d'effets

Dans le cas univarié, le modèle statistique choisi pour analyser la réponse correspond à celui qui a été présenté à la section II.3.1, autrement dit, une ANOVA à mesures répétées :

$$y_{ijk} = \mu + \rho_{i(j)} + \alpha_j + \beta_k + (\alpha\beta)_{jk} + \epsilon_{ijk} \text{ avec } i = 1, \dots, s; j = 1, \dots, a; k = 1, \dots, b$$

où y_{ijk} est la réponse observée pour le $i^{\text{ème}}$ sujet du traitement j et au temps k ; $s = 8, a = 2$ et $b = 3$. μ est la moyenne générale, α_j l'effet (fixe) du traitement j , β_k l'effet (fixe) du temps k et $(\alpha\beta)_{jk}$ l'interaction (fixe) entre le groupe j et le temps k et $\rho_{i(j)}$ est l'effet aléatoire lié au sujet i emboité dans le groupe j ($\sim iN(0, \sigma_p^2)$). Les erreurs ϵ_{ijk} sont supposées $\sim iN(0, \sigma_e^2)$.

Ce modèle peut être généralisé sous la forme d'un LMM qui, sous sa forme multivariée, est défini comme suit :

$$\mathbf{Y}^* = \mathbf{X}_0 \boldsymbol{\Theta}_0 + \mathbf{X}_T \boldsymbol{\Theta}_T + \mathbf{X}_G \boldsymbol{\Theta}_G + \mathbf{X}_{GT} \boldsymbol{\Theta}_{GT} + \mathbf{Z}_{S[G]} \boldsymbol{\Gamma}_{S[G]} + \mathbf{E}$$

où $0, T, G, GT$ et S correspondent respectivement à l'intercept, le temps, le groupe de traitement, l'interaction et au sujet.

Des LMMs sont ajustés en parallèle pour chacune des 14 CPs retenues à l'étape précédente afin d'obtenir des estimations REML pour les différents paramètres. La matrice de réponse $\mathbf{Y}^*$ peut ensuite être décomposée en matrices d'effets fixes et aléatoires de taille (46×14) :

$$\mathbf{Y}^* = \hat{\mathbf{M}}_0 + \hat{\mathbf{M}}_T + \hat{\mathbf{M}}_G + \hat{\mathbf{M}}_{GT} + \hat{\mathbf{M}}_{S[G]} + \hat{\mathbf{E}}$$

Notons que pour effectuer cette étape, la fonction `lmer` du package **lme4** a été utilisée. Par ailleurs, la normalité des résidus de chaque modèle ajusté a été vérifiée à l'aide de diagramme QQ-plot, disponibles en Annexe.

4.3 Étape 4 : Quantification de l'importance de chaque effet et significativité

Avant d'interpréter visuellement l'information contenue dans les matrices d'effets, il convient de mesurer l'importance relative de chaque terme du modèle et d'en attester la significativité.

La figure VI.6a montre les pourcentages de variation expliqués globalement par chaque effet. Tout comme pour l'ASCA+, l'effet *temps* est la source de variation principale du modèle avec 20,21%. Les parts de variabilité expliquées par le *traitement*, l'effet aléatoire *sujet* et l'*interaction* sont, quant à elles, moindres mais non-négligeables. Par ailleurs, bien qu'elle soit inférieure à ce qui a été observé avec l'ASCA (55% contre 60%), la variabilité résiduelle (intra-sujets) est toujours très importante ce qui indique qu'une large proportion de l'information présente dans les données n'est pas expliquée par notre modèle.

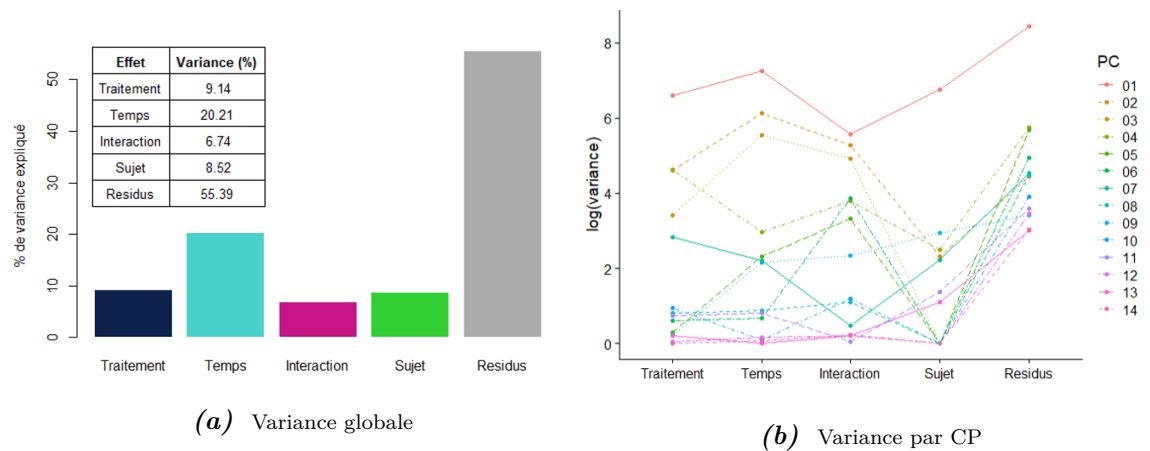


Figure VI.6 – Variance expliquée par chaque composante du modèle. La variance est exprimée en échelle logarithmique.

La figure VI.6b offre davantage d'information et permet de visualiser pour chaque effet, sur une échelle logarithmique, la part de variance expliquée par chacune des 14 CPs. Nous pouvons constater que les contributions des différents effets à leurs

variabilités globales respectives sont réparties entre plusieurs CPs. Les trois effets fixes sont principalement liés aux trois premières composantes bien que la 4^{ème} CP capture également une part importante de l'effet *traitement*. En ce qui concerne l'effet *sujet*, celui-ci est principalement lié à la CP 1 et dans une moindre mesure à la CP 9. Enfin, la variabilité résiduelle est retenue par toutes les CPs et majoritairement par la première.

En ce qui concerne la significativité des effets, la figure VI.7a représente les statistiques des tests de rapports de vraisemblance ((R)LLR) obtenues sur chacune des 14 CPs retenues. Ce graphique nous permet de visualiser quelles combinaisons effets-CPs contribuent davantage à la vraisemblance globale du modèle multivarié.

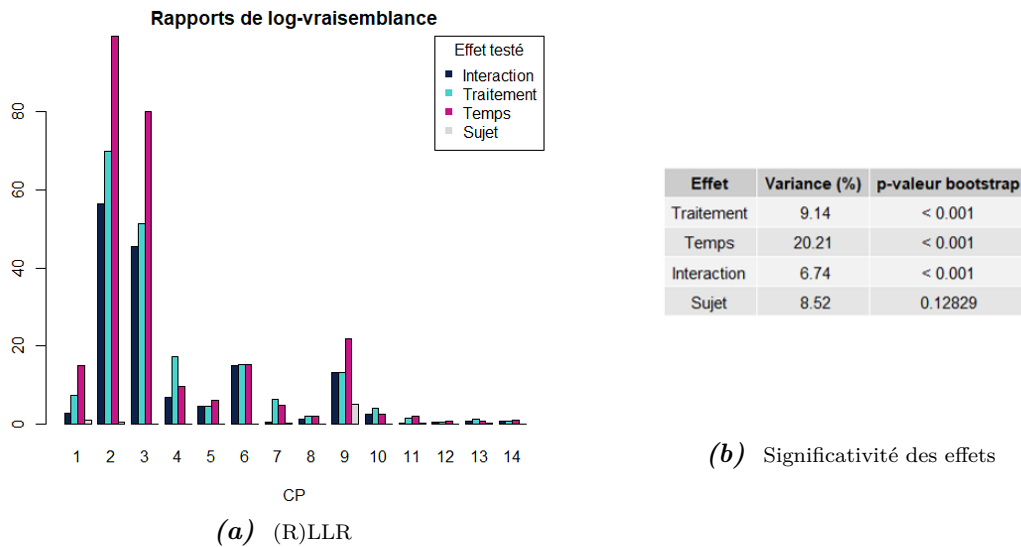


Figure VI.7 – Rapports de vraisemblance pour chaque effet sur chacune des CPs et p-valeurs obtenues sur base de la procédure bootstrap (1000 rééchantillonnages). Les distributions sous l'hypothèse nulle des LLRs pour chaque effet sont disponibles en Annexe.

Ce graphique peut être mis en parallèle avec ce qui a été observé à la figure précédente. En effet, nous observons que l'effet *temps* conduit à une augmentation importante de la vraisemblance sur les CPs 2 et 3. Les effets *traitement* et *interaction* contribuent également à leurs GLLRs respectives sur ces mêmes composantes mais à un niveau légèrement moins important. Les p-valeurs obtenues sur base de la procédure bootstrap, présentée à la section IV.2.5, nous confirment que ces trois effets fixes sont statistiquement significatifs pour le modèle multivarié (fig. VI.7b). Cependant, ce n'est pas le cas de l'effet aléatoire *sujet* qui montre une p-valeur supérieure à 0.05. En effet, l'inclusion de cet effet contribue peu à augmenter le GLLR. Par ailleurs, il est important de mentionner que pour 7 CPs sur 14, la composante de variance estimée est nulle et les modèles restreints associés donnent dès lors des log-vraisemblances identiques à celles des modèles complets.

Nous pouvons observer que les (R)LLRs de la CP1 sont peu élevés par rapport à ceux des CP 2 et CP 3. Or, celle-ci capture une part importante de la variabilité engendrée par les différents effets du modèle. Cependant, les effets des facteurs vont être masqués par la variance résiduelle qui est, elle aussi, principalement prise en compte par cette première composante. En outre, cela expliquerait pourquoi l'effet *sujet* est non significatif malgré les 8,52% de variance expliqués.

Un point qu'il semblait important de mentionner est que, pour cette étape, de nombreuses modifications ont dû être apportées au code développé et fourni par Martin (2020). En effet, afin de calculer les LLRs, il est nécessaire d'obtenir la vraisemblance du modèle sans l'effet à tester. Cependant, contrairement aux modèles étudiés jusqu'à présent avec la méthode LiMM-PCA, notre modèle ne comporte qu'un seul et unique effet aléatoire. Par ailleurs, comme mentionné au chapitre précédent, il n'est pas possible de modéliser un LMM avec la fonction `lmer` (ni `lme`) sans effet aléatoire.

Afin de contourner cette difficulté, plusieurs alternatives ont pu être trouvées :

1. la fonction `ranova` du package **lmerTest** : cette fonction, appliquée au résultat d'un LMM ajusté avec `lmer()`, fournit une table d'ANOVA pour les effets aléatoires du modèle avec les log-vraisemblances et les statistiques de tests RLLR associées.
2. la fonction `gls` : les résultats du LLM peuvent être directement comparés avec ceux d'un modèle `gls` comportant les mêmes effets fixes et utilisant la méthode REML.
3. le package **RLLRsim** : ce package implémente et fournit des tests (R)LLR exacts sur base de méthodes de simulation afin d'évaluer la significativité de composantes de variances dans des LMMs (Scheipl *et al.*, 2008). Les fonctions utiles `exactRLRT` et `exactLRT` peuvent traiter aussi bien des modèles ajustés avec `lmer` qu'avec `lme`.

Les alternatives 1 et 2 ont été utilisées pour obtenir les LLRs illustrés à la figure VI.7a et ont donné des résultats identiques. La significativité de l'effet *sujet* a, quant à elle, été attestée à l'aide du RLLRsim qui s'est révélé être une alternative équivalente et beaucoup plus rapide que les procédures de bootstrap qui nécessitent une forte intensité de calcul. Cette fonction pourrait dès lors être utilisée dans la LiMM-PCA pour tester n'importe quelle composante de variance.

4.4 Étape 5 : Visualisation des matrices d'effets par ACP

La dernière étape de la LiMM-PCA consiste à appliquer une ACP sur chacune des matrices d'effets afin de projeter les scores/observations sur un espace à dimensions réduites. Par ailleurs, les loadings sont retransformés pour visualiser le poids des variables initiales.

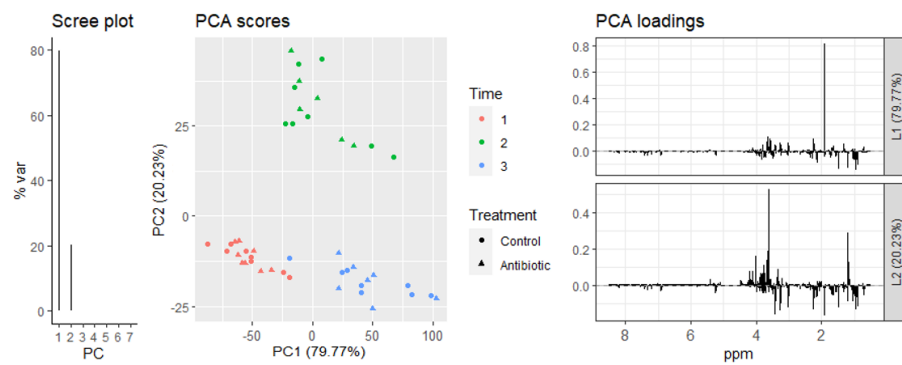
Sur base de ce qui a été présenté à section 2.6 du chapitre IV, le Tableau VI.1 montre quelle matrice doit être ajoutée à chaque matrice d'effet afin de visualiser la source de variation aléatoire adéquate. La matrice des résidus corrigée est ajoutée aux matrices des effets *temps*, *interaction* et *sujet* alors que la matrice liée au *traitement* est augmentée de la matrice *sujet*.

En ce qui concerne les facteurs de correction, conformément à ce qui a été vu à la section IV.2.6, les degrés de liberté associés à l'effet aléatoire *sujet* doivent être ajustés pour chacune des CPs sur base de la variance σ_p^2 estimée. Sans entrer dans les détails de la procédure, une des étapes nécessite d'inverser la matrice de covariance des effets aléatoires $\mathbf{G} = \sigma_p^2 I_{16}$. Cependant, comme mentionné précédemment, pour certaines composantes (7 CPs sur 14), la variance estimée est nulle. Les matrices $\mathbf{G}$ associées ne peuvent dès lors pas être inversées et, par conséquent, les dl effectifs ne peuvent être obtenus. Ce cas de figure devra faire l'objet de recherches futures afin d'être solutionné. Dans le cadre de ce mémoire, nous nous contenterons d'appliquer les dl donnés dans la table d'ANOVA à mesures répétées.

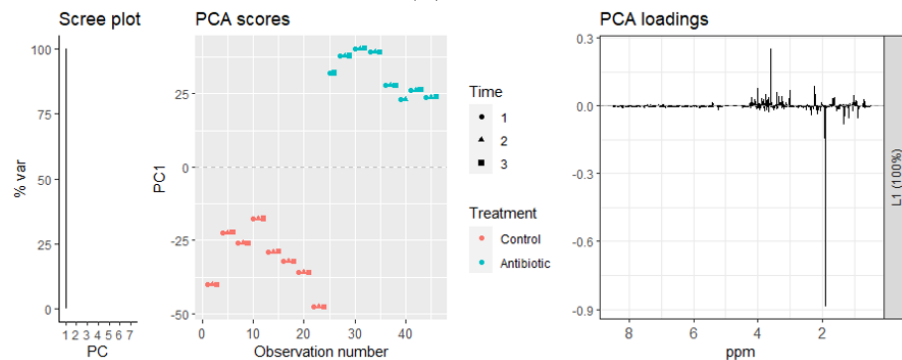
Tableau VI.1 – Matrices d'effets augmentées. a , b et s représentent respectivement le nombre de niveaux des effets traitement(2), temps (3) et le nombre de sujet par groupe (8). F_{dl_1, dl_2} représente le 95^{ème} quantile de la distribution F avec dl_1, dl_2 degrés de liberté.

Effet	$E(MS)$	ddl associés	Matrice ajoutée	Facteur de correction
Traitement	$E(MS)_G$	$dl_G = a - 1$	$\hat{M}_{S(G)}$	$\sqrt{F_{dl_G, dl_{S(G)}} \times \frac{df_G}{df_{S(G)}}}$
Temps	$E(MS)_T$	$dl_T = b - 1$	$\hat{E}$	$\sqrt{F_{dl_T, dl_E} \times \frac{df_T}{df_E}}$
Interaction	$E(MS)_{GT}$	$dl_{GT} = (a - 1)(b - 1)$	$\hat{E}$	$\sqrt{F_{dl_{GT}, dl_E} \times \frac{dl_{GT}}{dl_E}}$
Sujet	$E(MS)_{S(G)}$	$dl_{S(G)} = a(s - 1)$	$\hat{E}$	$\sqrt{F_{dl_{S(G)}, dl_E} \times \frac{dl_{S(G)}}{dl_E}}$
Résidus	$E(MS)_E$	$dl_E = a(s - 1)(b - 1)$	-	-

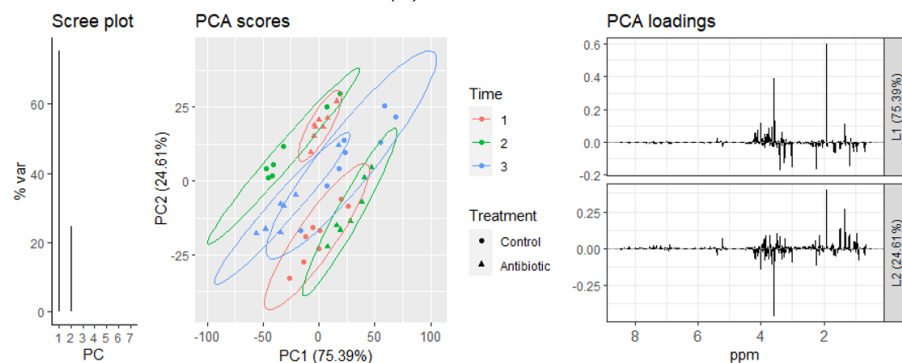
Les scores et des loadings propres aux différentes composantes de variance du modèle sont représentés à la figure VI.12. Le graphe des scores à la figure VI.8a montre une nette distinction entre les observations faites aux trois temps et notamment du T2 par rapport aux autres le long de la deuxième CP. Cela est en accord avec le fait que cet effet explique une plus grande part de variabilité dans le modèle et qu'il est significatif. Pour l'effet *traitement*, comme il a été vu pour l'APCA, 100 % de la variance est reprise dans la première composante et les observations faites dans chaque groupe sont donc bien séparées. Par ailleurs, dû au fait que la matrice d'effet *sujet* ait été ajoutée comme source de variation aléatoire, les scores sont regroupés par trois ce qui correspond aux observations faites sur un même individu aux différents temps. Il semblerait également qu'il y ait davantage de variation entre les individus du groupe contrôle. En ce qui concerne l'interaction, des ellipses ont été ajoutées afin de regrouper les observations d'une même combinaison temps-traitement. Nous pouvons constater que des clusters, certains plus homogènes que d'autres, tendent à se former mais que ceux-ci ne sont pas facilement discernables de part le fait que la variance expliquée par l'*interaction* est faible par rapport à la variabilité résiduelle. Néanmoins, on peut clairement apercevoir un effet opposé au temps 2, caractérisé par des pics à 3.6 ppm sur les graphes des loadings. Les scores pour l'effet *sujet* (fig. VI.8d) indiquent quant à eux que la première CP capture la majorité de la variance liée à cet effet. Cependant, les triangles reliant les observations faites sur un même individu se superposent et confirme la non-significativité de l'effet aléatoire du modèle. Enfin, nous pouvons remarquer que les résidus sont distribués aléatoirement autour de (0,0) ce qui justifie l'hypothèse de normalité, vérifiée pour chaque CP à l'étape 2 à l'aide de QQ-plot. Néanmoins, ils sont fortement dispersés ce qui témoigne de la présence d'autres sources de variations non expliquées par le modèle.



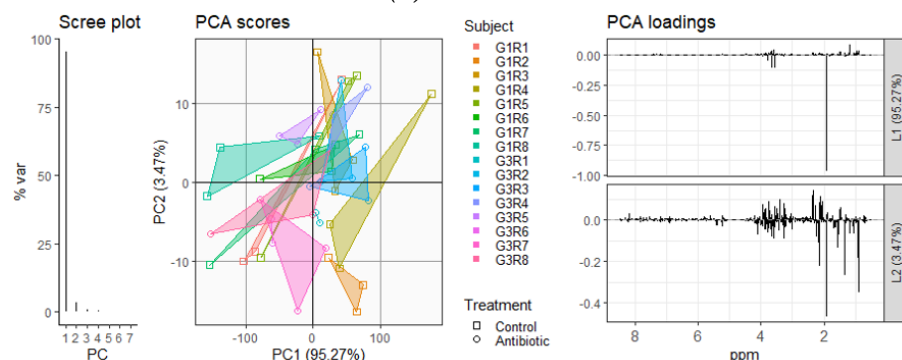
(a) Temps



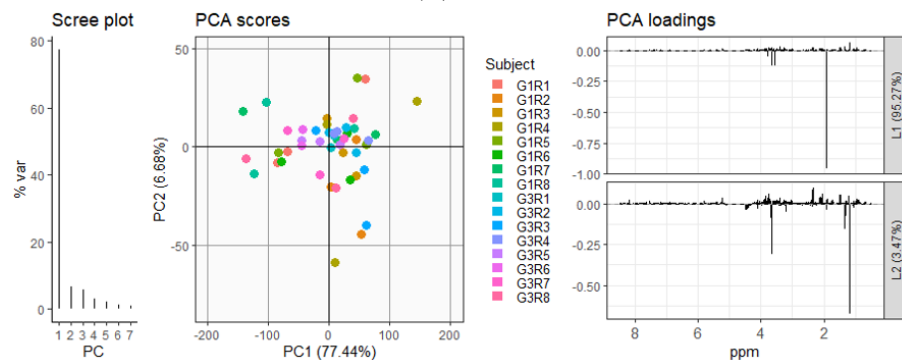
(b) Traitement



(c) Interaction



(d) Sujet



(e) Résidus

Figure VI.8 – ACP sur les matrices d'effets. Les scores sont augmentés tandis que les screeplot et loadings sont obtenues sur base des matrices pures.

L'ACP réalisée sur les matrices d'effets purs est peu informative car seuls les niveaux des facteurs sont représentés sur les graphes des scores. Cependant, dans le cas de données longitudinales, il peut être intéressant d'envisager les résultats sous un nouvel angle d'approche et de visualiser les effets fixes dans leur globalité sans que ceux-ci ne soient entachés d'autres sources de variations aléatoires. Pour cela, les matrices d'effets *temps*, *traitement* et *interaction* ont été additionnées entre elles et une ACP a été appliquée. Le graphe des scores (fig. VI.9) montre comment les niveaux moyens des combinaisons temps-traitement se situent les uns par rapport aux autres. Nous observons une dynamique semblable à celle observée à la section 2. Au T1, les points sont proches. Au T2, le groupe antibiotique se distingue des autres le long de la 2^{ème} composante. Enfin, au T3, nous observons un retour vers la situation de départ pour le groupe traité, sans doute liée à une restauration de la flore suite à l'arrêt du traitement. Cependant, la première CP sépare le T3-contrôle du reste du groupe. Nous pouvons observer sur le graphe des loadings que cette séparation est principalement liée à la présence d'un pic important juste en deçà de 2 ppm. Ce pic est également constaté sur les spectres initiaux ainsi que dans la majorité des graphes de loadings associés à la première constante. La variable correspondante a été isolée et est représentée à la figure VI.10a. Le graphique montre que la réponse moyenne augmente considérablement au fil du temps pour les individus du groupe contrôle. Cette constatation est surprenante et il serait intéressant de s'y attarder pour en connaître la cause. Est-ce une conséquence du développement naturel de la microflore (non soumise à la pression de l'antibiotique) ou bien une dérive liée à la prise d'échantillonnage ? De manière générale, nous constatons également une grande variabilité au niveau de cette réponse d'où sa présence récurrente sur les loadings (et notamment ceux pour l'effet *sujet* et les résidus).

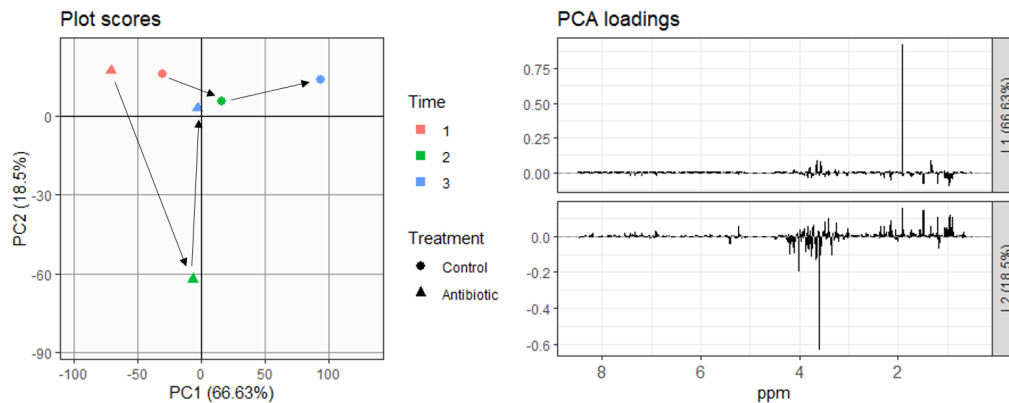


Figure VI.9 – ACP sur les matrices d'effets additionnées entre elles.

A l’opposé, les effets du *traitement* et de l’*interaction* sont clairement visibles sur la deuxième composante (figure VI.9) et un pic important apparait aux alentours de 3.6 ppm. La variable correspondante est représentée à la fig. VI.10b¹ Nous pouvons observer que la prise d’antibiotique induit une augmentation importante du métabolite au T2 et qu’un retour à la normale est constaté au T3. Le métabolite lié au descripteur à 3.6 ppm peut dès lors être considéré comme un biomarqueur dans le cadre de cette étude.

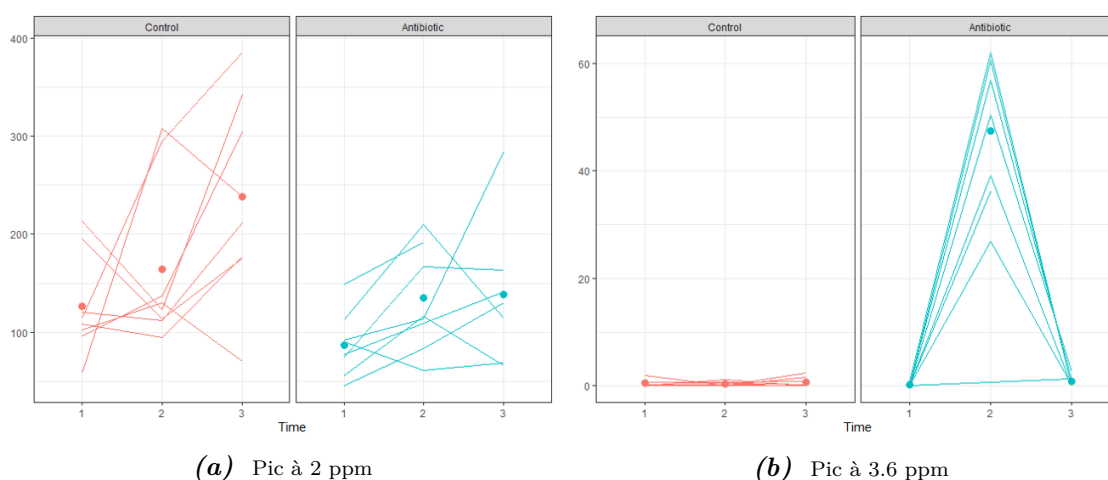


Figure VI.10 – Évolution de la réponse pour les variables liées aux pic observés à 2 ppm et 3.6 ppm en fonction du groupe de traitement. Les moyennes à chaque temps sont représentées par des points.

L’analyse LiMM-PCA a été réitérée sur une matrice renormalisée après retrait de la variable correspondante au pic à 2 ppm. Le Tableau VI.2 reprend et compare les pourcentages de variances expliqués par chaque effet. Comme attendu, la part de variabilité aléatoire (liée à l’effet *sujet* et aux résidus) ainsi que celle liée au *traitement* ont diminué au profit des effets *temps* et *interaction*. Par ailleurs, les scores de l’ACP sur la matrice d’interaction augmentée montrent une séparation nette des différents clusters par rapport au modèle précédent (non montré).

Tableau VI.2 – Comparaison des pourcentages de variances expliqués par chaque effet avec et sans le pic.

% variance	Traitement	Temps	Interaction	Sujet	Résidus
Avec pic	9.14	20.21	6.74	8.52	55.39
Sans pic	5.97	28.21	14.01	3.27	48.54

1. Notons que pour un des individus traités, la ligne est horizontale car la valeur au T2 est manquante.

5 Implémentation et analyse d'un modèle multi-varié à structure de covariance CS

La dernière partie de ce mémoire vise à étendre la méthode LiMM-PCA à l'analyse de modèles à structures de covariance. Le code fourni par M. Martin a dès lors été adapté en remplaçant notamment la fonction `lmer` par la fonction `glms` et en apportant les modifications subséquentes nécessaires. Il a ensuite été choisi d'appliquer cette nouvelle méthodologie sur les données *Choo* en considérant un modèle à composante symétrique (modèle CS) afin de pouvoir comparer les résultats à ceux obtenus pour le modèle mixte à effet aléatoire *sujet* (modèle EA).

La première étape consiste à ajuster des modèles à CS en parallèle sur chacune des CPs sur base de la fonction `glms` suivante :

```
glms(model = CP ~ time*treatment ,
correlation = corCompSymm(form =~ 1|Subject) , method='REML')
```

Le modèle multivarié est ensuite décomposé en matrices d'effets :

$$\mathbf{Y}^* = \hat{\mathbf{M}}_0 + \hat{\mathbf{M}}_T + \hat{\mathbf{M}}_G + \hat{\mathbf{M}}_{GT} + \hat{\mathbf{E}}$$

Une première constatation est que la matrice des résidus estimés est différente. En effet, dans le modèle EA, une matrice d'effet est obtenue pour l'effet *sujet* sur base des prédictions des coefficients aléatoires. La matrice des résidus devient alors

$$\hat{\mathbf{E}} = \mathbf{Y}^* - \hat{\mathbf{M}}_0 - \hat{\mathbf{M}}_T - \hat{\mathbf{M}}_G - \hat{\mathbf{M}}_{GT} - \hat{\mathbf{M}}_{S[G]}$$

et sera donc différente de celle estimée pour le modèle CS. Pour une observation, cette différence sera égale à la prédiction faite pour le sujet correspond dans le modèle EA. En effet, il a été vérifié que $\hat{\mathbf{E}}_{CS} = \hat{\mathbf{E}}_{EA} + \hat{\mathbf{M}}_{S[G]}$.

En ce qui concerne les matrices d'effets fixes, pour des raisons qui seront révélées plus loin, celles-ci sont similaires mais pas exactement identiques. Comparons tout d'abord les matrices de covariance $\mathbf{V}_i$ obtenues pour les modèles EA et CS. Pour rappel, celles-ci sont définies respectivement comme suit :

$$\begin{pmatrix} \sigma^2 + \sigma_p^2 & \sigma_p^2 & \sigma_p^2 \\ \sigma_p^2 & \sigma^2 + \sigma_p^2 & \sigma_p^2 \\ \sigma_p^2 & \sigma_p^2 & \sigma^2 + \sigma_p^2 \end{pmatrix} \text{ et } \begin{pmatrix} \sigma^2 & \sigma^2 \rho & \sigma^2 \rho \\ \sigma^2 \rho & \sigma^2 & \sigma^2 \rho \\ \sigma^2 \rho & \sigma^2 \rho & \sigma^2 \end{pmatrix}$$

Pour chacune des composantes, nous allons comparer la corrélation intra-sujet, c'est à dire $\frac{\sigma_p^2}{\sigma^2 + \sigma_p^2}$ pour le modèle EA et ρ pour le modèle CS. Les résultats sont représentés à la figure VI.11. Nous pouvons constater que les corrélations ne sont pas équivalentes pour toutes les CPs. En effet, une variance est toujours positive. Dès lors, la corrélation intra-sujet dans le modèle EA sera comprise exclusivement entre 0 et 1 alors que le modèle CS permet à ρ de prendre des valeurs négatives. Le fait que des corrélations négatives soient obtenues sur les CPs pour lesquelles la variance liée à l'effet aléatoire était nulle est à l'origine des différences, bien que minimes, observées au niveau des matrices d'effets fixes.

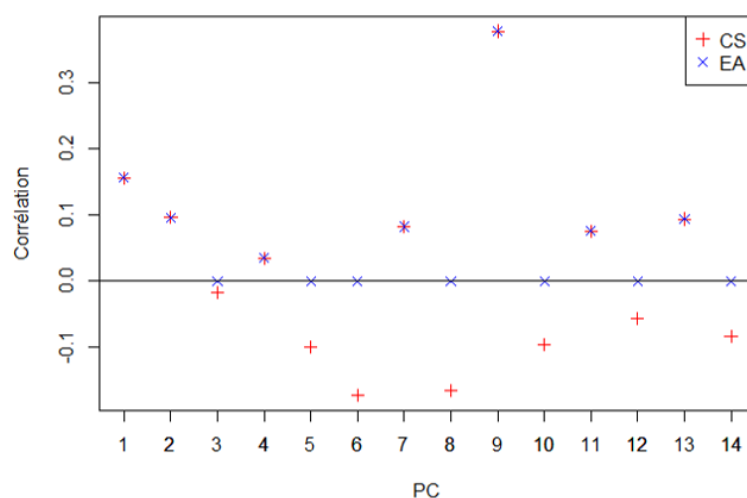


Figure VI.11 – Comparaison des corrélations intra-sujets des modèles EA et CS obtenues pour chaque CP.

Malgré les différences observées, les pourcentages de variances expliqués par les effets fixes des modèles sont similaires (Tableau VI.3). Par conséquence, la part de variabilité résiduelle dans le modèle CS est égale à la somme des variations liées aux sujets et aux résidus dans le modèle EA.

Tableau VI.3 – Comparaison des pourcentages de variance expliqués par chaque effet pour les deux modèles.

% variance	Traitement	Temps	Interaction	Sujet	Résidus
Modèle EA	9.14	20.21	6.74	8.52	55.39
Modèle CS	9.14	20.21	6.74	-	63.91

Enfin, les graphes des scores augmentés (fig.VI.12a) offre un schéma visuel global comparable au modèle EA pour les effets *temps* et *interaction* bien que la dispersion des observations engendrée par l'ajout des résidus soit légèrement différente. En ce qui concerne le traitement, la dissemblance constatée est liée au fait que, dans le modèle EA, c'est la matrice d'effet *sujet* qui avait été ajoutée et non celle des résidus. La figure VI.12b représente les scores augmentés mais sans facteur de correction afin de pouvoir faire la comparaison avec l'ASCA-E à la section 3². Ceux-ci montrent des schémas comparables et portent à croire que considérer une structure CS par rapport à une structure identité n'améliore pas considérablement le modèle à effets fixes croisés (ce qui, par ailleurs, serait en adéquation avec la non-significativité de l'effet *sujet* dans le modèle EA).

De manière générale, il semblerait qu'une structure à CS ne soit tout simplement pas la plus adaptée pour prendre en compte la dépendance qui existe entre les mesures faites dans le temps sur une même souris. Cette structure est, en effet, souvent peu réaliste car elle considère une corrélation constante entre les différents temps. Dans nos données, nous avons notamment pu constater que les observations au T1 étaient davantage liées à celles du T3 que celles du T2 pour les individus traités.

Il a été vu précédemment qu'un grand nombre d'autres structures de dépendance existaient. Se pose alors le problème du choix de celle qui ajuste au mieux les données observées. Dans cette optique, le test GLLR, implémenté sur base d'une procédure bootstrap en LiMM-PCA, pourrait être utilisé pour comparer deux modèles avec des structures différentes. Cependant, ce test nécessite que les modèles en question soient "emboîtés". En univarié, une méthode usuelle et plus générale consiste à choisir la structure qui minimise un critère d'ajustement tel que l'AIC ou le BIC. Ces critères ne sont pas applicables directement dans le cas de réponses à haute dimension mais plusieurs méthodes dérivées ont été proposées. Certaines ont fait leurs preuves, notamment dans la sélection d'une structure de dépendance appropriée, et pourraient être incluses dans la méthodologie LiMM-PCA (Sofer *et al.*, 2014).

2. Il est à préciser qu'augmenter les matrices en LiMM-PCA donne la possibilité de mieux visualiser les niveaux des effets par rapport à la variabilité aléatoire mais que les graphes ne peuvent pas servir à conclure de la significativité d'un effet

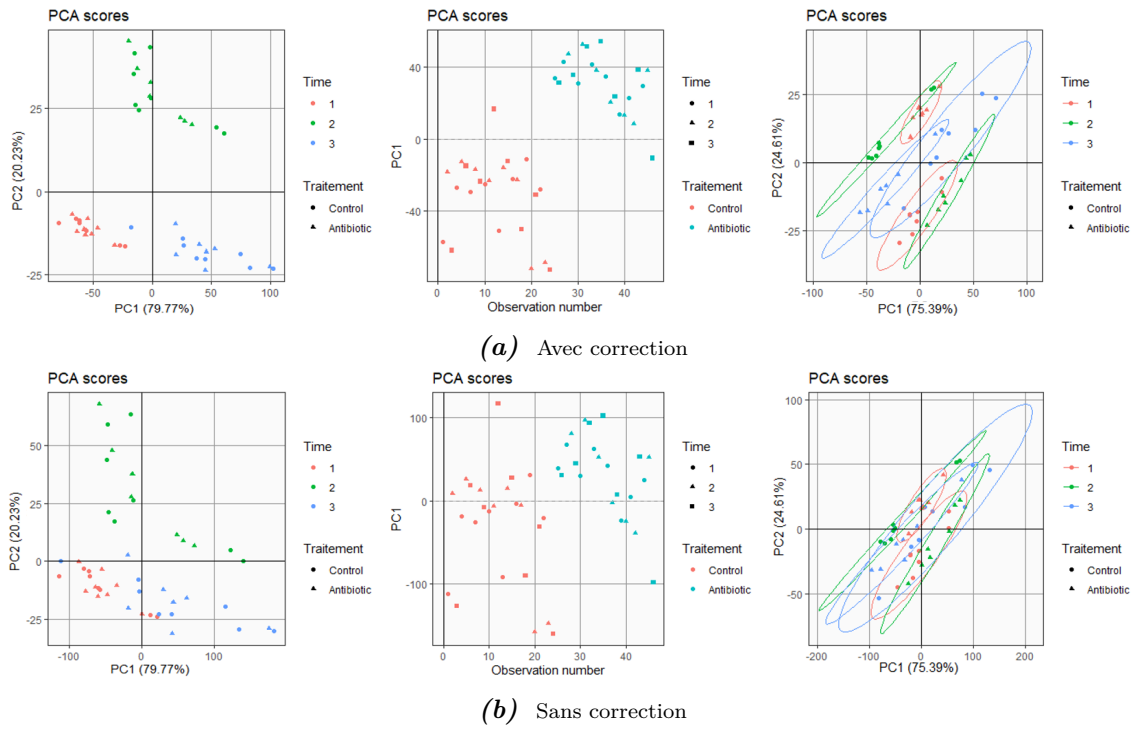


Figure VI.12 – ACP sur les matrices d'effets du modèle CS augmentés des résidus, avec ou sans facteur de correction.

Chapitre VII

Conclusions et perspectives

Dans le domaine des sciences -omiques, les études longitudinales suscitent un vif intérêt car elles offrent un aperçu dynamique et progressif de la réponse globale d'un système vivant à des stimuli. Ces expériences sont généralement conçues pour comparer les changements observés entre différents groupes et identifier des biomarqueurs capables de définir un état pathologique ou la réponse à un traitement. Les études -omiques génèrent des bases de données caractérisées par un nombre de variables corrélées souvent supérieur au nombre d'observations et dont l'analyse nécessite dès lors des outils statistiques spécifiques. Dans ce contexte, ces dernières années ont vu le développement de méthodes qui permettent l'analyse de données multivariées tout en tenant compte du design expérimental sous-jacent. Parmi ces méthodes, l'ASCA+ et APCA+ ont prouvé qu'elles pouvaient distinguer et analyser de manière efficace les sources de variabilité induites par des facteurs fixes. Cependant, la dimension temporelle des études longitudinales induit une structure dans les données car les observations faites sur un même individu ne peuvent pas être considérées comme indépendantes. La méthode LiMM-PCA, en se basant sur les LMMs, offre un cadre plus général et permet d'étendre l'utilisation de l'ASCA+/APCA+ à l'analyse de designs plus sophistiqués (eg. mesures répétées, hiérarchiques, multiniveaux). En incluant des effets aléatoires, cette classe de modèles permet en effet de prendre en compte ces structures complexes de manière appropriée.

La base de données utilisée dans le cadre de ce mémoire, appelée *Choo*, provient d'une étude qui vise à analyser l'impact de la prise d'un antibiotique sur le microbiote intestinal de souris. Les sujets ont été répartis en deux groupes (contrôle et traitement) et des échantillons de matières fécales ont été analysés par RMN à trois reprises : avant, pendant et après l'antibiothérapie. Les LMMs peuvent être utilisés afin de tenir compte de la nature groupée de ces données longitudinales soit 1) via l'introduction d'un effet aléatoire *sujet*, soit 2) en imposant une structure de corrélation entre les

termes d'erreurs relatifs à un sujet. Sur cette base, les objectifs de ce travail consistaient, d'une part, à appliquer et adapter la méthode LiMM-PCA au jeu de données *Choo* en considérant le premier modèle. D'autre part, à tenter d'étendre l'utilisation de la méthode à l'analyse du second type de modèle dit à structure de covariance.

Afin de répondre au mieux à ces objectifs, il était primordial dans un premier temps d'étudier en profondeur les concepts théoriques sous-jacents aux différentes méthodes statistiques utiles dans l'analyse de données longitudinales à une réponse (e.g. ANOVA à mesures répétées, LMMs). La compréhension de ces méthodes en univarié s'avérait en effet indispensable pour pouvoir étendre leur utilisation aux cas de réponses multiples.

Dans un deuxième temps, il a jugé important d'identifier les différents outils disponibles en R pour ajuster les modèles longitudinaux afin d'avoir une vue globale des perspectives d'extension de la méthodologie LiMM-PCA en pratique. L'étape d'ajustement du LLM multivarié dans cette procédure repose actuellement sur l'utilisation de la fonction `lmer` issue du package **lme4**. Cette fonction, récente et rapide, est souvent préférée lorsque plusieurs composantes de variances sont présentes dans le modèle considéré ainsi que dans le cas d'effets croisés. Elle pourra être utilisée pour ajuster le modèle à effet *sujet* aléatoire. Cependant, `lmer` ne permet pas d'imposer une structure de corrélation particulière entre les observations faites sur un même individu. Au contraire, la fonction `lme` du package **nlme** offre la possibilité de modéliser directement la structure de covariance au niveau des résidus et pourra donc être utilisée pour ajuster le deuxième type de modèle. Néanmoins, `lme` nécessite l'inclusion d'au minimum un effet aléatoire et dès lors qu'un seul niveau de regroupement est présent dans les données, une troisième fonction, `gls`, sera préférée. Cette dernière est, en effet, destinée aux modèles à structure de covariance sans effets aléatoires. Ces trois fonctions, qui utilisent les mêmes méthodes d'estimations (ML et REML), ont été comparées sur base d'un jeu de données univarié et il a été prouvé que pour des modèles comparables, elles donnaient des résultats identiques. En effet, un modèle à structure de covariance CS est équivalent à un modèle à effet *sujet* aléatoire en termes d'estimation des effets fixes, de vraisemblance et de matrice de covariance.

La deuxième partie de ce mémoire consiste en l'application des méthodes d'analyses multivariées. La méthodologie ASCA/APCA+ a tout d'abord été appliquée au jeu de données *Choo* en considérant un design expérimental à deux facteurs fixes croisés, le temps et le traitement. Les analyses ont été réalisées grâce aux fonctions fournies par le package en développement **LMWiRe**. Celles-ci ont montré que le facteur temps était la première source de variation du modèle suivi du traitement et de l'interaction et que ces trois effets étaient significatifs. Cependant, les résidus représentent 60.78% de la variabilité du modèle ce qui a pour conséquence que les niveaux des facteurs sont difficilement observables sur les graphes des scores en APCA. De fait, l'effet *sujet* ayant été omis du modèle, la variabilité biologique propre à chaque individu se retrouve dans la variabilité résiduelle. Par ailleurs, ce modèle est peu adapté aux données longitudinales car il ne prend pas en compte le fait que les mesures effectuées sur une même souris ne sont pas indépendantes. Notons que l'application de **LMWiRe** sur le jeu de données *Choo* a permis de mettre en lumière plusieurs coquilles (bugs) et d'apporter des améliorations notamment au niveau des tests de bootstrap et des graphes des scores en ASCA et ASCA-E, ceux-ci ne s'affichant pas correctement pour des effets à 2 niveaux.

Par la suite, la procédure LiMM-PCA a été appliquée aux données *Choo* en incluant l'effet *sujet* comme aléatoire. Tout comme pour l'ASCA/APCA, les effets *traitement*, *temps* et *l'interaction* se sont révélés significatifs. Par ailleurs, le fait de prendre en compte l'effet *sujet* permet de mieux différencier les niveaux des facteurs sur les graphes des scores. Cependant, la variabilité résiduelle reste élevée (55,39%) ce qui a tendance à masquer, sur certaines composantes, l'effet *sujet* qui n'est, quant à lui, pas significatif. L'ACP réalisée sur les matrices d'effets fixes combinées a permis de mettre en évidence l'effet d'interaction en montrant une dynamique temporelle différente pour les deux groupes. En effet, durant l'antibiothérapie, les individus traités montrent un profil métabolomique différent et un biomarqueur a notamment pu être mis en évidence (pic à 3.6 ppm). Un retour à la normale est ensuite constaté quelques jours après l'arrêt de la prise d'antibiotique, sans doute lié à une restauration subséquente de la microflore intestinale. Nous avons pu également constater la présence récurrente d'un pic important à 2 ppm sur les loadings correspondants aux CPs 1 des différents effets. La variable correspondante a été analysée visuellement et nous avons pu observer une augmentation étonnante de la réponse au fil du temps pour le groupe contrôle ainsi qu'une grande variabilité. Il serait intéressant d'identifier si ces constatations sont liées à un processus biologique ou à une erreur analytique.

Il est important de notifier que nous avons rencontré un certain nombre de limites techniques dans l'application de la méthodologie LiMM-PCA développée par Martin et Govaerts (2020) et certaines alternatives/adaptations ont dû être trouvées/apportées. La principale difficulté est liée au fait que notre modèle ne contient qu'un effet aléatoire. En effet, la méthode repose sur le calcul d'un rapport de vraisemblance global pour tester la significativité d'un effet. Il est donc nécessaire d'obtenir une valeur de vraisemblance pour un modèle multivarié ne comprenant pas cet effet. Or, la fonction `lmer` nécessite la présence d'au minimum une composante de variance et ne peut donc pas être utilisée pour calculer le GLLR lié à l'unique effet aléatoire *sujet*. En outre, les fonctions destinées à ajuster des modèles linéaires simples (e.g. `lm`, `glm`) ne constituent pas une solution car les méthodes d'estimations des paramètres ne sont pas comparables. Parmi les alternatives trouvées, l'utilisation des fonctions `ranova` et `gls` se sont montrées efficaces et comparables. Par ailleurs, le package `RLRsim` s'est révélé être une alternative rapide et équivalente aux procédures de bootstrap qui sont particulièrement exigeantes en termes de calcul et peuvent prendre jusqu'à plusieurs dizaines de minutes. La deuxième difficulté rencontrée implique le calcul des *dl* effectifs lors de l'étape de visualisation des effets dans un espace à dimension réduite. En cause, des variances nulles estimées pour l'effet *sujet* sur certaines CPs qui empêchent l'inversion des matrices de covariance $\mathbf{G}$ correspondantes et, par conséquent, l'obtention des *dl* ajustés. Une solution temporaire a pu être apportée mais ce cas de figure devra faire l'objet de recherches approfondies.

La dernière étape de ce mémoire visait à étudier la possibilité d'étendre la méthodologie LiMM-PCA à l'analyse de modèles à structures de covariance. Pour ce faire, la fonction `lmer` a été remplacée par la fonction `gls` et des adaptations subséquentes ont été apportées au code de base. L'analyse a ensuite porté sur un modèle à composante symétrique dans le but de pouvoir comparer les résultats avec ceux du modèle à effet *sujet* aléatoire. La différence majeure constatée se trouve au niveau des matrices des résidus estimés. En effet, dans le modèle EA, une matrice est obtenue pour l'effet *sujet* sur base des prédictions. L'inclusion d'un effet aléatoire apporte donc une information supplémentaire car il permet de visualiser les effets sujets individuellement et de faire de l'inférence. Ajuster un modèle CS semble donc avoir moins d'intérêt, d'autant plus que cette structure est peu réaliste car elle considère une corrélation constante entre les temps. Cependant, d'autres patterns peuvent être envisagées (eg. AR(1), Toeplitz) et c'est là que cette méthodologie trouve son intérêt dans l'analyse de données longitudinales. Néanmoins, comme le véritable modèle est rarement connu a priori, il sera nécessaire de mettre en place une procédure afin de sélectionner la structure de covariance la plus adéquate (e.g. sur base d'un critère d'information glo-

bal). Il serait intéressant alors de trouver une méthode permettant de visualiser et de quantifier l'information contenue dans la structure sélectionnée. Cette méthodologie présente encore de nombreuses lacunes et un travail supplémentaire sera nécessaire afin de déterminer son potentiel réel et sa valeur ajoutée par rapport aux autres méthodes.

En conclusion, ce mémoire avait pour objectif de mettre en évidence les possibilités d'utilisation et les avantages de la méthode LiMM-PCA dans l'analyse de données -omiques, générées sur base d'un design d'expérience longitudinal. L'application de la LiMM-PCA à un nouveau jeu de données a permis d'exposer les limites techniques de la méthodologie mais également d'y apporter des améliorations. Ces phases de débogage sont exigeantes mais indispensables si l'on veut généraliser l'utilisation de LiMM-PCA à tout type de design expérimental complexe et, à terme, implémenter cette méthode dans un package R. Pour finir, les aspects théoriques de la méthode pourraient être étendus afin d'analyser des plans d'expériences incluant également des variables continues. Une telle adaptation aurait un intérêt certain en études longitudinales car cela permettrait d'ajuster des modèles à *intercepts et pentes aléatoires*.

Bibliographie

- ANDERSON, M. et BRAAK, C. T. (2003). Permutation tests for multi-factorial analysis of variance. *Journal of statistical computation and simulation*, 73(2):85–113.
- BATES, D., MÄCHLER, M., BOLKER, B. et WALKER, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.
- BATES, D. M. (2010). lme4 : Mixed-effects modeling with r.
- BEYOĞLU, D. et IDLE, J. R. (2013). Metabolomics and its potential in drug development. *Biochemical pharmacology*, 85(1):12–20.
- BOCCARD, J. et RUDAZ, S. (2016). Exploring omics data from designed experiments using analysis of variance multiblock orthogonal partial least squares. *Analytica Chimica Acta*, 920:18–28.
- BONVALLOT, N. (2014). *Application de la métabolomique à l'étude du lien entre les expositions environnementales aux pesticides pendant la grossesse et le développement de l'enfant : approches épidémiologique et toxicologique*. Thèse de doctorat.
- BOUVERESSE, D. J.-R., PINTO, R. C., SCHMIDTKE, L., LOCQUET, N. et RUTLEDGE, D. N. (2011). Identification of significant factors by an extension of anova-pca based on multi-block analysis. *Chemometrics and Intelligent Laboratory Systems*, 106(2):173–182.
- BROWN, H. et PRESCOTT, R. (2014). *Applied mixed models in medicine*. John Wiley & Sons.
- CANLET, C., SHINTU, L. et ELENA-HERRMANN, B. (2012). La métabolomique par rmn et ses applications en biologie végétale, toxicologie et médecine. *l'actualité chimique*, (364-365).
- CHOO, J. M., KANNO, T., ZAIN, N. M. M., LEONG, L. E., ABELL, G. C., KEEBLE, J. E., BRUCE, K. D., MASON, A. J. et ROGERS, G. B. (2017). Divergent rela-

- tionships between fecal microbiota and metabolome following distinct antibiotic-induced disruptions. *mSphere*, 2(1):e00005–17.
- DAVISON, A. C. et HINKLEY, D. V. (1997). *Bootstrap methods and their application*, volume 1. Cambridge university press.
- de NOORD, O. E. et THEOBALD, E. H. (2005). Multilevel component analysis and multilevel pls of chemical process data. *Journal of Chemometrics : A Journal of the Chemometrics Society*, 19(5-7):301–307.
- DIGGLE, P., DIGGLE, P. J., HEAGERTY, P., LIANG, K.-Y., HEAGERTY, P. J., ZEGER, S. *et al.* (2002). *Analysis of longitudinal data*. Oxford University Press.
- EL GHAZIRI, A., QANNARI, E. M., MOYON, T. et ALEXANDRE-GOUABAU, M.-C. (2015). Aov-pls : a new method for the analysis of multivariate data depending on several factors. *Electronic Journal of Applied Statistical Analysis*, 8(2):214–235.
- FITZMAURICE, G., DAVIDIAN, M., VERBEKE, G. et MOLENBERGHS, G. (2008). *Longitudinal data analysis*. CRC press.
- FRANCESCHINI, S. (2019). *LMWiRe : R Routines for Linear Model With Wide Response*.
- GAŁECKI, A. et BURZYKOWSKI, T. (2013). Linear mixed-effects model. In *Linear Mixed-Effects Models Using R*, pages 245–273. Springer.
- GOVAERTS, B., MARTIN, M. et FRANCO, B. (2019). The family of linear regression, anova, general linear and linear mixed models for the chemist : a survival guide. Forthcoming discussion paper, Institute of Statistics, Biostatistics and Actuarial Sciences, UCLouvain, Belgium.
- GUISSET, S., MARTIN, M. et GOVAERTS, B. (2019). Comparison of parafasca, acof, and amopls approaches in the multivariate glm modelling of multifactorial designs. *Chemometrics and Intelligent Laboratory Systems*, 184:44–63.
- HALEKOH, U., HØJSGAARD, S. *et al.* (2014). A kenward-roger approximation and parametric bootstrap methods for tests in linear mixed models—the r package pbkrtest. *Journal of Statistical Software*, 59(9):1–30.
- HARRINGTON, P. d. B., VIEIRA, N. E., ESPINOZA, J., NIEN, J. K., ROMERO, R. et YERGEY, A. L. (2005). Analysis of variance–principal component analysis : A soft tool for proteomic discovery. *Analytica chimica acta*, 544(1-2):118–127.

- HEDEKER, D. et GIBBONS, R. D. (2006). *Longitudinal data analysis*, volume 451. John Wiley & Sons.
- JANSEN, J. J., BRO, R., HOEFSLOOT, H. C., van den BERG, F. W., WESTERHUIS, J. A. et SMILDE, A. K. (2008). Parafasca : Asca combined with parafac for the analysis of metabolic fingerprinting data. *Journal of Chemometrics : A Journal of the Chemometrics Society*, 22(2):114–121.
- JANSEN, J. J., HOEFSLOOT, H. C., van der GREEF, J., TIMMERMAN, M. E. et SMILDE, A. K. (2005a). Multilevel component analysis of time-resolved metabolic fingerprinting data. *Analytica chimica acta*, 530(2):173–183.
- JANSEN, J. J., HOEFSLOOT, H. C., van der GREEF, J., TIMMERMAN, M. E., WESTERHUIS, J. A. et SMILDE, A. K. (2005b). Asca : analysis of multivariate data obtained from an experimental design. *Journal of Chemometrics : A Journal of the Chemometrics Society*, 19(9):469–481.
- KARLOVICH, C., DUCHATEAU-NGUYEN, G., JOHNSON, A., McLOUGHLIN, P., NAVARRO, M., FLEURBAEY, C., STEINER, L., TESSIER, M., NGUYEN, T., WILHELMSEILER, M. *et al.* (2009). A longitudinal study of gene expression in healthy individuals. *BMC medical genomics*, 2(1):33.
- KEUN, H. C., EBBELS, T. M., BOLLARD, M. E., BECKONERT, O., ANTTI, H., HOLMES, E., LINDON, J. C. et NICHOLSON, J. K. (2004). Geometric trajectory analysis of metabolic responses to toxicity can define treatment specific profiles. *Chemical Research in Toxicology*, 17(5):579–587.
- KUTNER, M. H., NACHTSHEIM, C. J., NETER, J., LI, W. *et al.* (2005). *Applied linear statistical models*, volume 5. McGraw-Hill Irwin New York.
- LI, H., WOOD, C. L., GETCHELL, T. V., GETCHELL, M. L. et STROMBERG, A. J. (2004). Analysis of oligonucleotide array experiments with repeated measures using mixed models. *BMC bioinformatics*, 5(1):209.
- LIQUET, B., LÊ CAO, K.-A., HOCINI, H. et THIÉBAUT, R. (2012). A novel approach for biomarker selection and the integration of repeated measures experiments from two assays. *BMC bioinformatics*, 13(1):325.
- MARINI, F., de BEER, D., JOUBERT, E. et WALCZAK, B. (2015). Analysis of variance of designed chromatographic data sets : The analysis of variance-target projection approach. *Journal of Chromatography A*, 1405:94–102.

- MARTIN, M. (2020). *Uncovering informative content in metabolomics data : from pre-processing of 1H NMR spectra to biomarkers discovery in multifactorial designs*. Thèse de doctorat, UCL-Université Catholique de Louvain.
- MARTIN, M. et GOVAERTS, B. (2020). Limm-pca : combining asca+ and linear mixed models to analyse high dimensional designed data. Rapport technique.
- NAKAGAWA, S. et SCHIELZETH, H. (2013). A general and simple method for obtaining r^2 from generalized linear mixed-effects models. *Methods in ecology and evolution*, 4(2):133–142.
- PINHEIRO, J., BATES, D., DEBROY, S., SARKAR, D. et R CORE TEAM (2020). *nlme : Linear and Nonlinear Mixed Effects Models*. R package version 3.1-148.
- PINHEIRO, J. C. et BATES, D. M. (2000). Linear mixed-effects models : basic concepts and examples. *Mixed-effects models in S and S-Plus*, pages 3–56.
- POTTHOFF, R. F. et ROY, S. (1964). A generalized multivariate analysis of variance model useful especially for growth curve problems. *Biometrika*, 51(3-4):313–326.
- PUCHADES-CARRASCO, L. et PINEDA-LUCENA, A. (2015). Metabolomics in pharmaceutical research and development. *Current opinion in biotechnology*, 35:73–77.
- ROUSSEAU, R. (2011). Statistical contribution to the analysis of metabonomics data in 1h nmr spectroscopy. *Université Catholique de Liège*.
- SCHEIPL, F., GREVEN, S. et KÜCHENHOFF, H. (2008). Rlrsim : Testing for random effects or nonparametric regression functions in additive mixed models. *Thomas Achia, Atinuke Adebajji, John Owino, Anne Wangombe : Spatial Durbin Model for Poverty Mapping and Analysis 8 Daniel Adler, Jens Oehlschlägel, Oleg Nenadic, Walter Zucchini : Large atomic data in R : package 'ff' 9 Claudio Agostinelli : Robust Inference in Generalized Linear Models 10*, page 164.
- SMILDE, A. K., JANSEN, J. J., HOEFSLOOT, H. C., LAMERS, R.-J. A., VAN DER GREEF, J. et TIMMERMAN, M. E. (2005). Anova-simultaneous component analysis (asca) : a new tool for analyzing designed metabolomics data. *Bioinformatics*, 21(13):3043–3048.
- SOFER, T., DICKER, L. et LIN, X. (2014). Variable selection for high dimensional multivariate outcomes. *Statistica Sinica*, 24(4):1633.
- STRAUBE, J., GORSE, A.-D., HUANG, B. E., LÊ CAO, K.-A. *et al.* (2015). A linear mixed model spline framework for analysing time course ‘omics’ data. *PloS one*, 10(8):e0134540.

- THIEL, M., FÉRAUD, B. et GOVAERTS, B. (2017). Asca+ and apca+ : Extensions of asca and apca in the analysis of unbalanced multifactorial designs. *Journal of Chemometrics*, 31(6):e2895.
- TIMMERMAN, M. E. (2006). Multilevel component analysis. *British Journal of Mathematical and Statistical Psychology*, 59(2):301–320.
- VERBEKE, G. (1997). Linear mixed models for longitudinal data. *In Linear mixed models in practice*, pages 63–153. Springer.
- WOLFINGER, R. D. (1997). An example of using mixed models and proc mixed for longitudinal data. *Journal of biopharmaceutical statistics*, 7(4):481–500.
- ZWANENBURG, G., HOEFSLOOT, H. C., WESTERHUIS, J. A., JANSEN, J. J. et SMILDE, A. K. (2011). Anova–principal component analysis and anova–simultaneous component analysis : a comparison. *Journal of Chemometrics*, 25(10):561–567.

Annexe A

Normalité des résidus

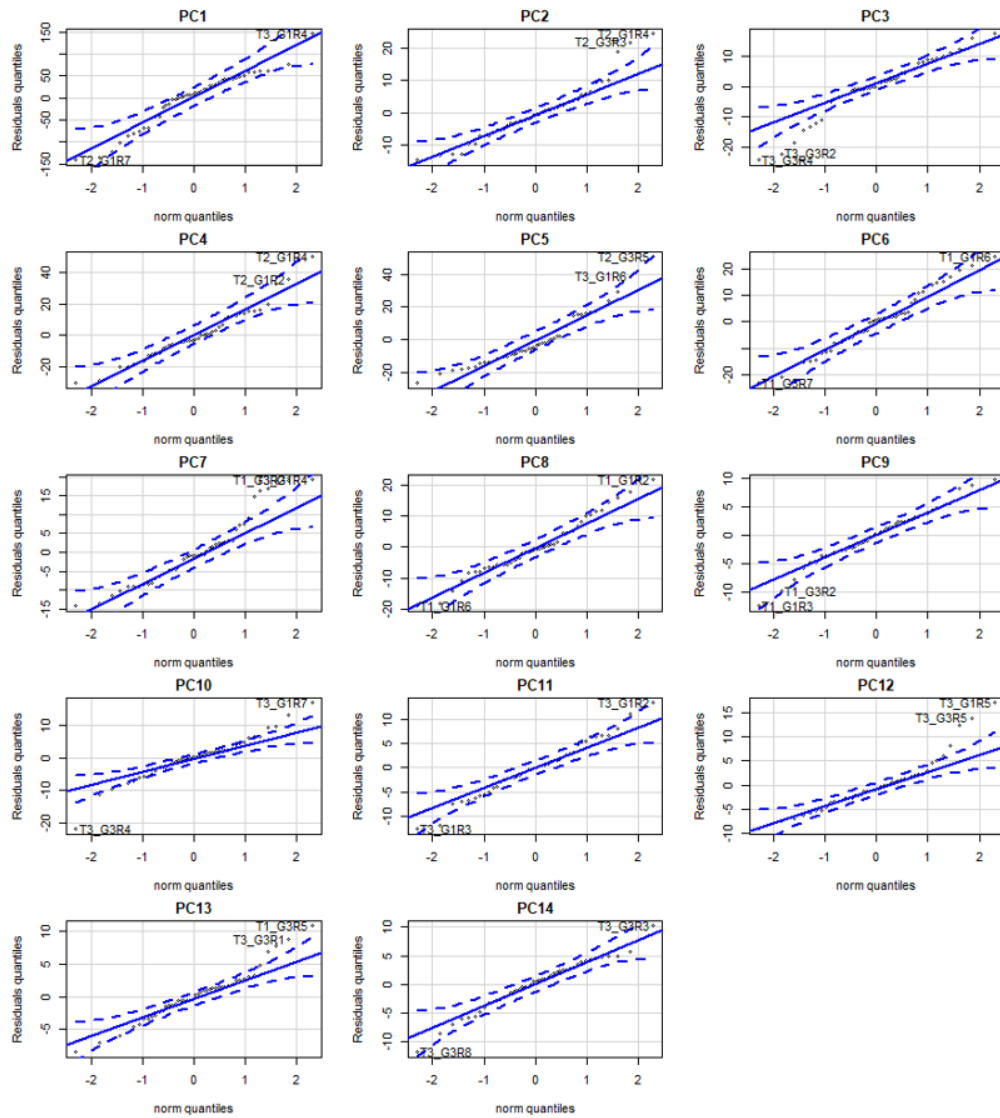


Figure A.1 – Diagrammes quantile-quantile des résidus des modèles ajustés sur chacune des 14 CPs.

Annexe B

Distributions des GLLR - procédure bootstrap

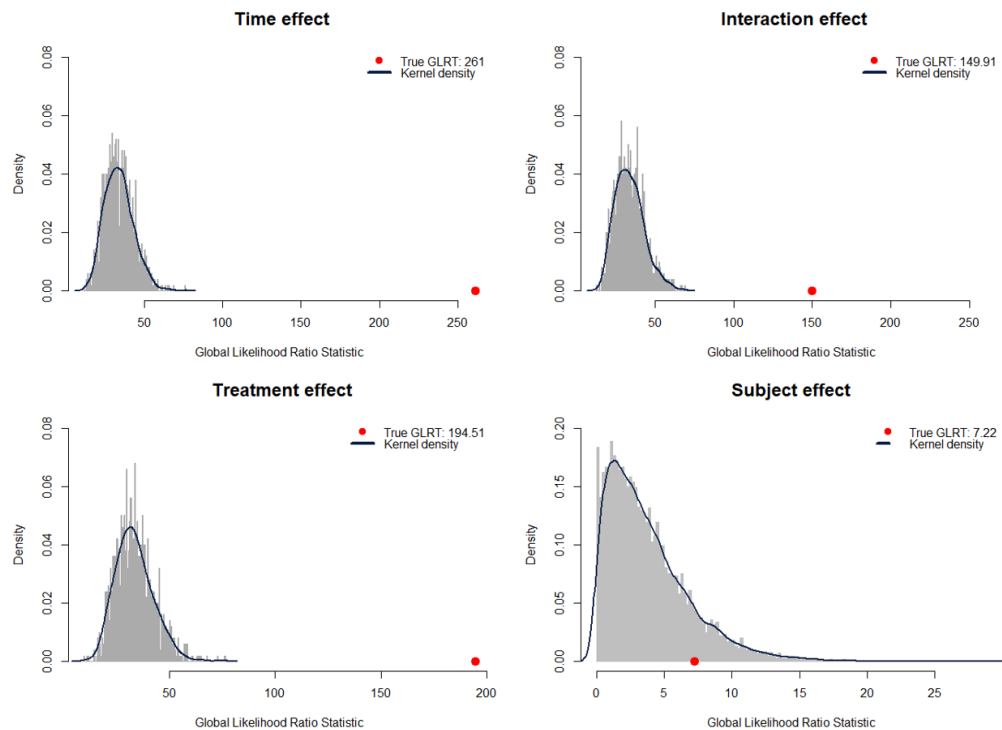


Figure B.1 – Distributions nulles des statistiques GLLR pour chaque effet obtenues par la procédure bootstrap.

