الجمهورية الجزائرية الديمقراطية الشعبية

Ministère de L'Enseignement Supérieur et de la Recherche Scientifique

UNIVERSITE SAAD DAHLAB DE BLIDA

Faculté des sciences

Département de Mathématiques

MEMOIRE DE MASTER

En Mathématiques

Option : Modélisation Stochastiques et Statistique

THÈME :

Analyse Statistique des données longitudinales

Réalisé par

Trea Fatma

Soutenu devant le Jury :

TAMI Omar	Université Blida 1	Président
FRIHI Redhouane	Université Blida 1	Examinateur
RASSOUL Abdelaziz	ENSH de Blida	Promoteur

Juillet 2023

DÉDICACES

Je dédie ce modeste mémoire :

je dédie ma graduation d'amour,de tendresse et du pouls qui vit dans mes veines , mon cher défunt père et ma chère mère , vous avez tout l'amour de moi.

De mes souhaits aujourd'hui est que mon défunt père soit présent avec moi a la cérémonie de remise des diplômes.

mes parents sont mon soutien dans la vie et ils ont un grand crédit aprés Dieu pour étudier et atteint ce stade .
je leur dédie la remise des diplômes.

TREA

REMERCIÉMENTS

En tout premier lieu, nous tenons à remercier ALLAH, le tout puissant et miséricordieux, qui m'a donné la volonté, la force et la santé de réaliser ce mémoire.

TABLE DES MATIÈRES

TABLE DES FIGURES

LISTE DES TABLEAUX

ملخص

تستخدم الدراسات الطولية في أبحاث ودراسات خدمات الصحة العقلية. النهج السائدة لتحليل البيانات الطولية هي نماذج التأثيرات المختلطة الخطية المعممة (MMLG) ومعادلات التقدير المعممة الموزونة (EEGW). على الرغم من أن كلا الفئتين من النماذج قد تم نشرهما على نطاق واسع وتطبيقهما على نطاق واسع ، إلا أن الاختلافات بين هذه الأساليب والقيود المفروضة عليها لم يتم تحديدها بشكل واضح وتوثيقها جيدًا. لسوء الحظ ، تحمل بعض الاختلافات والقيود آثارًا كبيرة على الإبلاغ عن نتائج البحث ومقارنتها وتفسيرها. في هذه الذاكرة ، نراجع كلا النهجين الرئيسيين لتحليل البيانات الطولية ونبرز أوجه التشابه والاختلافات الرئيسية بينهما. نحن نركز على المقارنة بين فئتي النماذج من حيث النموذج الافتراضات وتفسير معلمات النموذج وقابلية التطبيق والقيود باستخدام كل من البيانات الحقيقية والمحاكاة. نناقش التحذيرات والتحذيرات عند تطبيق النهجين المختلفين لبيانات الدراسة الحقيقية.

Résumé

Les études longitudinales sont utilisées dans la recherche sur la santé mentale et les études sur les services. Les approches dominantes pour l'analyse des données longitudinales sont les modèles linéaires généralisés à effets mixtes (GLMM) et les équations d'estimation généralisées pondérées (WGEE). Bien que les deux classes de modèles aient été largement publiées et largement appliquées, les différences et les limites de ces méthodes ne sont pas clairement définies et bien documentées. Malheureusement, certaines différences et limitations ont des implications importantes pour la communication, la comparaison et l'interprétation des résultats de la recherche. Dans ce mémoire, nous passons en revue les deux principales approches d'analyse de données longitudinales et soulignons leurs similitudes et leurs principales différences. Nous nous concentrons sur la comparaison des deux classes de modèles en termes de modèle hypothèses, interprétation des paramètres du modèle, applicabilité et limites, en utilisant à la fois des données réelles et simulées. Nous discutons des mises en garde et des mises en garde lors de l'application des deux approches différentes aux données d'études réelles.

Abstract

Longitudinal studies are used in mental health research and services studies. The dominant approaches for longitudinal data analysis are the generalized linear mixed-effects models (GLMM) and the weighted generalized estimating equations (WGEE). Although both classes of models have been extensively published and widely applied, differences between and limitations about these methods are not clearly delineated and well documented. Unfortunately, some of the differences and limitations carry significant implications for

reporting, comparing and interpreting research findings. In this thesis, we review both major approaches for longitudinal data analysis and highlight their similarities and major differences. We focus on comparison of the two classes of models in terms of model assumptions, model parameter interpretation, applicability and limitations, using both real and simulated data. We discuss caveats and cautions when applying the two different approaches to real study data.

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
ASCA	(ANOVA-Simultaneous Component Analysis)
APCA	(ANOVA-Principal Component Analysis)

INTRODUCTION GÉNÉRALE

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.

Les conceptions d'études longitudinales sont devenues de plus en plus populaires dans la recherche et la pratique dans toutes les disciplines. De telles conceptions capturent à la fois les différences entre les individus et la dynamique intra-sujet, offrant des possibilités d'étudier des changements biologiques, psychologiques et comportementaux complexes au fil du temps, tels que les effets causals du traitement et les mécanismes de changement. Étant donné que les conceptions d'études longitudinales créent des corrélations en série sur des évaluations répétées des mêmes sujets, les méthodes statistiques traditionnelles d'analyse de données transversales telles que la régression linéaire et logistique ne s'appliquent pas. De plus, étant donné que les études longitudinales sont généralement de longue durée, les données manquantes sont fréquentes.

Modèles spécialisés Malgré l'existence d'un corpus extrêmement important de littérature traitant du développement et de l'application des deux approches, les praticiens sont encore souvent confrontés à de nombreuses questions difficiles lors du choix et de l'application de tels modèles à des données d'étude réelles.

Par exemple, quelle approche est appliquée compte tenu des données d'une étude ?

Les deux approches produisent-elles des estimations et/ou des inférences identiques ? Si-non, comment approcher et interpréter ces différences ? Quels sont les avantages et les inconvénients associés à chaque approche ? Si certaines questions ont des réponses bien documentées dans la littérature, d'autres n'ont été abordées que récemment et attendent toujours des réponses.

Dans cette mémoire, nous donnons d'abord un aperçu des approches, puis discutons des différences majeures entre les deux classes de modèles. Contrairement à la littérature sur la discussion des deux méthodes, nous nous concentrons sur leurs implications pratiques, qui, selon nous, fournissent des conseils utiles aux praticiens pour sélectionner les bonnes approches pour leurs études et répondre efficacement à leurs questions d'étude. Et des méthodes doivent être utilisées pour résoudre les deux problèmes majeurs.

Les deux approches dominantes pour l'analyse des données longitudinales sont le modèle linéaire généralisé à effets mixtes (GLMM) et les équations d'estimation généralisées pondérées (WGEE).

Les deux méthodes sont dérivées de la même classe de modèles pour les données transversales, les modèles linéaires généralisés (GLM). Étant donné que différentes techniques sont utilisées pour étendre le GLM aux données longitudinales, le GLMM et le WGEE sont assez différents et certaines des différences ont des implications importantes pour Leur applicabilité et leur interprétation des résultats de l'étude.

CHAPITRE 1

APERÇUE SUR LES DONNÉES LONGITUDINALES

1.1 Généralités

Les données longitudinales sont connais courante dans de nombreux domaines, tels que la médecine, l'assurance, l'économie ou l'analyse des jeux vidéo. A titre d'exemple, le dossier médical partagé collecte, pour un grand nombre de patients et (dans certains pays) depuis plusieurs décennies, de nombreux marqueurs cliniques au cours du temps, ainsi que les temps d'évènements associés,

1.2 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 [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 and 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.
- elle 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 Molenberghs et al. [2014]. 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.

1.3 Généralités sur les données longitudinales

Pour préparer le terrain pour la discussion statistique qui suivra, il est utile de présenter une notation unifiée pour les divers aspects de la conception longitudinale. Nous indexons les N sujets de l'étude longitudinale comme

$$i = 1, 2, ..., N \quad \text{sujets.}$$

Pour une conception équilibrée dans laquelle tous les sujets disposent de données complètes et sont mesurés aux mêmes occasions. Nous indexons les occasions de mesure comme

$$j = 1, 2, ..., n \quad \text{observations.}$$

Ou dans le cas déséquilibré de nombres inégaux de mesures ou de points de temps différents pour différents sujets

$$j = 1, 2, ..., n_i \quad \text{observations pour le sujet i.}$$

Le nombre total d'observations est donné par

$$\sum_{i}^{N} n_i.$$

Les réponses répétées, ou les résultats, ou les mesures dépendantes pour le sujet i sont désignés par le vecteur $y_i = n_i \times 1$.
Les valeurs des p prédicteurs, ou covariables, ou variables indépendantes pour le sujet i à

l'occasion j sont notées (y compris un terme d'ordonnée à l'origine) :

$$x_{ij} = p \times 1.$$

Pour les prédicteurs indépendants du temps (entre sujets, par exemple le sexe), les valeurs de x_{ij} sont constantes pour $j = 1,...,n_i$. Pour les prédicteurs variant dans le temps (intra-sujet, par exemple l'âge), le x_{ij} peut prendre des valeurs spécifiques au sujet et au point temporel. Pour décrire l'ensemble de la matrice des prédicteurs du sujet i, nous utilisons la notation

$$X_i = n_i \times p.$$

1.3.1 Disposition des données

Il est également utile d'appliquer cette notation décrite précédemment pour décrire un ensemble de données longitudinales comme suit. Dans ce plan univarié, n varie selon

Sujet	Observation	Réponse	Covariables		
1	1	y_{11}	x_{111}	...	x_{11p}
1	2	y_{12}	x_{121}	...	x_{12p}
.	.	.	.	...	
1	n_1	y_{1n_1}	x_{1n_11}	...	x_{1n_1p}
.	.	.	.	...	
.	.	.	.	...	
N	1	y_{N1}	x_{N11}	...	x_{N1p}
N	2	y_{N2}	x_{N21}	...	x_{N2p}
.	.	.	.	...	
.	.	.	.	...	
N	n_N	y_{Nn_N}	$x_{Nn_{N1}}$	...	$x_{Nn_{Np}}$

le sujet et le nombre de lignes de données par sujet peut donc varier. En termes de covariables, si x_r, est invariant dans le temps (c'est-à-dire une variable inter-sujets), alors, pour un sujet donné i, les valeurs des covariables sont les mêmes dans le temps, à savoir, $X_{i1r} = x_{i2r} = x_{i3r} = ... = X_{in_ir}$.

1.4 Approches générales

Il existe plusieurs caractéristiques différentes des études longitudinales qui doivent être prises en compte lors de la sélection d'une analyse longitudinale appropriée.
Premièrement, il y a la forme du résultat ou de la mesure de la réponse. Si le résultat d'intérêt est continu et normalement distribué, des analyses beaucoup plus simples sont généralement possibles (par exemple, un modèle de régression linéaire à effets mixtes). En revanche, si le résultat est continu mais n'a pas de distribution normale (par exemple,

un décompte), des modèles non linéaires alternatifs (par exemple, un modèle de régression de Poisson à effets mixtes) peuvent être envisagés. Pour les résultats qualitatifs, tels que binaires (oui ou non), ordinaux (par exemple, triste, neutre, heureux) ou nominaux (républicain, démocrate, indépendant), des modèles non linéaires plus complexes sont également généralement requis.
Deuxièmement, le nombre de sujets N est une considération importante pour le choix d'une méthode d'analyse longitudinale. Les modèles plus avancés sont les modèles de régression à effets mixtes généralisés qui sont appropriés pour l'analyse de données longitudinales non équilibrées sont basés sur la théorie des grands échantillons et peuvent être inappropriés pour l'analyse d'études à petit N.
Troisièmement, le nombre d'observations par sujet n_i est également une considération importante lors du choix d'une méthode analytique. Pour $n_2 = 2$ pour tous les sujets, un score de changement simple peut être calculé et les données peuvent être analysées à l'aide de méthodes pour données transversales, telles que l'ANCOVA. Lorsque $n_i = n$ pour tous les sujets, la conception est dite équilibrée et les modèles traditionnels ANOVA ou MANOVA pour les mesures répétées peuvent être utilisés. Dans le cas le plus général où n_i varie d'un sujet à l'autre, des méthodes plus générales sont nécessaires (par exemple, des modèles de régression généralisés à effets mixtes).
Quatrièmement, le nombre et le type de covariables sont une considération importante pour la sélection du modèle pour $E(y_i)$.
Chacun de ces facteurs est important pour sélectionner un modèle analytique approprié pour l'analyse d'un ensemble particulier de données longitudinales.

1.5 Base des données et étude longitudinales

1.5.1 La métabolomique

La métabolomique est un domaine de recherche qui consiste à identifier et quantifier de manière exhaustive et non sélective l'ensemble des métabolites présents dans un système biologique (biofluides, tissus, cellules, organes, etc.) Vicente-Muñoz et al. [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 xénobiotiqueBonvallot et al. [2014].
Les techniques analytiques les plus couramment utilisées pour générer des données métabolomiques sont la spectrométrie de masse (MS) et la spectrométrie par réso- nance 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 Martin [2019].
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 Prud'homme et al. [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.

1.5.2 ANOVA

L'analyse de variance (ANOVA) est une famille de méthodes statistiques qui est principalement utilisée pour analyser l'influence statistique d'un ensemble de facteurs catégoriels sur une réponse quantitative, idéalement dans des études avec un plan expérimental équilibré. La variation totale observée dans la réponse d'intérêt est décomposée en variabilité expliquée par les différents effets d'un modèle choisi en fonction de la structure du plan expérimental sous-jacent. L'objectif principal de la méthode est de mesurer l'importance et la signification de chaque effet, mais plusieurs autres résultats, présentés dans cette section, peuvent en être dérivés.

1.5.3 ANOVA à mesures répétées (RM-ANOVA)

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 Kitayama 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 emboités 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 and 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 :

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

où :

y_{ijk} : est la réponse observée pour le ième 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

$(\alpha\beta)_{jk}$; est l'interaction entre le groupe j et le temps k.

Les $\rho_i(j)$ sont supposés être des variables aléatoires indépendantes $iN(0, \sigma^2 p)$ et les erreurs $\epsilon_{ijk} \sim \mathcal{N}(0, \sigma^2)$.

Par ailleurs, pour éviter une surparamétrisation du modèle, des contraintes sont imposées sur les paramètres :

$$\sum_j \alpha_j = 0, \quad \sum_k \beta_k = 0, \quad \sum_k (\alpha\beta)_{jk} = 0, \forall k \quad et \sum_j (\alpha\beta)_{jk} = 0 \forall j.$$

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 balancé, 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 :

$$\begin{aligned} SS_T &= \sum_{ijk} (y_{ijk} - \bar{y}_{...})^2 \\ &= bs \sum_j (\bar{y}_{.j.} - \bar{y}_{...})^2 + as \sum_k (\bar{y}_{..k} - \bar{y}_{...})^2 + s \sum_{jk} (\bar{y}_{.jk} - \bar{y}_{.j.} - \bar{y}_{..k} + \bar{y}_{...})^2 + b \sum_{ij} (\bar{y}_{ij.} - \bar{y}_{.j.}) + \\ &\quad \sum_{ijk} (y_{ijk} - \bar{y}_{ij.} - \bar{y}_{.jk} + \bar{y}_{.j.})^2 \\ &= SS_A + SS_B + SS_{AB} + SS_{S(A)} + SS_E \end{aligned} \tag{1.1}$$

où

- $\bar{y}_{...} = \sum_{i=1}^{s} \sum_{j=1}^{a} \sum_{k=1}^{s} \frac{y_{ijk}}{abs}$ 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.
- $\bar{y}_{.j_k} = \sum_{i=1}^{s} \frac{y_i jk}{s}$ est la moyenne estimée des sujets du groupe j au temps k.

- $\bar{y}_{ij.} = \sum_{k=1}^{b} \frac{y_i jk}{b}$ est la moyenne estimée des mesures du sujet i du groupe j.

Le tableau 1.1 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	$asb-1$			

TAB. 1.1 : 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 :**
  $H_0 : \beta_1 = \beta_2 = \cdots = \beta_b = 0$
  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 :**

  $H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_a = 0$
  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 :**
  $H_0 : \sigma_p^2 = 0$
  $H_1 : \sigma_p^2 > 0$

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

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

1.5.4 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 tem- porelle comme une variable distincte. Cette caractéristique ne lui permet cependant pas de décrire des tendances car elle ne reconnait pas l'ordre temporel des variables Hedeker and 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].

1.5.4.1 Modèles linéaire 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 2

MODÈLES LINÉAIRES MIXTES

Histoire et Motivation

La classe des modèles mixtes linéaires (LLM) fournit un cadre pour définir des modèles linéaires très généraux avec des facteurs fixes et aléatoires, estimer leurs paramètres et interpréter leurs résultats avec des outils d'inférence et de prédiction. Dans le monde statistique, les LMM ont maintenant pris le pas sur les modèles Anova traditionnels en raison de leur généralité et de leur capacité à traiter des conceptions déséquilibrées et des équations de modèles et des structures de covariance plus avancées. Bien que les premières versions de ce modèle aient été introduites très tôt par Fisher puis Henderson [39]. En chimiométrie, les LMM sont peu connus, malgré leur utilité dans de nombreuses situations, telles que l'analyse des composantes de la variance en chimie analytique, les expériences organisées en blocs, les mesures répétées ou les études longitudinales et l'analyse de plus en plus populaire des données expérimentales des sciences de la vie (ex. -études en omique). Par rapport aux modèles linéaires généraux et ANOVA, les modèles mixtes linéaires s'accompagnent de nombreux problèmes techniques d'estimation et d'inférence, ce qui dépasse le cadre de cet article. Cette section ne fournira donc que des principes généraux et des résultats sur cette famille de modèles, en restreignant sa présentation aux concepts et formules essentiels.

2.1 Modèles linéaire 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 [2019].

Ce chapitre 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 Humphreys et al. [2019], ainsi que des ouvrages des références suivant : "Applied mixed models in medicine" (Ansari et al. [2014]), "Longitudinal data analysis" (Hedeker and Gibbons [2006]) et "Linear mixed models for longitudinal data"Verbeke and Lesaffre [1997]

2.2 Définition et notation

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

$$Y = X\alpha + Z\beta + \epsilon \tag{2.1}$$

où X la matrice de modèle des effets.Elle est de dimension $n \times p$ où p est le nombre de paramètre fixes du modèle(terme constant inclus).Ces modèle matrice est organisées en blocs et les variable catégorielle peuvent être codées de différentes manière.la méthode *sum-to-Zero-coding* sera néanmoins utilisées.nous verrons plus en détaille comment cette matrice est construire dans le cadre de la méthode $ASCA$ au chapitre suivant.le deuxième bloc est lié aux effets aléatoire.la matrice de design Z est de dimension $n \times p$ où q où q correspond à la somme du nombre de niveaux des effets aléatoires.elle est construire à l'aide d'un codage fictif(avec des 0 est des 1) pour les variable catégorielles(aléatoire)du modèle et aucun intercept n'est inclus. β est un vecteur de variables aléatoires associées à chaque effet aléatoire. ϵ représente la variation résiduelle, c'est a dire celle qui n'est pas inclus dans aucune autre composante du modèle. les hypothèses principales pour les effets aléatoire sont $\beta \sim \mathcal{N}(0, \mathbf{G})$ et $\epsilon \sim \mathcal{N}(0; \mathbf{R})$ avec $\mathbf{G}$ et $\mathbf{R}$ les matrice de covariance de taille respective $q \times q$ et $n \times n$. par ailleurs,les effets aléatoire et les résidus sont supposés indépendants. le modèle linéaire mixte est également caractérisé par son espérance et sa variance qui s'écrivent :

$$E[Y] = X\alpha \tag{2.2}$$

et

$$V = var(\mathbf{Y}) = var(\mathbf{X}\alpha + \mathbf{Z}\beta + \epsilon) = \mathbf{Z}var(\beta)\mathbf{Z}' + var(\epsilon) = \mathbf{ZGZ}' + \mathbf{R} \tag{2.3}$$

on en déduit que la distribution marginale de $\mathbf{Y}$ suit une loi normale multivariée :
$\mathbf{y} \sim \mathcal{N}(\mathbf{X}\alpha, V)$.
Contrairement aux GLM, la matrice $\mathbf{V}$ n'est pas plus diagonale mais inclut des termes de covariance entre les observation liés aux termes aléatoires .en effets ,deux mesures effectuées sur un même sujet seront corrélées là où, pour des sujets différents,elle 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 Ansari et al. [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.

2.3 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 5.1 du chapitre I

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

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

$$\mathbf{Y} = \mathbf{X}\alpha + \mathbf{Z}\beta + \epsilon = \begin{pmatrix} 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & {}_1 & -1 & -1 & -1 \\ 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & -1 & 1 & 0 & 1 & 0 \\ 1 & -1 & 0 & 1 & 0 & 1 \\ 1 & -1 & -1 & -1 & -1 & -1 \\ 1 & -1 & 1 & 0 & 1 & 0 \\ 1 & -1 & 0 & 1 & 0 & 1 \\ 1 & -1 & -1 & -1 & -1 & -1 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \beta_1 \\ \beta_2 \\ \alpha\beta_{11} \\ \alpha\beta_{12} \end{pmatrix} + \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \rho_1(1) \\ \rho_2(1) \\ \rho_1(2) \\ \rho_2(2) \end{pmatrix} + \epsilon$$

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

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

où σ^2 p 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 :

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 :

La variance d'une seule observation dans un modèle à effet aléatoire est égale à la somme des composantes de variance : $var(y_{ijk}) = \sigma^2 + \sigma_p^2$.
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 Ansari et al. [2014].

2.4 Modèle a covariance de 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 R. Le modèle suivant à effets fixes, et sans l'effet sujet aléatoire, est alors considéré :

Compte tenu que le modèle n'inclut pas d'effets aléatoires, la matrice de covariance V est égale à la matrice des résidus R. En effet, $V = ZGZ' + R = R$. Cependant, la différence avec un modèle linéaire simple est que, dans un modèle à structure de covariance, la matrice 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, R est obtenue comme suit :

$$R = \begin{pmatrix} R_1 & 0 & 0 & 0 \\ 0 & R_2 & 0 & 0 \\ 0 & 0 & R_3 & 0 \\ 0 & 0 & 0 & R_4 \end{pmatrix}$$

Chaque bloc $R_i (i = 1, ..., 4)$ 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.

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

  $cov(\epsilon_{ij}, \epsilon_{ij'}) = \sigma_{jj'}$
  Nombre de paramètres : $b(b+1)/2$

- Composante symétrique : toutes les covariances sont égales.

$$R_i = \sigma^2 \begin{pmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{pmatrix}$$

  $cov(\epsilon_{ij}, \epsilon_{ij'}) = \rho\sigma^2$
  Nombre de paramètres : 2

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

$$R_i = \sigma^2 \begin{pmatrix} 1 & \rho & \rho^2 \\ \rho & 1 & \rho \\ \rho^2 & \rho & 1 \end{pmatrix}$$

$cov(\epsilon_{ij}, \epsilon_{ij'}) = \sigma^2 \rho^{|j-j'|}$
Nombre de paramètres : 2

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

$$R_i = \sigma^2 \begin{pmatrix} 1 & \rho_1 & \rho_2 \\ \rho_1 & 1 & \rho_1 \\ \rho_2 & \rho_1 & 1 \end{pmatrix}$$

$cov(\epsilon_{ij}, \epsilon_{ij'}) = \alpha^{|j-j'|}$
Nombre de paramètres : b

décrites dans la littérature (Brown and Prescott [2015]). 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 Balshi and Wolfinger [1997].

Pour des modèles emboités, 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.

2.5 Modèle a coefficients aléatoire

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_{jl}'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 :

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

Considérons toujours un modèle à quatre invididus pour lesquels il y trois observations à des temps différents. Les matrices Z et G s'écrivent

$$Z = \begin{pmatrix} 1 & t_{11} & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & t_{12} & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & t_{13} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & t_{21} & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & t_{22} & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & t_{23} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & t_{31} & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & t_{32} & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & t_{33} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & t_{41} \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & t_{42} \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & t_{43} \end{pmatrix}$$

$$G = \begin{pmatrix} G_1 & 0 & 0 & 0 \\ 0 & G_2 & 0 & 0 \\ 0 & 0 & G_3 & 0 \\ 0 & 0 & 0 & G_4 \end{pmatrix}$$

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

$$V = \begin{pmatrix} V_1 & 0 & 0 & 0 \\ 0 & V_2 & 0 & 0 \\ 0 & 0 & V_3 & 0 \\ 0 & 0 & 0 & V_4 \end{pmatrix} \quad \text{avec} \quad V_i = \begin{pmatrix} \sigma^2 + c_{i,11} & c_{i,12} & c_{i,13} \\ c_{i,12} & \sigma^2 + c_{i,22} & c_{i,23} \\ c_{i,13} & c_{i,23} & \sigma^2 + c_{i,33} \end{pmatrix}$$

où $i = 1, ..., 4$ et $c_{i,kk'} = \sigma_p^2 + (t_{ik} + t_{ik'})\sigma_{p,p} + t_{ik}t_{ik'}\sigma_p^2$. 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.

$$y_{ijk} = (\mu + p_{ij}) + \lambda_{tijk} + \epsilon_{ijk}$$

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.

2.6 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 $y_i : y \sim N(X_\alpha, V)$. Le logarithme de la fonction de vraisemblance est donné par :

$$l = \log(L) = k - \frac{1}{2}[\log|V| + (y - X\alpha)'V^{-1}(y - X\alpha)], \tag{2.4}$$

où k est une constante. Les valeurs des paramètres qui maximisent cette vraisem- blance 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 2.4 avec $\mathbf{V}$ égal à l'estimateur REML $\hat{\mathbf{V}}$:

$$\hat{\alpha} = (X'\hat{V}^{-1}X)^1 X'\hat{V}^{-1}y. \tag{2.5}$$

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

$$\hat{\beta} = GZ'\hat{V}^{-1}(y - X\hat{\alpha}). \tag{2.6}$$

Pour plus d'informations concernant les méthodes d'estimations en LLM, voir Ansari et al. [2014].

2.7 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 Verbeke and Lesaffre [1997]. Le test de rapport de vraisemblance (LRR) peut également être utilisé afin de tester la significativé des paramètres de variances ou des effets fixes en comparant deux modèles emboités. La statistique de test suivante est alors calculée :

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

où M_1 et M_0 sont les vraisemblance des modèles qui incluent ou excluent respectivement l'effet testé. Sous l'hypothèse nulle, cette statistique suit une loi chicarré χ^2 d'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 $\chi^2 : H_0 : 0.5\chi_0^2 + 0.5\chi_1^2$ Pinheiro and 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 and 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 and Lesaffre [1997].

2.8 Extensions des LMMs a ux réponses de grande dimension

2.8.1 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 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,Beier 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 2.1 La suite de cette section est consacrée à une présentation succincte des concepts principaux de cette approche décrite plus en détails dans Beier et al. [2017].

La méthode s'articule en deux grandes étapes. Dans un premier temps, la matrice de réponses 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.

2.8.1.1 Étape 1 : Modèle linéaire général et matrice du modèle

Le modèle GLM multivarié s'écrit comme suit :

$$Y = X\Theta + E$$

où $\mathbf{Y}$ est la matrice $n \times m$ des réponses observées, $\mathbf{X}$ la matrice du modèle de dimension $n \times p$, Θ la matrice $p \times m$ des paramètres et $\mathbf{E}$ est la matrice $n \times m$ des erreurs dont les

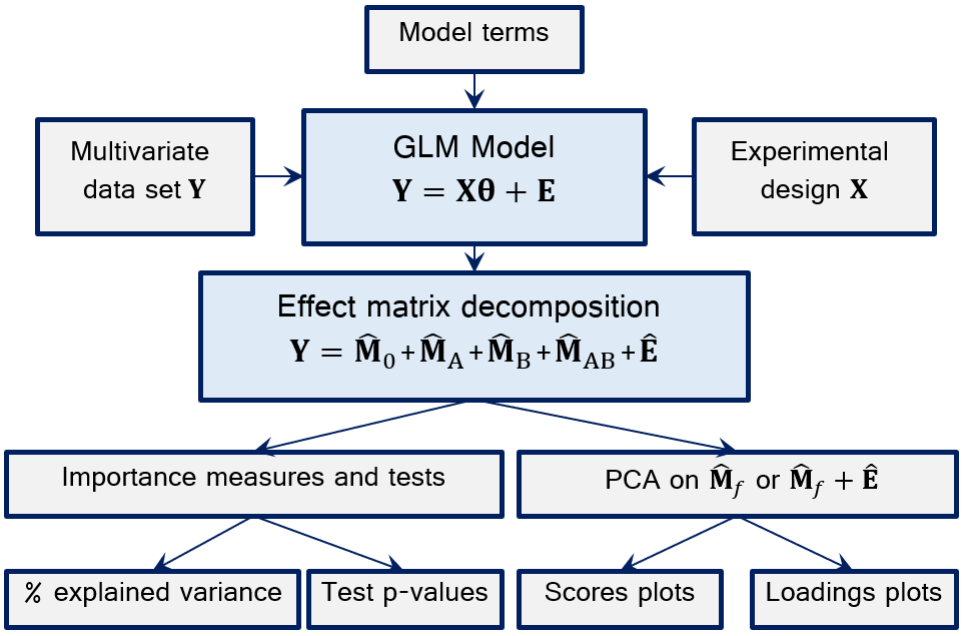


FIG. 2.1 : Etapes principales des méthodes ASCA+ et APCA+

colonnes sont supposées suivre une loi normale multivariée $N(0, \sigma^2 I_n)$. 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 : $X = (X_0|X_1|...|X_F)$. Pour un facteur A comportant a niveaux, une matrice Xa avec $a-1$ colonnes est crée. 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 :

$$X = \begin{array}{c} \begin{array}{cc|cc|cc} X_1 & X_A & \multicolumn{2}{c|}{X_B} & \multicolumn{2}{c}{X_{AB}} \end{array} \\ \left(\begin{array}{cc|cc|cc} 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & -1 & 1 & 0 & -1 & 0 \\ 1 & -1 & 0 & 1 & 0 & -1 \\ 1 & -1 & -1 & -1 & 1 & 1 \end{array}\right) \end{array}$$

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

2.8.1.2 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 :

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

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 X, $\hat{\Theta}$ est divisée en blocs par rapport à chaque effet : $\hat{\Theta} = (\hat{\Theta}'_0|\hat{\Theta}'_1|...|\hat{\Theta}'_F)$. Pour l'exemple considéré, la matrice des paramètres estimés est la suivante :

$$\hat{\Theta} = \left(\begin{array}{cccc} \hat{\mu}_{..1} & \hat{\mu}_{..2} & \cdots & \hat{\mu}_{..m} \\ \hline \hat{\alpha}_{11} & \hat{\alpha}_{12} & ... & \hat{\alpha}_{1m} \\ \hline \hat{\beta}_{11} & \hat{\beta}_{12} & ... & \hat{\beta}_{1m} \\ \hat{\beta}_{21} & \hat{\beta}_{22} & ... & \hat{\beta}_{2m} \\ \hline (\widehat{\alpha\beta})_{111} & (\widehat{\alpha\beta})_{112} & ... & (\widehat{\alpha\beta})_{11m} \\ (\widehat{\alpha\beta})_{121} & (\widehat{\alpha\beta})_{122} & ... & (\widehat{\alpha\beta})_{12m} \end{array}\right) = \left(\begin{array}{c} \hat{\Theta}_0 \\ \hline \hat{\Theta}_A \\ \hline \hat{\Theta}_B \\ \hline \hat{\Theta}_{AB} \end{array}\right)$$

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

$$\hat{M}_f = X_f \hat{\Theta}_f$$

Chaque matrice d'effet est de même dimension que **Y** (nm) et représente la part de **Y** expliquée par chaque effet du modèle (terme constant, effets principaux et interactions). Par exemple, pour l'effet B :

$$\hat{M}_B = X_B \hat{\Theta}_B = \left(\begin{array}{cc} 1 & 0 \\ 0 & 1 \\ -1 & -1 \\ 1 & 0 \\ 0 & 1 \\ -1 & -1 \end{array}\right) \left(\begin{array}{cccc} \hat{\beta}_{11} & \hat{\beta}_{12} & ... & \hat{\beta}_{1m} \\ \hat{\beta}_{21} & \hat{\beta}_{22} & ... & \hat{\beta}_{2m} \end{array}\right)$$

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

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

$$Y = \hat{M}_0 + \sum_{f=1}^{F} \hat{M}_f + \hat{E}; \tag{2.7}$$

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

2.8.1.3 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 **Y** de dimension nm tout en conservant un maximum d'information.

La matrice **Y** est alors décomposée en

Y=TP'

où

- **T** $(n \times r)$, la matrice des scores, contient les coordonnées des observations initiales dans le nouvel espace.
- **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'inter-prétation

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 y_l 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 :

$$ASCA : \hat{M}f = T_f P_f' \qquad APCA : \hat{M}f + \hat{E} = T_f P_f'$$

où T_f et 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] :

$$T_f^E = (\hat{M}_f + \hat{E})P_f'$$

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 T_f^E ne seradonc pas maximisée.

2.8.1.4 Étape 4 : Pourcentage de variation et tests de significa-tivité

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 et pour calculer le pourcentage de variance expliqué par un effet f :

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

où $\hat{E}_{full}$ et $\hat{E}_{/f}$ 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 and 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 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 $\hat{Y} = X\hat{\Theta}$ et les résidus $\hat{E} = Y - \hat{Y}$ du modèle sous H_0. Ces deux matrices seront utilisées dans la boucle bootstrap.
5. Pour un nombre fixé B d'itérations ($b = 1, ..., B$), répéter les étapes suivantes :
   - Échantillonner $\hat{\mathbf{E}}_\mathbf{b}$ avec remise parmi $\hat{\mathbf{E}}$
   - Générer une matrice de réponses bootstrap $\mathbf{Y_b} = \hat{\mathbf{Y}} + \hat{\mathbf{E}}_\mathbf{b}$
   - Estimer les modèles complet et restreint sur base de cette nouvelle matrice $\mathbf{Y_b}$
   - Calculer la statistique de test $\mathbf{F_b}$.
6. Calculer la $p-$valeur pour la statistique de test :
$$p_f^{boot} = \frac{\sum_{b=1}^{B} I(Fb \leq F_{obs}) + 1}{B + 1}$$

2.8.2 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 Antonelli et al. [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 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 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 2.2 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 Antonelli et al. [2020].

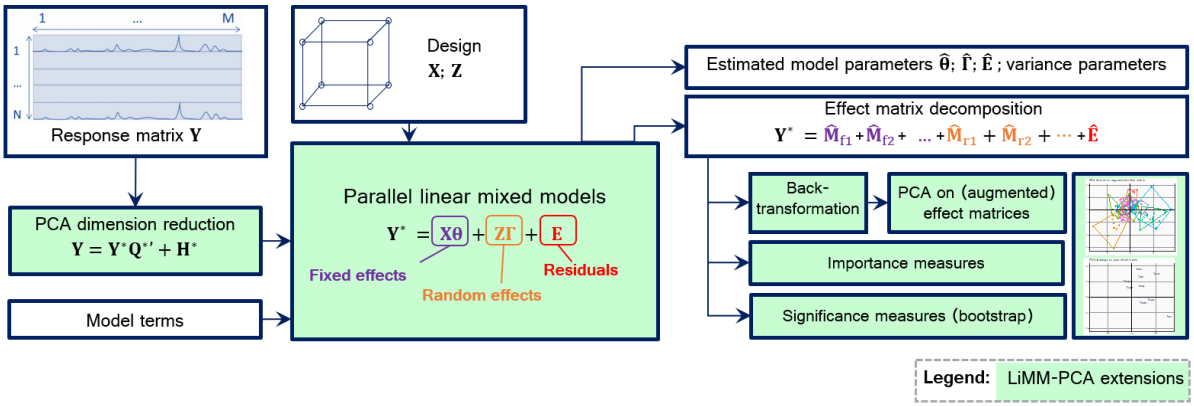


FIG. 2.2 : Etapes de la méthode LiMM-PCA. Les encadrés surlignés correspondent aux extensions apportées à la méthode ASCA+

2.8.2.1 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 Y^* de dimension $(n \times m^*)$ est telle que :

$$Y = Y^* Q^* + H^* \tag{2.8}$$

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, $\sum_{m^*} var(CP) \geq 99\%$). La matrice des loadings, 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.

2.8.2.2 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 : $\mathbf{X} = (\mathbf{X_0}|\mathbf{X_1}|\mathbf{X_2}|...|\mathbf{X_F})$ et $\mathbf{Z} = (\mathbf{Z_1}|\mathbf{Z_2}|...|\mathbf{Z_R})$. Les codages "sum deviation" et binaires sont utilisés pour coder respectivement les F effets fixes et les R effets aléatoires.

Pour $l = 1,...,m$, nous avons

$$y_l^* = X\theta_l + Z\gamma_l + \epsilon_l \tag{2.9}$$

où les paramètres fixes θ_l et les prédictions aléatoires γ_l sont définis et estimés par maximum de vraisemblance en accord avec ce qui a été présenté au chapitre 2. Par ailleurs, les composantes de variance, $\hat{\sigma}^2_{rl}(r = 1,...,R)$ et $\hat{\sigma}^2_{el}$, associées aux R effets aléatoires et aux résidus sont estimés. Les différents modèles estimés peuvent être écrits sous la forme multivariée suivante :

$$Y^* = X\hat{\Theta} + Z\hat{\Gamma} + \hat{E} \tag{2.10}$$

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

2.8.2.3 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 :

$$Y^* = \hat{M}_0 + \sum_{f=1}^{F} X_f \hat{\Theta}_f + \sum_{r=1}^{R} Z_r \hat{\Gamma}_r + \hat{E}$$

$$Y^* = \hat{M}_0 + \sum_{f=1}^{F} \hat{M}_f + \sum_{r=1}^{R} \hat{M}_r + \hat{E}$$

2.8.2.4 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 and 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 $y_l^*, j = 1, ..., m^*$:

- Effets aléatoires : conditionnellement à $\mathbf{X}$, la variance est estimée en additionnant les estimations des composantes de variance,

$$\widehat{var}(y_l^* | X) = \sum_{r=1}^{R} \hat{\sigma}_{rl}^2 + \hat{\sigma}_{\epsilon l}^2.$$

- Effets fixes : la variance pour un facteur est définie comme $\hat{\sigma}_{fj}^2 = var(X_f,\ \theta_{fj})$, ce qui équivaut à la variance de la l^{eme} colonne de la matrice d'effet correspondante, $\hat{\sigma}_{fj}^2 = var(\hat{M}_{fl})$

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,

$$\widehat{var}(Y^*) = \sum_{l=0}^{m^*} \widehat{var}(y_l^*) = \sum_{l=0}^{m^*} \left(\sum_{f=1}^{F} \hat{\sigma}_{fl}^2 + \sum_{r=1}^{R} \hat{\sigma}_{rl}^2 + \hat{\sigma}_{\epsilon l}^2 \right);$$

- Variance liée à l'effet g,

$$g \in \{1...R, 1...F\} : \widehat{var}_g = \sum_{l=1}^{m^*} \hat{\sigma}_{gl}^2$$

- Pourcentage de variance expliqué par l'effet g :

$$\%\widehat{var}\ g = \widehat{var(g)}/\widehat{var(Y^*)}$$

2.8.2.5 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 $\mathbf{L_{H_{0,j/g}}}$ les log-vraisemblances des modèles respectivement avec et sans l'effet $g \in \{1...R, 1...F\}$ testé. Le rapport global de vraisemblance pour cet effet s'obtient comme suit :

$$GLLR_g = \Lambda_g^{obs} = 2\left[\sum_{l=1}^{m^*} \left(\log\left(L_{H_1,l}\right) - \log\left(L_{H_{0,l/g}}\right) \right) \right]$$

Pour les effets fixes, sous H_0, la statistique suit une loi de chi-carré $X^2_{m^* \times k}$ 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 X^2. Cependant, Antonelli et al. [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 and Hinkley [1997]et Halekoh and Højsgaard [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 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 ($b = 1, ..., B$), répéter les étapes suivantes :
   - Générer une matrice de réponse bootstrap Y_b^* avec $Y_b^* = X\hat{\theta} + Z\Gamma_b + E_b$ où les matrices Γ_b et 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 X et$Z$ proviennent également du modèle nul.
   - Estimer les modèles complet et restreint sur base de cette nouvelle matrice Y_b^*.
   - Calculer la statistique $GLLR\Lambda_g^b$. 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 :

$$p_g^{boot} = \frac{\sum_{b=1}^{B} I(\Lambda_g^b \leq \Lambda_g^{obs} + 1}{B+1}.$$

2.8.2.6 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 $\hat{\mathbf{M}}_\mathbf{g}$:

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

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 $(1-\alpha)^{\psi}me$ 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,Antonelli et al. [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 3

ANALYSE DE DONNÉES LONGITUDINALES UNIVARIÉES

Implémentions 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 ré-sidus 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.

3.1 Outils disponibles en *R*

3.1.1 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 diagnostique : lme4 Bates [2010] ;Hell et al. [2015] et nlme Pinheiro and Bates [2000] ; Gałecki et al. [2013] ;Goyal 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 de Fisher F et de Student 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.

3.1.2 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 Hell 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'hétéroscédasticé 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 3.1.

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

TAB. 3.1 : 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-vraissemblance restreinte (method = "REML") que par maximisation de la log-vraissemblance (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.

3.2 Données univariées : de Orthodont

Ce jeu de données, introduit par Potthoff and 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 3.1 montre l'évolution au fil du temps de la distance pour les filles et les garçons.

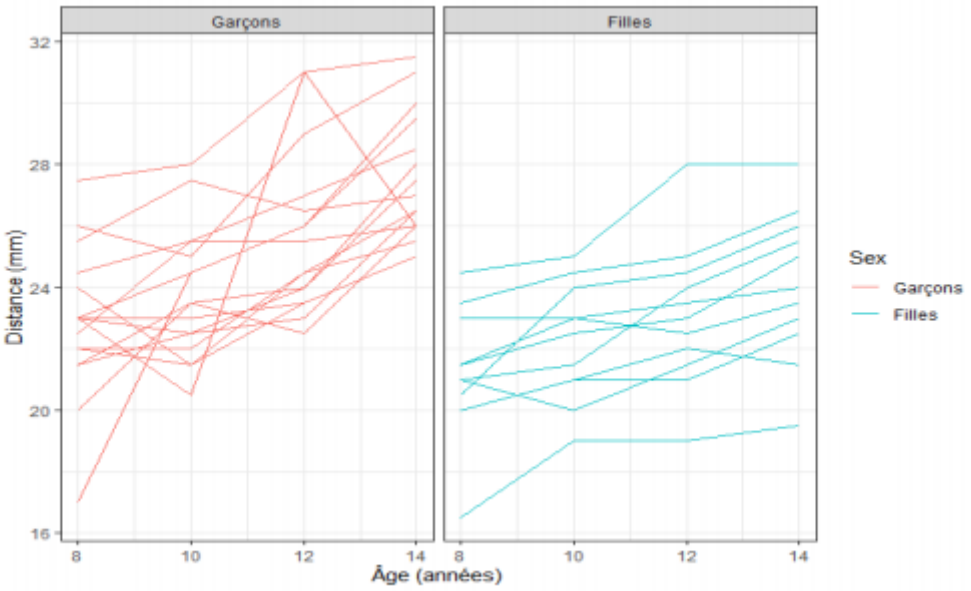


FIG. 3.1 : Croissance de la distance entre la glande pituitaire et la fosse ptérygo-maxillaire chez les enfants.

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 $V = ZGZ' + R$ où Z est la matrice de design des effets aléatoires, G la matrice de covariance de ces effets et R la matrice des résidus. V est bloc-diagonale et nous noterons 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 $V_i = R_i$.

3.2.1 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 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 3.2), 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			

$^{***}p < 0.001; ^{**}p0.01; ^{*}p$

TAB. 3.2 : 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".

3.2.2 lme/lmer vs gls

3.2.3 Composante symétrique

Au chapitre II, il a été vu que pour le modèle à intercepts aléatoires ajusté au point précédent, chaque bloc Vi de la matrice de covariance était obtenu comme suit

$$V_i = Z_i G_i Z_i' + R_i = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \sigma_p^2 \begin{pmatrix} 1 & 1 & 1 & 1 \end{pmatrix} + \begin{pmatrix} \sigma^2 & 0 & 0 & 0 \\ 0 & \sigma^2 & 0 & 0 \\ 0 & 0 & \sigma^2 & 0 \\ 0 & 0 & 0 & \sigma^2 \end{pmatrix}$$

$$V_i = \begin{pmatrix} \sigma^2+\sigma_p^2 & \sigma_p^2 & \sigma_p^2 & \sigma_p^2 \\ \sigma_p^2 & \sigma^2+\sigma_p^2 & \sigma_p^2 & \sigma_p^2 \\ \sigma_p^2 & \sigma_p^2 & \sigma^2+\sigma_p^2 & \sigma_p^2 \\ \sigma_p^2 & \sigma_p^2 & \sigma_p^2 & \sigma^2+\sigma_p^2 \end{pmatrix}$$

Nous pouvons constater que cette matrice a une structure à composante symé-trique (CS) avec une corrélation égale à $\frac{\rho=\sigma_p^2}{\sigma^2+\sigma_p^2}$. 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.

```
correlation = corCompSymm( form = ~1| Subject ), orthodont, method="REML")
```

Sur base des résultats obtenus pour les deux modèles, les matrices V_i respectives ont pu être construites (voir Tableau). Nous pouvons constater que celles-ci sont exactement similaires, tout comme les estimations des effets fixes et des log-vraisemblances reprises dans le Tableau 3.2.

3.2.4 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 3.4. 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, V_i a pu être obtenue sur base des matrices G_i, Z_i et R_i (Tableau 3.4). La matrice 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 $V_i = R_i$ 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 minimes par rapport à l'ordre de grandeur des estimations.

Modèle à effet aléatoire	Modèle à covariance pattern
Fonction	
`lme , lmer` effet sujet aléatoire	`gls` Stucture de corrélation à composante symétrique
Log-vraissemblance	
-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 $\rho = \frac{\sigma_p^2}{\sigma^2+\sigma_p^2} = 0.6245479$ $\mathbf{V_i} = \mathbf{Z_i G Z_i'} + \mathbf{R_i}$ $= \begin{pmatrix} \sigma^2+\sigma_p^2 & \sigma_p^2 & \sigma_p^2 & \sigma_p^2 \\ \sigma_p^2 & \sigma^2+\sigma_p^2 & \sigma_p^2 & \sigma_p^2 \\ \sigma_p^2 & \sigma_p^2 & \sigma^2+\sigma_p^2 & \sigma_p^2 \\ \sigma_p^2 & \sigma_p^2 & \sigma_p^2 & \sigma^2+\sigma_p^2 \end{pmatrix}$ $\sigma^2 + \sigma_p^2 = 5.260426$ $\sigma_p = 3.285388$ $= \begin{pmatrix} 5.2604 & 3.2854 & 3.2854 & 3.2854 \\ 3.2854 & 5.2604 & 3.2854 & 3.2854 \\ 3.2854 & 3.2854 & 5.2604 & 3.2854 \\ 3.2854 & 3.2854 & 3.2854 & 5.2604 \end{pmatrix}$	Variance : • sur les résidus $\sigma^2 = 5.260422$ Corrélation intra-sujet $\rho = 0.6245472$ $\mathbf{V_i} = \mathbf{Z_i G Z_i'} + \mathbf{R_i} = \mathbf{R_i}$ $= \begin{pmatrix} \sigma^2 & \sigma_\rho^2 & \sigma_\rho^2 & \sigma_\rho^2 \\ \sigma_\rho^2 & \sigma^2 & \sigma_\rho^2 & \sigma_\rho^2 \\ \sigma_\rho^2 & \sigma_\rho^2 & \sigma & \sigma_\rho^2 \\ \sigma_\rho^2 & \sigma_\rho^2 & \sigma_\rho^2 & \sigma^2 \end{pmatrix}$ $\sigma^2 = 5.260422$ $\sigma^2\rho = 3.285382$ $= \begin{pmatrix} 5.2604 & 3.2854 & 3.2854 & 3.2854 \\ 3.2854 & 5.2604 & 3.2854 & 3.2854 \\ 3.2854 & 3.2854 & 5.2604 & 3.2854 \\ 3.2854 & 3.2854 & 3.2854 & 5.2604 \end{pmatrix}$

TAB. 3.3 : Comparaison des résultats du modèle à effet sujet aléatoire et du modèle à composante symétrique.

<table>
<tr><th>Modèle à effet aléatoire</th><th>Modèle à covariance pattern</th></tr>
<tr><td colspan="2">Fonction</td></tr>
<tr><td><code>lme , lmer</code>
effet temps aléatoire
au sein de chaque sujet</td><td><code>gls</code>
Structure de corrélation générale (non structurée)
Paramètres de variance différents à chaque temps</td></tr>
<tr><td colspan="2">Log-vraissemblance</td></tr>
<tr><td>-212.56</td><td>-212.56</td></tr>
<tr><td colspan="2">Matrice de variance-covariance</td></tr>
<tr><td>$\mathbf{V_i = Z_i G Z_i' + Ri}$
$G_i = \begin{pmatrix} 4.7693 & 1.9806 & 0.9251 & 2.1451 \\ & 1.9806 & 2.7997 & 1.0195 & 2.5281 \\ & 0.9251 & 1.0195 & 2.9044 & 2.6032 \\ & 2.1451 & 2.5281 & 2.6032 & 4.2912 \end{pmatrix}$
$Z_i = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix}$
R_i est diagonale et $\sigma^2 = 0.4820$
$V_i = \begin{pmatrix} 5.4154 & 2.7168 & 3.9102 & 2.7102 \\ 2.7168 & 4.1848 & 2.9271 & 3.3172 \\ 3.9102 & 2.9271 & 6.4557 & 4.1307 \\ 2.7102 & 3.3172 & 4.1307 & 4.9857 \end{pmatrix}$</td><td>$\mathbf{V_i = R_i}$
$= \sigma^2 \begin{pmatrix} 1 & . & . & . \\ \rho_{12^{K_2}} & K_2^2 & . & . \\ \rho_{13^{K_3}} & \rho_{23^{K_2 K_3}} & K_3^2 & . \\ \rho_{13^{K_4}} & \rho_{24^{K_2 K_4}} & \rho_{34^{K_3 K_4}} & K_4^2 \end{pmatrix}$
$\rho_{ij} = \begin{pmatrix} 0.586 & & & \\ 0.661 & 0.563 & & \\ 0.522 & 0.726 & 0.728 & \end{pmatrix}$
$\sigma^2 = 5.4154$
Facteurs d'inflation$(K_1, ..., K_4)$: $1, 0.8791, 1.0918, 0.9595$
$= \begin{pmatrix} 5.4154 & 2.7168 & 3.9102 & 2.7102 \\ 2.7168 & 4.1848 & 2.9271 & 3.3172 \\ 3.9102 & 2.9271 & 6.4557 & 4.1307 \\ 2.7102 & 3.3172 & 4.1307 & 4.9857 \end{pmatrix}$</td></tr>
</table>

TAB. 3.4 : Comparaison des résultats du modèle à effet sujet aléatoire et du modèle à composante symétrique.

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.

3.3 Données multivariées : de 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 determinée 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 (fig3.2) :

- Avant le traitement (T1)

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

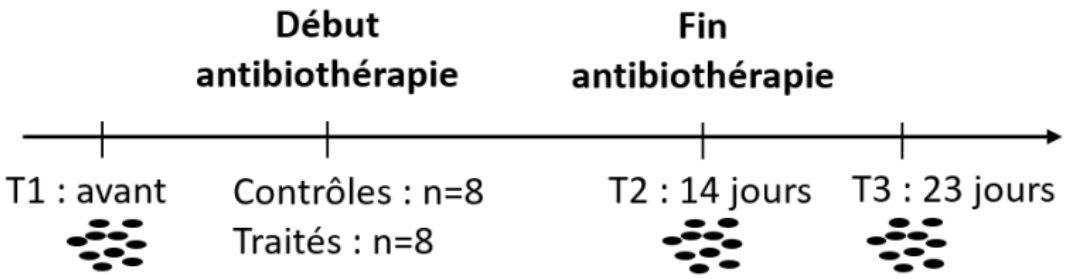


FIG. 3.2 : Design d'expérience sous-jacent aux données Choo

Chaque échantillon a ensuite été analysé par spectroscopie RMN. Les signaux bruts obtenus ont alors subi une série de transformations et de pré-traitements 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 inclues dans la base de données :

- la matrice outcomes $(n \times m)$ ou matrice spectrale $\mathbf{Y}$: 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 $0,5ppm$). 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 $(T1, T2ouT3)$, le traitement (contrôle ou non) ainsi que le sujet $(1\xi16)$. Notons que suite au retrait de deux observations, le design est non-balancé.

Notons que la fenêtre spectrale est comprise ici entre $4,5$ et $0,5$ ppm.

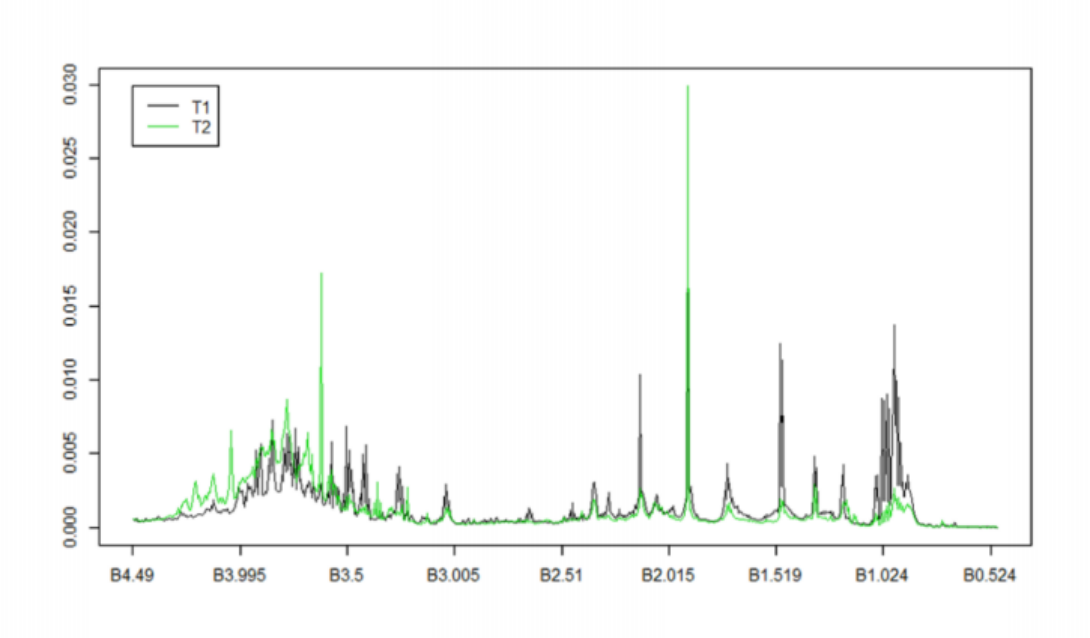


FIG. 3.3 : Profils spectraux1H-RMN d'un même individu aux temps T1 et T2

3.3.1 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 Figure 3.4. 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.

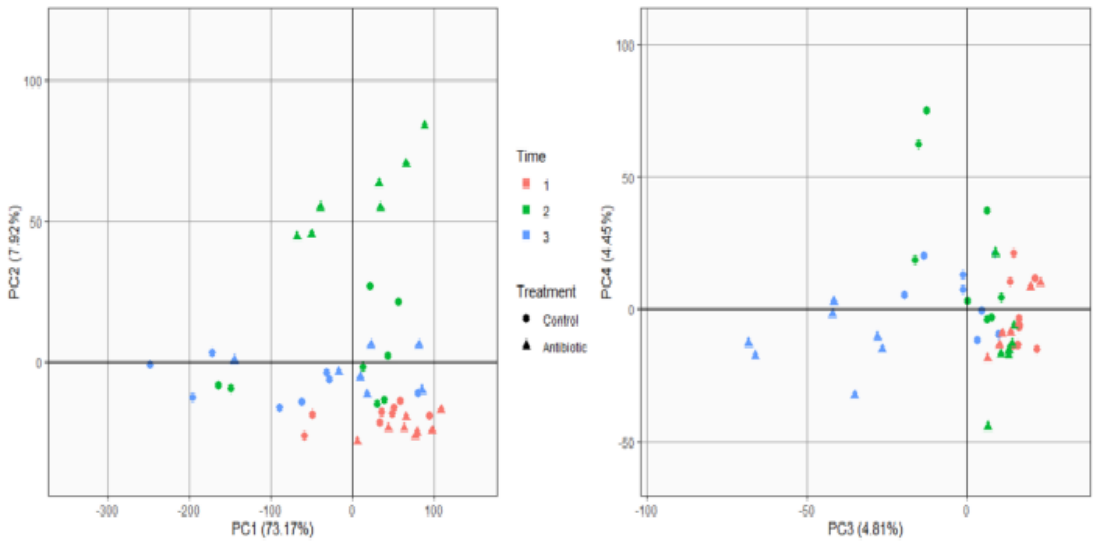


FIG. 3.4 : – Graphiques des scores sur les CPs 1-2 (à gauche) et 3-4 (à droite).

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.

3.3.2 ASCA+ et APCA+

Dans cette section, les différentes méthodes multivariées vont être appliquées à la matrice 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 débugger et d'améliorer le package LMWiRe développé par France. Thiel et al. [2023]
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.

3.3.2.1 Étape 1 : Modèle GLM

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

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

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 $\sim i\mathcal{N}(0, \sigma^2)$.
L'ANOVA 2 croisée peut être généralisée sous la forme d'un modèle GLM qui, sous sa forme multivariée, est défini comme suit :

$$Y = X\Theta + E = X_0\Theta_0 + X_T\Theta_T + X_G\Theta_G + X_{GT}\Theta_{GT} + E$$

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 X est de taille 46×6 et Θ de taille 6×1452.

3.3.2.2 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 Y peut être décomposée en matrices d'effets qui contiennent les moyennes des niveaux des facteurs :

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

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

3.3.2.3 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 $\hat{M}_T$, $\hat{M}_G$ et $\hat{M}_{GT}$ (liées respectivement au temps, au traitement et à l'interaction). La Figure 3.5 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.

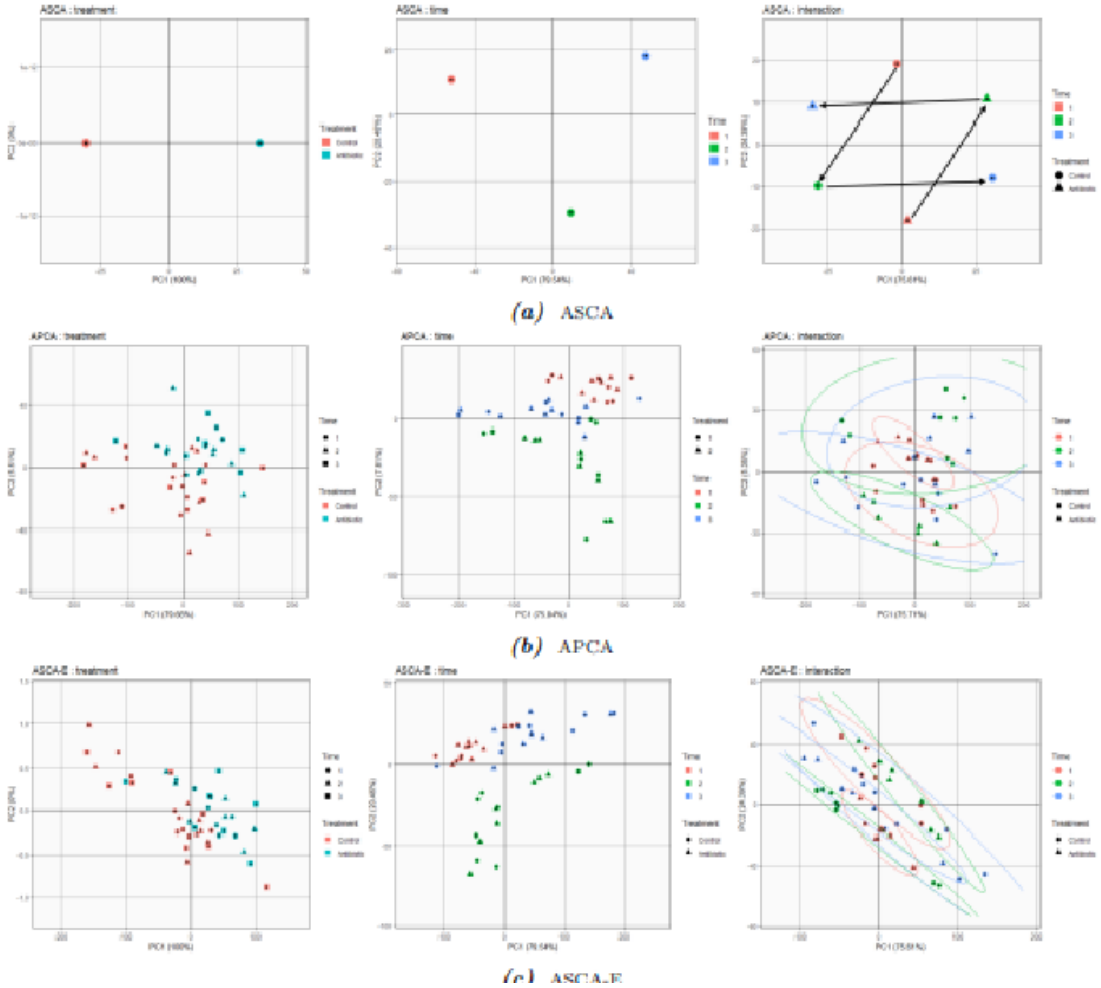


FIG. 3.5 : Graphique des scores pour l'ANOVA–principal component analysis (APCA), l'ANOVA–simultaneous component analysis (ASCA) et l'ASCA-E pour les effets temps, traitement et l'interaction.

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. 3.5.3b et 3.5.3c). 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. 3.6a. 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.

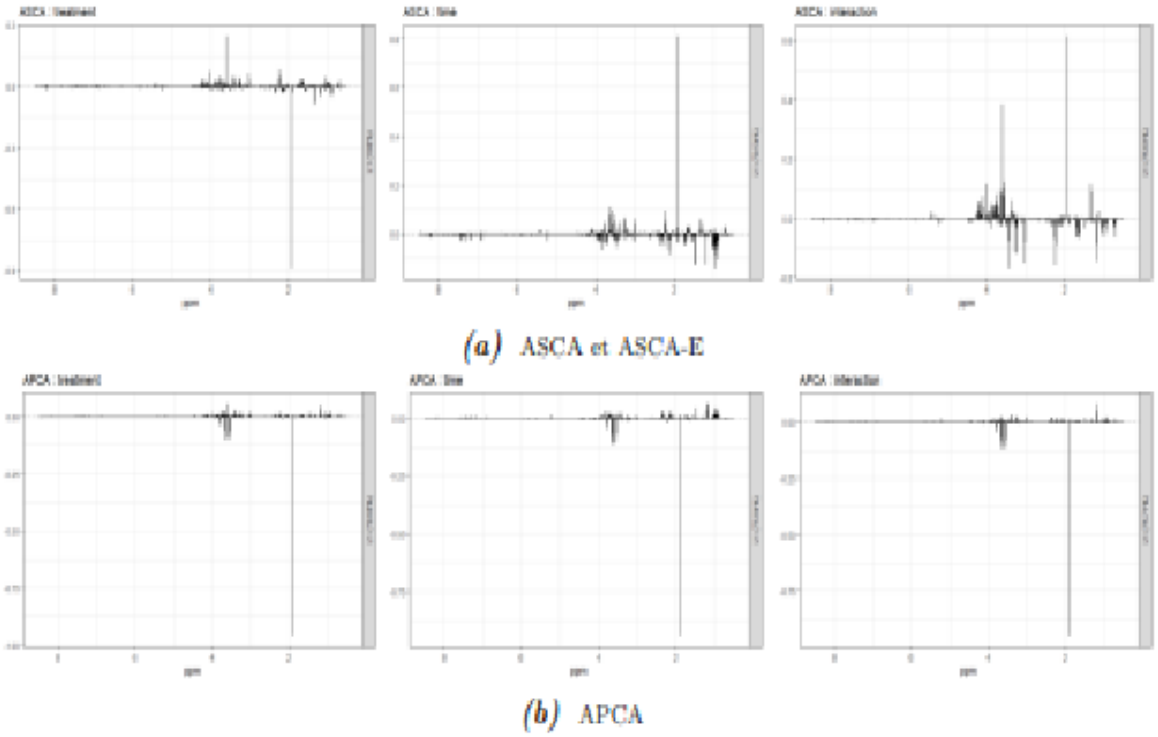


FIG. 3.6 : Graphiques des loadings pour la première composante principale en ASCA/ASCAE et APCA pour les effets temps, traitement et l'interaction.

3.3.2.4 Étape 4 : Pourcentage de variation et significativité

Les pourcentages de variation des facteurs sont repris à la Fig. 3.7. Les p-valeurs associées ont été obtenues à l'aide d'une procédure de boostrap (1000 rééchantillo- nages). 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.

Facteur	%	p-valeur
Traitement	10.02	0.023
Temps	21.37	<0.001
Interaction	7.42	0.042
Résidus	60.78	-

FIG. 3.7 : Pourcentage de variance et significativité de chaque facteur.

3.3.3 LiMM-PCA

3.3.3.1 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 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.

3.3.3.2 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 :

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

où y_{ijk} = est la réponse observée pour le sujet i du traitement j et au temps k ; $s = 8$, $a = 2$ et $b = 3$.
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 $\sim i\mathcal{N}(0, \sigma^2)$.
$\rho_i(j)$ est l'effet aléatoire lié au sujet i emboité dans le groupe $j \sim i\mathcal{N}(0, \sigma^2)$.
Les erreurs ϵ_{ijk} sont supposées $\sim i\mathcal{N}(0, \sigma^2)$.
Ce modèle peut être généralisé sous la forme d'un LMM qui, sous sa forme multivariée, est défini comme suit :

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

$$Y^* = X_0\Theta_0 + E = X_0\Theta_0 + X_T\Theta_T + X_G\Theta_G + X_{GT}\Theta_{GT} + Z_{S[G]}\Gamma_{S[G]} + E$$

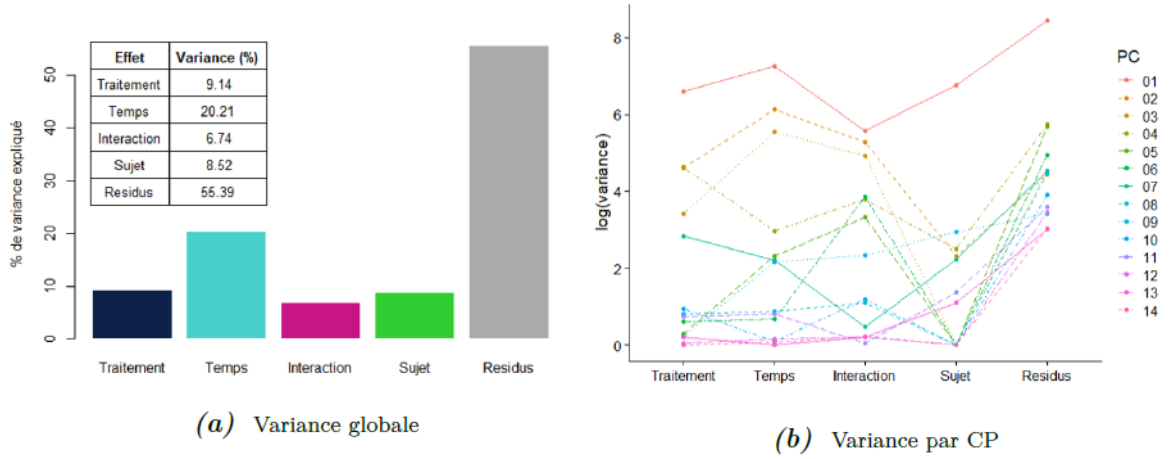
où 0, T, G et GT et S correspondent respectivement à l'intercept, le temps, le groupe de traitement, l'interaction et au sujet. Des LMMs sont ajustés en parallèle pour chacune des 14 CPs retenues à l'étape précédente afin d'obtenir des estimations REML pour les différents paramètres. La matrice de réponse Y^* peut ensuite être décomposée en matrices d'effets fixes et aléatoires de taille (46×14) :

$$Y^* = \hat{M}_0 + \hat{M}_T + \hat{M_{SG}} + \hat{M_{S[G]}} + \hat{E}$$

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.

3.3.3.3 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 3.8.a 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.



(a) Variance globale

(b) Variance par CP

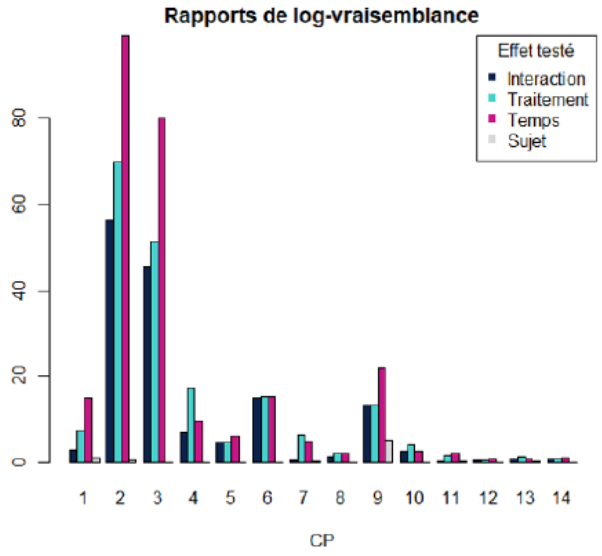
FIG. 3.8 : Variance expliquée par chaque composante du modèle. La variance est exprimée en échelle logarithmique.

La figure 3.8.b 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 4ème 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 3.9.a 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.

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. 3.9b). 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.



(a) (R)LLR

Effet	Variance (%)	p-valeur bootstrap
Traitement	9.14	< 0.001
Temps	20.21	< 0.001
Interaction	6.74	< 0.001
Sujet	8.52	0.12629

(b) Significativité des effets

FIG. 3.9 : Rapports de vraisemblance pour chaque effet sur chacune des CPs et p-valeurs obtenues sur base de la procédure bootstrap (1000 rééchantillonnages).

3.3.3.4 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 $G = \sigma_p^2 I16$. Cependant, comme mentionné précédemment, pour certaines composantes (7 CPs sur 14), la variance estimée est nulle. Les matrices 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). Fdl1,dl2 représente le 95ème quantile de la distribution F avec dl1, dl2 degrés de liberté.

Effet	E(MS)	ddl associés	Matrice ajoutée	Facteur de correction
Traitement	E(MS)G	dlG=a-1	$\hat{M}S(G)$	$\sqrt{F_{dlG,dlS(G)} \times \frac{dfG}{dfS(G)}}$
Temps	E(MS)T	dlT=b-1	$\hat{E}$	$\sqrt{F_{dlT,dlE} \times \frac{dfT}{dfE}}$
Interaction	E(MS)GT	dlGT = (a-1)(b-1)	$\hat{E}$	$\sqrt{F_{dlGT,dlE} \times \frac{dlGT}{dlE}}$
Sujet	E(MS)S(G)	dlS(G) = a(s-1)	$\hat{E}$	$\sqrt{F_{dlS(G),dlE} \times \frac{dlS(G)}{dlE}}$
Résidus	E(MS)E	$dlE = a(s-1)(b-1)$	-	-

TAB. 3.5 : Caption

Les scores et des loadings propres aux différentes composantes de variance du modèle sont représentés à la figure 3.12. 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. 3.10d) 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.

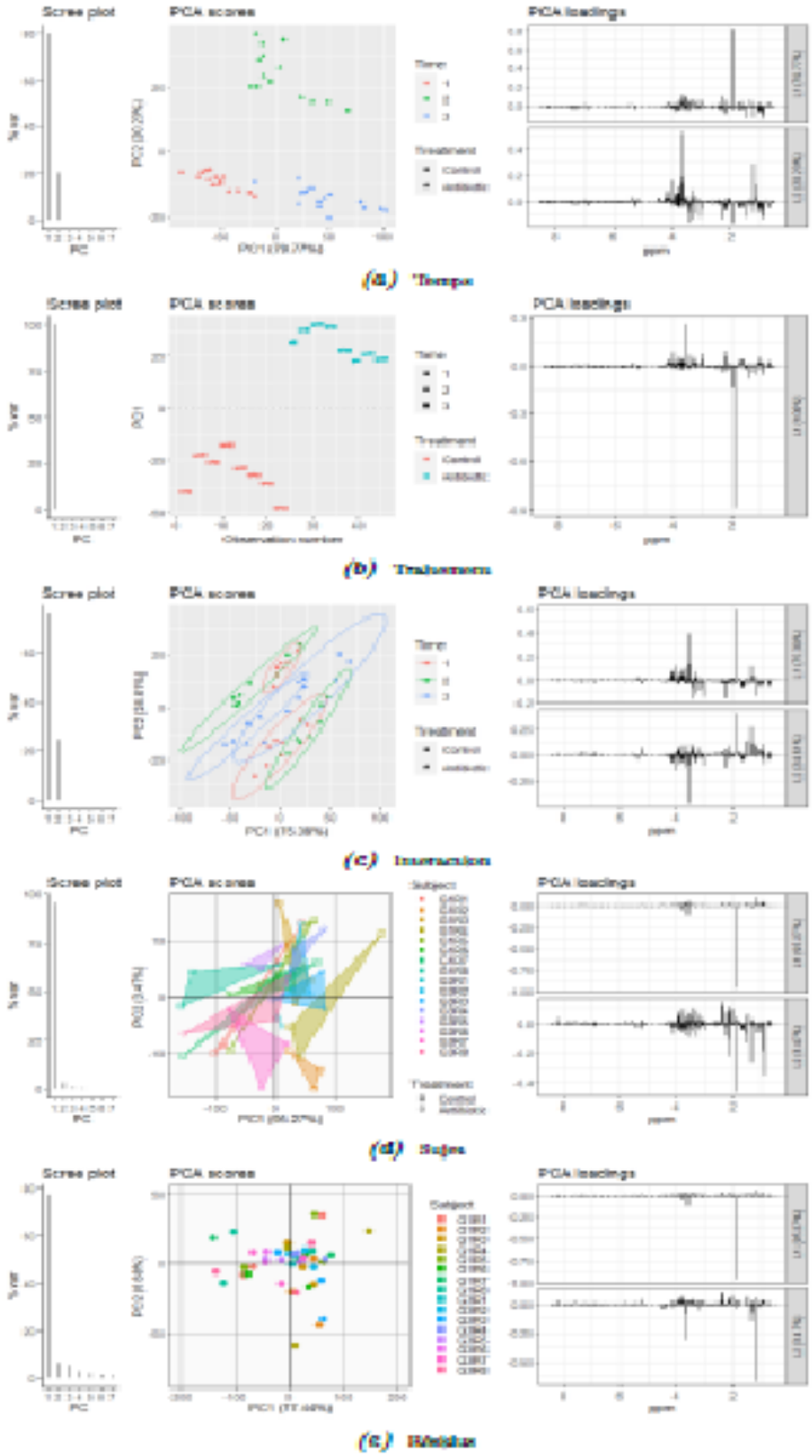


FIG. 3.10 : ACP sur les matrices d'effets. Les scores sont augmentés tandis que les screeplot et loadings sont obtenues sur base des matrices pures.

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. 3.11) 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 2ème 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 deça 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 3.12a.
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 connaitre 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).

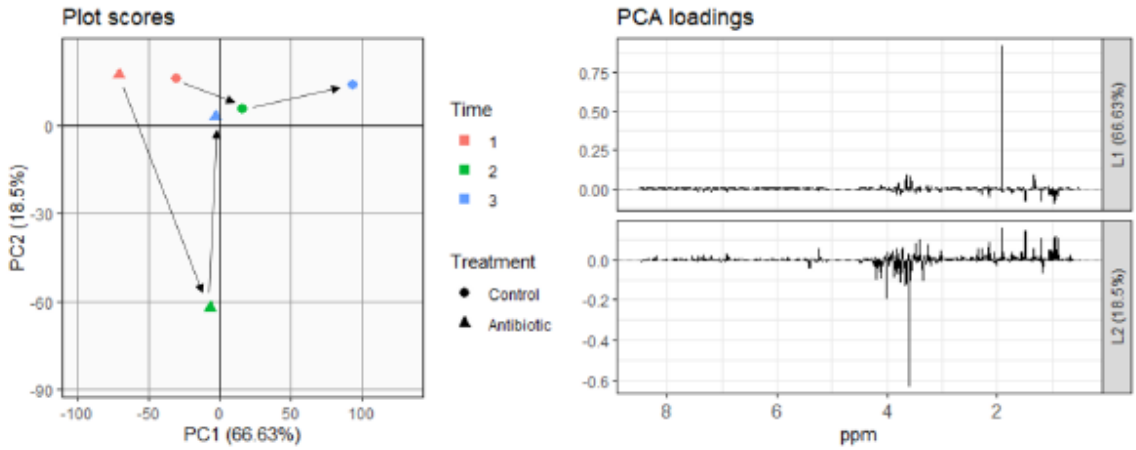
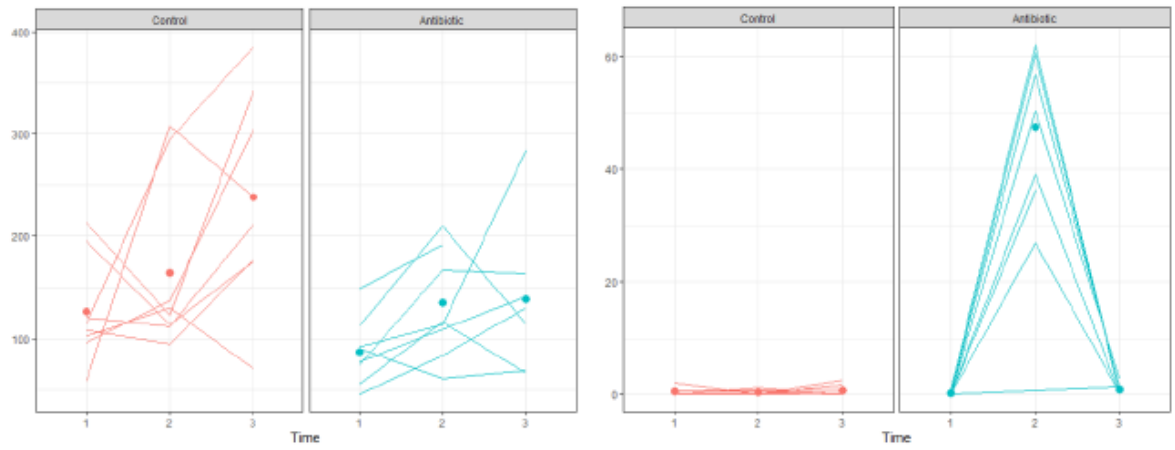


FIG. 3.11 : ACP sur les matrices d'effets additionnées entre elles.

A l'opposé, les effets du traitement et de l'interaction sont clairement visibles sur la deuxième composante (figure 3.11) et un pic important apparait aux alentours de 3.6 ppm. La variable correspondante est représentée à la fig. 3.12b1 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.



(a) Pic à 2 ppm *(b)* Pic à 3.6 ppm

FIG. 3.12 : 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 3.6 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).

% 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

TAB. 3.6 : Comparaison des pourcentages de variances expliqués par chaque effet avec et sans le pic.

3.4 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, compo-sante symmétrique, autorégressive d'ordre 1, etc.). Elle peut également permettre à la variance des résidus d'être différente pour différents groupes d'individus (hétéros-cédasticité : 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. 3.4

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 gls sera préférée pour ajuster un modèle à covariance pattern à des données longitudinales à un seul niveau de regroupement.

BIBLIOGRAPHIE

Marti Anderson and Cajo Ter Braak. Permutation tests for multi-factorial analysis of variance. *Journal of statistical computation and simulation*, 73(2) :85–113, 2003.

Morad Ansari, Gemma Poke, Quentin Ferry, Kathleen Williamson, Roland Aldridge, Alison M Meynert, Hemant Bengani, Cheng Yee Chan, Hülya Kayserili, Şahin Avci, et al. Genetic heterogeneity in cornelia de lange syndrome (cdls) and cdls-like phenotypes with observed and predicted levels of mosaicism. *Journal of medical genetics*, 51(10) : 659–668, 2014.

Alexandre Antonelli, RJ Smith, C Fry, Monique SJ Simmonds, Paul J Kersey, HW Pritchard, MS Abbo, C Acedo, J Adams, AM Ainsworth, et al. *State of the World's Plants and Fungi*. PhD thesis, Royal Botanic Gardens (Kew) ; Sfumato Foundation, 2020.

Thomas J Balshi and Glenn J Wolfinger. Immediate loading of brånemark implants in edentulous mandibles : a preliminary report. *Implant dentistry*, 6(2) :83–92, 1997.

Ed Bates. *The evolution of the European Convention on Human Rights : from its inception to the creation of a permanent Court of Human Rights*. Oxford University Press, 2010.

Sebastian Beier, Thomas Thiel, Thomas Münch, Uwe Scholz, and Martin Mascher. Misaweb : a web server for microsatellite prediction. *Bioinformatics*, 33(16) :2583–2585, 2017.

Nathalie Bonvallot, Marie Tremblay-Franco, Cecile Chevrier, Cecile Canlet, Laurent Debrauwer, Jean-Pierre Cravedi, and Sylvaine Cordier. Potential input from metabolomics for exploring and understanding the links between environment and health. *Journal of Toxicology and Environmental Health, Part B*, 17(1) :21–44, 2014.

Helen Brown and Robin Prescott. *Applied mixed models in medicine*. John Wiley & Sons,

2015.

Jocelyn M Choo, Tokuwa Kanno, Nur Masirah Mohd Zain, Lex EX Leong, Guy CJ Abell, Julie E Keeble, Kenneth D Bruce, A James Mason, and Geraint B Rogers. Divergent relationships between fecal microbiota and metabolome following distinct antibiotic-induced disruptions. *Msphere*, 2(1) :10–1128, 2017.

Anthony Christopher Davison and David Victor Hinkley. *Bootstrap methods and their application*. Number 1. Cambridge university press, 1997.

Peter Diggle. *Analysis of longitudinal data*. Oxford university press, 2002.

Garrett Fitzmaurice, Marie Davidian, Geert Verbeke, and Geert Molenberghs. *Longitudinal data analysis*. CRC press, 2008.

Andrzej Gałecki, Tomasz Burzykowski, Andrzej Gałecki, and Tomasz Burzykowski. *Linear mixed-effects model*. Springer, 2013.

Parag Goyal, Justin J Choi, Laura C Pinheiro, Edward J Schenck, Ruijun Chen, Assem Jabri, Michael J Satlin, Thomas R Campion Jr, Musarrat Nahid, Joanna B Ringel, et al. Clinical characteristics of covid-19 in new york city. *New England Journal of Medicine*, 382(24) :2372–2374, 2020.

Ulrich Halekoh and Søren Højsgaard. A kenward-roger approximation and parametric bootstrap methods for tests in linear mixed models–the r package pbkrtest. *Journal of Statistical Software*, 59 :1–32, 2014.

Donald Hedeker and Robert D Gibbons. *Longitudinal data analysis*. Wiley-Interscience, 2006.

Stefan W Hell, Steffen J Sahl, Mark Bates, Xiaowei Zhuang, Rainer Heintzmann, Martin J Booth, Joerg Bewersdorf, Gleb Shtengel, Harald Hess, Philip Tinnefeld, et al. The 2015 super-resolution microscopy roadmap. *Journal of Physics D : Applied Physics*, 48(44) : 443001, 2015.

Aelys M Humphreys, Rafaël Govaerts, Sarah Z Ficinski, Eimear Nic Lughadha, and Maria S Vorontsova. Global dataset shows geography and life form predict modern plant extinction and rediscovery. *Nature ecology & evolution*, 3(7) :1043–1047, 2019.

Noriyuki Kitayama, Viola Vaccarino, Michael Kutner, Paul Weiss, and J Douglas Bremner. Magnetic resonance imaging (mri) measurement of hippocampal volume in post-traumatic stress disorder : a meta-analysis. *Journal of affective disorders*, 88(1) :79–86, 2005.

Manon Martin. *Uncovering informative content in.* PhD thesis, Erasmus University Medical Center, The Netherlands, 2019.

Geert Molenberghs, Garrett Fitzmaurice, Michael G Kenward, Anastasios Tsiatis, and Geert Verbeke. *Handbook of missing data methodology.* CRC Press, 2014.

Shinichi Nakagawa and Holger Schielzeth. A general and simple method for obtaining r2 from generalized linear mixed-effects models. *Methods in ecology and evolution*, 4(2) : 133–142, 2013.

José C Pinheiro and Douglas M Bates. Linear mixed-effects models : basic concepts and examples. *Mixed-effects models in S and S-Plus*, pages 3–56, 2000.

Richard F Potthoff and SN Roy. A generalized multivariate analysis of variance model useful especially for growth curve problems. *Biometrika*, 51(3-4) :313–326, 1964.

Luc Prud'homme, Raymond Vienneau, Serge Ramel, and Nadia Rousseau. La légitimité de la diversité en éducation : réflexion sur l'inclusion. *Éducation et francophonie*, 39 (2) :6–22, 2011.

Michel Thiel, Baptiste Feraud, and Bernadette Govaerts. Asca+ and apca+ : Extensions of asca and apca in the analysis of unbalanced multifactorial designs. *Journal of Chemometrics*, 31(6) :e2895, 2017.

Michel Thiel, Nadia Benaiche, Manon Martin, Sébastien Franceschini, Robin Van Oirbeek, and Bernadette Govaerts. limpca : An r package for the linear modeling of high-dimensional designed data based on asca/apca family of methods. *Journal of Chemometrics*, 37(7) :e3482, 2023.

Geert Verbeke and Emmanuel Lesaffre. The effect of misspecifying the random-effects distribution in linear mixed models for longitudinal data. *Computational Statistics & Data Analysis*, 23(4) :541–556, 1997.

Sara Vicente-Muñoz, Inmaculada Morcillo, Leonor Puchades-Carrasco, Vicente Payá, Antonio Pellicer, and Antonio Pineda-Lucena. Nuclear magnetic resonance metabolomic profiling of urine provides a noninvasive alternative to the identification of biomarkers associated with endometriosis. *Fertility and sterility*, 104(5) :1202–1209, 2015.

Gooitzen Zwanenburg, Huub CJ Hoefsloot, Johan A Westerhuis, Jeroen J Jansen, and Age K Smilde. Anova–principal component analysis and anova–simultaneous component analysis : a comparison. *Journal of Chemometrics*, 25(10) :561–567, 2011.