Analyse longitudinale multivariée par modèles mixtes et application à l'épidémie de la malaria

Eric Houngla Adjakossa

THÈSE

présentée pour obtenir

LE GRADE DE DOCTEUR EN SCIENCES
DES UNIVERSITÉS PIERRE ET MARIE CURIE
ET D'ABOMEY-CALAVI

Spécialité : Mathématiques Appliquées

par

Eric Houngla ADJAKOSSA

Analyse longitudinale multivariée par modèles mixtes et application à l'épidémie de la malaria

dirigée par MM. Grégory NUEL et M. Norbert HOUNKONNOU

Rapporteurs :	M. Daniel	**COMMENGES**	Université de Bordeaux
	M. Romain L.	**GLELE KAKAÏ**	Université d'Abomey-Calavi

Soutenue le 03 avril 2017 à l'Université Pierre et Marie Curie devant le jury composé de

M. Daniel	**COMMENGES**	Université de Bordeaux	Président
M. Amaury	**LAMBERT**	Université Pierre et Marie Curie	Examinateur
M. Grégory	**NUEL**	Université Pierre et Marie Curie	Directeur
M. M. Norbert	**HOUNKONNOU**	Université d'Abomey-Calavi	Directeur
M. Romain L.	**GLELE KAKAÏ**	Université d'Abomey-Calavi	Rapporteur
M. David	**COURTIN**	Université Paris Descartes	Examinateur

"The only thing we have to fear is fear itself."

Franklin D. Roosevelt

Je dédie cette thèse à tous ceux qui croient en l'AMOUR

Remerciements

Je tiens, en premier lieu, à exprimer toute ma gratitude à mes deux directeurs de thèse Grégory NUEL et Mahouton Norbert HOUNKONNOU pour avoir accepté de m'accompagner et de me soutenir pendant les trois années qu'ont duré mes travaux de thèse. Lors de nos multiples rendez-vous, j'ai pu profiter de vos nombreuses capacités scientifiques. Un merci tout particulier à Grégory avec qui j'ai passé plus de temps pendant toutes ces années, et qui a su me rendre progressivement indépendant, en tout cas beaucoup plus que je ne l'étais en tout début de thèse. Très sincèrement merci.

Merci à tous les membres du jury pour l'honneur qu'ils me font d'être présents aujourd'hui. Je suis très reconnaissant envers M. Daniel COMMENGES et M. Romain Lucas GLÈLÈ KAKAÏ AGBIDINOUKOUN d'avoir pris le temps de lire de façon critique mon manuscrit de thèse en tant que rapporteurs, et ce malgré leur emploi du temps déjà bien chargé et leurs obligations personnelles.

Mes remerciements vont également à toute l'équipe de l'UMR 216 de l'IRD qui a généré et mis à notre disposition les données d'anticorps anti-palustres sur lesquelles j'ai travaillé. Remerciements particuliers à André GARCIA et à Gilles COTTRELL pour leur soutien depuis mon Master 2. Je vous en suis profondément reconnaissant. Un merci particulier à David COURTIN qui, depuis quelques années, dès que j'en exprime le besoin, trouve toujours le temps de m'expliquer très clairement différentes notions biologiques. Merci aussi d'avoir accepté d'être présent dans le jury, toujours pour m'accompagner jusqu'au bout. J'en suis sincèrement touché.

Toute ma gratitude va également à Amaury LAMBERT pour l'honneur qu'il me fait d'être membre de mon jury de thèse malgré son emploi du temps très chargé. Sincèrement merci.

Je tiens à exprimer ma gratitude aux Professeurs intervenus dans le Master STAFAV (Statistique pour l'Afrique Francophone et Application au Vivant) de Cotonou dont je suis un pur produit. Gratitude particulière aux Professeurs Jean-Marc BARDET, Olivier WINTENBERGER et Simplice DOSSOU-GBÉTÉ. Mes remerciements vont également aux autorités ainsi qu'à l'administration de la CIPMA chaire-UNESCO à Cotonou pour m'avoir beaucoup aidé dans les diverses démarches administratives. Je pense au Président Norbert HOUNKONNOU, au Dr BALOÏTCHA ainsi qu'aux secrétaires.

Je remercie le LPMA (Laboratoire de Probabilités et Modèles Aléatoires), mon laboratoire d'accueil à Paris, où j'ai pu bénéficier de bonnes conditions de travail. Je pense à Florence Deschamps, Josette SAMAN, Serena BENASSU, Altaïr PELISSIER et Khashayar DADRAS. Je remercie également tout le personnel de l'Ecole Doctorale de Sciences Mathématiques de Paris Centre 386, et plus particulièrement Jean-François VENUTI et Corentin LACOMBE.

Pendant ces années de thèse, les doctorants du LPMA ont joué un grand rôle dans ma vie et ont été pour moi une seconde famille. Un grand merci à mon "club" de *footing* composé de: Pierre Antoine, Alice, Sarah, Paul, Carlo et Nicolas le chef d'équipe. Je remercie également mes collègues de bureau avec qui j'ai passé d'agréables moments. Un merci tout particulier à Olga et Omar dont la silencieuse sympathie continue de m'épater. Un grand merci à Nelo qui n'a cessé de m'écouter patiemment, de m'encourager et de m'aider au besoin. Des moments comme le ciné club des thésards sont à jamais gravés en moi. Du fond du cœur, je vous exprime, à chacun et à tous, toute ma gratitude.

Je tiens à remercier tous les enseignants que j'ai connus depuis que j'ai mis pieds à l'école. Quelques uns parmi eux sont définitivement restés gravés dans ma mémoire. Merci de tout cœur à Barnabé HOUNGBEME qui a été mon maître en classe de CM2, et dont la rigueur et la pédagogie m'ont beaucoup apporté et ont semé en moi l'envie de réussir. Profonde gratitude aussi à Gustave GODONOU, mon professeur de français en classe de 3^e, dont l'acharnement à corriger la qualité de notre rédaction est aussi resté gravé en moi. Je me rappelle encore toutes ses stratégies pour nous aider; tellement cela m'a marqué ! Mes sincères remerciements à Cyrille HOUNKPATIN, mon professeur de mathématiques en classe de terminale, dont l'humilité et la pédagogie m'ont définitivement rendu confiant.

Mes remerciements vont aussi à toute la communauté des ECKistes, béninois comme français. Leur amour m'a constamment soutenu et nourri. Merci à Thérèse, Achille, Clovis, Estelle, Chimène, Bernadette, Dovis, Sonia, Marie-Agnès et Doune. Merci aussi aux familles AGLI, HOUNWANOU, KOUTON, HOUNGBO, NOUHOUN, de SOUZA (de Hugor) et ADEOCHUN (de Djahafar). Un merci tout particulier à Sabira EZIN que je ne finirai jamais de remercier pour son AMOUR. Profonde gratitude aussi à tous les Maîtres ECK qui m'ont comblé de leur soutien et de leur AMOUR. Merci de tout cœur.

Mon envie de faire de la recherche scientifique aurait pu être étouffée sans le secours de personnes au grand cœur comme Jacques SEGLA et Luc HESSOU. Je vous suis infiniment reconnaissant pour votre soutien déterminant sans lequel je ne serais peut-être pas en train de finir cette thèse en ce moment. Pour moi, vous êtes des exemples à suivre pour bâtir un Bénin plus grand, voire une Afrique plus grande.

Cette thèse, évidemment, n'aurait été possible sans l'assistance de mes parents. Merci à toi papa pour ta rigueur et ta discipline absolues sans lesquelles je ne serais pas là. Tu savais bien de quoi tu parlais quand tu me disais il y a quelques années: "Eric, tu ne peux plus jamais te laisser aller à la paresse". Merci à toi maman pour ta tendresse et tes nombreux sacrifices pour tes enfants. Puisse le temps nous donner la force de te témoigner concrètement et suffisamment notre gratitude. Merci à tous mes frères et sœurs. Merci à tous mes amis que je ne peux distinctement citer ici.

Enfin, il n'aurait pas été pensable que je vous oublie. Merci à Sandra, Shariy et Huson pour votre amour, votre soutien, votre patience et cette amitié totale. Merci pour ce qu'il y a de meilleur: l'AMOUR ici et maintenant !

Préambule

Contexte et objectifs

La première trace de paludisme est la présence d'ADN de *Plasmodium falciparum*[1] chez des momies datées de 3200 avant Jésus Christ (JC) [118]. Historiquement, l'origine du paludisme chez l'Homme serait due à une transmission venant du gorille, excluant ainsi une transmission due aux chimpanzés ou aux ancêtres de l'Homme lui-même [127]. De nombreuses références au paludisme existent dans les écrits védiques[2] (1600 avant JC) et sur les tablettes d'argile d'Ashurbanipal[3] (669 avant JC) en Mésopotamie [43]. Le terme "malaria" venu de l'italien (ancien) est une combinaison des mots *mala* signifiant "mauvais, insalubre" et "aria" signifiant "air", et est utilisé comme nom de maladie depuis le 17^{e} siècle. D'ailleurs, Hippocrate, vraisemblablement parmi les premiers à décrire certains aspects et symptômes de la maladie (fièvre, aspect saisonnier), a établi une relation avec les eaux stagnantes des marais [130]. Au 19^{e} siècle, le terme "paludisme" apparait venant du mot *palud*, lui-même dérivé du latin *palus* signifiant "marais".

Des travaux du français Charles Laveran en 1880 à ceux du britanique Ronald Ross en 1897, il a été prouvé que le paludisme est une maladie infectieuse à transmission vectorielle faisant intervenir trois acteurs. L'homme, jouant le rôle d'hôte, est infecté par un protozoaire parasite du genre *Plasmodium* qui lui a été transmis par la piqûre d'un moustique (femelle) du genre *Anopheles* [29]. La maladie se caractérise par des épisodes fébriles aigus et peut être mortelle. Des symptômes tels que maux de tête, fièvre, vomissements et frissons apparaissent au bout de sept jours ou plus après la piqûre infectante du moustique. L'intensité de la transmission du paludisme dépend de facteurs liés au parasite, au vecteur (moustique), à l'hôte humain et à l'environnement [36, 136, 176].

Les efforts mis en œuvre pour la maîtrise du développement du paludisme ont réduit sa distribution géographique mondiale de 53% en 1900 à 27% en 2002 [89]. En effet, le paludisme a commencé à régresser en Europe dès le 18^{e} siècle, puis en Amérique du Nord au milieu du 19^{e} siècle. Il a disparu en Angleterre, en France ainsi que dans de nombreux pays européens vers la fin du 19^{e} siècle. Il se maintenait par contre en Italie, en Grèce, en Allemagne et sévissait surtout dans les régions tropicales et subtropicales [130].

Malgré les efforts acharnés fournis depuis des décennies, le paludisme est resté un problème d'ampleur mondiale qui se concentre sur la santé et le potentiel économique des communautés les

[1] *Plasmodium falciparum* est une des espèces de *Plasmodium*, parasites qui causent le paludisme chez l'être humain. Il est le plus dangereux de ces parasites causant le paludisme car il entraîne le taux de mortalité le plus élevé.

[2] Le Veda, du sanskrit "vision ou connaissance", est un ensemble de textes qui, selon la tradition, ont été révélés par l'audition aux sages indiens nommés *Richi* [134].

[3] Tablettes réunies en faisceau par des bandelettes et portant des inscriptions à caractère prophétique [128].

plus pauvres de la planète. C'est pour cela qu'en 2000, la cible 6C des Objectifs du Millénaire pour le Développement (OMD) liée au paludisme appelait à avoir maîtrisé cette maladie en 2015 et commencé à inverser la tendance (de 2000). Malheureusement, en 2012, selon les estimations de l'Organisation Mondiale de la Santé (OMS), 207 millions de cas sont survenus et la maladie a tué quelques 627.000 personnes - pour la plupart, des enfants de moins de cinq ans vivant en Afrique. En moyenne, le paludisme tue un enfant chaque minute [138]. Le *Rapport sur le paludisme dans le monde 2015* résume les progrès accomplis en matière de contrôle et d'élimination de la maladie à la date-butoir de 2015 et mentionne que "malgré des progrès remarquables, il reste beaucoup à faire". La *Stratégie technique mondiale de lutte contre le paludisme 2016-2030*, approuvée par l'Assemblée mondiale de la Santé en mai 2015, définit des objectifs ambitieux pour 2030, notamment réduire d'au moins 90% l'incidence du paludisme et la mortalité associée. La version complète du *Rapport sur le paludisme dans le monde 2015* est disponible en anglais à l'adresse `http://www.who.int/malaria/publications/world-malaria-report-2015/report/en/`.

En Afrique où surviennent le plus de cas, l'objectif est de contrôler le paludisme par des approches intégrées combinant le diagnostic, un traitement efficace et précoce, la lutte anti-vectorielle et la chimioprophylaxie, notamment par le Traitement Préventif Intermittent (TPI) des enfants et des femmes enceintes [43]. Un vaccin antipaludique serait bien entendu un apport essentiel dans cette lutte [157].

Les deux populations majoritairement à risque restent les enfants de moins de cinq ans et les femmes enceintes. C'est pour cette raison que l'Institut de Recherche pour le Développement (IRD) et en particulier l'unité UMR 216 "Mère et enfant face aux infections tropicales" s'intéressent à ces populations à travers plusieurs suivis de cohortes en zone subsaharienne (Bénin, Sénégal) dont l'objectif est de comprendre les déterminants immunologiques, parasitologiques, génétiques et environnementaux de la survenue des infections palustres dans le but d'élaborer à terme un vaccin contre le paludisme. Les enquêtes épidémiologiques menées sur ces cohortes génèrent d'importants jeux de données comportant plusieurs dizaines de variables épidémiologiques, biologiques, entomologiques, environnementaux, etc. Ces données sont le plus souvent répétées dans le temps (longitudinales) et présentent des structures de corrélation particulières dans le temps et/ou dans l'espace qu'il faut prendre en compte dans les analyses au travers d'ajustement de modèles statistiques adaptés.

Spécifiquement, les données dont l'analyse est à l'origine de cette thèse proviennent des travaux d'un projet de l'UMR 216 dont l'objectif principal était d'étudier la construction de la réponse immunitaire humorale dirigée contre *P. falciparum* durant les dix huit premiers mois de vie chez 600 enfants vivant dans une zone rurale d'endémie palustre à Tori-Bossito dans le Sud-Ouest du Bénin. Ces données incluent sept antigènes des stades sanguins asexués de *P. falciparum*, retenus en leur qualité de candidats vaccins: MSP1, MSP2 (3D7 et FC27), MSP3, AMA1, GLURP (R0 et R2). Ces antigènes interviennent de façon déterminante dans le développement d'une immunité protectrice dans la mesure où ils induisent la production d'anticorps associés à la protection clinique, notamment les immunoglobulines (Ig)G de type cytophile IgG1 et IgG3 [43].

Dans les modèles statistiques utilisés classiquement pour analyser ces données en prenant en compte leur structure, chaque profile d'anticorps était en général expliqué par quelques uns ou tous les autres profiles d'anticorps et/ou en présence d'autres facteurs (environnement, anémie, etc.).

Plusieurs modèles indépendants étaient alors ajustés à ces données en vue d'apporter des réponses adéquates aux différentes questions. Or tous ces antigènes inter-agissent au même moment dans l'organisme de l'enfant et il semble plus approprié de procéder à l'analyse de la distribution jointe de ces profiles d'anticorps, malgré l'utilité incontestable des analyses séparées (facilité d'usage, résultats aisément interprétables).

Les objectifs de cette thèse étaient donc d'apporter un support en statistique ainsi qu'une profonde connaissance de la littérature dans le but d'améliorer la compréhension des méthodes d'analyses de données longitudinales multidimensionnelles. Ils étaient également de contribuer au passage aux développements méthodologiques en vue d'une analyse plus efficace de données multidimensionnelles multi-niveaux comme celles issues des suivis épidémiologiques de cohortes relatives à la lutte contre le paludisme.

Cette thèse a été effectuée dans un contexte de co-tutelle internationale entre le Bénin et la France, précisément les Universités Pierre et Marie Curie (UPMC) et d'Abomey-Calavi (UAC). Les séjours en France aussi bien qu'au Bénin ont été conjointement financés par le Service de Coopération et d'Action Culturelle (SCAC) de l'ambassade de France au Bénin et l'Institut de Recherche pour le Développement (IRD). L'ensemble de ce travail a été réalisé sous l'encadrement scientifique des Professeurs Grégory NUEL (LPMA/UPMC) et M. Norbert HOUNKONNOU (CIPMA/UAC).

Plan de la thèse

Cette thèse porte sur l'estimation et la sélection dans le modèle linéaire à effets mixtes et s'organise autour de cinq chapitres indépendants. Chaque chapitre s'ouvre sur un énoncé introductif des différents points qui y sont abordés et se conclut sur une synthèse.

Le chapitre 1 est une introduction générale qui présente le développement historique des méthodes d'analyse de données longitudinales et définit le modèle statistique linéaire à effets mixtes à partir d'un exemple simple. Il présente ensuite les principales méthodes d'estimation des paramètres du modèle, précise les différentes contributions scientifiques faites dans le cadre de cette thèse suivies des perspectives envisagées.

Le chapitre 2 présente un article de recherche publié dans PLOS ONE portant sur l'estimation des paramètres du modèle linéaire multidimensionnel à effets mixtes, utilisant l'algorithme EM. Les estimateurs EM présentés dans ce chapitre ont des expressions plus générales que celles rencontrées dans la littérature traitant des données longitudinales multi-variées. Ces estimateurs tels que présentés intègrent le cadre plus général d'analyse de données multidimensionnelles multi-niveaux. Dans ce chapitre, avons-nous également introduit un test de corrélation bi-varié permettant de tester la significativité de l'ensemble des corrélations entre les effets aléatoires de deux dimensions du modèle. Ce qui permettra de tester si deux composantes du vecteur de réponses (les variables dépendantes) sont à analyser séparément ou non. Ce test de corrélation bi-varié permettra également de construire un modèle multidimensionnel plus parcimonieux en terme de nombre de composantes de variance des effets aléatoires, à travers une procédure de sélection pas-à-pas ascendante.

Le chapitre 3 présente un autre article de recherche que nous avons soumis à une revue scientifique. Le travail exposé ici est une généralisation de la méthode d'estimation utilisée dans le package lme4 du logiciel R, proposé par Douglas Bates et ses collaborateurs. Au travers d'une étude de simulation, nous avons montré que la procédure proposée est beaucoup moins sensible au point de départ et plus rapide, comparée à l'estimation par l'algorithme EM. Aussi fournit-elle des estimations consistantes sur nos données simulées.

Le chapitre 4 présente une procédure de sélection d'effets fixes dans le modèle linéaire à effets mixtes. Cette procédure de sélection que nous avons appelée l'*adaptive ridge* itérative utilise une pénalité de type L_2 sur le vecteur des effets fixes en présence d'une matrice de poids itérativement mise à jour, et où nous utilisons la log-vraisemblance profilée des composantes de variance des effets aléatoires et des effets fixes, conditionnellement aux observations. Cette approche permet d'approcher les performances en matière de sélection d'une pénalité de type L_0.

Dans le chapitre 5, nous présentons davantage le cadre d'application qu'est le notre dans cette thèse avec les données d'anticorps anti-paludisme. Ici, les différentes illustrations de nos méthodes sur ces données sont également exposées.

Table des matières

CHAPITRE 1

INTRODUCTION

ll est courant que les observations ou les mesures faites dans plusieurs domaines d'application des mathématiques (biologie, sociologie, statistique, économie, etc.) fassent émerger une certaine structure de corrélation dont la compréhension et la prise en compte sont nécessaires pour l'analyse de l'ensemble des données engendrées. Ainsi parle-t-on de *données corrélées*. Les termes génériques *données corrélées* embracent une multitude de structures de données telles que : les observations multivariées, les données groupées, les mesures répétées, les données de type longitudinal et les données spatialement corrélées [193]. Dans cette thèse, nous nous focalisons sur l'analyse statistique de données de type *longitudinal multidimensionnel* qu'on appelle parfois données de type *longitudinal multivarié*. Ces données proviennent des études longitudinales dont la caractéristique essentielle est qu'un ou plusieurs attributs descripteurs des sujets (ou unités statistiques) à l'étude sont mesurés de façon répétée au cours du temps. Par exemple, on pourrait s'intéresser à l'évolution du poids et de la taille d'enfants nés dans un hôpital un jour donné, en les mesurant tous les trois mois pendant cinq ans. Ceci constitue un exemple de données longitudinales, et si les mesures (poids ou tailles) provenant d'un même enfant sont généralement corrélées, celles provenant de deux différents enfants peuvent ne pas l'être. Les données de type longitudinal peuvent être recueillies de façon prospective, les unités statistiques étant suivies progressivement dans le temps, ou de façon rétrospective, en extrayant plusieurs informations des archives relatives à chaque unité statistique. Les études longitudinales, contrairement aux études transversales où chaque unité statistique est observée (ou mesurée) une seule fois, ont l'avantage d'aider à distinguer les changements parmi les unités statistiques au fil du temps (effets du vieillissement par exemple) [45]. Dans les études longitudinales, bien que certaines questions puissent être résolues en analysant séparément différentes variables d'intérêt, d'autres quant à elles ne peuvent l'être qu'en analysant conjointement ces variables [194]. Ce qui revient à considérer un vecteur dont les composantes sont les différentes variables d'intérêt de type longitudinal : d'où la terminologie *données longitudinales multidimensionnelles* ou *multivariées*. Nous nous intéressons ici au cas où ces variables sont des nombres réels et aux méthodes statistiques permettant de les analyser. Parmi les modèles statistiques permettant d'analyser efficacement les données longitudinales multivariées, nous nous intéressons plus précisément au modèle linéaire à effets mixtes. La raison de cet intérêt est expliquée à travers le développement historique des modèles linéaires pour l'analyse de données longitudinales.

1.1 Développement historique des modèles linéaires pour l'analyse de données longitudinales

Un des plus anciens exemples d'analyse de données longitudinales fourni par la littérature nous vient des travaux de l'astrologue George Biddel Airy [2, Partie IV] qui portaient sur la modélisation de plusieurs niveaux d'erreurs contenues dans les mesures successives d'une même quantité supposée invariante, comme par exemple le diamètre angulaire apparent d'une planète ou encore la distance angulaire entre deux étoiles. Plusieurs mesures étaient faites quotidiennement et ce pendant plusieurs jours. Airy constata que même après avoir appliqué tous les types de correction d'erreur de mesure connus, il apparaissait parfois que les résultats obtenus différaient grandement d'un jour à l'autre. C'est pourquoi il émit l'hypothèse de l'existence de deux types d'erreur suivant le modèle

$$y_{ij} = \mu + \gamma_i + \varepsilon_{ij}, \tag{1.1}$$

où y_{ij} est la j^{e} mesure du i^{e} jour et μ la moyenne générale ou la "vraie" valeur de l'objet mesuré. La quantité γ_i représente, quant à elle, l'erreur (constante) commise sur toutes les mesures effectuées le i^{e} jour, et ε_{ij} l'erreur due aux conditions particulières d'observation (atmosphère, instruments de mesure, etc.) dans lesquelles la j^{e} mesure du i^{e} jour a été effectuée. Plus tard, Henry Scheffé [162], proposant une alternative pour le modèle d'analyse de la variance tel qu'introduit par R. A. Fisher [60], dans le but d'analyser des données issues de dispositifs en blocs randomisés, introduisit les termes d'effet fixe et d'effet aléatoire. De façon plus explicite, selon Scheffé, le modèle d'erreurs d'Airy (équation (1.1)) s'écrit :

$$y_{ij} = \mu + \gamma_i + \varepsilon_{ij}, \text{ avec } \gamma_i \sim \mathcal{N}(0, \sigma_\gamma^2), \quad \varepsilon_{ij} \sim \mathcal{N}(0, \sigma_\varepsilon^2) \text{ et } \gamma_i \perp \varepsilon_{ij}, \tag{1.2}$$

où γ_i s'interprète désormais comme étant *l'effet aléatoire* du i^{e} jour sur la valeur mesurée y_{ij}, et ε_{ij} le terme d'erreur (ou résiduel) du modèle. Le terme γ_i est appelé *effet* parce qu'il représente une déviation par rapport à la moyenne générale μ, la moyenne des mesures effectuées le i^{e} jour étant $\mu + \gamma_i$. Les jours pendant lesquels les mesures ont été effectuées peuvent être considérés comme un échantillon aléatoire de l'ensemble J des jours pour lesquels on voudrait inférer des informations. Par exemple, J pourrait être l'ensemble des jours depuis la naissance du système solaire jusqu'à son extinction. En supposant que J est infini et que l'estimation et l'inférence sont faites sur J et non uniquement sur les jours où des mesures ont été effectuées, l'effet du jour i est aléatoire (par exemple, $\gamma_i \sim \mathcal{N}(0, \sigma_\gamma^2)$) et non fixe. Henry Scheffé introduisit ainsi les concepts de modèle à effets fixes, de modèle à effets aléatoires et de modèle à effets mixtes [162] pour l'analyse de données longitudinales et plus généralement, l'analyse de données à structure hiérarchique, induisant une structure de correlation plus ou moins complexe à prendre en compte. Le modèle est à effets fixes si le seul effet aléatoire est le terme d'erreur. Il est à effets aléatoires si en dehors de la moyenne générale (qui n'est pas aléatoire) et du terme résiduel, tous les autres termes sont des effets aléatoires. Le modèle est à effets mixtes s'il incorpore à la fois des effets fixes et des effets aléatoires. Le modèle à effets mixtes a donc plusieurs sources de variabilité (effets aléatoires et terme résiduel) et est parfois appelé *modèle hiérarchique* [113], *modèle multi-*

niveaux [76] ou encore *modèle à composantes de variance* [88, 162, 168] (car la variance de y_{ij} se décompose : $\sigma_y^2 = \sigma_\gamma^2 + \sigma_\varepsilon^2$).

Dans plusieurs domaines (économie, éducation, sociologie, biologie, etc.), le recours aux données longitudinales a pris une ampleur considérable pour plusieurs raisons dont la principale est la possibilité d'étudier, à la fois, l'aspect dynamique et transversal d'un problème d'intérêt au sein d'une population [64]. En économie par exemple, on utilise l'appellation *données de panel* pour désigner des données de type longitudinal. Ici, un *panel* désigne un groupe d'individus suivis de façon répétée au cours du temps. Le terme *étude de panel* a été utilisé pour la première fois par Lazarsfeld et Fiske [107] qui étudiaient l'effet de la publicité radiodiffusée sur la vente d'un produit commercialisé. Ils entreprirent d'examiner si les personnes achetant leur produit étaient plus susceptibles d'entendre l'annonce. Pour ce faire, ils interrogea à plusieurs reprises un groupe (le panel) de personnes. Considérées comme des séries chronologiques (temporelles) transversales groupées, les données de panel étaient initialement analysées (en économie) par une stratégie en deux étapes :

1) estimer des paramètres transversales en ajustant un modèle (standard) de régression aux données transversales à temps fixé ;

2) utiliser les méthodes propres à l'analyse des séries chronologiques pour modéliser les paramètres transversales estimés.

Bien qu'utile dans certains contextes, cette approche est complètement inappropriées dans d'autres. Par exemple, Edward W. Frees [64] a régressé le taux de divorce (en 1965 et 1975) des cinquante Etats des Etats Unis d'Amérique sur l'aide financière apportées aux familles ayant des enfants à charge. Pour cet exemple, chaque Etat représente une unité statistique observée deux fois (en 1965 et en 1975) au cours du temps, la variable d'intérêt étant le taux de divorce, et l'aide financière, une variable devant aider à mieux comprendre le taux de divorce. Les pentes transversales estimées sont de -0.95% pour 1965 et de -1.0% pour 1975. L'extrapolation de ces pentes négatives donne un résultat complètement différent de celui de l'estimateur de la dynamique du taux de divorce qui est de 2.9%. Afin de contourner ce problème, Theil et Goldberger [184] proposent plutôt une analyse directe (en une seule étape) et fournissent ainsi, en économie, les premières discussions concernant les avantages de l'analyse simultanée de l'aspect transversal et de l'aspect temporel (série chronologique) des données de panel.

Les outils dédiés au traitement de séries chronologiques unidimensionnelles sont bien adaptés à l'analyse de la dynamique des relations entre les observations d'un même sujet. Toutefois, ces méthodes ne sont pas appropriées pour prendre en compte les relations entre différents sujets. Par contre, les méthodes d'analyse dédiées aux séries chronologiques multidimensionnelles peuvent prendre en compte les relations inter-sujets, mais sont utiles dans le cas de données de panel pour un nombre limité de sujets. Aussi, le nombre d'observations par sujet doit-il être au moins de l'ordre de la trentaine afin d'espérer une inférence relativement fiable [64]. Ce qui est vu comme une importante limite du recours aux méthodes de traitement des séries chronologiques pour l'analyse des données de panel. Ainsi, l'approche consistant à considérer les données de panel comme étant des séries chronologiques répétées a très vite montré ses limites et n'a pas connu un grand succès [64].

L'utilisation des données de panel (en économie) est généralement destinée à l'analyse de la dynamique comportementale des sujets au fil du temps. Celle-ci revient à l'analyse de la nature des termes latents de perturbation qui gouvernent les relations (économiques) à l'étude, et qui sont communs aux sujets. Ainsi, le problème central d'intérêt est celui de l'hétérogénéité latente relative aux sujets [133]. L'hétérogénéité entre sujets peut s'expliquer par la dynamique du profil des observations ou par le fait que les observations partagent un même phénomène sous-jacent, non observable, et qui induit une corrélation significative entre les données d'un même sujet. Ainsi, l'hétérogénéité s'interprète ici par le fait que les observations venant du même sujet ont plus tendance à se "ressembler" que celles venant de différents sujets. Finalement, cette hétérogénéité se modélise de deux manières [64]. La première consiste à distinguer les paramètres qui sont communs à tous les sujets de ceux qui sont spécifiques à chaque sujet. Cette approche est connue en économie sous l'appellation de modèle à effets fixes. La seconde méthode ajoute une hypothèse distributionnelle aux paramètres spécifiques aux sujets. Ce qui revient au modèle à effets mixtes ou au modèle à effets aléatoires décrit plus haut.

Bien qu'introduit depuis le 19^{e} siècle et utilisé dans plusieurs domaines, ce n'est que dans les années 1980 que l'utilité du modèle linéaire à effets mixtes a été fortement mise en lumière par les travaux de Laird et Ware [105] qui en ont proposé une formulation plus générale et plus flexible pouvant naturellement prendre en compte plusieurs particularités des données dont principalement leur aspect déséquilibré[(1)]. La formulation proposée par Laird et Ware [105] est la suivante :

$$y_i = X_i\beta + Z_i\gamma_i + \varepsilon_i, \tag{1.3}$$

avec

$$\gamma_i \sim \mathcal{N}(0, \Gamma), \quad \varepsilon_i \sim \mathcal{N}(0, \Sigma_i) \quad \text{et} \quad \gamma_i \perp \varepsilon_i. \tag{1.4}$$

Dans les équations (1.3) et (1.4), y_i est le vecteur des observations du i^{e} sujet, β appartient à $\mathbb{R}^p$ et est le vecteur des effets fixes, X_i est une matrice de covariables de taille $n_i \times p$, où n_i est le nombre d'observations du i^{e} sujet. γ_i est le vecteur, de taille q', des effets aléatoires relatifs au i^{e} sujet et Z_i est une matrice de design basée sur des covariables et de taille $n_i \times q'$. ε_i est le terme résiduel du modèle. Γ et Σ_i sont des matrices définies positives de taille $q' \times q'$ et $n_i \times n_i$, respectivement. Σ_i ne dépend de i que par sa taille. Les sujets étant indépendants les uns des autres, γ_i est indépendant de $\gamma_{i'}$ pour tout $i \neq i'$. L'un des avantages de cette formulation est qu'elle ne fait pas de restrictions sur les matrices X et Z, et permet une estimation efficace des paramètres du modèle par les méthodes basées sur le maximum de vraisemblance. Ainsi, Laird et Ware [105] ont montré comment l'algorithme EM (Expectation-Maximization) de Dempster, Laird et Rubin [44] peut être utilisé pour estimer les paramètres du modèle dans le cas de données longitudinales. Ce qui représentait, à cette époque, une vraie avancée puisque l'ajustement du modèle, dans un cadre général, à un jeu de données longitudinales posait toujours problème [61]. Peu de temps après, Jennrich et Schluchter [96] proposèrent de recourir à plusieurs méthodes alternatives d'estimation des paramètres du modèle dont principalement l'algorithme de Newton-Raphson et celui des scores de Fisher. Depuis les travaux de Jennrich et Schluchter [96], l'estimation et l'inférence dans le modèle à effets mixtes sont généralement basées sur le maximum de vraisemblance (Maximum

[(1)]Est dit déséquilibré, un jeu de données longitudinales dont les sujets n'ont pas le même nombre d'observations et/ou les observations répétées des sujets n'ont pas été prises à la même date.

Likelihood - ML - en anglais) et le maximum de vraisemblance restreinte (Restricted Maximum Likelihood - REML - en anglais). Dans la suite, nous nous focaliserons sur l'estimation des paramètres du modèle linéaire à effets mixtes en dimension k, avec $k \geq 1$. Dans la section suivante, nous allons nous consacrer à la revue des méthodes d'estimation des paramètres du modèle en dimension 1.

1.2 Modèle linéaire à effets mixtes : différentes méthodes d'estimation des paramètres

Nous allons définir plus explicitement le modèle linéaire à effets mixtes avant de présenter les différentes méthodes d'estimation de ses paramètres.

1.2.1 Définition du modèle

Nous faisons ici l'option d'introduire la définition du modèle linéaire à effets mixtes à travers un exemple simple. Supposons que l'on cherche à comprendre l'effet de la qualité nutritionnelle des enfants d'un pays (d'une population) P, pendant leur trois premières années de vie, sur leur croissance.

TABLE 1.1 : Exemple de données longitudinales

					Variables d'intérêt	
sujet	age	sexe	MScoreN	ScoreN	poids	taille
1	0	F	63.76	38.16	3.14	47.82
2	4	M	100.88	41.46	4.87	64.02
1	6	F	60.98	41.37	8.43	73.21
3	0	M	93.24	48.76	2.82	44.93
2	7	M	101.95	44.79	8.03	89.54
2	10	M	99.24	48.17	10.08	92.14
1	16	F	85.02	44.79	13.96	86.12
3	9	M	88.38	47.91	8.47	86.42

On recrute aléatoirement n enfants à la naissance que l'on suit jusqu'à l'âge de 36 mois. Tous les mois, on mesure leur poids (en kg), leur taille (en cm), leur score de nutrition (ScoreN) aussi bien que le score de nutrition de leur mère (MScoreN). Pour faire plus simple, intéressons-nous à trois des enfants dont une partie des données est présentée dans le tableau 1.1. L'enfant (le sujet) 1 est une fille qui pèse 3.14 kg et mesure 47.82 cm à la naissance, et a été vue le sixième et le seizième mois, d'après les informations contenues dans le tableau 1.1. Le poids et la taille sont *a priori* corrélés, et on pourrait commencer par expliquer le poids de l'enfant (le sujet) i au j^{e} mois avec le modèle linéaire ci-après :

$$\text{poids}_{ij} = \beta_0 + \beta_1 \text{taille}_{ij} + \varepsilon_{ij}; \quad i = 1, \ldots, n; \quad j = 1, \ldots, n_i, \tag{1.5}$$

où β_0 est le poids moyen des enfants de la population P, β_1 mesurant l'effet de la taille sur le poids et ε_{ij} le terme d'erreur. Les termes poids_{ij} et taille_{ij} désignent respectivement la j^{e} mesure du poids et de la taille du sujet i. On pourrait faire l'hypothèse que $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$. Ici, $n = 3$ sujets,

$n_1 = n_2 = 3$ et $n_3 = 2$. Le nombre total d'observations est $N = \sum_{i=1}^{n} n_i = 8$. Puisque les enfants sont nés et grandissent chacun dans des conditions spécifiques, l'idée d'avoir un β_0 et un effet global β_1 de la taille sur le poids peut être complètement erronée, et il pourrait être plus pertinent d'introduire dans le modèle, un effet spécifique par sujet comme suit :

$$\text{poids}_{ij} = \beta_0 + \gamma_i^0 + (\beta_1 + \gamma_i^t)\text{taille}_{ij} + \varepsilon_{ij}; \quad i = 1, \ldots, n; \quad j = 1, \ldots, n_i. \tag{1.6}$$

Par ailleurs, puisque l'on s'intéresse à tous les enfants de la population P, et non aux seuls n enfants recrutés dans l'échantillon, on fait l'hypothèse :

$$\begin{pmatrix} \gamma_i^0 \\ \gamma_i^t \end{pmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{pmatrix} \eta_1^2 & \rho\eta_1\eta_2 \\ \rho\eta_1\eta_2 & \eta_2^2 \end{pmatrix}\right), \quad i = 1, \ldots, n. \tag{1.7}$$

Le terme γ_i^0 (intercepte aléatoire) mesure ainsi l'écart spécifique induit par le sujet i par rapport au poids moyen β_0 de la population P. β_1 mesure l'effet moyen global de la taille sur le poids et $\beta_1 + \gamma_i^t$ mesure l'effet spécifique de la taille du sujet i sur son poids ; γ_i^t étant la pente aléatoire induit par le sujet i, suivant la variable "taille". En prenant en compte, dans le modèle, le sexe et la qualité nutritionnelle de l'enfant, on a par exemple pour $i = 1, \ldots, n$; et $j = 1, \ldots, n_i$,

$$\begin{aligned} \text{poids}_{ij} \quad = \quad & \beta_0 + \gamma_i^0 + (\beta_1 + \gamma_i^t)\text{taille}_{ij} + \gamma_i^s \text{ScoreN}_{ij} + \\ & \beta_2 \mathbb{1}_{\text{sexe=M}}(\text{sexe}_i) + \beta_3 \mathbb{1}_{\text{sexe=F}}(\text{sexe}_i) + \varepsilon_{ij}, \end{aligned} \tag{1.8}$$

avec

$$\gamma_i = \begin{pmatrix} \gamma_i^0 \\ \gamma_i^t \\ \gamma_i^s \end{pmatrix} \sim \mathcal{N}(\mathbf{0}, \Gamma) \quad \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2). \tag{1.9}$$

Ce qui s'écrit plus simplement à l'échelle individuelle en regroupant parties fixes et parties aléatoires comme suit :

$$\begin{aligned} \text{poids}_i \quad = \quad & \beta_0 \mathbb{1}_{n_i} + \beta_1 \text{taille}_i + \beta_2 \mathbb{1}_{\text{sexe=M}}(\text{sexe}_i)\mathbb{1}_{n_i} + \beta_3 \mathbb{1}_{\text{sexe=F}}(\text{sexe}_i)\mathbb{1}_{n_i} \\ & + \gamma_i^0 \mathbb{1}_{n_i} + \gamma_i^t \text{taille}_i + \gamma_i^s \text{ScoreN}_i + \varepsilon_i, \quad i = 1, \ldots, n, \end{aligned}$$

où, $\mathbb{1}_{n_i} = (1, \ldots, 1)^\top \in \mathbb{R}^{n_i}$, $\varepsilon_i = (\varepsilon_{i1}, \ldots, \varepsilon_{in_i})^\top$, et taille_i et ScoreN_i sont des vecteurs qui contiennent les mesures successives du poids et du score de nutrition du sujet i, respectivement. Pour des questions d'identifiabilité du modèle, on met des contraintes sur les paramètres β_2 et β_3 qui mesurent l'effet des modalités de la variable "sexe" (variable non numérique) sur le poids, exactement comme dans un modèle linéaire simple par exemple. Si nous faisons l'option d'une

contrainte de type "cellule de référence" sur les modalités du sexe, où la modalité de référence est "sexe = F" (c'est à dire $\beta_3 = 0$), matriciellement, le modèle s'écrit :

$$\text{poids} = X\beta + Z\gamma + \varepsilon; \quad \gamma \sim \mathcal{N}(\mathbf{0}, \Gamma), \quad \varepsilon \sim \mathcal{N}(\mathbf{0}, \Sigma). \tag{1.10}$$

Ainsi, β, Γ et Σ sont les paramètres (à estimer) du modèle. Ici, $\beta = (\beta_0, \beta_1, \beta_2)^\top$, puisque nous avons fait l'option d'une contrainte d'identifiabilité de type "cellule de référence" sur les modalités du sexe, où la modalité de référence est "sexe = F" dans le modèle exprimé par l'équation (1.10). Sous cette contrainte, β_2 s'interprète plutôt comme l'effet différentiel du sexe sur le poids, toutes choses étant égales par ailleurs. En considérant le jeu de données fourni par le tableau 1.1, $\gamma = (\gamma_1, \gamma_2, \gamma_3)^\top$ avec $\gamma_i, i = 1, \ldots, 3$ définis dans l'équation (1.9). Aussi a-t-on :

$$X = \begin{pmatrix} 1 & 47.82 & 0 \\ 1 & 64.02 & 1 \\ 1 & 73.21 & 0 \\ 1 & 44.93 & 1 \\ 1 & 89.54 & 1 \\ 1 & 92.14 & 1 \\ 1 & 86.12 & 0 \\ 1 & 86.42 & 1 \end{pmatrix},$$

$$Z = \begin{pmatrix} 1 & 47.82 & 38.16 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 64.02 & 41.46 & 0 & 0 & 0 \\ 1 & 73.21 & 41.37 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 44.93 & 48.76 \\ 0 & 0 & 0 & 1 & 89.54 & 44.79 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 92.14 & 48.17 & 0 & 0 & 0 \\ 1 & 86.12 & 44.79 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 86.42 & 47.91 \end{pmatrix}.$$

Si le tableau 1.1 était ordonné par sujet, on aurait pour la matrice Z :

$$Z = \begin{pmatrix} 1 & 47.82 & 38.16 & . & . & . & . & . & . \\ 1 & 73.21 & 41.37 & . & . & . & . & . & . \\ 1 & 86.12 & 44.79 & . & . & . & . & . & . \\ . & . & . & 1 & 64.02 & 41.46 & . & . & . \\ . & . & . & 1 & 89.54 & 44.79 & . & . & . \\ . & . & . & 1 & 92.14 & 48.17 & . & . & . \\ . & . & . & . & . & . & 1 & 44.93 & 48.76 \\ . & . & . & . & . & . & 1 & 86.42 & 47.91 \end{pmatrix},$$

où les zéros sont représentés par des points. Le modèle linéaire à effets mixtes peut alors se définir comme suit.

Définition 1. *Soit † une variable aléatoire numérique, expression d'un attribut descripteur des unités statistiques d'une population, et y un vecteur de $\mathbb{R}^N$ dont les composantes sont des réalisations*

de † sur un échantillon de taille $n \leqslant N$. On suppose que y est une réalisation d'un vecteur aléatoire $\mathcal{Y}$. Soit γ un vecteur aléatoire de $\mathbb{R}^q$ de réalisation non observable. X et Z deux matrices de dimensions respectives $N \times p$ et $N \times q$. On appelle modèle linéaire à effets mixtes, le modèle statistique défini par :

$$\gamma \sim \mathcal{P}_\theta, \quad \theta \in \Theta, \tag{1.11}$$

$$\mathcal{Y}|\gamma \sim \mathcal{Q}_{\theta'}, \quad \theta' \in \Theta', \tag{1.12}$$

$$\mathbb{E}\left[\mathcal{Y}|\gamma\right] = X\beta + Z\gamma, \tag{1.13}$$

où $\mathcal{P}_\theta$ et $\mathcal{Q}_{\theta'}$ sont des lois de probabilité absolument continues par rapport à la mesure de Lebesgue. $\beta \in \mathbb{R}^p$ est appelé le vecteur des effets fixes et γ le vecteur des effets aléatoires.

La quasi-totalité des travaux effectués sur le modèle à effets mixtes (linéaire ou non) choisit la loi gaussienne pour $\mathcal{P}_\theta$ [102, 115, 153, 195] pour des raisons de simplicité de calculs, et dans le cas de la version linéaire du modèle, $\mathcal{Q}_{\theta'}$ est aussi choisie comme étant une loi gaussienne [8, 13, 191, 202]. Dans cette thèse, nous nous conformons à ce choix, et notre contribution, relativement à la version multidimensionnelle du modèle, reste dans ce cadre précis. Dans ce cadre, et avec les notations de la définition 1, le modèle s'écrit :

$$\mathcal{Y} = X\beta + Z\gamma + \varepsilon \tag{1.14}$$

$$\gamma \sim \mathcal{N}(\mathbf{0}, \Gamma), \quad \varepsilon \sim \mathcal{N}(\mathbf{0}, \Sigma), \quad \gamma \perp \varepsilon. \tag{1.15}$$

Le terme résiduel du modèle ε est supposé indépendant du vecteur des effets aléatoires pour des raisons techniques. β, Γ et Σ sont les paramètres à estimer du modèle.

1.2.2 Estimation des paramètres du modèle linéaire à effets mixtes

Les méthodes d'estimation des paramètres du modèle peuvent être scindées en deux groupes. Le premier regroupe des "méthodes purement statistiques" et le second les méthodes standards d'optimisation connues en analyse numérique sous l'appellation de méthodes du gradient et incluent : l'algorithme de Newton-Raphson, celui de Marquardt, l'algorithme des scores de Fisher, les méthodes du Quasi-Newton, la descente du gradient, la "coordinate descente" du gradient, etc. Nous nous intéressons ici aux méthodes les plus utilisées du premier groupe que nous exposerons. Les méthodes d'estimation des paramètres du modèle linéaire à effets mixtes que nous exposons ici sont : la méthode dite d'estimation empirique de Bayes, l'algorithme EM (Expectation-Maximization), et celle implémentée par Douglas Bates et ses co-auteurs sous le package lme4 [14] du logiciel R [146]. Cette dernière méthode peut être vue comme une amélioration de la méthode d'estimation empirique de Bayes. La plupart de ces méthodes sont basées sur la maximisation de la vraisemblance des paramètres du modèle. Pour le modèle à effets mixtes, l'estimation des paramètres de variance est restée pendant longtemps un véritable problème. Plusieurs stratégies d'estimation, non basées sur la maximisation de la vraisemblance, proposées dans la littérature ont montré leurs limites [88]. Par exemple, pour un modèle d'analyse de la variance, en présence de données équilibrées, l'estimation des paramètres de variance était faite en posant une égalité

entre les carrés moyens du tableau d'analyse de la variance et leur espérances. Henderson [92] a également développé une méthode analogue, largement utilisée, pour des données déséquilibrées. Progressivement, l'intérêt porté aux méthodes s'appuyant sur la maximisation de la vraisemblance pour l'estimation des paramètres de variance a pris de l'ampleur, car elles présentent plusieurs caractéristiques intéressantes : consistence, normalité asymptotique et efficience (dans le sens décrit par Miller [126]). Par ailleurs, la méthode du maximum de vraisemblance est toujours bien définie, même en présence de contraintes sur certains paramètres de variance. Aussi obtient-on aisément les estimateurs de maximum de vraisemblance et la matrice d'information de Fisher pour n'importe quelle paramétrisation du modèle [88]. Malgré ses intéressantes propriétés, l'approche du maximum de vraisemblance pour l'estimation des composantes de variance est restée longtemps sans être utilisée en pratique puisque, à l'exception des cas simples, le calcul des estimateurs du maximum de vraisemblance exige le recours aux méthodes numériques qui résolvent les problèmes d'optimisation non linéaire à contraintes. Il a fallu attendre l'avènement des ordinateurs électroniques pour surmonter ce problème et accéder à une utilisation plus courante de l'approche. Cependant, même après l'avènement des ordinateurs, il était important d'écrire des algorithmes efficaces pour le calcul des estimateurs de maximum de vraisemblance de composantes de variance.

Dans cette section, nous considérons le modèle linéaire à effets mixtes défini par les équations (1.14) et (1.15), avec y_{obs} une réalisation observée de $\mathcal{Y}$.

Estimation empirique de Bayes

La méthode d'estimation empirique de Bayes fait historiquement partie des premières méthodes les plus utilisées pour l'estimation des paramètres du modèle linéaire à effets mixtes. L'approche classique d'estimation est celle du maximum de vraisemblance pour les effets fixes et les paramètres de variance, où les effets aléatoires sont obtenus par l'utilisation de la version étendue du théorème de Gauss-Markov pour l'"estimation" d'effets aléatoires, proposée par Harville [87]. Soit θ le vecteur de tous les paramètres de variance du modèle. Ces paramètres de variance sont ceux provenant de Γ et de Σ. Les paramètres à estimer sont désormais β et θ, en plus d'une réalisation des effets aléatoires conditionnellement à y_{obs}. L'algorithme de l'estimation empirique de Bayes est le suivant :

1. On initialise θ par une valeur arbitraire θ_0.
2. Avec θ, on écrit la matrice de précision de y_{obs}
$$W = \left(\Sigma + Z\Gamma Z^\top\right)^{-1}. \tag{1.16}$$
3. L'estimateur de β est celui des moindres carrés généralisés
$$\widehat{\beta} = \left(X^\top W X\right)^{-1} X^\top W y_{\text{obs}}. \tag{1.17}$$

4. Une réalisation de γ, conditionnellement à y_{obs} s'écrit

$$\widetilde{\gamma} = \mathbb{E}\left[\gamma | y_{\text{obs}}, \widehat{\beta}, \theta\right] = \Gamma Z^\top W(y_{\text{obs}} - X\widehat{\beta}). \tag{1.18}$$

5. La vraisemblance s'écrit finalement comme fonction uniquement de θ, facile à optimiser. Avec la valeur $\widehat{\theta}$ de θ qui maximise la vraisemblance, on obtient $\widehat{\beta}$, $\widehat{\Gamma}$ et $\widehat{\Sigma}$.

Cette méthode tire son nom de l'expression de $\widetilde{\gamma}$ et n'est en fait qu'une approche unifiée basée sur une combinaison d'estimation par maximum de vraisemblance et de Bayes empirique.

Les estimateurs du maximum de vraisemblance du vecteur des composantes de variance, θ, ne tiennent pas compte de la perte en degré de liberté resultant de l'estimation des effets fixes du modèle, et sont donc biaisés [88, 104]. Patterson et Thompson [139] ont alors proposé l'estimation non biaisée de composantes de variance par la méthode du REML. Cette méthode, vue comme une alternative efficace à l'estimation basée sur la maximisation de la vraisemblance, consiste à maximiser l'intégrale de la vraisemblance ; laquelle intégrale est calculée relativement au vecteur des effets fixes.

Estimateurs de maximum de vraisemblance au sens de l'algorithme EM

L'algorithme EM [44] est un algorithme statistique itératif qui permet d'approcher plus aisément les estimateurs du maximum de vraisemblance des paramètres d'un modèle statistique qu'on ajuste à un jeu de données incomplètes. Dans le cas du modèle linéaire à effets mixtes, les effets aléatoires sont non observés et les données observées, y, peuvent être considérées comme étant des observations incomplètes. Considérons le modèle linéaire à effets mixtes défini par les équations (1.14) et (1.15), et posons ϕ le vecteur de tous ses paramètres. A l'itération d'ordre $p+1$, la mise en oeuvre de l'algorithme EM pour calculer la valeur ϕ^{p+1} du paramètre à partir de ϕ^p se fait en deux étapes comme suit :

Etape E (Espérance) : calcul de

$$Q(\phi|\phi^p) = \mathbb{E}\left[\log f(y,\gamma)|y,\phi^p\right], \tag{1.19}$$

où f est la fonction densité de la loi jointe de y et γ.

Etape M (Maximisation) : calcul de ϕ^{p+1} tel que

$$\phi^{p+1} = \arg\max_{\phi} Q(\phi|\phi^p). \tag{1.20}$$

Les estimateurs de maximum de vraisemblance au sens de l'algorithme EM s'écrivent.

$$\widehat{\beta} = \left(X^\top \Sigma^{-1} X\right)^{-1} X^\top \Sigma^{-1} \left(y - Z\mathbb{E}[\gamma|y,\phi^p]\right), \tag{1.21}$$

$$\widehat{\Gamma} = \mathbb{V}[\gamma|y,\phi^p] + \mathbb{E}[\gamma|y,\phi^p]\mathbb{E}[\gamma|y,\phi^p]^\top, \tag{1.22}$$

et

$$\widehat{\Sigma} = Z\mathbb{V}[\gamma|y,\phi^p]Z^\top + (y - X\beta - Z\mathbb{E}[\gamma|y,\phi^p])(y - X\beta - Z\mathbb{E}[\gamma|y,\phi^p])^\top, \tag{1.23}$$

où,

$$\mathbb{E}[\gamma|y,\phi^p] = \text{Cov}(\gamma, y|\phi^p)\mathbb{V}[y|\phi^p]^{-1}(y - \mathbb{E}[y|\phi^p]),$$

$$\mathbb{V}[\gamma|y,\phi^p] = \Gamma - \text{Cov}(\gamma, y|\phi^p)\mathbb{V}[y|\phi^p]^{-1}\text{Cov}(\gamma, y|\phi^p)^\top,$$

$$\mathbb{E}[y|\phi^p] = X\beta, \quad \mathbb{V}[y|\phi^p] = Z\Gamma Z^\top + \Sigma \text{ et } \text{Cov}(\gamma, y|\phi^p) = \Gamma Z^\top.$$

Ces estimateurs s'obtiennent simplement en recherchant les points critiques de la fonction $\phi \mapsto Q(\phi|\phi^p)$, puis en considérant que $\begin{pmatrix}\gamma \\ y\end{pmatrix}$ est un vecteur gaussien.

Estimation des paramètres selon l'approche de Bates et ses co-auteurs [14]

Dans cette section, nous décrivons la stratégie d'optimisation utilisée sous le package lme4 [14] du logiciel R pour obtenir la valeur des estimateurs des paramètres d'un modèle linéaire à effets mixtes, relativement à un jeu de données. Cette stratégie exige que les résidus soient homosédastiques de façon à ce que le modèle s'écrive

$$(\mathcal{Y}|\gamma) \sim \mathcal{N}(X\beta + Z\gamma, \sigma^2 I_n), \tag{1.24}$$

$$\gamma \sim \mathcal{N}(0, \Gamma). \tag{1.25}$$

Nous rappelons que X est une matrice $n \times p$, Z est une matrice $n \times q$, I_n est la matrice identité d'ordre n, Γ est une matrice $q \times q$, $\beta \in \mathbb{R}^p$, $\sigma \in \mathbb{R}_+$ et $\gamma \in \mathbb{R}^q$ est le vecteur des effets aléatoires.

Les paramètres du modèle sont désormais : β, σ et Γ.

Γ étant la matrice de variance-covariance des effets aléatoires, elle doit être définie semi-positive. Il est alors plus pratique d'exprimer le modèle en terme de facteur de covariance relatif, Λ_θ en utilisant la décomposition de Cholesky ci-après

$$\Gamma = \sigma^2 \Lambda_\theta \Lambda_\theta^\top, \tag{1.26}$$

où θ est le vecteur de composantes de variance qui génère Λ_θ. Bien que q, la dimension de Γ, soit parfois très grande, la dimension de θ est généralement faible (moins de 10). L'égalité (1.26) permet de définir la variable aléatoire $\mathcal{U}$ dénommée *variable sphérique des effets aléatoires* comme suit

$$\mathcal{U} \sim \mathcal{N}(0, \sigma^2 I_q), \tag{1.27}$$

$$\gamma = \Lambda_\theta \mathcal{U}. \tag{1.28}$$

Ecriture de la vraisemblance y_{obs} étant une réalisation observée de $\mathcal{Y}$, la vraisemblance des paramètres du modèle, conditionnellement à y_{obs} s'écrit

$$L(\theta, \beta, \sigma^2 | y_{obs}) = \int_{\mathbb{R}^q} f_{\mathcal{Y},\mathcal{U}}(y_{obs}, u) du. \tag{1.29}$$

$f_{\mathcal{Y},\mathcal{U}}(y, u)$, la densité jointe de $\mathcal{Y}$ et de $\mathcal{U}$ s'écrit

$$\begin{aligned} f_{\mathcal{Y},\mathcal{U}}(y, u) &= f_{\mathcal{Y}|\mathcal{U}}(y|u) f_{\mathcal{U}}(u) \\ &= \frac{\exp(-\frac{1}{2\sigma^2}||y - X\beta - Z\Lambda_\theta u||^2)}{(2\pi\sigma^2)^{n/2}} \frac{\exp(-\frac{1}{2\sigma^2}||u||^2)}{(2\pi\sigma^2)^{q/2}} \\ &= \frac{\exp(-\left[||y - X\beta - Z\Lambda_\theta u||^2 + ||u||^2\right]/(2\sigma^2))}{(2\pi\sigma^2)^{(n+q)/2}}. \end{aligned} \tag{1.30}$$

Par ailleurs,

$$f_{\mathcal{Y},\mathcal{U}}(y_{obs}, u) = f_{\mathcal{U}|\mathcal{Y}}(u|y_{obs}) \int f_{\mathcal{Y},\mathcal{U}}(y_{obs}, u) du \tag{1.31}$$

Ainsi, $f_{\mathcal{Y},\mathcal{U}}(y_{obs}, u)$ et $f_{\mathcal{U}|\mathcal{Y}}(u|y_{obs})$ atteignent leur maximum au même point $\mu_{\mathcal{U}|\mathcal{Y}=y_{obs}}$. Et donc,

$$\begin{aligned} \begin{pmatrix} \mu_{\mathcal{U}|\mathcal{Y}=y_{obs}} \\ \hat{\beta}_\theta \end{pmatrix} &= \arg\max_{u,\beta} f_{\mathcal{Y},\mathcal{U}}(y_{obs}, u) \\ &= \arg\min_{u,\beta}(||y_{obs} - X\beta - Z\Lambda_\theta u||^2 + ||u||^2) \end{aligned} \tag{1.32}$$

Remarque La détermination simultanée de l'espérance conditionnelle des effets aléatoires $\mu_{\mathcal{U}|\mathcal{Y}=y_{obs}}$ (qui est aussi le mode de la distribution conditionnelle des effets aléatoires) et du paramètre β, prend la forme d'un problème des moindres carrés pénalisés. Cette remarque permet d'obtenir les expressions analytiques de $\mu_{\mathcal{U}|\mathcal{Y}=y_{obs}}$ et de β, ce qui permet ensuite d'obtenir une écriture explicite et simple à optimiser de la vraisemblance $L(\theta, \beta, \sigma^2|y_{obs})$ relativement à une valeur fixée de θ. En itérant le processus suivant les valeurs de θ, on obtient l'estimateur $\widehat{\theta}$ du maximum de vraisemblance de θ ; ce qui fournit l'estimateur de Λ_θ.

Notations Dans la suite, nous adoptons les notations ci-après :

$$g(u, \beta) = \|y_{obs} - X\beta - Z\Lambda_\theta u\|^2 + \|u\|^2 \tag{1.33}$$

$$r_\theta^2 = \min_{u,\beta}(\|y_{obs} - X\beta - Z\Lambda_\theta u\|^2 + \|u\|^2) \tag{1.34}$$

et $\mu_{\mathcal{U}|\mathcal{Y}=y_{obs}} = \mu$ pour raison de simplicité.

$g(u,\beta)$ **en fonction de** $\mu_{\mathcal{U}|\mathcal{Y}=y_{obs}}$ **et de** $\hat{\beta}_\theta$ **:**

$$\begin{aligned}\begin{pmatrix}\mu_{\mathcal{U}|\mathcal{Y}=y_{obs}}\\ \hat{\beta}_\theta\end{pmatrix} &= \underset{u,\beta}{\arg\min}\, g(u,\beta)\\ &= \underset{u,\beta}{\arg\min}\left\|\begin{pmatrix}y_{obs}\\0\end{pmatrix}-\begin{pmatrix}Z\Lambda_\theta & X\\ I_q & 0\end{pmatrix}\begin{pmatrix}u\\\beta\end{pmatrix}\right\|^2 \end{aligned} \tag{1.35}$$

$$\Longrightarrow \quad \begin{pmatrix}Z\Lambda_\theta & X\\ I_q & 0\end{pmatrix}^\top\begin{pmatrix}y_{obs}\\0\end{pmatrix}=\begin{pmatrix}Z\Lambda_\theta & X\\ I_q & 0\end{pmatrix}^\top\begin{pmatrix}Z\Lambda_\theta & X\\ I_q & 0\end{pmatrix}\begin{pmatrix}\mu\\\hat{\beta}_\theta\end{pmatrix} \text{ Eq. normale} \tag{1.36}$$

$$\Longleftrightarrow \quad \begin{pmatrix}Z\Lambda_\theta & X\\ I_q & 0\end{pmatrix}^\top\left[\begin{pmatrix}y_{obs}\\0\end{pmatrix}-\begin{pmatrix}Z\Lambda_\theta & X\\ I_q & 0\end{pmatrix}\begin{pmatrix}\mu\\\hat{\beta}_\theta\end{pmatrix}\right]=\begin{pmatrix}0\\0\end{pmatrix} \tag{1.37}$$

$$\Longleftrightarrow \quad \begin{cases}(Z\Lambda_\theta)^\top(y_{obs}-Z\Lambda_\theta\mu-X\hat{\beta}_\theta)+\mu=0\\ X^\top(y_{obs}-Z\Lambda_\theta\mu-X\hat{\beta}_\theta)=0\end{cases} \tag{1.38}$$

Par ailleurs,

$$(1.36) \implies \begin{pmatrix}(Z\Lambda_\theta)^\top Z\Lambda_\theta+I_q & (Z\Lambda_\theta)^\top X\\ X^\top Z\Lambda_\theta & X^\top X\end{pmatrix}\begin{pmatrix}\mu\\\hat{\beta}_\theta\end{pmatrix}=\begin{pmatrix}(Z\Lambda_\theta)^\top y_{obs}\\ X^\top y_{obs}\end{pmatrix} \tag{1.39}$$

On définit le facteur L_θ de Cholesky tel que

$$L_\theta L_\theta^\top=(Z\Lambda_\theta)^\top Z\Lambda_\theta+I_q. \tag{1.40}$$

g est de classe C^∞. En posant $x=\begin{pmatrix}u\\\beta\end{pmatrix}$ et $\hat{x}=\begin{pmatrix}\mu\\\hat{\beta}_\theta\end{pmatrix}$, on a :

$$g(x)=g(\hat{x})+g'(\hat{x})(x-\hat{x})+\frac{1}{2}(x-\hat{x})^\top g"(\hat{x})(x-\hat{x})+\cdots+\ldots \tag{1.41}$$

$\frac{\partial g}{\partial u}(\hat{x})=2L_\theta L_\theta^\top(\mu-\mu)=0\,;\quad \frac{\partial g}{\partial\beta}(\hat{x})=-2X^\top(y_{obs}-Z\Lambda_\theta\mu-X\hat{\beta}_\theta)=0$ d'après (1.38). Ainsi, $g'(\hat{x})=(0,0)$.

$\frac{\partial g}{\partial u^2}(x)=2L_\theta L_\theta^\top\,;\quad \frac{\partial g}{\partial u\partial\beta}(x)=2X^\top Z\Lambda_\theta\,;\quad \frac{\partial g}{\partial\beta\partial u}(x)=2(Z\Lambda_\theta)^\top X\,;\quad \frac{\partial g}{\partial\beta^2}(x)=2X^\top X$

Ainsi,

$$g"(\hat{x})=2\begin{pmatrix}L_\theta L_\theta^\top & (Z\Lambda_\theta)^\top X\\ Z\Lambda_\theta X^\top & X^\top X\end{pmatrix}=2M \tag{1.42}$$

et $g^{(n)}(x)=\mathbf{0}\quad\forall n\geq 3$.
D'où

$$\begin{aligned} g(x) &= g(\hat{x}) + (x-\hat{x})^\top M(x-\hat{x}) \\ &= r_\theta^2 + (x-\hat{x})^\top M(x-\hat{x}) \end{aligned} \tag{1.43}$$

En faisant la décomposition de Cholesky de M ci-après

$$M = \begin{pmatrix} L_\theta^\top & R_{ZX} \\ 0 & R_X \end{pmatrix}^\top \begin{pmatrix} L_\theta^\top & R_{ZX} \\ 0 & R_X \end{pmatrix}, \tag{1.44}$$

on a :

$$\begin{aligned} g(u,\beta) &= r_\theta^2 + \begin{pmatrix} u-\mu \\ \beta-\hat{\beta}_\theta \end{pmatrix}^\top \begin{pmatrix} L_\theta & 0 \\ R_{ZX}^\top & R_X^\top \end{pmatrix} \begin{pmatrix} L_\theta^\top & R_{ZX} \\ 0 & R_X \end{pmatrix} \begin{pmatrix} u-\mu \\ \beta-\hat{\beta}_\theta \end{pmatrix} \\ &= r_\theta^2 + \begin{pmatrix} u-\mu \\ \beta-\hat{\beta}_\theta \end{pmatrix}^\top \left[\begin{pmatrix} L_\theta \\ R_{ZX}^\top \end{pmatrix} \begin{pmatrix} L_\theta^\top & R_{ZX} \end{pmatrix} + \begin{pmatrix} 0 \\ R_X^\top \end{pmatrix} \begin{pmatrix} 0 & R_X \end{pmatrix} \right] \begin{pmatrix} u-\mu \\ \beta-\hat{\beta}_\theta \end{pmatrix} \\ &= r_\theta^2 + \left\| \begin{pmatrix} L_\theta^\top & R_{ZX} \end{pmatrix} \begin{pmatrix} u-\mu \\ \beta-\hat{\beta}_\theta \end{pmatrix} \right\|^2 + \left\| \begin{pmatrix} 0 & R_X \end{pmatrix} \begin{pmatrix} u-\mu \\ \beta-\hat{\beta}_\theta \end{pmatrix} \right\|^2 \qquad (1.45) \\ &= r_\theta^2 + \left\| L_\theta^\top(u-\mu) + R_{ZX}(\beta-\hat{\beta}_\theta) \right\|^2 + \left\| R_X(\beta-\hat{\beta}_\theta) \right\|^2 \qquad (1.46) \end{aligned}$$

Expression de la vraisemblance à θ fixé

$$\begin{aligned} L(\theta,\beta,\sigma^2|y_{obs}) &= \int_{\mathbb{R}^q} \frac{\exp\left(g(u,\beta)\right)}{(2\pi\sigma^2)^{(n+q)/2}} du \\ &= \frac{\exp\left[-\frac{r_\theta^2 + \left\|R_X(\beta-\hat{\beta}_\theta)\right\|^2}{2\sigma^2}\right]}{(2\pi\sigma^2)^{n/2}} \int_{\mathbb{R}^q} \frac{1}{(2\pi\sigma^2)^{q/2}} \\ &\quad \exp\left[-\frac{\left\| L_\theta^\top(u-\mu) + R_{ZX}(\beta-\hat{\beta}_\theta) \right\|^2}{2\sigma^2} \right] du \end{aligned} \tag{1.47}$$

En posant $v = L_\theta^\top(u-\mu) + R_{ZX}(\beta-\hat{\beta}_\theta)$, on a :

$$\begin{aligned} dv &= \left| \frac{Dv}{Du} \right| du \\ &= \left| L_\theta^\top \right| du \end{aligned}$$

et

$$
\begin{aligned}
L(\theta,\beta,\sigma^2|y_{obs}) &= \frac{\exp\left[-\frac{r_\theta^2+\left\|R_X(\beta-\hat{\beta}_\theta)\right\|^2}{2\sigma^2}\right]}{(2\pi\sigma^2)^{n/2}|L_\theta|}\underbrace{\int_{\mathbb{R}^q}\frac{1}{(2\pi\sigma^2)^{q/2}}\exp\left[-\frac{\|v\|^2}{2\sigma^2}\right]dv}_{=1}\\
&= \frac{\exp\left[-\frac{r_\theta^2+\left\|R_X(\beta-\hat{\beta}_\theta)\right\|^2}{2\sigma^2}\right]}{(2\pi\sigma^2)^{n/2}|L_\theta|}. \qquad (1.48)
\end{aligned}
$$

Estimateur de la déviance du modèle à θ fixé $\hat{\beta}_\theta$ est l'estimateur du maximum de vraisemblance de β à θ fixé, et en remplaçant β par $\hat{\beta}_\theta$ dans l'égalité (1.48), on obtient l'estimateur ci-après de la vraisemblance

$$
\tilde{L}(\theta,\sigma^2|y_{obs}) = \frac{\exp\left[-\frac{r_\theta^2}{2\sigma^2}\right]}{(2\pi\sigma^2)^{n/2}|L_\theta|}. \qquad (1.49)
$$

Ce qui fournit l'estimateur de la déviance ci-après

$$
\begin{aligned}
\tilde{l}(\theta,\sigma^2|y_{obs}) &= -2\log\left(\tilde{L}(\theta,\sigma^2|y_{obs})\right)\\
&= \frac{r_\theta^2}{\sigma^2}+n\log(2\pi\sigma^2)+\log(|L_\theta|^2). \qquad (1.50)
\end{aligned}
$$

Vu comme fonction de σ^2, $\tilde{l}(\theta,\sigma^2|y_{obs})$ est continument dérivable et on a :

$$
\frac{\partial(\tilde{l}(\theta,\sigma^2|y_{obs}))}{\sigma^2} = -\frac{r_\theta^2}{\sigma^4}+\frac{n}{\sigma^2}
$$

$$
\frac{\partial(\tilde{l}(\theta,\sigma^2|y_{obs}))}{\sigma^2} = 0 \iff \sigma^2 = \frac{r_\theta^2}{n}.
$$

Ainsi,

$$
\widehat{\sigma^2} = \frac{r_\theta^2}{n} \qquad (1.51)
$$

est l'estimateur du maximum de vraisemblance de σ^2 à θ fixé. Ce qui fournit, en remplaçant σ^2 par $\widehat{\sigma^2}$ dans l'égalité (1.50), l'estimateur de la déviance du modèle à θ fixé ci-après :

$$
\tilde{l}(\theta|y_{obs}) = \log(|L_\theta|^2)+n\left[1+\log\left(\frac{2\pi r_\theta^2}{n}\right)\right]. \qquad (1.52)
$$

Finalement,

$$
\hat{\theta} = \arg\min_\theta \tilde{l}(\theta|y_{obs}) \qquad (1.53)
$$

Récapitulatif de la méthode La méthode d'estimation des paramètres d'un modèle linéaire mixte, utilisée sous le package lme4 du logiciel R se résume en ces quelques étapes qui suivent :

1. On fixe une valeur du paramètre de composante de variance θ, puis on engendre Λ_θ
2. on calcule L_θ solution de l'équation (1.40)
3. on calcule $\hat{\beta}_\theta$ et μ solutions de l'équation (1.39)
4. on calcule $r_\theta^2 = \|y_{obs} - X\hat{\beta}_\theta - Z\Lambda_\theta\mu\|^2 + \|\mu\|^2$ issu des équations (1.32) et (1.34)
5. on calcule $\tilde{l}(\theta|y_{obs})$ dont l'expression est donnée par l'égalité (1.52)
6. on reprend les opérations 1. à 5. jusqu'à retrouver la valeur $\hat{\theta}$ de θ qui minimise $\tilde{l}(\theta|y_{obs})$.

$\widehat{\beta}_{\hat{\theta}}$ est l'estimateur du maximum de vraisemblance des effets fixes, et $\tilde{\gamma} = \Lambda_{\hat{\theta}}\mu$ est le mode de la distribution des effets aléatoires, conditionnellement à y_{obs}. $\tilde{\gamma}$ est considéré comme étant la réalisation de γ, conditionnellement à y_{obs}. $\widehat{\sigma^2}$ est l'estimateur du maximum de vraisemblance du paramètre de variance résiduelle, tandis que $\widehat{\Gamma} = \widehat{\sigma^2}\Lambda_{\hat{\theta}}\Lambda_{\hat{\theta}}^\top$ est l'estimateur du maximum de vraisemblance de la matrice de variance-covariance des effets aléatoires.

Critère REML pour l'estimation non biaisée des composantes de variance Le critère REML (Restricted Maximum Likelihood) est la vraisemblance basée sur les composantes de variance et permet d'obtenir un estimateur non biaisé de ces composantes. Il s'écrit

$$L_R(\theta, \sigma^2|y_{obs}) = \int_{\mathbb{R}^p} L(\theta, \beta, \sigma^2|y_{obs})d\beta. \tag{1.54}$$

En utilisant l'égalité (1.48) et en posant $t = R_X(\beta - \hat{\beta}_\theta)$, on a $dt = |R_X|d\beta$ et

$$\begin{aligned} L_R(\theta, \sigma^2|y_{obs}) &= \frac{1}{|L_\theta|}\int_{\mathbb{R}^p} \frac{\exp\left[-\frac{r_\theta^2 + \left\|R_X(\beta-\hat{\beta}_\theta)\right\|^2}{2\sigma^2}\right]}{(2\pi\sigma^2)^{n/2}}d\beta \qquad (1.55)\\ &= \frac{\exp(-r_\theta^2/(2\sigma^2))}{|L_\theta||R_X|(2\pi\sigma^2)^{(n-p)/2}} \underbrace{\int_{\mathbb{R}^p} \frac{\exp\left[-\frac{\|t\|^2}{2\sigma^2}\right]}{(2\pi\sigma^2)^{p/2}}dt}_{=1}\\ &= \frac{\exp(-r_\theta^2/(2\sigma^2))}{|L_\theta||R_X|(2\pi\sigma^2)^{(n-p)/2}} \qquad (1.56) \end{aligned}$$

$$\begin{aligned} \frac{\partial\left(-2\log(L_R)\right)}{\partial(\sigma^2)} = 0 \quad &\Longleftrightarrow \quad -\frac{r_\theta^2}{\sigma^4} + (n-p)\frac{1}{\sigma^2} = 0\\ &\Longrightarrow \quad \widehat{\sigma^2}_R = \frac{r_\theta^2}{n-p} \qquad (1.57) \end{aligned}$$

Et donc,

$$\begin{aligned}\tilde{l}_R(\theta|y_{obs}) &= -2\left(\log(L_R)\right) \\ &= \log\left(|L_\theta|^2\right) + \log\left(|R_X|^2\right) + (n-p)\left[1 + \log\left(\frac{2\pi r_\theta^2}{n-p}\right)\right]\end{aligned} \tag{1.58}$$

puis,

$$\hat{\theta}_R = \arg\min_\theta \tilde{l}_R(\theta|y_{obs}). \tag{1.59}$$

1.3 Contributions méthodologiques de la thèse

Dans cette thèse, nos contributions sont de deux ordres. Le premier est relatif à l'estimation des paramètres de la version multidimensionnelle du modèle linéaire à effets mixtes, et comprend deux articles scientifiques qui représentent les chapitres 2 et 3 du présent manuscrit. Le second quant à lui s'appuie sur une version modifiée de la méthode d'estimation utilisée dans le package lme4 pour proposer une procédure de sélection d'effets fixes dans le modèle linéaire à effets mixtes.

En ce qui concerne la définition de la version multidimensionnelle du modèle, nous pensons, pour des raisons de simplicité et pour faciliter la lecture, qu'il serait plus utile de se limiter à la définition de la version bi-dimensionnelle, puisqu'une généralisation en dimension supérieur devient naturelle et intuitive. En dimension $k = 2$, le modèle s'écrit

$$\begin{aligned} y_1 &= X_1\beta_1 + Z_1\gamma_1 + \varepsilon_1 \\ y_2 &= X_2\beta_2 + Z_2\gamma_2 + \varepsilon_2 \end{aligned} \tag{1.60}$$

$$\boldsymbol{\gamma} = \begin{pmatrix}\gamma_1 \\ \gamma_2\end{pmatrix} \sim \mathcal{N}\left(\mathbf{0}; \mathbf{\Gamma} = \begin{pmatrix}\Gamma_1 & \Gamma_{12} \\ \Gamma_{12}^\top & \Gamma_2\end{pmatrix}\right), \tag{1.61}$$

$$\boldsymbol{\varepsilon} = \begin{pmatrix}\varepsilon_1 \\ \varepsilon_2\end{pmatrix} \sim \mathcal{N}\left(\mathbf{0}; \mathbf{\Sigma} = \begin{pmatrix}\Sigma_1 & \Sigma_{12} \\ \Sigma_{12}^\top & \Sigma_2\end{pmatrix}\right); \quad \boldsymbol{\gamma} \perp \boldsymbol{\varepsilon}. \tag{1.62}$$

Pour $k \in \{1, 2\}$, β_k et γ_k désignent respectivement le vecteur des effets fixes et le vecteur des effets aléatoires, et ε_k est la composante résiduelle. X_k est une matrice de covariables et Z_k une matrice de *design* faite de covariables. dim(X_k)= $N \times p_k$ et dim(Z_k)= $N \times q_k$, avec N le nombre total d'observations composant le jeu de données auquel le modèle est ajusté. p_k et q_k sont respectivement le nombre d'effets fixes et le nombre d'effets aléatoires dans la dimension k du modèle. Les symboles en gras représentent les paramètres qui ont des composantes dans plus d'une dimension du modèle. Par exemple, Σ_1 intervient seulement dans la dimension 1 du modèle alors que $\mathbf{\Sigma}$ intervient dans les deux dimensions du modèle de par ses composantes.

Dans cette thèse, nos contributions, relativement à la version multidimensionnelle du modèle ont été faites sous l'hypothèse d'indépendance entre les variables réponses conditionnellement

aux effets aléatoires. C'est à dire que les résidus de deux différentes dimensions du modèle sont indépendants. Ce qui se traduit dans la version bi-dimensionnelle par

$$\Sigma_{12} = \mathbf{0}. \tag{1.63}$$

Généralement dans la littérature, cette hypothèse est faite pour des raisons techniques relatives au calcul des estimateurs des paramètres de variance résiduelle (voir par exemple, [172], [161], [56] et [6]). Aussi peut-on toujours faire cette hypothèse parce qu'on aura choisi de concentrer toute la dépendance entre réponses marginales dans les effets aléatoires.

Dans nos travaux, nous avons procédé à une revue de littérature des différentes stratégies d'analyse et modèles ajustés aux données longitudinales multidimensionnelles. En effet, les différentes approches utilisées dans la littérature peuvent être scindées en deux groupes à savoir : 1) celles qui spécifient la distribution jointe de toutes les m variables de réponse $y_1, \ldots, y_m$ (également noté sous la forme d'un vecteur $\boldsymbol{y} = (y_1^\top, \ldots, y_m^\top)^\top$) sans recourir à des structures latentes et 2) celles qui font une modélisation basée sur des structures latentes.

Dans le premier groupe, on retrouve trois différentes approches. D'abord une spécification directe de la structure de corrélation de $\boldsymbol{y}$ où l'on peut citer par exemple les travaux de Galecki [68] et ceux de O'Brien [135] qui ont factorisé la matrice de covariance de $\boldsymbol{y}$ en se servant du produit de Kronecker afin d'avoir des modèles plus parcimonieux dans le cas d'analyse de données longitudinales équilibrées. D'autres modèles comme les copules [132, 175] utilisés par exemple dans les travaux de Lambert [106] peuvent être aussi classés dans ce premier sous-groupe d'approches au même titre que certains modèles de séries chronologiques comme les ARIMA (*Auto Regressive Integrated Moving Average*) et les VAR (*Vector Auto Regressive*) utilisés par exemple par MaCurdy [121] et Tschacher [188] dans leurs travaux. Ensuite une analyse sans la modélisation explicite de la structure de corrélation de $\boldsymbol{y}$ où l'on peut principalement citer les GEE (*Generalized Estimating Equations*) proposés par Liang [110] en tant qu'extension du modèle linéaire généralisé pour l'analyse de données longitudinales. Des exemples d'utilisation des GEE dans le contexte d'analyse de données longitudinales multidimensionnelles incluent les travaux de Prentice [143], ceux de Liang [111] et de Rochon [156]. Enfin dans ce premier groupe, retrouve-t-on aussi les modèles d'analyses marginales conditionnelles où l'on évite une spécification directe de la distribution jointe de $\boldsymbol{y}$ (que nous notons $f(\boldsymbol{y})$) en la factorisant (par exemple en dimension 2, on peut bien écrire $f(y_1, y_2) = f(y_1|y_2)f(y_2)$ où y_2 joue le rôle de covariable).

Le second groupe d'approches quant à lui se subdivise en deux sous-groupes dont le premier consiste en une réduction de la dimension de $\boldsymbol{y}$ et le second, aux modèles connus sous le nom de modèles à effets mixtes. Les stratégies basées sur la réduction de la dimension de $\boldsymbol{y}$ consistent à utiliser une méthode de type analyse en composantes principales afin de résumer l'information contenue dans $\boldsymbol{y}$ en un petit nombre de facteurs latents pertinents auxquels il est ensuite ajusté des modèles unidimensionnels standards et indépendants (voir par exemple [49, 123, 137]). En ce qui concerne le modèle à effets mixtes, plusieurs auteurs dont Reinsel [151], Maccallum [120] et Beckett [18] ont proposé son utilisation dans le cadre d'analyse de données longitudinales multidimensionnelles. Bien qu'apparaissant comme étant le modèle le plus souvent utilisé en matière d'analyse de données longitudinales et plus généralement de données multi-niveaux, son utilisation dans le cas multidimensionnel se heurte à des problèmes numériques souvent dus à une dimension

trop élevée du vecteur de tous les effets aléatoires marginaux. Pour contourner ce problème, Fieuws et ses co-auteurs [57] ont proposé comme estimateurs des paramètres du modèle joint, la moyenne des estimateurs issus de tous les $m(m-1)/2$ modèles bi-dimensionnels. Toutes ces approches, leurs forces aussi bien que leurs faiblesses ont été amplement discutées avec davantage de détails dans le deuxième chapitre du présent manuscrit.

Les outils logiciels dédiés à l'ajustement de modèles linéaires à effets mixtes multidimensionnels sont vraiment rares ; ce qui fait que parfois, certains auteurs organisent les données de façon à pouvoir utiliser des outils initialement destinés aux analyses unidimensionnelles. C'est le cas par exemple des travaux de Thiébaut et ses co-auteurs [185] qui ont utilisé la procédure Proc MIXED du logiciel SAS afin d'ajuster un modèle bi-dimensionnel à effets mixtes à des profils longitudinaux de biomarqueurs relatifs au VIH SIDA. Récemment, le programme SabreR [39] du logiciel R, destiné à ajuster des modèles linéaires mixtes à au plus trois dimensions et avec seulement des intercepts aléatoires a été enlevé du dépôt. Schafer et Yucel [161] dont nous avons également mentionné les travaux au chapitre 2, ont mis au point un programme de calculs nommé mlmmm [203] basé sur leur méthode, sous le logiciel R.

Dans cette thèse, nos contributions sont relatives à l'utilisation du modèle linéaire à effets mixtes multidimensionnel et à une amélioration de la qualité de l'estimation de ses paramètres.

Après une bonne lecture de la littérature, on peut remarquer que l'estimation des paramètres du modèle linéaire à effets mixtes multidimensionnel se fait le plus souvent par l'algorithme EM, surtout dans le cadre d'analyse de données longitudinales. C'est pour cela que dans un premier temps, en utilisant l'algorithme EM, nous proposons des estimateurs écrits sous une forme plus générale que celle proposée dans la littérature (voir par exemple [6], [161] et [172]). Ces estimateurs ont une forme plus générale en ce sens qu'ils sont utilisables à la fois pour des données longitudinales et des données multi-niveaux. Par la suite, nous proposons une version multidimensionnelle de la méthode d'estimation proposée par Bates et ses co-auteurs [16]. Bien que restreinte au cas où les résidus dimensionnels sont homoscédastiques et indépendants des effets aléatoires, cette stratégie d'estimation s'est montrée plus robuste par rapport au point de départ (de l'algorithme) et fournit des estimations plus consistantes (sur nos simulations) que celles obtenues avec l'algorithme EM, surtout en ce qui concerne les paramètres de variance des effets aléatoires.

Par ailleurs, nous proposons une procédure de sélection d'effets fixes dans la version unidimensionnelle du modèle. Laquelle procédure est basée sur une pénalisation itérée de type L_2 de la log-vraisemblance profilée de certains paramètres du modèle en vue d'approcher les performances d'une pénalité L_0. Il importe ici de signaler que ce travail est le prélude d'un autre à venir et qui consistera à faire de la sélection de variables dans la version multidimensionnelle du même modèle. De façon un peu plus détaillée, nous présentons dans les paragraphes qui suivent un résumé de nos différentes contributions.

Contributions relatives au chapitre 2 : Analyse longitudinale multi-variée et test de corrélation bi-varié

Dans le contexte de l'estimation des paramètres du modèle linéaire multidimensionnel à effets mixtes, nous proposons des expressions plus générales des estimateurs en recourant à l'algorithme EM, exactement comme nous avons commencé par le faire en dimension $k = 1$ (voir

Section 1.2.2). Ces estimateurs ont des expressions plus générales en ce sens qu'ils sont aussi utilisables tels quels dans le cadre plus général de l'estimation des paramètres d'un modèle linéaire multidimensionnel multi-niveaux. Aussi est-il important de rappeler qu'un jeu de données longitudinales peut-être vu comme étant un jeu de données multi-niveaux. Ces estimateurs du maximum de vraisemblance obtenus via l'algorithme EM s'écrivent en dimension 2 comme suit. Pour $k \in \{1, 2\}$,

$$\widehat{\beta}_k = \left(X_k^\top \Sigma_k^{-1} X_k\right)^{-1} X_k^\top \Sigma_k^{-1} \left(y_k - Z_k \mathbb{E}[\gamma_k | \boldsymbol{y}, \phi^p]\right), \tag{1.64}$$

$$\widehat{\boldsymbol{\Gamma}} = \mathbb{V}[\boldsymbol{\gamma} | \boldsymbol{y}, \phi^p] + \mathbb{E}[\boldsymbol{\gamma} | \boldsymbol{y}, \phi^p] \mathbb{E}[\boldsymbol{\gamma} | \boldsymbol{y}, \phi^p]^\top, \tag{1.65}$$

$$\widehat{\Sigma}_k = Z_k \mathbb{V}[\gamma_k | \boldsymbol{y}, \phi^p] Z_k^\top + (y_k - X_k \beta_k - Z_k \mathbb{E}[\gamma_k | \boldsymbol{y}, \phi^p])(y_k - X_k \beta_k - Z_k \mathbb{E}[\gamma_k | \boldsymbol{y}, \phi^p])^\top, \tag{1.66}$$

où,

$$\mathbb{E}[\gamma_k | \boldsymbol{y}, \phi^p] = \text{Cov}(\gamma_k, \boldsymbol{y} | \phi^p) \mathbb{V}[\boldsymbol{y} | \phi^p]^{-1} (\boldsymbol{y} - \mathbb{E}[\boldsymbol{y} | \phi^p]), \tag{1.67}$$

$$\mathbb{V}[\gamma_k | \boldsymbol{y}, \phi^p] = \boldsymbol{\Gamma}_k - \text{Cov}(\gamma_k, \boldsymbol{y} | \phi^p) \mathbb{V}[\boldsymbol{y} | \phi^p]^{-1} \text{Cov}(\gamma_k, \boldsymbol{y} | \phi^p)^\top \tag{1.68}$$

et

$$\mathbb{V}[\boldsymbol{y} | \phi^p] = \begin{pmatrix} Z_1 \Gamma_1 Z_1^\top + \Sigma_1 & Z_1 \Gamma_{12} Z_2^\top \\ Z_2 \Gamma_{12}^\top Z_1^\top & Z_2 \Gamma_2 Z_2^\top + \Sigma_2 \end{pmatrix}, \quad \text{Cov}(\boldsymbol{\gamma}, \boldsymbol{y} | \phi^p) = \begin{pmatrix} Z_1 \Gamma_1 & Z_1 \Gamma_{12} \\ Z_2 \Gamma_{12}^\top & Z_2 \Gamma_2 \end{pmatrix}^\top, \tag{1.69}$$

$$\text{Cov}(\gamma_1, \boldsymbol{y} | \phi^p) = \begin{pmatrix} Z_1 \Gamma_1 \\ Z_2 \Gamma_{12}^\top \end{pmatrix}^\top, \quad \text{Cov}(\gamma_2, \boldsymbol{y} | \phi^p) = \begin{pmatrix} Z_1 \Gamma_{12} \\ Z_2 \Gamma_2 \end{pmatrix}^\top. \tag{1.70}$$

Nous avons procédé à une investigation empirique de la précision de ces estimateurs en calculant l'erreur quadratique moyenne de plusieurs estimations faites sur 1000 replications de jeux de données longitudinales simulées sous le modèle bi-dimensionnel ci-contre. En identifiant les sujets par i, et pour $i = 1, \ldots, n$

$$\begin{aligned} y_{1i} &= X_{1i} \beta_1 + Z_{1i} \gamma_{1i} + \varepsilon_{1i} \\ y_{2i} &= X_{2i} \beta_2 + Z_{2i} \gamma_{2i} + \varepsilon_{2i}, \end{aligned} \tag{1.71}$$

avec

$$\boldsymbol{\gamma}_i = \begin{pmatrix} \gamma_{1i} \\ \gamma_{2i} \end{pmatrix} \sim \mathcal{N}\left(\mathbf{0}; \bar{\boldsymbol{\Gamma}} = \begin{pmatrix} \bar{\Gamma}_1 & \bar{\Gamma}_{12} \\ \bar{\Gamma}_{12}^\top & \bar{\Gamma}_2 \end{pmatrix}\right) \tag{1.72}$$

TABLE 1.2 : **Erreur quadratique moyenne des EM-estimateurs avec l'intervalle de confiance à 95% calculés sur 1000 replications pour différentes valeurs du nombre n de sujets et du nombre N d'observations.**

Paramètre	n	$N = 600$	$N = 1000$	$N = 3000$
β_1	50	1.27 (0.01 - 4.64)	1.10 (0.00 - 3.83)	0.71 (0.00 - 2.57)
	60	1.38 (0.00 - 5.56)	0.99 (0.00 - 3.90)	0.65 (0.00 - 2.44)
	100	1.31 (0.00 - 4.79)	0.87 (0.00 - 3.41)	0.46 (0.00 - 1.58)
	300	1.64 (0.00 - 5.93)	0.80 (0.00 - 3.23)	0.29 (0.00 - 1.13)
β_2	50	2.13 (0.01 - 7.22)	1.73 (0.00 - 6.21)	1.08 (0.00 - 4.17)
	60	2.12 (0.00 - 8.06)	1.63 (0.01 - 6.66)	0.95 (0.00 - 3.52)
	100	1.88 (0.00 - 7.14)	1.29 (0.00 - 4.76)	0.62 (0.00 - 2.24)
	300	2.30 (0.00 - 8.54)	1.23 (0.00 - 4.57)	0.43 (0.00 - 1.65)
σ_1	50	0.03 (0.00 - 0.10)	0.10 (0.00 - 0.07)	0.01 (0.00 - 0.02)
	60	0.04 (0.00 - 0.14)	0.02 (0.00 - 0.07)	0.01 (0.00 - 0.02)
	100	0.04 (0.00 - 0.14)	0.02 (0.00 - 0.08)	0.01 (0.00 - 0.02)
	300	0.06 (0.00 - 0.22)	0.03 (0.00 - 0.11)	0.01 (0.00 - 0.02)
σ_2	50	0.05 (0.00 - 0.18)	0.03 (0.00 - 0.12)	0.01 (0.00 - 0.03)
	60	0.06 (0.00 - 0.21)	0.03 (0.00 - 0.12)	0.01 (0.00 - 0.03)
	100	0.06 (0.00 - 0.24)	0.04 (0.00 - 0.15)	0.01 (0.00 - 0.03)
	300	0.09 (0.00 - 0.36)	0.04 (0.00 - 0.18)	0.01 (0.00 - 0.05)
$\bar{\mathbf{\Gamma}}$	50	400.91 (1.52 - 1274.32)	670.46 (2.36 - 2536.73)	489.07 (2.02 - 1840.60)
	60	706.45 (2.58 - 2497.42)	620.34 (2.70 - 2293.58)	408.79 (0.85 - 1597.77)
	100	701.50 (3.79 - 2747.70)	477.55 (2.08 - 1839.26)	283.25 (1.97 - 1023.77)
	300	798.28 (3.65 - 2603.65)	547.83 (3.03 - 1721.46)	199.54 (0.81 - 736.08)

et

$$\boldsymbol{\varepsilon}_i = \begin{pmatrix} \varepsilon_{1i} \\ \varepsilon_{2i} \end{pmatrix} \sim \mathcal{N}\left(\mathbf{0}; \bar{\mathbf{\Sigma}} = \begin{pmatrix} \sigma_1^2 \mathbf{I}_{N_{1i}} & 0 \\ 0 & \sigma_2^2 \mathbf{I}_{N_{2i}} \end{pmatrix}\right) \tag{1.73}$$

N_{1i} et N_{2i} sont les dimensions de y_{1i} et y_{2i}, respectivement.

Les résultats des estimations faites sur ces jeux de données de différentes tailles ($n \in \{50, 60, 100, 300\}$, $N \in \{600, 1000, 3000\}$) prouvent empiriquement que les estimateurs des effets fixes et des paramètres de variance résiduelle s'estiment beaucoup mieux (asymptotiquement consistants) que les paramètres de variance des effets aléatoires (voir Tableau 1.2). Ces résultats suggèrent aussi, en ce qui concerne l'analyse de données longitudinales, un bon compromis entre le nombre n de sujets et le nombre N d'observations en vue d'avoir des estimations d'une qualité relativement acceptable, surtout lorsque N n'est pas très grand.

Dans ce chapitre qui a fait l'objet d'une publication scientifique dans PloS one (voir [1]), nous avons également construit un test de rapport de vraisemblance basé sur ces EM-estimateurs afin de tester la significativité de la dépendance entre deux dimensions du modèle (y_k et $y_{k'}$, $k \neq k'$). Dans le cadre de l'analyse de données longitudinales, ce test de corrélation bi-varié a montré de bonnes performances empiriques avec des données simulées. Ce test peut en effet détecter une corrélation de signal faible ($\rho = 0.3$ avec un AUC$= 0.81$, voir Figure 1.1), même avec des jeux de données de taille modeste ($n = 60$ sujets, avec $N = 600$ observations). En utilisant ce test de corrélation bi-varié, nous avons illustré, par des simulations, certaines des conséquences du choix d'analyser séparément plutôt que conjointement deux variables réponses. Deux variables dépendantes qui sont prouvées non significativement corrélées après le test, seront donc analysées

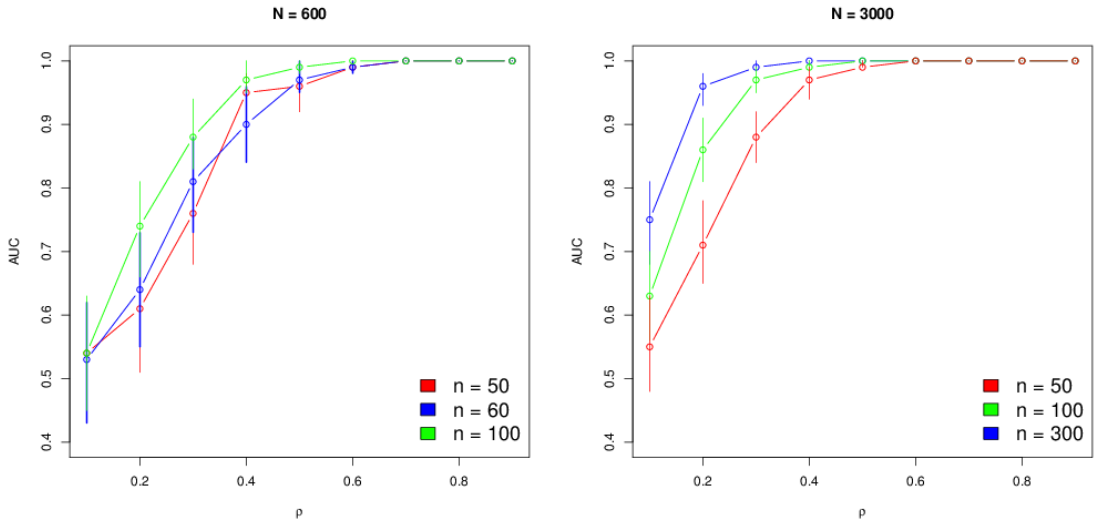

FIGURE 1.1 : **Analyse empirique de la puissance du test de corrélation bi-varié.** Les valeurs des AUC des courbes ROC avec leurs intervalles de confiance sont calculées, utilisant différentes valeurs du coefficient de corrélation linéaire ρ, du nombre de sujets (n) et du nombre d'observations (N). Le panel de gauche pour $N = 600$ et $n \in \{50, 60, 100\}$. Le panel de droite pour $N = 3000$ et $n \in \{50, 100, 300\}$.

séparément. Nous pensons que cette stratégie pourrait être utilisée dans une procédure de sélection pas-à-pas du modèle conjoint faisant intervenir toutes les variables dépendantes dans le but d'avoir un modèle conjoint plus parcimonieux en terme de nombre de paramètres de variance à estimer pour les effets aléatoires.

Contributions relatives au chapitre 3 : Estimateurs consistants pour les paramètres du modèle linéaire multidimensionnel à effets mixtes

Dans le chapitre 2, nous avons remarqué que les EM-estimateurs des composantes de variance des effets aléatoires sont ceux qui ont des erreurs quadratiques les plus élevées, même dans le cas d'homoscédasticité des résidus marginaux (dimensionnels) où nous sommes restés pour simuler et estimer. Dans le chapitre 3, nous restons strictement sous l'hypothèse d'homoscédasticité des résidus marginaux pour proposer une stratégie d'estimation qui permet d'avoir des estimations empiriquement consistantes. Cette approche que nous proposons est en fait une généralisation (extension) de la méthode d'estimation proposée par Bates et ses co-auteurs [14], et que nous avons détaillée à la Section 1.2.2. Cette méthode qui pourrait être qualifiée d'*estimation à la loupe*, surtout en ce qui concerne les paramètres de variance des effets aléatoires, n'est finalement qu'une réécriture habile de la vraisemblance vue comme fonction, uniquement, des paramètres de variance sur lesquels le processus d'optimisation est itéré jusqu'à convergence.

Comme expliqué un peu plus haut, nous allons nous restreindre à la dimension $k = 2$, puisqu'une écriture générale paraît trop lourde mais relativement simple à obtenir quand on part du

cas $k = 2$. Posons $\boldsymbol{\beta} = (\beta_1^\top, \beta_2^\top)^\top$, $\boldsymbol{\sigma} = (\sigma_1^2, \sigma_2^2)^\top$, $\boldsymbol{\theta} = (\theta_1^\top, \theta_2^\top)^\top$, où σ_1^2 et σ_2^2 sont les variances résiduelles dans les deux dimensions respectivement, et θ_1 et θ_2 sont les vecteurs respectifs des composantes de variances des effets aléatoires dans les deux dimensions 1 et 2 du modèle. Notons en suite $Y_{\boldsymbol{\sigma}} = \left(\sqrt{\sigma_2^2} y_1^\top, \sqrt{\sigma_1^2} y_2^\top\right)^\top$, $X_{\boldsymbol{\sigma}} = \begin{pmatrix} \sqrt{\sigma_2^2} X_1 & \mathbf{0} \\ \mathbf{0} & \sqrt{\sigma_1^2} X_2 \end{pmatrix}$, $Z_{\boldsymbol{\sigma\theta}} = \begin{pmatrix} \sqrt{\sigma_2^2} Z_1 \Lambda_{\theta_1} & \mathbf{0} \\ \mathbf{0} & \sqrt{\sigma_1^2} Z_2 \Lambda_{\theta_2} \end{pmatrix}$. Soit u_1 et u_2 les variables sphériques des effets aléatoires $\gamma_1 \in \mathbb{R}^{q_1}$ et $\gamma_2 \in \mathbb{R}^{q_2}$ de chacune des dimensions, et $q = q_1 + q_2$.

$$\gamma_1 = \Lambda_{\theta_1} u_1, \quad \gamma_2 = \Lambda_{\theta_2} u_2 \tag{1.74}$$

tel que

$$\boldsymbol{u} = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} \sim \mathcal{N}(\mathbf{0}, \Sigma_{\boldsymbol{u}}), \quad \text{avec} \quad \Sigma_{\boldsymbol{u}} = \begin{pmatrix} \sigma_1^2 \mathrm{I}_{q_1} & \sigma_1\sigma_2 \boldsymbol{\rho} \\ \sigma_1\sigma_2 \boldsymbol{\rho}^\top & \sigma_2^2 \mathrm{I}_{q_2} \end{pmatrix}. \tag{1.75}$$

$\boldsymbol{\rho} = \mathrm{diag}(\rho, \ldots, \rho)$, où ρ est la matrice qui contient les corrélations entre les effets aléatoires marginaux γ_1 et γ_2. Notons $\mu_{\boldsymbol{\mathcal{U}}|\boldsymbol{\mathcal{Y}}=\boldsymbol{y}}$ l'espérance de $\boldsymbol{\mathcal{U}}$ conditionnellement à $\boldsymbol{\mathcal{Y}} = \boldsymbol{y}$. La log-vraisemblance de $\boldsymbol{\beta}$, $\boldsymbol{\sigma}$, $\boldsymbol{\theta}$ et ρ conditionnellement à $\boldsymbol{y}$ s'écrit

$$\begin{aligned} \ell(\boldsymbol{\beta}, \boldsymbol{\theta}, \rho, \boldsymbol{\sigma}|\boldsymbol{y}) \;=\; & -\frac{r(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\theta},\rho,\boldsymbol{\sigma}}, \mu_{\boldsymbol{\mathcal{U}}|\boldsymbol{\mathcal{Y}}=\boldsymbol{y}}) + \left\| R_X(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_{\boldsymbol{\theta},\rho,\boldsymbol{\sigma}}) \right\|^2}{2\sigma_1^2\sigma_2^2} - \frac{N-q}{2}\log(\sigma_1^2\sigma_2^2) \\ & -\frac{1}{2}\log(|\Sigma_{\boldsymbol{u}}|) - \frac{1}{2}\log(|L_{\boldsymbol{\theta},\rho,\boldsymbol{\sigma}}|^2) \end{aligned} \tag{1.76}$$

où, $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\theta},\rho,\boldsymbol{\sigma}}$ et $\mu_{\boldsymbol{\mathcal{U}}|\boldsymbol{\mathcal{Y}}=\boldsymbol{y}}$ satisfont

$$\begin{pmatrix} X_{\boldsymbol{\sigma}}^\top X_{\boldsymbol{\sigma}} & X_{\boldsymbol{\sigma}}^\top Z_{\boldsymbol{\sigma\theta}} \\ Z_{\boldsymbol{\sigma\theta}}^\top X_\sigma & Z_{\boldsymbol{\sigma\theta}}^\top Z_{\boldsymbol{\sigma\theta}} + \sqrt{\sigma_1^2\sigma_2^2}\Sigma_{\boldsymbol{u}}^{-1} \end{pmatrix} \begin{pmatrix} \widehat{\boldsymbol{\beta}}_{\boldsymbol{\theta},\rho,\boldsymbol{\sigma}} \\ \mu_{\mathcal{U}|\mathcal{Y}=\boldsymbol{y}} \end{pmatrix} = \begin{pmatrix} X_{\boldsymbol{\sigma}}^\top \\ Z_{\boldsymbol{\sigma\theta}}^\top \end{pmatrix} Y_{\boldsymbol{\sigma}}, \tag{1.77}$$

$$r(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\theta},\rho,\boldsymbol{\sigma}}, \mu_{\boldsymbol{\mathcal{U}}|\boldsymbol{\mathcal{Y}}=\boldsymbol{y}}) = \left\| Y_{\boldsymbol{\sigma}} - X_{\boldsymbol{\sigma}}\widehat{\boldsymbol{\beta}}_{\boldsymbol{\theta},\rho,\boldsymbol{\sigma}} - Z_{\boldsymbol{\sigma\theta}}\mu_{\boldsymbol{\mathcal{U}}|\boldsymbol{\mathcal{Y}}=\boldsymbol{y}} \right\|^2 + \sigma_1^2\sigma_2^2 \mu_{\boldsymbol{\mathcal{U}}|\boldsymbol{\mathcal{Y}}=\boldsymbol{y}}^\top \Sigma_{\boldsymbol{u}}^{-1} \mu_{\boldsymbol{\mathcal{U}}|\boldsymbol{\mathcal{Y}}=\boldsymbol{y}}, \tag{1.78}$$

$L_{\boldsymbol{\theta},\rho,\boldsymbol{\sigma}}$ vérifie

$$L_{\boldsymbol{\theta},\rho,\boldsymbol{\sigma}} L_{\boldsymbol{\theta},\rho,\boldsymbol{\sigma}}^\top = Z_{\boldsymbol{\sigma\theta}}^\top Z_{\boldsymbol{\sigma\theta}} + \sqrt{\sigma_1^2\sigma_2^2}\Sigma_{\boldsymbol{u}}^{-1}, \tag{1.79}$$

et R_X vérifie

$$\begin{pmatrix} X_{\boldsymbol{\sigma}}^\top X_{\boldsymbol{\sigma}} & X_{\boldsymbol{\sigma}}^\top Z_{\boldsymbol{\sigma\theta}} \\ Z_{\boldsymbol{\sigma\theta}}^\top X_\sigma & L_{\boldsymbol{\theta},\rho,\boldsymbol{\sigma}} L_{\boldsymbol{\theta},\rho,\boldsymbol{\sigma}}^\top \end{pmatrix} = \begin{pmatrix} R_X & \mathbf{0} \\ R_{ZX} & L_{\boldsymbol{\theta},\rho,\boldsymbol{\sigma}}^\top \end{pmatrix}^\top \begin{pmatrix} R_X & \mathbf{0} \\ R_{ZX} & L_{\boldsymbol{\theta},\rho,\boldsymbol{\sigma}}^\top \end{pmatrix}. \tag{1.80}$$

En résolvant l'équation $\frac{\partial \ell(\boldsymbol{\beta},\boldsymbol{\theta},\rho,\boldsymbol{\sigma}|\boldsymbol{y})}{\partial \beta} = 0$, on obtient $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}_{\boldsymbol{\theta},\rho,\boldsymbol{\sigma}}$ qu'on re-injecte dans l'expression de $\ell(\boldsymbol{\beta}, \boldsymbol{\theta}, \rho, \boldsymbol{\sigma}|\boldsymbol{y})$ pour obtenir

TABLE 1.3 : **Erreur quadratique moyenne des estimateurs avec son intervalle de confiance à 95% estimés sur 100 replications de données simulées en prenant $n \in \{50, 60, 100, 300\}$ et $N \in \{600, 1000, 3000\}$.**

Parameter	n	$N = 600$	$N = 1000$	$N = 3000$
β_1	50	2.43 (0.11 - 7.11)	1.89 (0.22 - 4.89)	1.02 (0.07 - 2.44)
	60	2.57 (0.14 - 7.87)	2.13 (0.26 - 5.55)	0.77 (0.10 - 2.09)
	100	2.27 (0.16 - 5.61)	1.55 (0.14 - 5.17)	0.71 (0.04 - 1.85)
	300	3.16 (0.14 - 10.54)	1.70 (0.10 - 4.55)	0.51 (0.04 - 1.36)
β_2	50	5.50(0.09 - 16.35)	3.26 (0.02 - 11.64)	2.06 (0.09 - 5.98)
	60	5.06 (0.12 - 15.09)	3.22 (0.02 - 10.24)	1.78 (0.10 - 5.62)
	100	4.33 (0.03 - 13.17)	2.37 (0.02 - 6.89)	1.06 (0.02 - 3.60)
	300	4.58 (0.18 - 15.92)	2.43 (0.05 - 7.39)	0.90 (0.04 - 2.88)
σ_1	50	0.03 (0.00 - 0.14)	0.02 (0.00 - 0.09)	0.00 (0.00 - 0.02)
	60	0.04 (0.00 - 0.11)	0.02 (0.00 - 0.08)	0.00 (0.00 - 0.02)
	100	0.03 (0.00 - 0.13)	0.01 (0.00 - 0.06)	0.00 (0.00 - 0.01)
	300	0.05 (0.00 - 0.23)	0.03 (0.00 - 0.13)	0.00 (0.00 - 0.02)
σ_2	50	0.04 (0.00 - 0.15)	0.03 (0.00 - 0.08)	0.00 (0.00 - 0.02)
	60	0.04 (0.00 - 0.18)	0.03 (0.00 - 0.12)	0.01 (0.00 - 0.03)
	100	0.06 (0.00 - 0.18)	0.03 (0.00 - 0.11)	0.01 (0.00 - 0.03)
	300	0.13 (0.00 - 0.45)	0.04 (0.00 - 0.15)	0.01 (0.00 - 0.05)
$\bar{\mathbf{\Gamma}}$	50	0.90 (0.12 - 2.41)	0.62 (0.06 - 1.41)	0.45 (0.04 - 1.14)
	60	1.07 (0.06 - 2.64)	0.68 (0.08 - 1.98)	0.25 (0.03 - 0.66)
	100	0.90 (0.05 - 2.40)	0.47 (0.03 - 1.29)	0.21 (0.02 - 0.71)
	300	1.45 (0.19 - 4.39)	0.69 (0.06 - 1.94)	0.17 (0.02 - 0.57)

$$\begin{aligned}\tilde{\ell}(\boldsymbol{\theta}, \rho, \boldsymbol{\sigma}|\boldsymbol{y}) &= -\frac{r(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\theta},\rho,\boldsymbol{\sigma}}, \mu_{\boldsymbol{\mathcal{U}}|\boldsymbol{\mathcal{Y}}=\boldsymbol{y}})}{2\sigma_1^2\sigma_2^2} - \frac{N-q}{2}\log(\sigma_1^2\sigma_2^2) \\ &\quad -\frac{1}{2}\log(|\Sigma_{\boldsymbol{u}}|) - \frac{1}{2}\log(|L_{\boldsymbol{\theta},\rho,\boldsymbol{\sigma}}|^2). \end{aligned} \tag{1.81}$$

Les estimateurs du maximum de vraisemblance $\widehat{\boldsymbol{\beta}}$, $\widehat{\boldsymbol{\sigma}}$, $\widehat{\boldsymbol{\theta}}$, $\widehat{\rho}$ de $\boldsymbol{\beta}$, $\boldsymbol{\sigma}$, $\boldsymbol{\theta}$ et ρ satisfont à

$$\left(\widehat{\boldsymbol{\theta}}, \widehat{\rho}, \widehat{\boldsymbol{\sigma}}\right) = \underset{\boldsymbol{\theta},\rho,\boldsymbol{\sigma}}{\arg\max}\, \tilde{\ell}(\boldsymbol{\theta}, \rho, \boldsymbol{\sigma}|\boldsymbol{y}) \quad \text{et} \quad \widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}_{\widehat{\boldsymbol{\theta}},\widehat{\rho},\widehat{\boldsymbol{\sigma}}} \tag{1.82}$$

Soit p, p_1 et p_2 tels que $\dim(X_1) = N \times p_1$, $\dim(X_2) = N \times p_2$ et $p = p_1 + p_2$. Avec les mêmes notations, le critère REML à optimiser s'écrit

$$\mathscr{L}(\boldsymbol{\sigma}, \boldsymbol{\theta}, \rho|\boldsymbol{y}) = \frac{\exp\left[-\frac{r(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\theta},\rho,\boldsymbol{\sigma}}, \mu_{\mathcal{U}|\mathcal{Y}=\boldsymbol{y}})}{2\sigma_1^2\sigma_2^2}\right](\sigma_1^2\sigma_2^2)^{\frac{p+q-N}{2}}}{(2\pi)^{(2N-p)/2}|\Sigma_{\boldsymbol{u}}|^{1/2}|L_{\boldsymbol{\theta},\rho,\boldsymbol{\sigma}}||R_X|} \tag{1.83}$$

Sur 100 jeux de données simulées, nous avons investigué empiriquement les performances de cette stratégie d'estimation en calculant l'erreur quadratique moyenne de chaque estimateur sur les 100 réplications. Les résultats obtenus, contenus dans le tableau 1.3 montrent clairement une nette amélioration de cette erreur à tous les niveaux, comparativement aux EM-estimateurs. Quand on regarde particulièrement les paramètres de variance des effets aléatoires, on constate que la présente méthode est de loin plus précise que l'algorithme EM. Avec des jeux de données

TABLE 1.4 : **Erreur quadratique moyenne des estimateurs avec son intervalle de confiance à 95% estimés sur 100 replications de données simulées en prenant $(n, N) \in \{(500, 7000), (600, 8000), (800, 10000), (1000, 15000)\}$.**

Parameter	$n = 500, N = 7000$	$n = 600, N = 8000$	$n = 800, N = 10000$	$n = 1000, N = 15000$
β_1	0.22 (0.01 - 0.62)	0.17 (0.01 - 0.48)	0.16 (0.01 - 0.47)	0.11 (0.00 - 0.32)
β_2	0.32 (0.00 - 1.01)	0.34 (0.01 - 1.16)	0.25 (0.02 - 0.67)	0.21 (0.00 - 0.69)
σ_1	0.00 (0.00 - 0.00)	0.00 (0.00 - 0.01)	0.00 (0.00 - 0.00)	0.00 (0.00 - 0.00)
σ_2	0.00 (0.00 - 0.01)	0.00 (0.00 - 0.01)	0.00 (0.00 - 0.01)	0.00 (0.00 - 0.00)
$\mathbf{\Gamma}$	0.09 (0.00 - 0.25)	0.07 (0.00 - 0.19)	0.06 (0.00 - 0.19)	0.03 (0.00 - 0.09)

de plus grandes tailles (voir tableau 1.4), il devient très clair que cette méthode permet d'obtenir des estimateurs asymptotiquement consistants.

Contributions relatives au chapitre 4 : Procédure de sélection d'effets fixes dans le modèle linéaire à effets mixtes : Approcher les performances d'une pénalité de type L_0

La procédure de sélection des effets fixes que nous proposons dans ce chapitre se base sur l'écriture de la vraisemblance proposée par Bates et ses co-auteurs [14] avec une légère modification, en présence d'un terme de pénalité de type *adaptive ridge*. L'exposé que nous faisons ici concerne seulement les modifications apportées à la méthode, quant à la façon d'écrire la vraisemblance. Puisqu'il est question de faire de la sélection parmi les effets fixes, nous avons réécrit la vraisemblance comme fonction à la fois des effets fixes et des paramètres de variance ; ce qui n'est pas le cas de l'approche initiale où la vraisemblance apparaît comme fonction uniquement des paramètres de variance des effets aléatoires, en présence d'une mise à jour itérative de "l'expression" des effets fixes et des effets aléatoires. Avec les mêmes notations qu'à la Section 1.2.2, on a

$$\|y - X\beta - Z\Lambda_\theta u\|^2 + \|u\|^2 = \left\| \begin{pmatrix} y - X\beta \\ \mathbf{0} \end{pmatrix} - \begin{pmatrix} Z\Lambda_\theta \\ I_q \end{pmatrix} u \right\|^2 \tag{1.84}$$

$$\begin{aligned} \mu_{\mathcal{U}|\mathcal{Y}=y} &= \arg\min_u \left\| \begin{pmatrix} y - X\beta \\ \mathbf{0} \end{pmatrix} - \begin{pmatrix} Z\Lambda_\theta \\ I_q \end{pmatrix} u \right\|^2 \\ &\Longrightarrow \quad \begin{pmatrix} Z\Lambda_\theta \\ I_q \end{pmatrix}^\top \begin{pmatrix} Z\Lambda_\theta \\ I_q \end{pmatrix} \mu_{\mathcal{U}|\mathcal{Y}=y} = \begin{pmatrix} Z\Lambda_\theta \\ I_q \end{pmatrix}^\top \begin{pmatrix} y - X\beta \\ \mathbf{0} \end{pmatrix}, \end{aligned} \tag{1.85}$$

et $\|y - X\beta - Z\Lambda_\theta u\|^2 + \|u\|^2$ se réécrit comme suit

$$\|y - X\beta - Z\Lambda_\theta u\|^2 + \|u\|^2 \quad = \quad \|y - X\beta - Z\Lambda_\theta \mu_{\mathcal{U}|\mathcal{Y}=y}\|^2 + \|\mu_{\mathcal{U}|\mathcal{Y}=y}\|^2 +$$

$$
\begin{aligned}
& (u-\mu_{\mathcal{U}|\mathcal{Y}=y})^\top \underbrace{\Big((Z\Lambda_\theta)^\top Z\Lambda_\theta + I_q\Big)}_{=L_\theta^\top L_\theta}(u-\mu_{\mathcal{U}|\mathcal{Y}=y}) \\
= \quad & \|y - X\beta - Z\Lambda_\theta\mu_{\mathcal{U}|\mathcal{Y}=y}\|^2 + \|\mu_{\mathcal{U}|\mathcal{Y}=y}\|^2 + \big\|L_\theta(u-\mu_{\mathcal{U}|\mathcal{Y}=y})\big\|^2 .
\end{aligned} \tag{1.86}
$$

Ce qui fait que la vraisemblance s'écrit

$$
L(\beta,\sigma^2,\theta|y) = |L_\theta|^{-1}(2\pi\sigma^2)^{-n/2}\exp\left[-\frac{\|y-X\beta-Z\Lambda_\theta\mu_{\mathcal{U}|\mathcal{Y}=y}\|^2 + \|\mu_{\mathcal{U}|\mathcal{Y}=y}\|^2}{2\sigma^2}\right]. \tag{1.87}
$$

Le critère à optimiser pour la sélection d'effets fixes s'écrit

$$
\mathcal{C}(\beta,\sigma^2,\theta|y) = \log\big(L(\beta,\sigma^2,\theta|y)\big) - \frac{\lambda}{2}\beta^\top W\beta, \tag{1.88}
$$

où, W est une matrice de poids à mettre à jour au cours de la procédure de sélection. $\frac{\partial\mathcal{C}(\beta,\sigma^2,\theta|y)}{\partial\sigma^2} = 0 \implies \widehat{\sigma^2} = \frac{\|y-X\beta-Z\Lambda_\theta\mu_{\mathcal{U}|\mathcal{Y}=y}\|^2+\|\mu_{\mathcal{U}|\mathcal{Y}=y}\|^2}{n}$. Cette expression de $\widehat{\sigma^2}$ re-injectée dans celle de $\mathcal{C}(\beta,\sigma^2,\theta|y)$ nous fournit

$$
-2\tilde{\mathcal{C}}(\beta,\theta|y) = \log|L_\theta|^2 + n\left[1+\log\left(\frac{2\pi(\|y-X\beta-Z\Lambda_\theta\mu_{\mathcal{U}|\mathcal{Y}=y}\|^2+\|\mu_{\mathcal{U}|\mathcal{Y}=y}\|^2)}{n}\right)\right] + \lambda\beta^\top W\beta \tag{1.89}
$$

qui est le critère qui sera minimisé pour faire la sélection.

CHAPITRE 2

MULTIVARIATE LONGITUDINAL ANALYSIS WITH BIVARIATE CORRELATION TEST

Dans le cadre de l'ajustement d'un modèle linéaire à effets mixtes à des données longitudinales multidimensionnelles, nous proposons ici des expressions plus générales (que celles rencontrées dans la littérature) des estimateurs du maximum de vraisemblance des paramètres, utilisant l'algorithme EM. Ces estimateurs sont aussi directement utiles dans le cadre plus général de l'analyse de données multidimensionnelles multi-niveaux. L'idée générale de l'analyse jointe de plusieurs variables statistiques par un modèle multidimensionnel à effets mixtes, permettant de modéliser la distribution jointe de tous les effets aléatoires spécifiques à chaque dimension du modèle, est toujours applicable quel que soit le nombre de variables dépendantes. Cependant, la dimension du vecteur de tous les effets aléatoires du modèle peut très vite croitre avec le nombre de variables à l'analyse, entrainant ainsi de sérieux problèmes numériques d'optimisation. Dans ce chapitre qui a fait l'objet d'une publication dans PloS One, nous proposons de tester la dépendance entre deux dimensions du modèle par un test de rapport de vraisemblance qui porte sur la significativité globale des corrélations entre les effets aléatoires de ces dimensions. Cette méthode nous permettra de construire un modèle plus parcimonieux en terme de nombre de paramètres de variance relatifs aux effets aléatoires, à travers une procédure de sélection pas-à-pas ascendante. Ce test de corrélation bi-dimensionnel construit ici a montré ses performances dans les études de simulations que nous avons menées, mais aussi son utilité dans l'analyse de données réelles.

Multivariate Longitudinal Analysis with Bivariate Correlation Test

Eric Houngla Adjakossa[1,2 *], Ibrahim Sadissou[3,4], Mahouton Norbert Hounkonnou[2], Gregory Nuel[1]

1 Laboratoire de Probabilités et Modèles Aléatoires /Université Pierre et Marie Curie, Case courrier 188 - 4, Place Jussieu 75252 Paris cedex 05 France
2 University of Abomey-Calavi, 072 B.P. 50 Cotonou, Republic of Benin
3 Laboratoire de Biologie et de Physiologie Cellulaires /University of Abomey-Calavi, Cotonou, Republic of Benin
4 Centre d'Etude et de Recherche sur le Paludisme Associé à la Grossesse et à l'Enfance (CERPAGE), Cotonou, Republic of Benin

*** ericadjakossah@gmail.com**

Abstract

In the context of multivariate multilevel data analysis, this paper focuses on the multivariate linear mixed-effects model, including all the correlations between the random effects when the dimensional residual terms are assumed uncorrelated. Using the EM algorithm, we suggest more general expressions of the model's parameters estimators. These estimators can be used in the framework of the multivariate longitudinal data analysis as well as in the more general context of the analysis of multivariate multilevel data. By using a likelihood ratio test, we test the significance of the correlations between the random effects of two dependent variables of the model, in order to investigate whether or not it is useful to model these dependent variables jointly. Simulation studies are done to assess both the parameter recovery performance of the EM estimators and the power of the test. Using two empirical data sets which are of longitudinal multivariate type and multivariate multilevel type, respectively, the usefulness of the test is illustrated.

2.1 Introduction

In statistical studies, one often needs to analyze data with nested sources of variability: e.g., pupils in classes, employees in companies, repeated measurements in subjects, etc. [140] referred to these type of data as grouped data which are also named multilevel data, hierarchical data or nested data in the literature [72, 177, 210]. In the analysis of such data, it is usually illuminating to take account of the variability associated with each level of nesting. There is variability, e.g., between pupils but also between classes. The measurements related to a specific subject (level of nesting) can be correlated, while observations from different subjects are usually independent, and one may draw wrong conclusions if either of these sources of variability is ignored [206]. A series of works

in statistical literature focus on the analysis of univariate multilevel data (or univariate grouped data) where a single outcome of interest is analyzed [46, 110, 114, 129, 193, 204]. Such analyses are generally simple to deal with due to the availability of many software packages conceived to perform them [16, 117, 142]. In practice, many scientific questions of interest require to focus on multiple outcomes, all arising from the same multilevel study, leading to the so-called multivariate multilevel data. For example, to answer some questions of interest, [57] analyzed hearing threshold data (in the Baltimore Longitudinal Study on Aging) [174] which consisted in the longitudinal recording of 22 variables. [185] also studied the joint evolution of HIV RNA and CD4+ T lymphocytes in a cohort of HIV-1 infected patients treated with highly active antiretroviral treatment, by jointly analyzing both markers. [182] used multivariate multilevel regression analysis to investigate individual level determinants of self rated health and happiness, as well as the extent of community level covariation in health and happiness. [190] also used multivariate multilevel analysis to jointly model three commonly used indicators of fear of crime which are: feeling unsafe alone at home after dark, feeling unsafe walking alone after dark and worry about becoming a victim of crime. A variety of works were devoted to joint modeling during the last few decades (see e.g., [26, 32, 59, 183, 198]).

These analyses often require a specification of the joint density of all outcomes or, at least, the correlation structure of the data and therefore can lead to the parsimony and/or computation (optimization) problems as well as to numerical difficulties in statistical inference, when the dimension of these outcomes increases. Many analysis strategies were proposed in the statistical literature to circumvent these problems. These strategies generally consist in reducing the dimensionality of the multivariate vector of outcomes and/or in using a small number of latent variables to model correlations within these data. Joint analysis of multivariate multilevel data then requires a trade-off between the increase of the computational complexity and the gain in information.

In this work, we focus on the multivariate linear mixed-effects model, including all the correlations between the random effects along with the independent marginal (dimensional) residuals. The correlations between two dependent variables are then those from the random effects related to these dependent variables. The class of mixed-effects models considered here assumes that both the random effects and the errors (residuals) follow Gaussian distributions. These models are intended for the analysis of multivariate multilevel data in which the dependent variables are continuous.

We use the EM algorithm to estimate the parameters of the model but here, we have two novelties: 1) we suggest a general expression of EM-based estimators which can help in analyzing multivariate longitudinal data as well as the multivariate multilevel data, not of the longitudinal type, and 2) we test the significance of the correlations between the random effects of two dependent variables, using the likelihood ratio test which allows to decide if some dependent variables are significantly correlated or not. By using this bivariate correlation test, the novelty here is the illustration, through empirical data, of some of the consequences of performing separate analyses when a joint analysis is required. Two dependent variables which are found to be uncorrelated after this test will be analyzed with two independent models (or analyzed separately). This strategy may be considered as a way toward the obtaining of a more parsimonious model in high dimension without losing much information. It may also be used in a joint modeling selection procedure.

The paper is organized as follows. In Section 2, contributions of previous works are briefly presented. We also present in this section the EM-based estimators of the parameters of the multivariate linear mixed model. Simulations studies are done in Section 3 where we also discuss the power of the likelihood ratio test which allows to test the significance of the correlation between two response variables. Two illustrations on empirical data are also done in Section 3. The first, concerning bivariate two-level data, is about a study on the effects of school differences on pupils' progress in Dutch language and arithmetics in the Netherlands. The second illustration concerns a longitudinal study on the immune response to malaria of infants in Benin.

2.2 Materials and Methods

2.2.1 Previous works

In this Section, we briefly recall the framework of the multivariate multilevel analysis (see for instance, [10, 194]). We can basically distinguish two main approaches to model such data: those which specify the joint distribution of all outcomes without the use of latent structures, and the models using latent structures. We denote by $y_1, \ldots, y_m$ the m dependent vectors of interest, and $\boldsymbol{y} = (y_1^\top, \ldots, y_m^\top)^\top$.

Modeling methods without latent structures

The first approach which is that of the modeling without latent structures comprises three sub-approaches consisting in a) direct specification of the correlation structure of $\boldsymbol{y}$, b) analysis without explicit modeling of the correlation structure of $\boldsymbol{y}$ and c) conditional models.

In the case of direct specification of Cov($\boldsymbol{y}$), [68] and [135] factorized the covariance matrix of $\boldsymbol{y}$ by using the Kronecker product in order to have more parsimonious models in the context of fully balanced data. With the same idea of having a parsimonious structure, [30] specified the intra-outcome and inter-outcome correlations, respectively, as follows: $\text{Corr}(y_k(t), y_k(s)) = \exp(\alpha|s-t|^{\theta_k})$ and $\text{Corr}(y_k(t), y_{k'}(s)) = \exp\left((\alpha|s-t|+1)^{\theta_{kk'}}\right)$, with t and s indicating the time, and k and k' indicating the dimension. Although these models are useful, they are often too restrictive and may not be realistic in many applications, especially when the data, for example, in the longitudinal studies are unbalanced (i.e. the number of available measurements per subject and the time points at which the measurements were taken often differ from one subject to another). Another class of joint models, specifying directly the joint distribution of $\boldsymbol{y}$, and whose application is often not straightforward, due to unbalanced data structures is the so-called copula model [132, 175]. Denoting by $F_i, i = 1, \ldots, m$ the cumulative distribution function of the ith component of $\boldsymbol{y}$, y_i, a copula model is defined by an m-dimensional cumulative distribution function $C(u_1, \ldots, u_m)$ with uniform marginals such that $F(y_1, \ldots, y_m) = C(u_1, \ldots, u_m)$ with $(U_1, \ldots, U_m) = (F_1(y_1), \ldots, F_m(y_m))$, where F is the joint cumulative distribution of $\boldsymbol{y}$. That is, the joint distribution of $\boldsymbol{y}$ can be written in terms of its marginal distributions and a copula which describes the dependence structure between its components. While the construction of copulas is mathematically elegant, parameters estimation is often not feasible, especially in high-dimensional situations [194]. One of the rare applications of the copula-based modeling in the multivariate mul-

tilevel data analysis framework was proposed by [106], who studied the hemodynamic effect of a new antidepressant on the diastolic blood pressure, the systolic blood pressure and the heart rate of 10 healthy volunteers. They separately modeled, at first, each longitudinal series of response and used a copula to relate the marginal distributions of these responses at each observation time. In a second step, at each observation time, the conditional (on the past) distributions of each response were related using another copula describing the relationship between the corresponding variables. One of the advantages of this approach is that there is no need to use the same family of distributions for all response variables. As [121] used ARIMA process to model the error structure of earnings in a longitudinal data analysis context, time series models can also be used for modeling multivariate multilevel data in order to describe the dynamic dependence between variables and perform forecasting. The most commonly used multivariate time series model, the vector autoregressive (VAR) model which is relatively easy to estimate, is found to be similar to the multivariate multiple linear regression [187] where the errors for different response variables on the same trial are set to be correlated [99]. Other examples of VAR modeling include [188] and [189], but one drawback of the model is that the number of parameters can become very large, potentially leading to estimation problems [93].

Regarding analysis without explicit specification of $\text{Cov}(\boldsymbol{y})$, [110] proposed an extension of generalized linear models to the analysis of longitudinal data, where they introduced a class of estimating equations called generalized estimating equations (GEE). GEE estimation ensures consistent estimates of the regression parameters without specifying the joint distribution of a subject's observations. That is, GEE replaces $\mathbb{V}[\boldsymbol{y}]$ by the so-called working covariance matrix $W(\alpha)$ which depends on an unknown vector α to estimate. The related working correlation matrix, $R(\alpha)$, is also considered. Incorrect choice of $W(\alpha)$ does not affect the consistency of the regression parameters' estimators [110]. [111] discussed the use of GEEs with multivariate discrete variables, where focus was on the modeling of the marginal (dimensional) means of these variables and their pairwise associations. The extension of the GEE method to mixed continuous-discrete responses was discussed by [205] and [143]. [156] also avoided the need of explicit modeling of the covariance structure of bivariate longitudinal responses by using SUR [206] and GEE. As pointed out by [40], ambiguities concerning the definition of the working covariance matrix can result in a breakdown of the GEE-based estimation. For example in the longitudinal data analysis, if the true structure of correlation is equicorrelation, $(R_i^*)_{jk} = \rho$, and that the working structure is autoregressive, $(R_i)_{jk} = \alpha^{|j-k|}$, there is no solution for $\widehat{\alpha}$ when $-1/2 \leqslant \rho < -1/3$ [40]. This can be viewed as the major drawback of the GEE method since it can lead to the misspecification of within-subject associations in the context of longitudinal data analysis, for instance. Examples of procedures which bypass the need to explicitly model the underlying covariance structure of $\boldsymbol{y}$ include [79, 80, 156]. These procedures, generally, consist in regressing each component of $\boldsymbol{y}$ on relevant covariates of interest, followed by combination of these regression coefficients into a single global estimate of the covariates effect [10].

One way to avoid the direct specification of the joint distribution of $\boldsymbol{y}$ is to factorize it, leading to the so-called conditional models [73]. For two responses, the joint density $f(y_1, y_2)$ can be written as follows:

$$f(y_1, y_2) = f(y_1|y_2)f(y_2) = f(y_2|y_1)f(y_1) \tag{2.1}$$

The choice of the conditioning response is of course arbitrary and requires very careful reflection about plausible associations between components of $\boldsymbol{y}$. For example, in the specification of a conditional model such as $f(y_1|y_2)$, y_2 plays the role of covariate and different choices can lead to completely opposite results and conclusions [208]. In a clinical trial, for example, none of these factorizations will be of interest due to the conditioning on a post-randomization outcome which may partially attenuate the treatment effect on the other [194]. Another drawback of conditional models is that they do not directly lead to marginal inferences. Suppose that scientific interest would be in a comparison of the rate of longitudinal change in average of y_1 and y_2. The factorization $f(y_1, y_2) = f(y_1|y_2)f(y_2)$ directly allows for inferences about the marginal evolution of y_2, but the marginal expectation of y_1 requires computation of $\mathbb{E}[y_1] = \mathbb{E}\left[\mathbb{E}[y_1|y_2]\right]$, which, depending on the actual models, may be far from straightforward [194].

Modeling methods using latent structures

The second approach regarding models using latent structures can also be split in two sub-approaches including the strategy based on the reduction of the dimensionality of $\boldsymbol{y}$ and the mixed-effect models. The general idea of reducing the dimension of $\boldsymbol{y}$ is to use principal-component type analysis, or a summary function, to first reduce the dimensionality of $\boldsymbol{y}$ and then, use standard univariate multilevel models for the analysis of the principal factors or the retained summaries of $\boldsymbol{y}$ [49, 58, 84, 123, 137]. Although it is useful, simple to understand and easy to compute, this strategy of dimension reduction has some drawbacks such as the loss of information as discussed by [10] and [194]. [10] used this approach and retained the first principal-component only which explains 31% of the total variation in their data. They found out that the summary function does not have any physical significance and the inference results cannot be interpreted in terms of the effect of the covariates on the original (response) variables. They also found that the method fails to explore the association of the components of $\boldsymbol{y}$ along time, in the case of longitudinal studies. Furthermore, the method is not applicable in situations where all the components of $\boldsymbol{y}$ are not measured at the same time point [10], although a possible extension might be the use of functional principal components [148].

Regarding the mixed-effect models, [57], [151], [120], [152] and [18] proposed the use of random-effects models for multivariate longitudinal data. They pointed out that the main disadvantage of joining separate mixed models by allowing their model-specific random effects to be correlated is the increase of the dimension of the total vector of random effects with the number of outcomes, leading to computational problems. To circumvent these problems, [57] noted that all parameters in the joint model can be estimated by fitting all the bivariate models, based on

$$f(y_s, y_t) = \int\int f(y_s|\gamma_s)f(y_t|\gamma_t)f(\gamma_s, \gamma_t)d\gamma_s d\gamma_t$$

for all $m(m-1)/2$ pairs $(y_s, y_t), 1 \leq s < t \leq m$, resulting from the main multivariate model. Estimators for the main parameters are obtained by averaging over the results from fitting the $m(m-1)/2$ pairwise models. They then showed that the pseudo-likelihood theory can be used to derive the asymptotic distribution of these estimators, and used SAS procedures for mixed models [117] based on the Newton-Raphson algorithm to fit their models, following the approach in [185].

In some multilevel studies, focus is not to directly model $\boldsymbol{y}$, but a few number of latent variables which cannot be quantified directly (e.g., depression and anxiety), but through measurements of $\boldsymbol{y}$. In such situations, analysis may be conducted in two steps: the first produces the obtaining of the latent variables and the second proceeds to the joint analysis of these latent variables. For example, [6] proposed a latent factor linear mixed model to capture the joint trend over time of latent variables. the authors reduced, indeed, the high-dimensional responses to low-dimensional latent factors by the factor analysis model, and then used the multivariate linear mixed model to study the longitudinal trends of these latent factors, where the estimates have been done using the EM algorithm. To deal with missing values in multivariate longitudinal analysis using multivariate linear mixed-effects model, [161] proposed multiple imputations using Markov chain Monte Carlo, where they used EM algorithm for the parameters estimation. Here, the authors sped up the EM algorithm by analytically integrating the random effects out of the likelihood function, avoiding to treat them as missing data. [172] used EM based modeling to estimate the parameters of the multivariate linear mixed model under a SAS macro program encoded in IML.

Although the EM algorithm is known to be slow, one of the biggest advantages of this method is that it is not computationally expensive, even with a large number of response variables. In this context, our contribution is the writing of the EM-based estimators in a more general form than those used in [6], [161] and [172]. The expressions of the EM-based estimators used in this paper can easily perform any analysis in the framework of the multivariate multilevel data analysis using multivariate linear mixed-effects model.

Another technique somewhat close to those discussed in [6] is the structural equations-based techniques. For example, [19] developed linear structural equations with latent variables approach. Considering $\boldsymbol{y} = (y_1^\top, y_2^\top)^\top$, this approach can be expressed as follows: $y_i = \mu_i + G_i\eta_i; i = 1, 2$ and $\beta\eta_1 = \gamma\eta_2$, where $\eta_i, i = 1, 2$ are the latent variables, $\beta(m \times m)$ and $\gamma(m \times n)$ are coefficient matrices governing the linear relations of all variables involved in the m structural equations. $G_i, i = 1, 2$ are known matrices. The parameters of the model may be estimated by gradient and quasi-Newton methods, or a Gauss-Newton algorithm that obtains least-squares, generalized least-squares, or maximum likelihood estimates. One modeling strategy which fuses together mixed-effects model and VAR model in order to analyze multivariate multilevel data is the so-called multilevel-VAR method. For example, [25] used the multilevel-VAR model in the context of network inference in psychopathology, where they used the population standard deviation of the person-specific random effects to construct a network representing individual variability. Examples of multilevel-VAR modeling include [67] and [93].

State space models [83] which are useful to investigate the dynamical properties of latent variables can also be used to analyze multivariate multilevel data. For example, [119] introduced an extension of the basic state space model which is flexible and general in the sense of it is applicable to any time series for multiple systems.

Methods for estimating the connectivity maps containing heterogeneity may also be applied to analyze multivariate multilevel data. [70] presented the Group Iterative Multiple Model Estimation (GIMME) approach, which addresses the issue of heterogeneity (the need for individual-level maps) in effective connectivity mapping while capitalizing on shared information to arrive at group inferences. Unlike mixed-effects models, GIMME allows for the structure of the connectivity maps to be unique across individuals [70].

One can also use a nonparametric function f to handle the relationship between the components of $\boldsymbol{y}$ and the covariates [55, 153, 201]. This strategy requires also to have sufficient data per subject, in the case of multivariate longitudinal data. Other estimation strategies implemented under softwares and discussed by [42] can perhaps be extended to the multivariate analysis case, when necessary.

Let us finally point out that the software packages which can easily and accurately analyze (jointly) the data of multivariate multilevel type are extremely rare, and one arranges the data and manipulates packages primarily designed for fitting univariate models to handle their analysis. The SabreR [39] package, under the R software [145], which has been devoted to jointly fitting up to three mixed-effects models, with random intercepts only, has been recently removed from the depot. These facts prove by themselves that the analysis of multivariate multilevel data in a single framework is a challenging task. Bayesian-based approaches can be implemented using packages like R2WinBUGS [181] under the R software, and are useful but very time consuming and require a good expertise from the user who can easily be discouraged.

2.2.2 Model and notations

The model discussed here is the multivariate linear mixed-effects model (or the multivariate linear multilevel model), including all the correlations between the random effects, but the marginal residual terms are assumed to be uncorrelated. For a more general multivariate linear mixed-effects model, the dependent variables are assumed to be correlated, conditional on the random-effects. That is, the marginal residual terms are correlated. In this paper, as in many other works (see for example, [172], [161], [56] and [6]), we assume that conditional on the random-effects, the dependent variables are uncorrelated. In the context of using EM algorithm in estimating the model parameters, this assumption allows to derive the EM-based estimators for the residual variance parameters. If the dimensional residual terms are assumed to be correlated, the EM-based estimators of theirs variance parameters are not easy to deal with and we don't treat this case here. This model assumes that both the random effects and the residuals follow Gaussian distribution, and is intended for the analysis of multivariate multilevel data in which the dependent variables are continuous. For the sake of simplicity we focus on the bivariate case ($m = 2$) in most of the paper, but the generalization to higher dimensions ($m > 2$) is straightforward. The model is as follows:

$$\begin{aligned} y_1 &= X_1\beta_1 + Z_1\gamma_1 + \varepsilon_1 \\ y_2 &= X_2\beta_2 + Z_2\gamma_2 + \varepsilon_2 \end{aligned} \tag{2.2}$$

$$\boldsymbol{\gamma} = \begin{pmatrix} \gamma_1 \\ \gamma_2 \end{pmatrix} \sim \mathcal{N}\left(\mathbf{0}; \boldsymbol{\Gamma} = \begin{pmatrix} \Gamma_1 & \Gamma_{12} \\ \Gamma_{12}^\top & \Gamma_2 \end{pmatrix}\right), \tag{2.3}$$

$$\boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix} \sim \mathcal{N}\left(\mathbf{0}; \boldsymbol{\Sigma} = \begin{pmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{pmatrix}\right); \quad \boldsymbol{\gamma} \perp \boldsymbol{\varepsilon} \tag{2.4}$$

For $k \in \{1, 2\}$, β_k and γ_k denote respectively the fixed effects and the random effects vector of covariates, while ε_k is the residual component. X_k is a matrix of covariates and Z_k a covariates-based design matrix. dim(X_k)= $N_k \times p_k$ and dim(Z_k)= $N_k \times q_k$, where N_k is the total number of observations in the dimension k of the model. p_k and q_k are, respectively, the number of fixed effect related covariates and the number of random effect related covariates in the dimension k of the model. If N_k is a constant N for any k, the index k will be removed and N will denote the total number of observations in all dimensions of the model. The bold symbols represent parameters of multiple dimensions (i.e. Σ_1 concerns dimension 1 of the model while $\mathbf{\Sigma}$ concerns both dimensions).

Another way to easily understand the model is to express it using the levels of the covariate related to the random-effects. This expression (subject-based version) of the model is, generally, used in the framework of longitudinal data analysis, and lead to EM-based estimators (expressions) which are a particular case of the estimators expressions obtained in Equations 2.17, 2.18 and 2.19 (for example, see [172]). Denoting by n the total number of subjects involved in the longitudinal study, the model can be expressed as follows:

denoting by i a subject, for $i = 1, \ldots, n$

$$\begin{aligned} y_{1i} &= X_{1i}\beta_1 + Z_{1i}\gamma_{1i} + \varepsilon_{1i} \\ y_{2i} &= X_{2i}\beta_2 + Z_{2i}\gamma_{2i} + \varepsilon_{2i} \end{aligned} \tag{2.5}$$

with

$$\boldsymbol{\gamma}_i = \begin{pmatrix} \gamma_{1i} \\ \gamma_{2i} \end{pmatrix} \sim \mathcal{N}\left(\mathbf{0}; \bar{\mathbf{\Gamma}} = \begin{pmatrix} \bar{\Gamma}_1 & \bar{\Gamma}_{12} \\ \bar{\Gamma}_{12}^\top & \bar{\Gamma}_2 \end{pmatrix}\right) \tag{2.6}$$

and

$$\boldsymbol{\varepsilon}_i = \begin{pmatrix} \varepsilon_{1i} \\ \varepsilon_{2i} \end{pmatrix} \sim \mathcal{N}\left(\mathbf{0}; \bar{\mathbf{\Sigma}} = \begin{pmatrix} \sigma_1^2 \mathbf{I}_{N_{1i}} & 0 \\ 0 & \sigma_2^2 \mathbf{I}_{N_{2i}} \end{pmatrix}\right) \tag{2.7}$$

N_{1i} and N_{2i} are the dimensions of y_{1i} and y_{2i}, respectively. Here, we assume that the marginal residuals are homoscedastic ($\mathbb{V}[\varepsilon_{ki}] = \sigma_k^2 I_{N_{ki}}, k = 1, 2$), but the residual covariance matrices can be of full form as in Equation 2.4. In order to make clear the relation between the model described by Equations 2.2, 2.3 and 2.4, and its version expressed by Equations 2.5, 2.6 and 2.7, we propose below a detailed example.

Detailed example We place ourselves in the case of longitudinal data where we observe two response variables y_1 and y_2 which are respectively the weight (kg) and the size (cm) of infants according to the score (V_2) of the quality of their food as well as the quality score (V_1) of their mothers' food. Infants are $n = 3$ girls (sex = F) and boys (sex = M) who are monitored over time. The dataset is presented by Table 2.1.

Suppose that the model at each of two dimensions has one random intercept by subject (infant) and one random slope by subject in the direction of the infant's age (in months). For example, considering an identifiability constraint covering the sex variable whose level F is the reference, the bivariate linear mixed model can be written as follows:

Table 2.1: Example of data

					Response variables	
subject	age	sex	V_1	V_2	y_1	y_2
1	0	F	63.76	38.16	3.14	47.82
2	4	M	100.88	41.46	4.87	64.02
1	6	F	60.98	41.37	8.43	73.21
3	0	M	93.24	48.76	2.82	44.93
2	7	M	101.95	44.79	8.03	89.54
2	10	M	99.24	48.17	10.08	92.14
1	16	F	NA	44.79	13.96	86.12
3	9	M	88.38	47.91	8.47	86.42

$$
\begin{aligned}
y_1 &= \underbrace{(\mathbb{1}\ \mathrm{V}_1\ \text{sex=M})}_{=X_1}\beta_1 + Z_1\gamma_1 + \varepsilon_1 \\
y_2 &= \underbrace{(\mathbb{1}\ \mathrm{V}_2\ \text{sex=M})}_{=X_2}\beta_2 + Z_2\gamma_2 + \varepsilon_2
\end{aligned} \tag{2.8}
$$

where, explicitly

$$
X_1 = \begin{pmatrix} 1 & 63.76 & 0 \\ 1 & 100.88 & 1 \\ 1 & 60.98 & 0 \\ 1 & 93.24 & 1 \\ 1 & 101.95 & 1 \\ 1 & 99.24 & 1 \\ 1 & 88.38 & 1 \end{pmatrix}; X_2 = \begin{pmatrix} 1 & 38.16 & 0 \\ 1 & 41.46 & 1 \\ 1 & 41.37 & 0 \\ 1 & 48.76 & 1 \\ 1 & 44.79 & 1 \\ 1 & 48.17 & 1 \\ 1 & 44.79 & 0 \\ 1 & 47.91 & 1 \end{pmatrix} \tag{2.9}
$$

$$
Z_1 = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 4 & 0 & 0 \\ 1 & 6 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 7 & 0 & 0 \\ 0 & 0 & 1 & 10 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 9 \end{pmatrix}; Z_2 = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 4 & 0 & 0 \\ 1 & 6 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 7 & 0 & 0 \\ 0 & 0 & 1 & 10 & 0 & 0 \\ 1 & 16 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 9 \end{pmatrix} \tag{2.10}
$$

$$
\dim(X_1) = 7 \times 3 \text{ and } \dim(X_2) = 8 \times 3; \quad \dim(Z_1) = 7 \times 6 \text{ and } \dim(Z_2) = 8 \times 6
$$

In the present example we have $\dim(X_1) \neq \dim(X_2)$ and $\dim(Z_1) \neq \dim(Z_2)$ due to the presence of the NA (Not Available) within the values of the variable V_1. Removing information

related to this NA in the dimension 1 of the model does not affect its dimension 2.

$$\mathbf{\Gamma} = \begin{pmatrix} \Gamma_1 & \Gamma_{12} \\ \Gamma_{12}^\top & \Gamma_2 \end{pmatrix}$$

with,

$$\Gamma_1 = \begin{pmatrix} \eta_1^2 & \rho_\eta\eta_1\eta_2 & . & . & . & . \\ \rho_\eta\eta_1\eta_2 & \eta_2^2 & . & . & . & . \\ . & . & \eta_1^2 & \rho_\eta\eta_1\eta_2 & . & . \\ . & . & \rho_\eta\eta_1\eta_2 & \eta_2^2 & . & . \\ . & . & . & . & \eta_1^2 & \rho_\eta\eta_1\eta_2 \\ . & . & . & . & \rho_\eta\eta_1\eta_2 & \eta_2^2 \end{pmatrix} \tag{2.11}$$

$$\Gamma_2 = \begin{pmatrix} \tau_1^2 & \rho_\tau\tau_1\tau_2 & . & . & . & . \\ \rho_\tau\tau_1\tau_2 & \tau_2^2 & . & . & . & . \\ . & . & \tau_1^2 & \rho_\tau\tau_1\tau_2 & . & . \\ . & . & \rho_\tau\tau_1\tau_2 & \tau_2^2 & . & . \\ . & . & . & . & \tau_1^2 & \rho_\tau\tau_1\tau_2 \\ . & . & . & . & \rho_\tau\tau_1\tau_2 & \tau_2^2 \end{pmatrix} \tag{2.12}$$

$$\Gamma_{12} = \begin{pmatrix} \rho_{\eta_1\tau_1}\eta_1\tau_1 & \rho_{\eta_1\tau_2}\eta_1\tau_2 & . & . & . & . \\ \rho_{\eta_2\tau_1}\eta_2\tau_1 & \rho_{\eta_2\tau_2}\eta_2\tau_2 & . & . & . & . \\ . & . & \rho_{\eta_1\tau_1}\eta_1\tau_1 & \rho_{\eta_1\tau_2}\eta_1\tau_2 & . & . \\ . & . & \rho_{\eta_2\tau_1}\eta_2\tau_1 & \rho_{\eta_2\tau_2}\eta_2\tau_2 & . & . \\ . & . & . & . & \rho_{\eta_1\tau_1}\eta_1\tau_1 & \rho_{\eta_1\tau_2}\eta_1\tau_2 \\ . & . & . & . & \rho_{\eta_2\tau_1}\eta_2\tau_1 & \rho_{\eta_2\tau_2}\eta_2\tau_2 \end{pmatrix} \tag{2.13}$$

ρ_η, ρ_τ, $\rho_{\eta_1\tau_1}$, $\rho_{\eta_1\tau_2}$, $\rho_{\eta_2\tau_1}$, and $\rho_{\eta_2\tau_2}$ lie in $[-1, 1]$. All other parameters involved in Γ_1, Γ_2 and Γ_{12} are positive real numbers.

Referring to the subject-based version of the model,

$$\bar{\Gamma}_1 = \begin{pmatrix} \eta_1^2 & \rho_\eta\eta_1\eta_2 \\ \rho_\eta\eta_1\eta_2 & \eta_2^2 \end{pmatrix}, \quad \bar{\Gamma}_2 = \begin{pmatrix} \tau_1^2 & \rho_\tau\tau_1\tau_2 \\ \rho_\tau\tau_1\tau_2 & \tau_2^2 \end{pmatrix}, \quad \bar{\Gamma}_{12} = \begin{pmatrix} \rho_{\eta_1\tau_1}\eta_1\tau_1 & \rho_{\eta_1\tau_2}\eta_1\tau_2 \\ \rho_{\eta_2\tau_1}\eta_2\tau_1 & \rho_{\eta_2\tau_2}\eta_2\tau_2 \end{pmatrix},$$

$$\mathbb{V}(\gamma_1) = \bar{\Gamma}_1, \quad \mathbb{V}(\gamma_2) = \bar{\Gamma}_2, \text{ and } \mathrm{Cov}(\gamma_1, \gamma_2) = \bar{\Gamma}_{12}. \tag{2.14}$$

Then,

$$\Gamma_1 = \mathrm{diag}(\bar{\Gamma}_1, \ldots, \bar{\Gamma}_1), \quad \Gamma_2 = \mathrm{diag}(\bar{\Gamma}_2, \ldots, \bar{\Gamma}_2), \text{ and } \Gamma_{12} = \mathrm{diag}(\bar{\Gamma}_{12}, \ldots, \bar{\Gamma}_{12}). \tag{2.15}$$

2.2.3 EM estimation

Let θ be the vector of unknown parameters in $\beta_1, \beta_2, \mathbf{\Gamma}, \Sigma_1, \Sigma_2$. The EM algorithm requires an initial value of θ and some expressions (estimators) to update until convergence. In the next two subsections we provide these estimators, their initial values and the stopping criterion.

EM-based estimators of parameters

Theorem 1. *Suppose that $\boldsymbol{y} = \left(y_1^\top, y_2^\top\right)^\top$ satisfies the model based on Eq. (2.2, 2.3 and 2.4) and θ the vector of its unknown parameters while θ_{old} is the previous value of θ provided by the EM algorithm. Let $f(\boldsymbol{y}, \boldsymbol{\gamma}|\theta)$ be the joint density function of $\boldsymbol{y}$ and $\boldsymbol{\gamma}$ given θ, and $Q(\theta|\theta_{old}) = \mathbb{E}\left[\log f(\boldsymbol{y}, \boldsymbol{\gamma}|\theta)|\boldsymbol{y}, \theta_{old}\right]$. Let M be the mapping $\theta_{old} \mapsto M(\theta_{old}) = \widehat{\theta}$ such that:*

$$M(\theta_{old}) = \arg\max_{\theta} Q(\theta|\theta_{old}) \tag{2.16}$$

Then, the EM-based estimator of θ, i.e. $\widehat{\theta}$, is expressed through:
for $k \in \{1, \ldots, m\}$,

$$\widehat{\beta}_k = \left(X_k^\top \Sigma_k^{-1} X_k\right)^{-1} X_k^\top \Sigma_k^{-1} \left(y_k - Z_k \mathbb{E}[\gamma_k|y, \theta_{old}]\right), \tag{2.17}$$

$$\widehat{\mathbf{\Gamma}} = \mathbb{V}[\boldsymbol{\gamma}|\boldsymbol{y}, \theta_{old}] + \mathbb{E}[\boldsymbol{\gamma}|\boldsymbol{y}, \theta_{old}]\mathbb{E}[\boldsymbol{\gamma}|\boldsymbol{y}, \theta_{old}]^\top, \tag{2.18}$$

$$\widehat{\Sigma}_k = Z_k \mathbb{V}[\gamma_k|\boldsymbol{y}, \theta_{old}] Z_k^\top + (y_k - X_k\beta_k - Z_k\mathbb{E}[\gamma_k|\boldsymbol{y}, \theta_{old}])(y_k - X_k\beta_k - Z_k\mathbb{E}[\gamma_k|\boldsymbol{y}, \theta_{old}])^\top, \tag{2.19}$$

where,

$$\mathbb{E}[\gamma_k|\boldsymbol{y}, \theta_{old}] = Cov(\gamma_k, \boldsymbol{y}|\theta_{old})\mathbb{V}[\boldsymbol{y}|\theta_{old}]^{-1}(\boldsymbol{y} - \mathbb{E}[\boldsymbol{y}|\theta_{old}]), \tag{2.20}$$

$$\mathbb{V}[\gamma_k|\boldsymbol{y}, \theta_{old}] = \mathbf{\Gamma}_k - Cov(\gamma_k, \boldsymbol{y}|\theta_{old})\mathbb{V}[\boldsymbol{y}|\theta_{old}]^{-1} Cov(\gamma_k, \boldsymbol{y}|\theta_{old})^\top \tag{2.21}$$

and

$$\mathbb{V}[\boldsymbol{y}|\theta_{old}] = \begin{pmatrix} Z_1\Gamma_1 Z_1^\top + \Sigma_1 & Z_1\Gamma_{12}Z_2^\top \\ Z_2\Gamma_{12}^\top Z_1^\top & Z_2\Gamma_2 Z_2^\top + \Sigma_2 \end{pmatrix}, \quad Cov(\boldsymbol{\gamma}, \boldsymbol{y}|\theta_{old}) = \begin{pmatrix} Z_1\Gamma_1 & Z_1\Gamma_{12} \\ Z_2\Gamma_{12}^\top & Z_2\Gamma_2 \end{pmatrix}^\top, \tag{2.22}$$

$$Cov(\gamma_1, \boldsymbol{y}|\theta_{old}) = \begin{pmatrix} Z_1\Gamma_1 \\ Z_2\Gamma_{12}^\top \end{pmatrix}^\top, \quad Cov(\gamma_2, \boldsymbol{y}|\theta_{old}) = \begin{pmatrix} Z_1\Gamma_{12} \\ Z_2\Gamma_2 \end{pmatrix}^\top. \tag{2.23}$$

Proof. For $k \in \{1, \ldots, m\}$, $\widehat{\beta}_k$, $\widehat{\Sigma}_k$ and $\widehat{\mathbf{\Gamma}}$ optimize the quantity:

$$Q(\theta|\theta_{\text{old}}) = \mathbb{E}\left[\log f(\boldsymbol{y}, \boldsymbol{\gamma}|\theta)|\boldsymbol{y}, \theta_{\text{old}}\right] \tag{2.24}$$

where $f(\boldsymbol{y}, \boldsymbol{\gamma}|\theta)$ is the joint density function of the observed data $\boldsymbol{y}$ and the random effect $\boldsymbol{\gamma}$. In the case of $m = 2$, we have:

$$\begin{aligned} f(\boldsymbol{y}, \boldsymbol{\gamma}|\theta) &= f(\boldsymbol{y}|\boldsymbol{\gamma}, \theta) f(\boldsymbol{\gamma}|\theta) \\ &= f(y_1|\gamma_1, \theta) f(y_2|\gamma_2, \theta) f(\boldsymbol{\gamma}|\theta) \\ &= (2\pi)^{-(N_1+N_2+q)/2} |\Sigma_1|^{-1/2} |\Sigma_2|^{-1/2} |\mathbf{\Gamma}|^{-1/2} \exp\Big\{ -\frac{1}{2}\, \boldsymbol{\gamma}^\top \mathbf{\Gamma}^{-1} \boldsymbol{\gamma} \\ &\quad -\frac{1}{2}(y_1 - X_1\beta_1 - Z_1\gamma_1)^\top \Sigma_1^{-1}(y_1 - X_1\beta_1 - Z_1\gamma_1) \\ &\quad -\frac{1}{2}(y_2 - X_2\beta_2 - Z_2\gamma_2)^\top \Sigma_2^{-1}(y_2 - X_2\beta_2 - Z_2\gamma_2) \,\Big\} \end{aligned} \tag{2.25}$$

Since f is a multivariate Gaussian, using the dominated convergence theorem and the derivative under the integral sign, the differential of $Q(\theta|\theta_{\text{old}})$ yields:

$$\begin{aligned} \mathrm{d}Q(\theta|\theta_{\text{old}}) &= \mathbb{E}\,[\, -\frac{1}{2}\,\mathrm{tr}\left\{\Sigma_1^{-1}\mathrm{d}\Sigma_2 + \Sigma_2^{-1}\mathrm{d}\Sigma_2 + \mathbf{\Gamma}^{-1}\mathrm{d}\mathbf{\Gamma}\right\} \\ &\quad -\frac{1}{2}\,\mathrm{tr}\,\{\, -2(y_1 - X_1\beta_1 - Z_1\gamma_1)^\top \Sigma_1^{-1} X_1 \mathrm{d}\beta_1 \\ &\quad -\Sigma_1^{-1}(y_1 - X_1\beta_1 - Z_1\gamma_1)(y_1 - X_1\beta_1 - Z_1\gamma_1)^\top \Sigma_1^{-1}\mathrm{d}\Sigma_1 \,\} \\ &\quad -\frac{1}{2}\,\mathrm{tr}\,\{\, -2(y_2 - X_2\beta_2 - Z_2\gamma_2)^\top \Sigma_2^{-1} X_2 \mathrm{d}\beta_2 \\ &\quad -\Sigma_2^{-1}(y_2 - X_2\beta_2 - Z_2\gamma_2)(y_2 - X_2\beta_2 - Z_2\gamma_2)^\top \Sigma_2^{-1}\mathrm{d}\Sigma_2 \,\} \\ &\quad -\frac{1}{2}\,\mathrm{tr}\left\{-\boldsymbol{\gamma}^\top \mathbf{\Gamma}^{-1}\mathrm{d}\mathbf{\Gamma}\mathbf{\Gamma}^{-1}\boldsymbol{\gamma}\right\} |\boldsymbol{y}, \theta_{\text{old}}\,] \end{aligned} \tag{2.26}$$

$$\begin{aligned} &= -\frac{1}{2}\,\mathrm{tr}\left\{\Sigma_1^{-1}\mathrm{d}\Sigma_2 + \Sigma_2^{-1}\mathrm{d}\Sigma_2 + \mathbf{\Gamma}^{-1}\mathrm{d}\mathbf{\Gamma}\right\} \\ &\quad + \mathrm{tr}(y_1 - X_1\beta_1 - Z_1\mathbb{E}\left[\gamma_1|\boldsymbol{y}, \theta_{\text{old}}\right])^\top \Sigma_1^{-1} X_1 \mathrm{d}\beta_1 \\ &\quad +\frac{1}{2}\,\mathrm{tr}\left\{\Sigma_1^{-1}\mathbb{E}[(y_1 - X_1\beta_1 - Z_1\gamma_1)^\top (y_1 - X_1\beta_1 - Z_1\gamma_1)|\boldsymbol{y}, \theta_{\text{old}}]\Sigma_1^{-1}\mathrm{d}\Sigma_1\right\} \\ &\quad + \mathrm{tr}(y_2 - X_2\beta_2 - Z_2\mathbb{E}\left[\gamma_2|\boldsymbol{y}, \theta_{\text{old}}\right])^\top \Sigma_2^{-1} X_2 \mathrm{d}\beta_2 \\ &\quad +\frac{1}{2}\,\mathrm{tr}\left\{\Sigma_2^{-1}\mathbb{E}[(y_2 - X_2\beta_2 - Z_2\gamma_2)^\top (y_2 - X_2\beta_2 - Z_2\gamma_2)|\boldsymbol{y}, \theta_{\text{old}}]\Sigma_2^{-1}\mathrm{d}\Sigma_2\right\} \\ &\quad + \mathrm{tr}\,\frac{1}{2}\left\{\mathbf{\Gamma}^{-1}\mathbb{E}[\boldsymbol{\gamma}\boldsymbol{\gamma}^\top|\boldsymbol{y}, \theta_{\text{old}}]\mathbf{\Gamma}^{-1}\mathrm{d}\mathbf{\Gamma}\right\} \end{aligned} \tag{2.27}$$

Partial derivatives of $Q(\theta|\theta_{\text{old}})$ yield:
for $k \in \{1, \ldots, m\}$,

$$\frac{\partial Q(\theta|\theta_{\text{old}})}{\partial \beta_k} = (y_k - X_k\beta_k - Z_k\mathbb{E}\left[\gamma_k|\boldsymbol{y}, \theta_{\text{old}}\right])^\top \Sigma_k^{-1} X_k,$$

$$\frac{\partial Q(\theta|\theta_{\text{old}})}{\partial \Sigma_k} = -\frac{1}{2}\Sigma_k^{-1} + \Sigma_k^{-1}\mathbb{E}[(y_k - X_k\beta_k - Z_k\gamma_k)^\top (y_k - X_k\beta_k - Z_k\gamma_k)|\boldsymbol{y}, \theta_{\text{old}}]\Sigma_k^{-1}$$

and

$$\frac{\partial Q(\theta|\theta_{\text{old}})}{\partial \boldsymbol{\Gamma}} = \frac{1}{2}\left(\boldsymbol{\Gamma}^{-1}\mathbb{E}[\boldsymbol{\gamma}\boldsymbol{\gamma}^\top|\boldsymbol{y}, \theta_{\text{old}}]\boldsymbol{\Gamma}^{-1} - \boldsymbol{\Gamma}^{-1}\right).$$

We then get EM-based estimators by setting these partial derivatives equal to zero. $\mathbb{E}[\gamma_k|\boldsymbol{y}, \theta_{\text{old}}]$ and $\mathbb{V}[\gamma_k|y, \theta_{\text{old}}]$ are straightforward to get since $(\gamma_k^\top, \boldsymbol{y}^\top)^\top$ is a multivariate Gaussian. □

Initialization and stopping criterion of the algorithm Various ways exist for obtaining starting values for $\widehat{\beta}_k$, $\widehat{\boldsymbol{\Gamma}}$, $\widehat{\Sigma}_k$, for $k = 1, \ldots, m$. Taking inspiration from [103] and [172], we have separately fitted each dimension of the model by using the lme4 package [15] under the R software and have used marginal estimated parameters to initialize $\widehat{\beta}_k$ and $\widehat{\Sigma}_k$. We then keep the expected random effects $\tilde{\boldsymbol{\gamma}} = \widehat{\mathbb{E}}[\boldsymbol{\gamma}|\boldsymbol{y}]$ to initialize $\widehat{\boldsymbol{\Gamma}}$ by

$$\frac{1}{n-1}\sum_{i=1}^{n}\tilde{\boldsymbol{\gamma}}_i\tilde{\boldsymbol{\gamma}}_i^\top \tag{2.28}$$

The stopping criterion is related to the relative error of the components of θ as follows:

$$\max_j \left|\frac{\theta_j^{(r)} - \theta_j^{(r+1)}}{\theta_j^{(r+1)}}\right| < \text{tol} \tag{2.29}$$

where (r) is the iteration index and θ_j the jth component of θ. tol $= 10^{-5}$ seems to work well in practice.

2.2.4 Test of the significance of $\widehat{\text{Cor}}(\gamma_1, \gamma_2)$

After the calculation of $\boldsymbol{\Gamma}$ on dataset, we sometimes need to investigate if the correlation between marginal random effects is statistically significant, by testing H_0: Cor$(\gamma_1, \gamma_2) = \mathbf{0}$ against H_1 : Cor$(\gamma_1, \gamma_2) \neq \mathbf{0}$. The result of this test can help to decide if the bivariate analysis is justified or not. We perform the likelihood ratio (LR) test to choose between H_0 and H_1. We calculated S, the statistic of the likelihood ratio test.

$$S = -2\log\left(\frac{L(\theta|\text{data}, H_0)}{L(\theta|\text{data}, H_1)}\right) \tag{2.30}$$

where $L(\theta|\text{data}, H_0)$ and $L(\theta|\text{data}, H_1)$ are the likelihood of θ under H_0 and H_1, respectively. Under suitable and standard conditions, $S \sim \chi^2(df)$, asymptotically, under H_0 [200]. With df the difference in the number of parameters between $L(\theta|\text{data}, H_0)$ and $L(\theta|\text{data}, H_1)$.

2.3 Results and Discussion

2.3.1 Simulation studies

In this section, simulation studies are used to investigate the computational properties of the EM-based estimators. For the sake of simplicity, these simulation studies are conducted using simulated bivariate longitudinal data sets. Through these studies, we pursue two objectives: the first is to assess the accuracy of parameter estimates and the second is to analyze the power of the likelihood ratio test performed via these EM-based estimators. In the following paragraph, we explain how we choose the parameters that have been used to simulate the working longitudinal data sets.

The working data sets We suppose that we are following up a sample of subjects where the goal is to evaluate how the growth of the weight and the height of the individuals of this population are jointly explained by the sex, the score of nutrition (Nscore) and the age. We randomly choose through a uniform distribution the score of nutrition between 20 and 50, and the age between 18 and 37, using the R software. All the analysis in this paper are done using the R software. The subject's sex is also randomly chosen. The model under which we simulate the data sets is the following:
n indicating the total number of subjects, for $i = 1, \ldots, n$

$$\begin{aligned} \text{weight}_i &= (\mathbb{1}_{n_i}, \text{sex}_i, \text{Nscore}_i, \text{age}_i)\beta_1 + (\mathbb{1}_{n_i}, \text{Nscore}_i)\gamma_{1i} + \varepsilon_{1i} \\ \text{height}_i &= (\mathbb{1}_{n_i}, \text{sex}_i, \text{Nscore}_i, \text{age}_i)\beta_2 + (\mathbb{1}_{n_i}, \text{Nscore}_i)\gamma_{2i} + \varepsilon_{2i} \end{aligned} \tag{2.31}$$

with

$$\gamma_i = \begin{pmatrix} \gamma_{1i} \\ \gamma_{2i} \end{pmatrix} \sim \mathcal{N}\left(\mathbf{0}, \bar{\mathbf{\Gamma}}\right), \varepsilon_{1i} \sim \mathcal{N}\left(0, \sigma_1^2 \mathrm{I}_{n_i}\right), \varepsilon_{2i} \sim \mathcal{N}\left(0, \sigma_2^2 \mathrm{I}_{n_i}\right), \gamma_i \perp \varepsilon_{1i} \perp \varepsilon_{2i} \tag{2.32}$$

The random effect related to the dependent variable 'weight' or 'height' is a vector composed by one random intercept and one random slope in the direction of the covariate 'Nscore'. The total number of observations is denoted by N.

We randomly choose $\beta_1, \beta_2, \sigma_1$ and σ_2 whose values are in the first column of Table 2.2. $\bar{\mathbf{\Gamma}}$ is also randomly chosen such that it is positive definite, with the following form:

$$\bar{\mathbf{\Gamma}} = \begin{pmatrix} \eta_1^2 & \rho_\eta \eta_1 \eta_2 & \rho\eta_1\tau_1 & \rho\eta_1\tau_2 \\ \rho_\eta \eta_1 \eta_2 & \eta_2^2 & \rho\eta_2\tau_1 & \rho\eta_2\tau_2 \\ \rho\eta_1\tau_1 & \rho\eta_2\tau_1 & \tau_1^2 & \rho_\tau\tau_1\tau_2 \\ \rho\eta_1\tau_2 & \rho\eta_2\tau_2 & \rho_\tau\tau_1\tau_2 & \tau_2^2 \end{pmatrix} \tag{2.33}$$

The covariance between the random effects γ_1 and γ_2 is set, intentionally,

$$Cov(\gamma_1, \gamma_2) = \rho \begin{pmatrix} \eta_1\tau_1 & \eta_1\tau_2 \\ \eta_2\tau_1 & \eta_2\tau_2 \end{pmatrix} \tag{2.34}$$

Table 2.2: **Comparative table of true values of parameters and estimates based on** 1000 **replications using true values of parameters.**

Parameter	Value	Empirical mean	Empirical Sd	Bias
β_1	50.67	50.669	0.763	0.000
	-4.80	-4.779	0.811	0.021
	14.00	14.012	0.345	0.012
	2.70	2.700	0.016	0.000
β_2	13.20	13.263	1.077	0.063
	-2.80	-2.796	1.186	0.003
	27.00	27.000	0.068	0.000
	1.70	1.699	0.019	0.000
σ_1	5.80	5.796	0.062	0.003
σ_2	7.60	7.602	0.082	0.002

in order to be able to decrease or increase the correlation between the marginal random effects γ_1 and γ_2, by changing the value of ρ, without losing the positive definiteness of $\bar{\mathbf{\Gamma}}$. This property of $\bar{\mathbf{\Gamma}}$ will be used to assess the power of the likelihood ratio test through simulations, by changing the value of ρ. We simulate 1000 data sets with $\rho = 0.8$, in order to assess the accuracy of estimates using the EM-based estimators. With $\rho = 0.8$, the randomly chosen $\bar{\mathbf{\Gamma}}$ is

$$\bar{\mathbf{\Gamma}} = \begin{pmatrix} 27.77 & 18.80 & 41.70 & 4.93 \\ 18.80 & 36.00 & 47.47 & 5.62 \\ 41.70 & 47.47 & 97.81 & 8.91 \\ 4.93 & 5.62 & 8.91 & 1.37 \end{pmatrix} \tag{2.35}$$

Empirical accuracy of the estimates The 1000 data sets simulated in order to assess the accuracy of the estimates performed using the EM-based estimators contain $N = 5000$ observations provided by $n = 300$ independent subjects.

The mean and the standard deviation of the 1000 estimates are presented, respectively, in the second and the third column of the Table 2.2. The bias of the parameter estimates, which is the absolute difference between the true value of the parameter and the mean of the 1000 estimates, is calculated as measure of performance. These bias are contained in the forth column of the Table 2.2.

$\widehat{\mathbb{E}}[\widehat{\bar{\mathbf{\Gamma}}}]$, $\widehat{\sigma}_{\widehat{\bar{\mathbf{\Gamma}}}}$ and Bias$(\widehat{\bar{\mathbf{\Gamma}}})$ (Equations 2.36, 2.37) contain, respectively, the empirical mean, the empirical standard deviation and the empirical bias of $\bar{\mathbf{\Gamma}}$.

$$\widehat{\mathbb{E}}[\widehat{\bar{\mathbf{\Gamma}}}] = \begin{pmatrix} 27.86 & 18.71 & 41.19 & 4.93 \\ 18.71 & 35.73 & 47.09 & 5.58 \\ 41.19 & 47.09 & 95.93 & 8.87 \\ 4.93 & 5.58 & 8.87 & 1.36 \end{pmatrix}; \quad \widehat{\sigma}_{\widehat{\bar{\mathbf{\Gamma}}}} = \begin{pmatrix} 5.32 & 2.99 & 6.37 & 0.63 \\ 2.98 & 2.86 & 5.11 & 0.50 \\ 6.37 & 5.11 & 13.73 & 0.97 \\ 0.63 & 0.50 & 0.97 & 0.11 \end{pmatrix} \tag{2.36}$$

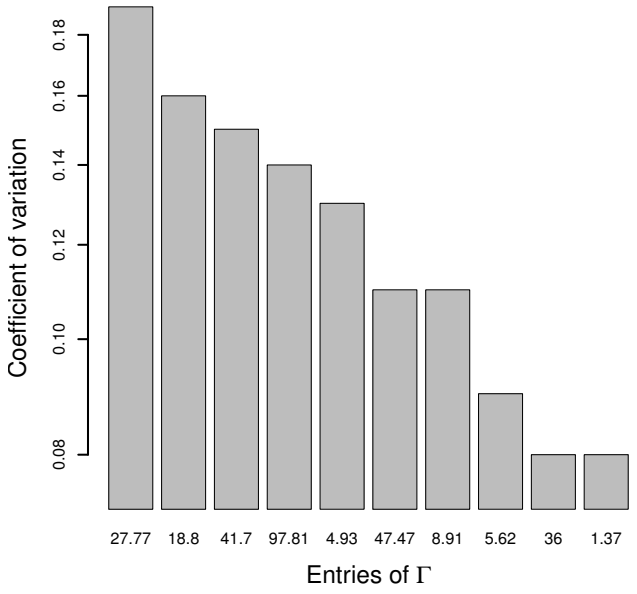

Figure 2.1: **Coefficients of variation of entries of $\mathbf{\Gamma}$**. $N = 5000$ observations and $n = 300$ subjects.

$$\text{Bias}(\widehat{\bar{\mathbf{\Gamma}}}) \begin{pmatrix} 0.085 & 0.090 & 0.515 & 0.001 \\ 0.089 & 0.265 & 0.382 & 0.045 \\ 0.514 & 0.382 & 1.881 & 0.039 \\ 0.001 & 0.045 & 0.039 & 0.007 \end{pmatrix} \tag{2.37}$$

The bias contained in the estimates of β_k and σ_k ranges from 0.000 to 0.063 (Table 2.2), and the bias contained in the estimates of the entries of $\bar{\mathbf{\Gamma}}$ ranges from 0.001 to 1.881 (Equation 2.37). These results show that $\widehat{\beta}_k$ and $\widehat{\sigma}_k$ (i.e. $\widehat{\Sigma}_k$) seem unbiased when $\widehat{\bar{\mathbf{\Gamma}}}$ is biased.

The estimates of $\bar{\mathbf{\Gamma}}$ appear to be poorer than the estimates of all other parameters. In order to investigate which entries of $\bar{\mathbf{\Gamma}}$ are particularly poorly estimated, we calculate the coefficients of variation (CV) of these entries. The CV computed here is obtained by dividing the standard deviation of the estimates by the true value of each entry of $\bar{\mathbf{\Gamma}}$. The CVs give an idea of the variability of estimates around the true values and enable to compare these variabilities between them. A particularly large value of CV could lead us to suspect that the corresponding input is particularly poorly estimated. Here, the CV ranges from 0.08 to 0.19, and is represented by the Fig 2.1 for more visibility. Given these CV values, it seems that none of the entries of $\bar{\mathbf{\Gamma}}$ is particularly poorly estimated.

Deep investigation on the estimates' accuracy Here, we compute the Mean Square Error (MSE) of the EM-based estimators with $N = 600, 1000$ and $N = 3000$ across $n = 50, 60, 100$ and 300 to investigate how both values of n and N affect the quality of the estimates. For each value of n

Table 2.3: **Mean Square Error of EM-based estimator with** 95% **CI estimated on** 1000 **replications for various values of** n **and** N**.**

Parameter	n	$N = 600$	$N = 1000$	$N = 3000$
β_1	50	1.27 (0.01 - 4.64)	1.10 (0.00 - 3.83)	0.71 (0.00 - 2.57)
	60	1.38 (0.00 - 5.56)	0.99 (0.00 - 3.90)	0.65 (0.00 - 2.44)
	100	1.31 (0.00 - 4.79)	0.87 (0.00 - 3.41)	0.46 (0.00 - 1.58)
	300	1.64 (0.00 - 5.93)	0.80 (0.00 - 3.23)	0.29 (0.00 - 1.13)
β_2	50	2.13 (0.01 - 7.22)	1.73 (0.00 - 6.21)	1.08 (0.00 - 4.17)
	60	2.12 (0.00 - 8.06)	1.63 (0.01 - 6.66)	0.95 (0.00 - 3.52)
	100	1.88 (0.00 - 7.14)	1.29 (0.00 - 4.76)	0.62 (0.00 - 2.24)
	300	2.30 (0.00 - 8.54)	1.23 (0.00 - 4.57)	0.43 (0.00 - 1.65)
σ_1	50	0.03 (0.00 - 0.10)	0.10 (0.00 - 0.07)	0.01 (0.00 - 0.02)
	60	0.04 (0.00 - 0.14)	0.02 (0.00 - 0.07)	0.01 (0.00 - 0.02)
	100	0.04 (0.00 - 0.14)	0.02 (0.00 - 0.08)	0.01 (0.00 - 0.02)
	300	0.06 (0.00 - 0.22)	0.03 (0.00 - 0.11)	0.01 (0.00 - 0.02)
σ_2	50	0.05 (0.00 - 0.18)	0.03 (0.00 - 0.12)	0.01 (0.00 - 0.03)
	60	0.06 (0.00 - 0.21)	0.03 (0.00 - 0.12)	0.01 (0.00 - 0.03)
	100	0.06 (0.00 - 0.24)	0.04 (0.00 - 0.15)	0.01 (0.00 - 0.03)
	300	0.09 (0.00 - 0.36)	0.04 (0.00 - 0.18)	0.01 (0.00 - 0.05)
$\bar{\mathbf{\Gamma}}$	50	400.91 (1.52 - 1274.32)	670.46 (2.36 - 2536.73)	489.07 (2.02 - 1840.60)
	60	706.45 (2.58 - 2497.42)	620.34 (2.70 - 2293.58)	408.79 (0.85 - 1597.77)
	100	701.50 (3.79 - 2747.70)	477.55 (2.08 - 1839.26)	283.25 (1.97 - 1023.77)
	300	798.28 (3.65 - 2603.65)	547.83 (3.03 - 1721.46)	199.54 (0.81 - 736.08)

and N, we simulate 1000 data sets on which we estimate the model parameters and compute the MSE of these estimates.

Without surprise, Table 2.3 shows that the quality of estimates is clearly improved when both n and N grow. Estimations performed on dataset containing $N = 3000$ observations are more accurate than those performed with $N = 600$, observing the maximum value of the MSE in each case. For $N = 600$, information contained in Table 2.3 shows that the MSE related to $n = 60$ (60 subjects) are better than those related to $n = 300$. This result suggests a good tradeoff between the number of subjects and the total number of observations in order to have accurate estimates, especially if the number of observations is not very high. Once again, it appears that $\widehat{\mathbf{\Gamma}}$ (Table 2.3) has the highest MSE for all values of n and N.

The bivariate likelihood ratio test Considering the random effects covariance matrix $\bar{\mathbf{\Gamma}}$ (see Equation 3.44), the related correlation matrix is

$$\begin{pmatrix} 1.00 & 0.59 & 0.80 & 0.80 \\ 0.59 & 1.00 & 0.80 & 0.80 \\ 0.80 & 0.80 & 1.00 & 0.77 \\ 0.80 & 0.80 & 0.77 & 1.00 \end{pmatrix}. \tag{2.38}$$

That is, the matrix of the correlations between the marginal random effects (i.e., the random effects related to the two dependent variables) is

$$\text{Cor}(\gamma_1, \gamma_2) = \begin{pmatrix} 0.80 & 0.80 \\ 0.80 & 0.80 \end{pmatrix} \tag{2.39}$$

whereas the estimate (on one of the previous simulated data) of this matrix, $\text{Cor}(\gamma_1, \gamma_2)$, of the correlations between the marginal random effects, is

$$\widehat{\text{Cor}}(\gamma_1, \gamma_2) = \begin{pmatrix} 0.77 & 0.78 \\ 0.90 & 0.74 \end{pmatrix} \tag{2.40}$$

If we decide to test H_0: $\text{Cor}(\gamma_1, \gamma_2) = \mathbf{0}$ against H_1: $\text{Cor}(\gamma_1, \gamma_2) \neq \mathbf{0}$ in the case of these simulated data, we must know the distribution of the LR statistic S. In order to approximate the distribution of S, under H_0, we proceed to an extensive simulation study in the next paragraph.

Empirical distribution of S under H_0 In this paragraph, our goal is to investigate about the empirical law of the LR statistic S, under H_0, when the size N of the data set increases. The simulated data sets used in this paragraph are also of bivariate longitudinal type, with N the total number of observations coming from n subjects. We choose N as an arithmetic sequence ranging from 50 to 2000, where the common difference is 50. We choose $n = N/5$ as it is sufficient to have two observations per subject for fitting the model. When $N/n = 1$, the random-effects parameters and the residual variance are unidentifiable [140].

The expected (standard) asymptotic distribution of S, under H_0, is a $\chi^2(4)$. This may be explained by the fact that $\text{Cov}(\gamma_1, \gamma_2)$ and its transpose, $\text{Cov}(\gamma_1, \gamma_2)^\top$, contain four entries, respectively, and $\bar{\mathbf{\Gamma}}$ contains $\text{Cov}(\gamma_1, \gamma_2)$ and $\text{Cov}(\gamma_1, \gamma_2)^\top$. Therefore, the difference between the number of entries of $\bar{\mathbf{\Gamma}}$ which need to be estimated with $L(\theta|\text{data}, H_0)$ and $L(\theta|\text{data}, H_1)$, respectively, is $df = 4$. Precisely, the parameters of interest are $\rho_{\eta_1\tau_1}$, $\rho_{\eta_1\tau_2}$, $\rho_{\eta_2\tau_1}$ and $\rho_{\eta_2\tau_2}$ (see Equation 2.14).

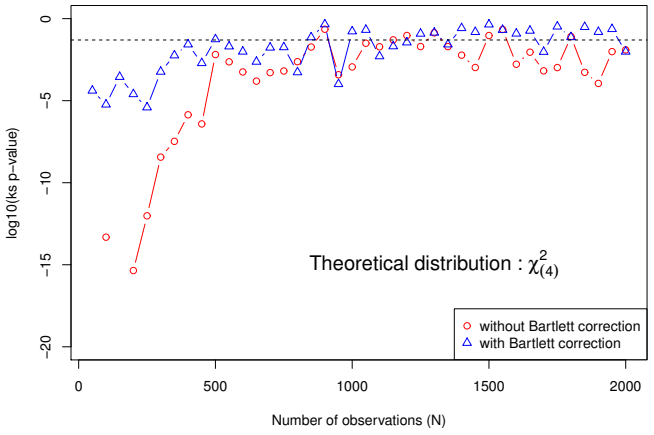

Figure 2.2: Empirical analysis of the asymptotic distribution of the LR statistic S under H_0, using longitudinal data sets (200 replications) with size $N \in \{50, 100, 150, 200, \ldots, 2000\}$ coming from $n \in \{10, 20, 30, \ldots, 400\}$ subjects. An asymptotic distribution of $\chi^2(4)$ is assumed and the Kolmogorov-Smirnov test's p-value (at log10 scale) is ploted against the total number of observations of the data set that has served to compute the LR statistic S. The blue curve is obtained by applying the empirical Bartlett correction to S and the red curve is obtained without correction. The horizontal dashed line represents $\log 10(0.05)$.

Fig 2.2 assumes an asymptotic distribution of $\chi^2(4)$ and plots the Kolmogorov-Smirnov test's p-value (at log10 scale) against the total number of observations of the data set that has served to compute the LR statistic S. The blue curve is obtained by applying the empirical Bartlett correction to S and the red curve is obtained without correction. The horizontal dashed line represents $\log 10(0.05)$. The empirical Bartlett corrected S, say $\widehat{S}_B$, can be expressed as $\widehat{S}_B = df \times S/\widehat{\mathbb{E}}[S|H_0]$. This Bartlett correction is applied in order to avoid the small size distortion of the $\chi^2(df)$ distribution, when performing the LR test using a data set of small size [12]. Fig 2.2 thus helps to investigate how the LR distribution performs in finite and small dimension. It also helps to investigate, in the case of this bivariate correlation test, how the Bartlett correction helps to avoid the small size distortion of the chi-square approximation. As the total number of observations increases, the curves (red and blue) reach the dashed line, gradually. Assuming the $\chi^2(4)$ distribution of S, it seems important to work with a data set containing at least 500 observations coming from at least 2 subjects, and to apply the Bartlett correction in order to avoid the breakdown of the procedure.

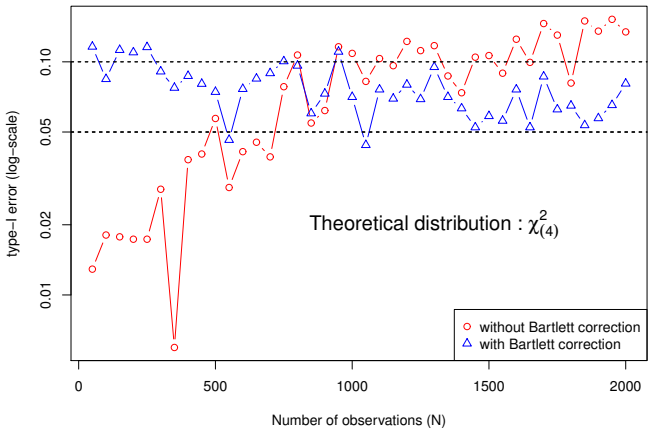

Figure 2.3: Empirical analysis of the asymptotic distribution of the LR statistic S under H_0, using longitudinal data sets (200 replications) with size $N \in \{50, 100, 150, 200, \ldots, 2000\}$ coming from $n \in \{10, 20, 30, \ldots, 400\}$ subjects. An asymptotic distribution of $\chi^2(4)$ is assumed and the type I error (at log10 scale) is ploted against the total number of observations of the data set that has served to compute the LR statistic S. The blue curve is obtained by applying the empirical Bartlett correction to S and the red curve is obtained without correction. The horizontal dashed lines represent the significance levels of 5% and 10%, respectively.

The type I error is generally controlled by the significance level of 10% (red and blue curves of Fig 2.3). It is clear that the control is almost full with the Bartlett correction (blue curve of Fig 2.3).

By simulating 1000×3000 realizations of $\chi^2(4)$ distribution, we plot the red sheath represented in Fig 2.4. This sheath corresponds to the minimum and the maximum of the simulated $\chi^2(4)$ realizations. The blue curve inside the red sheath represents the empirical LR statistics obtained from the 3000 simulated data sets under H_0. This figure (Fig 2.4) shows that the asymptotic

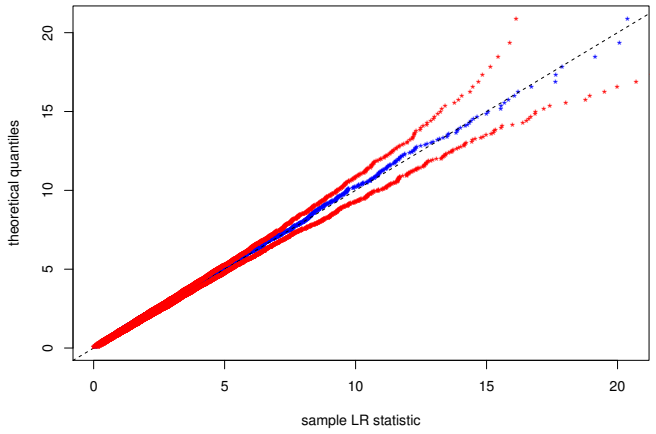

Figure 2.4: Empirical analysis of the asymptotic distribution of the LR statistic S under H_0, using 3000 (replications) simulated longitudinal data sets (under H_0) of size $N = 15000$ coming from $n = 500$ subjects. The minimum and the maximum of 1000×3000 simulated realizations of $\chi^2(4)$ are used to construct the red sheath. The blue curve represents the LR statistics related to the bivariate correlation test.

distribution of LR statistic related to the bivariate correlation test is not violated, since the blue curve does not go out of the red sheath.

Empirical power of the bivariate correlation test In order to analyze the power of this likelihood ratio test performed with EM-based estimates, we calculate S on data sets which have been simulated under H_0 and H_1, respectively, leading to what we named $S0$ and $S1$ vectors containing the resulting values of S. We then plot a ROC curve with $S0$ and $S1$, where $S0$ is the vector of the cases while S1 contains the controls. We calculate $S0$ and $S1$ in different situations where we have changed the value of ρ in the following configuration:

$$Cov(\gamma_1, \gamma_2) = \rho \begin{pmatrix} \eta_1\tau_1 & \eta_1\tau_2 \\ \eta_2\tau_1 & \eta_2\tau_2 \end{pmatrix} \tag{2.41}$$

We maintain fixed $\eta_1 = 5.27$, $\eta_2 = 6.00$, $\tau_1 = 9.89$, $\tau_2 = 1.17$ and change ρ ($\in \{0.1, 0.2, 0.3, \ldots, 0.9\}$). The number of subjects (n) and the total number of observations (N) have also been modified throughout these simulation studies. In each case, the estimated Area Under Curve (AUC) of the ROC curve with its confidence interval have been recorded to produce Fig 2.5.

With $n = 50$ subjects, we detect, indeed, a correlation of 0.6 when the total number of observations is $N = 3000$; in contrast, if the total number of observations is $N = 600$, we perfectly detect a correlation of 0.7.

Unsurprisingly, confidence intervals of AUC are also more accurate with $N = 3000$ than with $N = 600$. With a sufficient number of observations and subjects, weak correlations are easily detected. For example, we perfectly detect a correlation of 0.2 with $N = 3000$ and $n = 300$ where

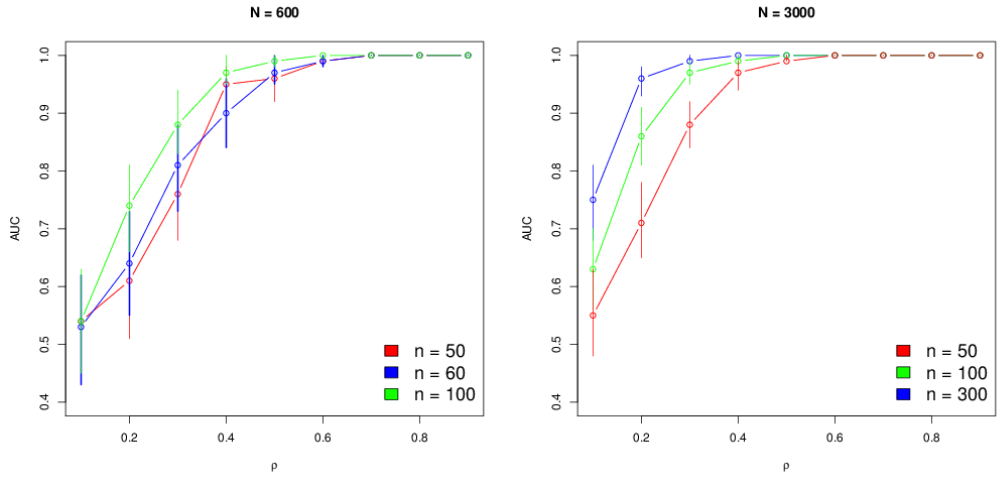

Figure 2.5: **Empirical analysis of the power of the correlation test.** AUC values of ROC curves with their confidence interval computed for different ρ, number of subjects (n) and observations (N). Left panel for $N = 600$, $n = 50, 60, 100$. Right panel for $N = 3000$, $n = 50, 100, 300$.

AUC $= 0.96(0.93 - 0.98)$ according to Fig 2.5. However, we detect quite well a correlation of 0.3 with $N = 600$ and $n = 60$ where AUC $= 0.81(0.73 - 0.88)$.

In the case where estimates are of a higher quality (because they are performed on data sets having a sufficient number of observations $N = 5000$ and subjects $n = 300$), we plot ROC curves with low values of ρ (0.1, 0.2 and 0.3). We then show in Fig 2.6, the estimated AUC and its 95% confidence interval.

Fig 2.6 shows that EM-based estimators lead to a good power of the bivariate correlation test, when we have a sufficient number of observations and subjects in the longitudinal study case. This goodness of the power of the bivariate correlation test persists when the correlation between marginal random effects is low (about 0.2).

2.3.2 Applications on real data sets

In this section we analyze two data sets by using the likelihood ratio test through the EM-based estimators presented above. The first dataset is of multivariate multilevel type and the second is, specifically, of longitudinal multivariate type.

Application to education data in the Netherlands The data used here are named 'bdf' under the package nlme[142] of the R software. These data contain $N = 3776$ Grade eight students (aged about eleven years) in $n = 208$ elementary schools in the Netherlands [22]. These pupils were tested twice (with an interval of one year between grade seven and grade eight) on their proficiency in Dutch language and arithmetics, where the goal was to investigate which characteristics of schools can account for the differences in the effectiveness of schools with regard to pupils' progress in language and arithmetics. Most of the previous analyses of this dataset were concerned with investigating how the language test score depends on the pupil's intelligence, his

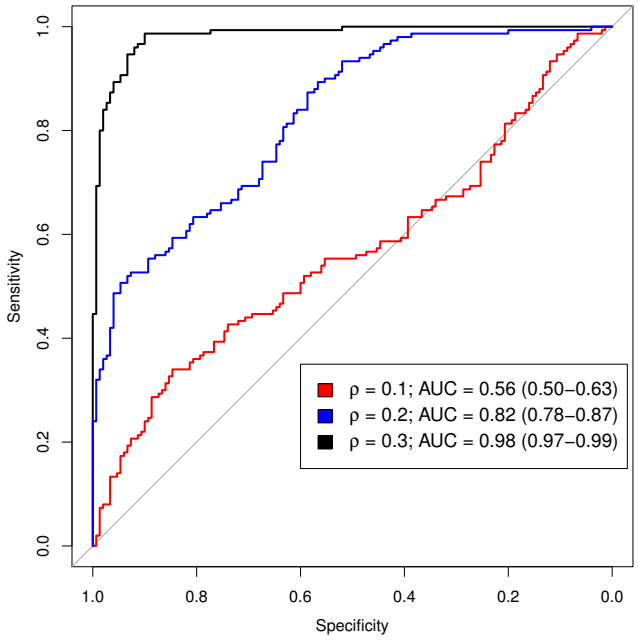

Figure 2.6: **Analysis of the power of the likelihood ratio test performed under EM-estimators.** ROC curves with $\rho \in \{0.1, 0.2, 0.3\}$. $N = 5000$ observations, $n = 300$ subjects, 95% CI on AUC.

family's socio-economic status and on related class or school variables. By fitting two independent (separate) models, [22] found that variables in Table 2.4 have a significant effect on post-test scores (language post-test and arithmetic post-test). These variables are: socio-economic status, intelligence score, age, gender and nationality. They also found a significant random slope related to the language pre-test and to the gender in the language post-test model.

Based on these results from [22] and some of their data ($n = 131$ schools, $N = 2287$ pupils; age and ethnicity are not present), we have fitted the bivariate linear mixed-effect model where post-test scores are the response variables and covariates are the pre-test scores, socio-economic status, intelligence score, gender and minority (a factor indicating if the pupil is a member of a

Table 2.4: **Modeling of covariates on post-test achievement in language and arithmetic from [22]**

Model	**Language (post)**	**Arithmetic (post)**
Language pre-test	0.567	-
Arithmetic pre-test	-	0.413
Socio-economic status	0.143	0.132
Intelligence score	0.124	0.270
Age	−0.069	−0.081
Gender (female)	0.187	−0.103
Ethnicity (foreign)	n.s.	−0.105

Table 2.5: **Model configuration**

	Variables	
Model parts	**Language (post)**	**Arithmetic (post)**
Fixed effects	Language (pre) Socio-eco. status Intelligence score Gender Minority	Arithmetic (pre) Socio-eco. status Intelligence score Gender Minority
Random effects (school level)	1 Language (pre)	1 Arithmetic (pre)

Table 2.6: **Estimated fixed effects and residual standard deviations in the joint bivariate model fitted to school data.**

	Response variables			
	Language (post)		**Arithmetic (post)**	
Covariates	**Estimate**	**p-value**	**Estimate**	**p-value**
Intercept	4.690		−1.456	
Language (pre)	0.795	0.000	-	-
Arithmetic (pre)	-	-	0.807	0.000
Socio-eco. status	0.101	0.000	0.089	0.000
Intelligence score	0.474	0.000	0.810	0.000
Gender (female)	1.778	0.000	−0.522	0.003
Minority (yes)	−0.302	0.589	−0.545	0.194
σ_1 and σ_2	5.356	4.066		

minority group). Random intercepts and random slopes related to pre-test scores are integrated to the model on the school level in the configuration shown by the Table 2.5.

Table 2.6 contains estimated fixed effects and residual standard deviations of the model.

The estimated covariance matrix $\widehat{\mathbf{\Gamma}}$ of the random effects is:

$$\widehat{\mathbf{\Gamma}} = \begin{pmatrix} 15.15 & -0.54 & 18.20 & -0.27 \\ -0.54 & 0.02 & -0.65 & 0.01 \\ 18.20 & -0.65 & 30.69 & -0.48 \\ -0.27 & 0.01 & -0.48 & 0.01 \end{pmatrix}$$

The null hypothesis, H_0, that the arithmetic post-test score and the language post-test score are independent, is rejected with a p-value of 1.436×10^{-7}. This result justifies a joint analysis of

post-scores conditionally on the covariates present in the model and is therefore a supplementary information obtained from the data. The estimated correlation matrix of the random effects is:

$$\widehat{\rho} = \begin{pmatrix} 1.00 & -1.00 & 0.84 & -0.82 \\ -1.00 & 1.00 & -0.84 & 0.82 \\ 0.84 & -0.84 & 1.00 & -1.00 \\ -0.82 & 0.82 & -1.00 & 1.00 \end{pmatrix}$$

Table 2.6 shows that covariates which are significant in the independent models are also significant in the joint model. The Minority covariate is neither significant in the joint model, nor significant in the independent models. $\widehat{\mathbf{\Gamma}}_{ij}$ and $\widehat{\rho}_{ij}$ identify the item which is at the intersection of the row i and the column j of the matrices $\widehat{\mathbf{\Gamma}}$ and $\widehat{\rho}$, respectively. These matrices are filled from the top to the bottom in the order of $(\text{Intercept}_{y_1}, \text{Slope}_{y_1}, \text{Intercept}_{y_2}, \text{Slope}_{y_2})$.

There is a clear inter-school variability with respect to the post-test scores ($\widehat{\mathbf{\Gamma}}_{11} = 15.15$ and $\widehat{\mathbf{\Gamma}}_{33} = 30.69$). Everything else being equal, schools that have good scores in arithmetics also have good scores in language ($\widehat{\rho}_{31} = 0.84$). The schools in which the differential effect of the pre-test score in arithmetics on the post-test score is strongly negative are in average above the average post-test score in language ($\widehat{\rho}_{41} = -0.82$); same as with the language pre-test score ($\widehat{\rho}_{32} = -0.84$). This confirms that the scores in language and in arithmetics vary in the same direction in schools. The schools in which the differential effect of the pre-test score (in arithmetics or language) on the post-test score is strongly negative are in average above the average post score ($\widehat{\rho}_{21} = -1$ and $\widehat{\rho}_{43} = -1$), and vice versa. These schools have strived to bring the level of all pupils above the average. In contrast, pupils with a good initial level maintain their level without becoming excellent. The differential effect of the pre-test score has a very weak variability (arithmetics score: $\widehat{\mathbf{\Gamma}}_{22} = 0.02$; language score: $\widehat{\mathbf{\Gamma}}_{44} = 0.01$) and this implies that the pre-test score explains about 0.15% (in arithmetics) and 0.03% (in language) of inter-school variability of post-test scores.

We have fitted the bivariate model without random slopes (with random intercept only) to investigate if it fits more to the data than the model with random slopes, due to the weak variability of these random slopes. The results are presented in Table 2.7, where the estimated fixed effects and their significance generally remain the same.

The estimated covariance matrix, related to results contained in Table 2.7 of random effects is

$$\widehat{\Gamma} = \begin{pmatrix} 5.15 & 5.41 \\ 5.41 & 7.07 \end{pmatrix} \tag{2.42}$$

which indicates a correlation of $\widehat{\rho} = 0.895$ between the random marginal intercepts, confirming a strong positive correlation between post-test scores in arithmetics and language. With a p-value of 8.505×10^{-5}, the likelihood ratio test indicates that the data are more compatible with the model incorporating random intercepts and random slopes at a time.

Fixed effects seem very strong and do not significantly change between the independent and bivariate models. In contrast, a posterior distribution of random effects changes significantly between the independent model and the joint bivariate model. For example, we plot the joint distribution of random effects conditional on the data concerning School 47 in the education dataset under the

Table 2.7: **Results of the model with random intercepts only.**

	Response variables			
	Language (post)		**Arithmetic (post)**	
Covariates	**Estimate**	**p-value**	**Estimate**	**p-value**
Intercept	4.698		−1.446	
Language (pre)	0.789	0.000	-	-
Arithmetic (pre)	-	-	0.789	0.000
Socio-eco. status	0.103	0.000	0.093	0.000
Intelligence score	0.480	0.000	0.809	0.000
Gender (female)	1.788	0.000	−0.526	0.002
Minority (yes)	−0.391	0.489	−0.498	0.249
σ_1 and σ_2	5.384	4.091		

independent model and the joint bivariate model. Fig 2.7a shows the joint posterior distribution of random intercepts under the independent model whereas Fig 2.7b presents the same posterior distribution under the joint bivariate model. A clear difference appears between these two distributions. We notice the same difference between distributions of random intercepts and slopes as shown in Fig 2.7c and Fig 2.7d as well as the joint distribution of random slopes in Fig 2.7e and Fig 2.7f. The joint bivariate model seems to fit more to the present data and we retain it for their analysis.

Application to malaria immune response data in Benin The data come from a study which was conducted in nine villages (Avamé centre, Gbédjougo, Houngo, Anavié, Dohinoko, Gbétaga, Tori Cada Centre, Zébè and Zoungoudo) of Tori Bossito area (Southern Benin), where *P. falciparum* is the most common species in the study area (95%) [47] from June 2007 to January 2010. The aim of this study was to evaluate the determinants of malaria incidence in the first months of life of child in Benin. Details of the follow-up procedures have been published elsewhere [108].

Data description Mothers ($n = 620$) were enrolled at delivery and their newborns were actively followed-up during the first year of life. One questionnaire was conducted to gather information on women's characteristics (age, parity, use of Intermittent Preventive Treatment during pregnancy (IPTp) and bed net possession) and on the course of their current pregnancy. Maternal peripheral blood as well as cord blood were collected into Vacutainer® EDTA (Ethylene diaminetetraacetic acid) tubes. At birth, newborn's weight and length were measured by midwives and gestational age was estimated using the Ballard method [9].

During the follow-up of newborns, axillary temperature was measured weekly. Symptomatic malaria cases, defined as fever (> 37.5°C) with TBS and/or RDT positive, were treated with an artemisinin-based combination therapy as recommended by the Benin National Malaria Control Program. Systematically, TBS were made every month to detect asymptomatic infections. Every three months, venous blood was sampled to quantify the level of antibody against malaria promised candidate vaccine antigens. The environmental risk of exposure to malaria was modeled for each

Table 2.8: **Variables present in the analyzed dataset**

Variable	Description
id	Child ID
conc.Y	concentration of Y
conc_CO.Y	Measured concentration of Y in the umbilical cord blood
conc_M3.Y	Predicted concentration of Y in the child's peripheral blood at 3 months
ap	Placental apposition
hb	Hemoglobin level
inf_trim	Number of malaria infections in the previous 3 months
pred_trim	Quarterly average number of mosquitoes child is exposed to
nutri_trim	Quarterly average nutrition scores

child, derived from a statistical predictive model based on climatic, entomological parameters, and characteristics of children's immediate surroundings as reported by [37].

Concerning the antibody quantification, two recombinant *P. falciparum* antigens were used to perform IgG subclass (IgG1 and IgG3) antibody quantification by Enzyme-Linked ImmunoSorbent Assay (ELISA) standard methods developed for evaluating malaria vaccines by the African Malaria Network Trust (AMANET [`www.amanet148trust.org`]). Protocol was described in detail [38].

Data analysis For our analysis, we use some of the data and we rename the proteins used in the study described above, for confidentiality reasons (some important findings are yet to be published). Thus, the proteins we use here, are named A1, A2, B and C, and are related to the antigens IgG1 and IgG3 as mentioned above in the description of the study. A1 and A2 are different domains of the same protein A, and C and D are two different proteins. Information contained in the multivariate longitudinal dataset of malaria are described in the Table 2.8, where Y denotes a protein which is one of the following:

$$\text{IgG1\_A1}, \text{IgG3\_A1}, \text{IgG1\_A2}, \text{IgG3\_A2}, \text{IgG1\_B}, \text{IgG3\_B}, \text{IgG1\_C}, \text{IgG3\_C} \tag{2.43}$$

The aim of the analysis of these data is to evaluate the effect of the malaria infection on the child's immune (against malaria). Since the antigens which characterize the child's immune status interact together in the human body, we analyze the characteristics of the joint distribution of these antigens, conditional on the malaria infection and other factors of interest. The dependent variables are then provided by conc.Y (Table 2.8) which describes the level of the protein Y in the children at 3, 6, 9, 12, 15 and 18 months. All other variables in the Table 2.8 are covariates. We then have eight dependent variables which describe the longitudinal profile (in the child) of the proteins listed in Equation 3.48.

In the models that we fit to these data, we specify one random intercept by child and one random slope by child in the direction of the malaria infection. The illustration we do here is to jointly analyze each of the 28 pairs of proteins, in order to investigate if some profiles of proteins are independent, conditional on the configuration of the fitted model. After performing the bivariate

correlation test on all 28 bivariate models, the obtained p-values, with a Bonferroni correction, range from 4.16×10^{-33} to 0.932. The p-value 0.932 is the only one which is not significant. This p-value corresponds to the pair of proteins $(\text{IgG3\_A1}, \text{IgG1\_B})$.

To investigate the general configuration of these proteins, in terms of correlations, we build their hierarchical cluster tree using $-\log(\text{p-value})$ as dissimilarity. This hierarchical cluster tree is presented by the Fig 2.8.

The branch related to the IgG1 is different from the branch related to the IgG3. In other words, IgG1_A1, IgG1_A2, IgG1_B and IgG1_C are on the same branch which is different from the branch containing IgG3_A1, IgG3_A2, IgG3_B and IgG3_C (Fig 2.8). Relatively to both IgG1 and IgG3, A1 and A2 go together, and B and C also go together. These results are biologically very consistent, since A1 and A2 are domains of the same protein, and B and C are two different proteins. On the cluster (Fig 2.8), it also appears that the proteins IgG3_A1 and IgG1_B which are not significantly correlated (according to our bivariate test) are distant. Statistically, the model which may be used to jointly analyze these 8 protein profiles is not probably the model which contains all the 27 significant correlations, avoiding overfitting problems. Based on the results provided by the bivariate correlation test, it may be useful to perform a regularization procedure in the fitting of the full eight-variate model.

2.4 Conclusion

In the context of the multivariate linear mixed-effects model, we have suggested the more general expressions of the EM-based estimators than those used in the literature to analyze multivariate longitudinal data. These estimators fit the framework of the multivariate multilevel data analysis which, obviously, englobes the multivariate longitudinal data analysis framework. We also have built a likelihood ratio test based on these EM estimators to test the independence of two dimensions of the model. Furthermore, the simulation studies have validated the power of this test and have shown that this is an extremely sensitive test. In the context of longitudinal data, it allows to detect a modest correlation signal with a very small sample ($\rho = 0.3$, AUC$= 0.81$, with $n = 60$ subjects and $N = 600$ observations). In the simulation studies, the empirical distribution of the likelihood ratio statistic fits the $\chi^2(4)$. The asymptotic properties of likelihood ratio statistics, under nonstandard conditions, have been shown by [31] and [171]. These works have been generalized by [196] to cover a large class of estimation problems which allow sampling from non identically distributed random variables. The asymptotic distribution of the LR statistic derived by [196] is a mixture of chi-squared distributions. In the context of likelihood ratio tests for variance components in linear mixed-effects models, [74] used the results of [196] to prove that the proposed mixture of chi-squared distributions is the actual asymptotic distribution of such LR used as test statistics for null variance components with one or two random effects. Based on these works, Further theoretical investigations may be done to properly find out the asymptotic distribution of the likelihood ratio statistic in the case of this bivariate correlation test. Finally, we have illustrated the usefulness of the test on two different real-life data. The first dataset, which is of multivariate multilevel type, concerns the effects of school and classroom characteristics on pupils' progress in Dutch language and arithmetics, where the scores in language and arithmetics are the two response

variables which have been considered. Our method has yielded results that are consistent both with information in existing publications and with a conceptual understanding of the phenomenon. On this dataset, we have highlighted a joint effect between the scores in arithmetics and language within schools in the Netherlands. The second dataset, which is of longitudinal multivariate type, concerns a study of the effect of the malaria infection on the child's immune response in Benin. By jointly analyzing all the pairs of protein profiles of interest, we have plotted a hierarchical cluster tree of these proteins, using the bivariate correlation test. Information contained in this hierarchical cluster tree is consistent with the biological literature related to this issue.

The model as it is written is easily extendable to more dimensions despite a sparsity problem in choosing the parameterization of the covariance matrix or the precision matrix. Probably we could use this two-dimensional dependence test to structure a larger covariance matrix. The bivariate correlation test can help to construct iteratively, using a stepwise procedure, a parsimonious joint model containing all the components of $\boldsymbol{y}$. This stepwise procedure may consist in adding to the constructing model, at each step, the significant correlation between two dependent variables. Using a model selection strategy, the model which fits more to the data will be retained. It could possibly be advantageous to turn to graphical LASSO type approaches to make a penalized estimation of this covariance (or precision) matrix. We could also resort to the rapid optimization methods such as that implemented in the lme4 [15] package, given the slow pace of the EM algorithm. It would be useful to assess the interest of this method compared to some heuristics such as the one which consists in setting one marginal response variable as a covariate of the other(s).

Acknowledgements

We warmly thank the SCAC (Service de Coopération et d'Actions Culturelles) of the France Embassy in Benin, as well as the IRD (Institut de Recherche pour le Développement) for their financial support in the realization of this work.

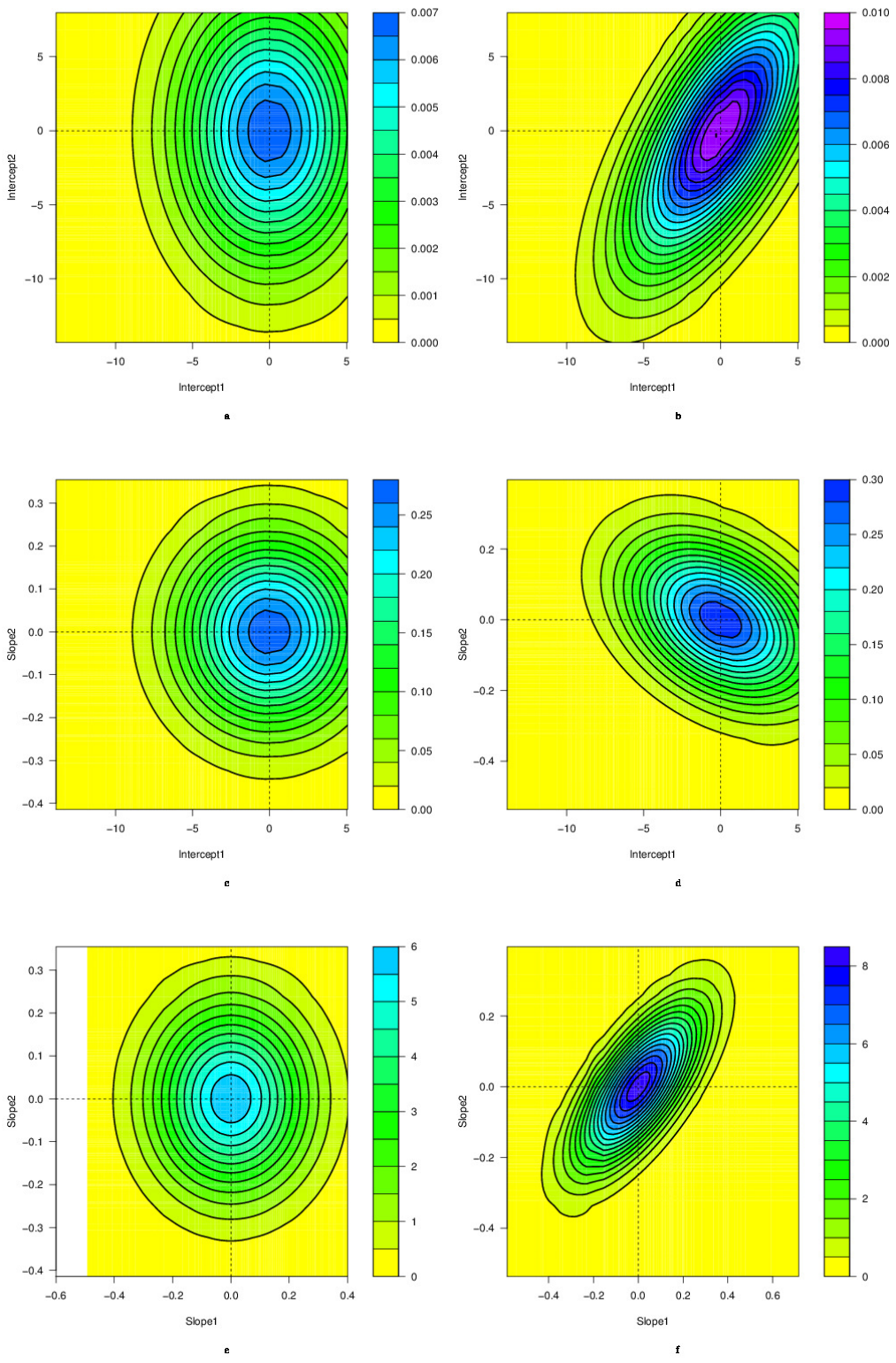

Figure 2.7: **Posterior distributions of random intercepts conditional on the data related to School** 47 **in the education dataset.** Left panels assume independence across the two dimensions while right panels assume dependence. Top panels for the joint distribution of the random intercepts, middle panels for the joint distribution of random intercept in first dimension and random slope in the second dimension, bottom panels for the joint distribution of the random slopes.

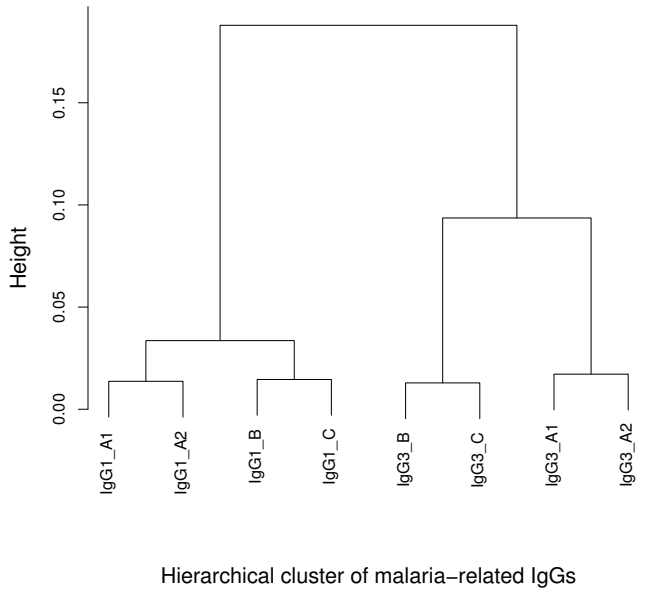

Figure 2.8: **Hierarchical cluster tree on malaria-related proteins**

CHAPTER 3

PROFILED DEVIANCE FOR THE MULTIVARIATE LINEAR MIXED-EFFECTS MODEL FITTING

Dans ce chapitre, nous avons généralisé l'approche utilisée par Bates et al. (2014) dans le package lme4 du logiciel R pour estimer les paramètres d'un modèle linéaire multidimensionnel à effets mixtes. Cette généralisation est relative au cas où les résidus marginaux sont homoscédastiques avec la possibilité d'avoir des variances différentes. Les critères ML et REML sont explicitement écrits, mais pour des raisons de rapidité algorithmique, nous avons plutôt cherché à minimiser la déviance relativement aux paramètres de variance, avec une mise à jour itérative du vecteur des effets fixes jusqu'à convergence. Une étude de simulations a été menée et il en ressort que les estimations obtenues sont consistantes pour tous les paramètres du modèle. Ainsi la présente méthode surclasse l'estimation par l'algorithme EM, surtout en ce qui concerne les paramètres de variance. Ce travail a fait l'objet d'un article de recherche soumis dans une revue scientifique.

Profiled deviance for the multivariate linear mixed-effects model fitting

Eric Houngla Adjakossa[1,2 *], Mahouton Norbert Hounkonnou[2], Gregory Nuel[1]

1 Laboratoire de Probabilités et Modèles Aléatoires /Université Pierre et Marie Curie, Case courrier 188 - 4, Place Jussieu 75252 Paris cedex 05 France
2 University of Abomey-Calavi, 072 B.P. 50 Cotonou, Republic of Benin

*** ericadjakossah@gmail.com**

abstract

This paper focuses on the multivariate linear mixed-effects model, including all the correlations between the random effects when the marginal residual terms are assumed uncorrelated and homoscedastic with possibly different standard deviations. The random effects covariance matrix is Cholesky factorized to directly estimate the variance components of these random effects. This strategy enables a consistent estimate of the random effects covariance matrix which, generally, has a poor estimate when it is grossly (or directly) estimated, using the estimating methods such as the EM algorithm. By using simulated data sets, we compare the estimates based on the present method with the EM algorithm-based estimates. We provide an illustration by using the real-life data concerning the study of the child's immune against malaria in Benin (West Africa).

3.1 Introduction

Linear mixed-effects model [62, 86, 90, 105, 192] has become a popular tool for analyzing univariate multilevel data which arise in many areas (biology, medicine, economy, etc), due to its flexibility to model the correlation contained in these data, and the availability of reliable and efficient software packages for fitting it [16, 82, 116, 141]. Univariate multilevel data are referred to as observations (or measurements) of a single variable of interest on several levels (school in a village which, in turn, is in a town), while multivariate multilevel data are characterized by multiple variables of interest measured at multiple levels. Examples include exam or test scores recorded for students across time, and multiple items at a single occasion for students in more than one school. Multivariate extension of the (single response variable-based) linear mixed-effects model is, indeed, having increasing popularity as flexible tool for the analysis of multivariate multilevel data [97, 159, 161, 197].

For the linear mixed-effects model, many methods for obtaining the estimates of the fixed and the random effects have been proposed in the literature. These methods include Henderson's mixed model equations [91], approaches proposed by [75] as well as techniques based on two-stage regression, Bayes estimation, etc. For details, see [168, Section 7.4c] and [155]. Concerning

the variance parameters estimation in linear mixed-effects model, the discussed methods in the literature include the ANOVA method for balanced data which uses the expected mean squares approach [169, 170]. For unbalanced data, [149] proposed the minimum norm quadratic estimation (MINQUE) method, where the resulting estimates are translation invariant under unbiased quadratic forms of the observations. [109] gave another method of estimating variance parameters using extended quasi-likelihood, i.e. gamma-log generalized linear models. For more details on these parameters' estimation methods in the linear mixed-effects model, see the paper of [81]. Beside all the methods cited earlier, come the Maximum Likelihood (ML) and the Restricted Maximum Likelihood (REML) methods. ML and REML methods are the most popular estimation methods in the linear mixed-effects model [114]. The main attraction of these methods is that they can handle a much wider class of variance models than simple variance components [81].

In the multivariate linear mixed-effects model, ML and REML estimates are frequently approached through iterative schemes such as EM algorithm [6, 44, 125, 161, 172]. This avoid the difficulties related to the direct calculating of the parameters' likelihood, since the random effects are not observed, without ignoring the flexible computationally of these algorithms. Despite the existence of valid theorems which show the asymptotic convergence of the sequences produced by these algorithms toward ML estimates [44], in practice this may not always work exactly as expected.

In this paper, we focus on the multivariate linear mixed-effects model, including all the correlations between the random effects while the marginal residuals are assumed independent homoscedastic with possibly different standard deviation. The class of multivariate mixed-effects models considered here assumes that the random effects and the residuals follow Gaussian distributions, and the dependent variables are continuous. In this model, our approach consists in directly calculating the likelihood of the model's parameters. This likelihood is used to obtain the ML estimates or the REML estimates through the provided REML criterion. This strategy may explain the high quality of the estimates of both fixed effects parameters and random effects' variance parameters as well as residual variance parameters. This approach may be viewed as a generalization of the approach proposed by [16] under the R software [147] package named lme4.

3.2 Multivariate linear mixed-effects model

For the sake of simplicity we focus on the bivariate case ($d = 2$) in most of the paper, but the generalization to higher dimensions ($d > 2$) is straightforward. Thus, in dimension 2, the model is the following:

$$\begin{aligned} y_1 &= X_1\beta_1 + Z_1\gamma_1 + \varepsilon_1, \\ y_2 &= X_2\beta_2 + Z_2\gamma_2 + \varepsilon_2, \end{aligned} \tag{3.1}$$

where

$$\boldsymbol{\gamma} = \begin{pmatrix} \gamma_1 \\ \gamma_2 \end{pmatrix} \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Gamma} = \begin{pmatrix} \Gamma_1 & \Gamma_{12} \\ \Gamma_{12}^\top & \Gamma_2 \end{pmatrix}\right), \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix} \sim \left(\mathbf{0}, \begin{pmatrix} \sigma_1^2 \mathrm{I}_N & 0 \\ 0 & \sigma_2^2 \mathrm{I}_N \end{pmatrix}\right). \tag{3.2}$$

For the sake of simplicity, we write $\boldsymbol{\gamma} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Gamma})$ to mean that $\boldsymbol{\gamma}$ is a realization of a random vector which is $\mathcal{N}(\mathbf{0}, \boldsymbol{\Gamma})$ distributed. For $k \in \{1, 2\}$, β_k and γ_k denote respectively the fixed effects and the random effects vector of covariates, while ε_k is the marginal residual component in the dimension k of the model. X_k is a matrix of covariates and Z_k a covariates-based matrix of design. $\dim(X_k) = N \times p_k$ and $\dim(Z_k) = N \times q_k$, where N is the total number of observations. p_k and q_k are, respectively, the number of fixed effect related covariates and the number of random effect related covariates in the dimension k of the model. $\boldsymbol{y} = (y_1^\top, y_2^\top)^\top$ is the vector of marginal observed response variables of the model. We assume that $\boldsymbol{y}$ is a realization of a random vector $\boldsymbol{\mathcal{Y}}$ and belongs to $\mathbb{R}^{2N}$. The bold symbols represent parameters, or vectors, of multiple dimensions (i.e. Γ_1 concerns dimension 1 of the model while $\boldsymbol{\Gamma}$ concerns both dimensions).

Γ_1 and Γ_2 are the variance-covariance matrices of γ_1 and γ_2, respectively. Γ_1 and Γ_2 must be, indeed, positive semidefinite. It is then convenient to express the model in terms of the relative covariance factors, Λ_{θ_1} and Λ_{θ_2}, which are $q_1 \times q_1$ and $q_2 \times q_2$ matrices, respectively. Λ_{θ_1} is a block diagonal matrix. Each element in the diagonal of Λ_{θ_1} is a lower triangular matrix whose nonzero entries are the components of the vector θ_1. That is, θ_1 generates the symmetric $q_1 \times q_1$ variance-covariance matrix Γ_1, according to

$$\Gamma_1 = \sigma_1^2 \Lambda_{\theta_1} \Lambda_{\theta_1}^\top. \tag{3.3}$$

Same as θ_2 which generates Γ_2 according to

$$\Gamma_2 = \sigma_2^2 \Lambda_{\theta_2} \Lambda_{\theta_2}^\top. \tag{3.4}$$

In Equations 3.3 and 3.4, σ_1^2 and σ_2^2 are the same marginal residual variances used in the model expression (see Equation 3.2). Using the variance-component parameters, θ_1 and θ_2, the marginal random effects, γ_1 and γ_2, are expressed as

$$\gamma_1 = \Lambda_{\theta_1} u_1, \quad \gamma_2 = \Lambda_{\theta_2} u_2, \tag{3.5}$$

such that

$$\boldsymbol{u} = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} \sim \mathcal{N}\left(\mathbf{0}, \Sigma_{\boldsymbol{u}}\right), \quad \text{with} \quad \Sigma_{\boldsymbol{u}} = \begin{pmatrix} \sigma_1^2 \mathrm{I}_{q_1} & \sigma_1 \sigma_2 \boldsymbol{\rho} \\ \sigma_1 \sigma_2 \boldsymbol{\rho}^\top & \sigma_2^2 \mathrm{I}_{q_2} \end{pmatrix}. \tag{3.6}$$

In Equation 3.6, $\boldsymbol{\rho}$ is a block diagonal matrix and $\boldsymbol{u}$ is a realization of a random vector $\boldsymbol{\mathcal{U}}$. The diagonal elements of $\boldsymbol{\rho}$, say ρ, are matrices which contain the correlations between γ_1 and γ_2. For example, if $\gamma_1 = (\gamma_1^{\mathrm{I}}, \gamma_1^{\mathrm{S}})^\top$ and $\gamma_2 = (\gamma_2^{\mathrm{I}}, \gamma_2^{\mathrm{S}})^\top$, with I = Intercept and S = Slope,

$$\rho = \begin{pmatrix} \mathrm{corr}(\gamma_1^{\mathrm{I}}, \gamma_2^{\mathrm{I}}) & \mathrm{corr}(\gamma_1^{\mathrm{I}}, \gamma_2^{\mathrm{S}}) \\ \mathrm{corr}(\gamma_1^{\mathrm{S}}, \gamma_2^{\mathrm{I}}) & \mathrm{corr}(\gamma_1^{\mathrm{S}}, \gamma_2^{\mathrm{S}}) \end{pmatrix} \quad \text{and} \quad \boldsymbol{\rho} = \mathrm{diag}(\rho, \ldots, \rho). \tag{3.7}$$

The bivariate linear mixed-effects model is then re-expressed as:

$$\begin{aligned} y_1 &= X_1\beta_1 + Z_1\Lambda_{\theta_1}u_1 + \varepsilon_1, \\ y_2 &= X_2\beta_2 + Z_2\Lambda_{\theta_2}u_2 + \varepsilon_2, \end{aligned} \tag{3.8}$$

with

$$\boldsymbol{u} = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} \sim \mathcal{N}(\mathbf{0}, \Sigma_{\boldsymbol{u}}), \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix} \sim \left(\mathbf{0}, \begin{pmatrix} \sigma_1^2\mathrm{I}_N & 0 \\ 0 & \sigma_2^2\mathrm{I}_N \end{pmatrix} \right). \tag{3.9}$$

Then the parameters which will be estimated are $\beta_1, \beta_2, \sigma_1^2, \sigma_2^2, \theta_1, \theta_2$ and ρ.

3.3 Parameters' estimates

In this Section, we first provide the likelihood of the model's parameters and then give the REML criterion which will be optimized for the obtaining of the parameters' REML estimates.

3.3.1 ML criterion

The ML criterion is the log-likelihood of the model's parameters which is displayed through the following theorem

Theorem 3.3.1. *Suppose that $\boldsymbol{y} = (y_1^\top, y_2^\top)^\top$ satisfies the bivariate linear mixed-effects model expressed by Equations (3.8 and 3.9), where $\beta_1, \beta_2, \sigma_1^2, \sigma_2^2, \theta_1, \theta_2, \rho$ are the parameters which need to be estimated, and $\boldsymbol{\beta} = (\beta_1^\top, \beta_2^\top)^\top$, $\boldsymbol{\sigma} = (\sigma_1^2, \sigma_2^2)^\top$, $\boldsymbol{\theta} = (\theta_1^\top, \theta_2^\top)^\top$. Denoting by $Y_{\boldsymbol{\sigma}} = \left(\sqrt{\sigma_2^2}y_1^\top, \sqrt{\sigma_1^2}y_2^\top\right)^\top$, $X_{\boldsymbol{\sigma}} = \begin{pmatrix} \sqrt{\sigma_2^2}X_1 & \mathbf{0} \\ \mathbf{0} & \sqrt{\sigma_1^2}X_2 \end{pmatrix}$, $Z_{\boldsymbol{\sigma\theta}} = \begin{pmatrix} \sqrt{\sigma_2^2}Z_1\Lambda_{\theta_1} & \mathbf{0} \\ \mathbf{0} & \sqrt{\sigma_1^2}Z_2\Lambda_{\theta_2} \end{pmatrix}$, and $\mu_{\boldsymbol{\mathcal{U}}|\boldsymbol{\mathcal{Y}}=\boldsymbol{y}}$ the conditional mean of $\boldsymbol{\mathcal{U}}$ given that $\boldsymbol{\mathcal{Y}} = \boldsymbol{y}$, the log-likelihood of $\boldsymbol{\beta}$, $\boldsymbol{\sigma}$, $\boldsymbol{\theta}$ and ρ given $\boldsymbol{y}$ is expressed as*

$$\begin{aligned} \ell(\boldsymbol{\beta}, \boldsymbol{\theta}, \rho, \boldsymbol{\sigma}|\boldsymbol{y}) \;=\; & -\frac{r(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\theta},\rho,\boldsymbol{\sigma}}, \mu_{\boldsymbol{\mathcal{U}}|\boldsymbol{\mathcal{Y}}=\boldsymbol{y}}) + \left\| R_X(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_{\boldsymbol{\theta},\rho,\boldsymbol{\sigma}}) \right\|^2}{2\sigma_1^2\sigma_2^2} - \frac{N-q}{2}\log(\sigma_1^2\sigma_2^2) \\ & -\frac{1}{2}\log(|\Sigma_{\boldsymbol{u}}|) - \frac{1}{2}\log(|L_{\boldsymbol{\theta},\rho,\boldsymbol{\sigma}}|^2), \end{aligned} \tag{3.10}$$

where $q = q_1 + q_2$, $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\theta},\rho,\boldsymbol{\sigma}}$ *and* $\mu_{\boldsymbol{\mathcal{U}}|\boldsymbol{\mathcal{Y}}=\boldsymbol{y}}$ *satisfy*

$$\begin{pmatrix} X_{\boldsymbol{\sigma}}^\top X_{\boldsymbol{\sigma}} & X_{\boldsymbol{\sigma}}^\top Z_{\boldsymbol{\sigma\theta}} \\ Z_{\boldsymbol{\sigma\theta}}^\top X_\sigma & Z_{\boldsymbol{\sigma\theta}}^\top Z_{\boldsymbol{\sigma\theta}} + \sqrt{\sigma_1^2\sigma_2^2}\Sigma_{\boldsymbol{u}}^{-1} \end{pmatrix} \begin{pmatrix} \widehat{\boldsymbol{\beta}}_{\boldsymbol{\theta},\rho,\boldsymbol{\sigma}} \\ \mu_{\mathcal{U}|\mathcal{Y}=\boldsymbol{y}} \end{pmatrix} = \begin{pmatrix} X_{\boldsymbol{\sigma}}^\top \\ Z_{\boldsymbol{\sigma\theta}}^\top \end{pmatrix} Y_{\boldsymbol{\sigma}}, \tag{3.11}$$

$$r(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\theta},\rho,\boldsymbol{\sigma}}, \mu_{\boldsymbol{\mathcal{U}}|\boldsymbol{\mathcal{Y}}=\boldsymbol{y}}) = \left\| Y_{\boldsymbol{\sigma}} - X_{\boldsymbol{\sigma}}\widehat{\boldsymbol{\beta}}_{\boldsymbol{\theta},\rho,\boldsymbol{\sigma}} - Z_{\boldsymbol{\sigma\theta}}\mu_{\boldsymbol{\mathcal{U}}|\boldsymbol{\mathcal{Y}}=\boldsymbol{y}} \right\|^2 + \sigma_1^2\sigma_2^2 \mu_{\boldsymbol{\mathcal{U}}|\boldsymbol{\mathcal{Y}}=\boldsymbol{y}}^\top \Sigma_{\boldsymbol{u}}^{-1} \mu_{\boldsymbol{\mathcal{U}}|\boldsymbol{\mathcal{Y}}=\boldsymbol{y}}, \tag{3.12}$$

$L_{\boldsymbol{\theta},\rho,\boldsymbol{\sigma}}$ *satisfies*

$$L_{\boldsymbol{\theta},\rho,\boldsymbol{\sigma}} L_{\boldsymbol{\theta},\rho,\boldsymbol{\sigma}}^\top = Z_{\boldsymbol{\sigma\theta}}^\top Z_{\boldsymbol{\sigma\theta}} + \sqrt{\sigma_1^2\sigma_2^2}\Sigma_{\boldsymbol{u}}^{-1}, \tag{3.13}$$

and R_X *satisfies*

$$\begin{pmatrix} X_{\boldsymbol{\sigma}}^\top X_{\boldsymbol{\sigma}} & X_{\boldsymbol{\sigma}}^\top Z_{\boldsymbol{\sigma\theta}} \\ Z_{\boldsymbol{\sigma\theta}}^\top X_\sigma & L_{\boldsymbol{\theta},\rho,\boldsymbol{\sigma}} L_{\boldsymbol{\theta},\rho,\boldsymbol{\sigma}}^\top \end{pmatrix} = \begin{pmatrix} R_X & \mathbf{0} \\ R_{ZX} & L_{\boldsymbol{\theta},\rho,\boldsymbol{\sigma}}^\top \end{pmatrix}^\top \begin{pmatrix} R_X & \mathbf{0} \\ R_{ZX} & L_{\boldsymbol{\theta},\rho,\boldsymbol{\sigma}}^\top \end{pmatrix}. \tag{3.14}$$

Proof. Denoting by $f_{\mathcal{X}}(.)$ the density function of any random vector $\mathcal{X}$,

$$f_{\boldsymbol{\mathcal{Y}}}(\boldsymbol{y}) = \int_{\mathbb{R}^{q_1+q_2}} f_{\boldsymbol{\mathcal{Y}},\boldsymbol{\mathcal{U}}}(\boldsymbol{y},\boldsymbol{u}) d\boldsymbol{u}, \tag{3.15}$$

where

$$\begin{aligned} f_{\boldsymbol{\mathcal{Y}},\boldsymbol{\mathcal{U}}}(\boldsymbol{y},\boldsymbol{u}) &= f_{\boldsymbol{\mathcal{Y}}|\boldsymbol{\mathcal{U}}}(\boldsymbol{y}|\boldsymbol{u}) f_{\boldsymbol{\mathcal{U}}}(\boldsymbol{u}) = f_{\mathcal{Y}_1|\mathcal{U}_1}(y_1|u_1) f_{\mathcal{Y}_2|\mathcal{U}_2}(y_2|u_2) f_{\boldsymbol{\mathcal{U}}}(\boldsymbol{u}) \\ &= (2\pi\sigma_1^2)^{-\frac{N}{2}} (2\pi\sigma_2^2)^{-\frac{N}{2}} (2\pi)^{-\frac{q_1+q_2}{2}} |\Sigma_{\boldsymbol{u}}|^{-\frac{1}{2}} \exp\left(-\frac{\|y_1 - X_1\beta_1 - Z_1\Lambda_{\theta_1}u_1\|^2}{2\sigma_1^2} \right. \\ &\quad \left. -\frac{\|y_2 - X_2\beta_2 - Z_2\Lambda_{\theta_2}u_2\|^2}{2\sigma_2^2} - \frac{1}{2}\boldsymbol{u}^\top \Sigma_{\boldsymbol{u}}^{-1}\boldsymbol{u} \right). \end{aligned} \tag{3.16}$$

Let us denote by $\widetilde{\Sigma}$ the matrix such that

$$\Sigma_{\boldsymbol{u}}^{-1} = \widetilde{\Sigma}^\top \widetilde{\Sigma}. \tag{3.17}$$

It then comes that $\boldsymbol{u}^\top \Sigma_{\boldsymbol{u}}^{-1} \boldsymbol{u} = \|\widetilde{\Sigma}\boldsymbol{u}\|^2$ and

$$\begin{aligned} & \frac{\|y_1 - X_1\beta_1 - Z_1\Lambda_{\theta_1}u_1\|^2}{\sigma_1^2} + \frac{\|y_2 - X_2\beta_2 - Z_2\Lambda_{\theta_2}u_2\|^2}{\sigma_2^2} + \boldsymbol{u}^\top \Sigma_{\boldsymbol{u}}^{-1}\boldsymbol{u} \\ = & \frac{\|\sqrt{\sigma_2^2}(y_1 - X_1\beta_1 - Z_1\Lambda_{\theta_1}u_1)\|^2 + \|\sqrt{\sigma_1^2}(y_2 - X_2\beta_2 - Z_2\Lambda_{\theta_2}u_2)\|^2 + \|\sqrt{\sigma_1^2\sigma_2^2}\widetilde{\Sigma}\boldsymbol{u}\|^2}{\sigma_1^2\sigma_2^2} \\ = & \left\| \begin{pmatrix} \sqrt{\sigma_2^2}y_1 \\ \sqrt{\sigma_1^2}y_2 \\ \mathbf{0}_{q_1+q_2} \end{pmatrix} - \begin{pmatrix} \sqrt{\sigma_2^2}X_1 & \mathbf{0}_{Np_2} & \sqrt{\sigma_2^2}Z_1\Lambda_{\theta_1} & \mathbf{0}_{Nq_2} \\ \mathbf{0}_{Np_1} & \sqrt{\sigma_1^2}X_2 & \mathbf{0}_{Nq_1} & \sqrt{\sigma_1^2}Z_2\Lambda_{\theta_2} \\ & \mathbf{0}_{q_1+q_2,p_1+p_2} & & \sqrt{\sigma_1^2\sigma_2^2}\widetilde{\Sigma} \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta} \\ \boldsymbol{u} \end{pmatrix} \right\|^2 \end{aligned} \tag{3.18}$$

$$= \left\| Y_\Lambda - Z_{X\Lambda} \begin{pmatrix} \boldsymbol{\beta} \\ \boldsymbol{u} \end{pmatrix} \right\|^2 \tag{3.19}$$

$$= g(\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{\theta}, \rho, \boldsymbol{\sigma}). \tag{3.20}$$

$$\begin{pmatrix} \widehat{\boldsymbol{\beta}}_{\boldsymbol{\theta},\rho,\boldsymbol{\sigma}} \\ \mu_{\boldsymbol{\mathcal{U}}|\boldsymbol{\mathcal{Y}}=\boldsymbol{y}} \end{pmatrix} = \underset{\boldsymbol{u},\boldsymbol{\beta}}{\arg\min}\, g(\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{\theta}, \rho, \boldsymbol{\sigma}) \iff Z_{X\Lambda}^\top Z_{X\Lambda} \begin{pmatrix} \widehat{\boldsymbol{\beta}}_{\boldsymbol{\theta},\rho,\boldsymbol{\sigma}} \\ \mu_{\boldsymbol{\mathcal{U}}|\boldsymbol{\mathcal{Y}}=\boldsymbol{y}} \end{pmatrix} = Z_{X\Lambda}^\top Y_\Lambda \text{ (normal eq.)}, \tag{3.21}$$

with

$$Z_{X\Lambda}^\top Z_{X\Lambda} = \begin{pmatrix} X_{\boldsymbol{\sigma}}^\top X_{\boldsymbol{\sigma}} & X_{\boldsymbol{\sigma}}^\top Z_{\boldsymbol{\sigma\theta}} \\ Z_{\boldsymbol{\sigma\theta}}^\top X_\sigma & Z_{\boldsymbol{\sigma\theta}}^\top Z_{\boldsymbol{\sigma\theta}} + \sqrt{\sigma_1^2\sigma_2^2}\Sigma_{\boldsymbol{u}}^{-1} \end{pmatrix} \quad \text{and} \quad Z_{X\Lambda}^\top Y_\Lambda = \begin{pmatrix} X_{\boldsymbol{\sigma}}^\top \\ Z_{\boldsymbol{\sigma}}^\top \end{pmatrix} Y_{\boldsymbol{\sigma}}. \tag{3.22}$$

By setting $p = p_1 + p_2$, $dim(Z_{X\Lambda}) = (2N+q) \times (p+q)$ and $S = Im(Z_{X\Lambda})$ is a subspace of $\mathbb{R}^{2N+q}$. $Y_\Lambda \in \mathbb{R}^{2N+q}$ and $Z_{X\Lambda} \begin{pmatrix} \widehat{\boldsymbol{\beta}}_{\boldsymbol{\theta},\rho,\boldsymbol{\sigma}} \\ \mu_{\boldsymbol{\mathcal{U}}|\boldsymbol{\mathcal{Y}}=\boldsymbol{y}} \end{pmatrix}$ is the orthogonal projection of Y_Λ on S. Then,

$$Z_{X\Lambda} u \perp \left[Y_\Lambda - Z_{X\Lambda} \begin{pmatrix} \widehat{\boldsymbol{\beta}}_{\boldsymbol{\theta},\rho,\boldsymbol{\sigma}} \\ \mu_{\boldsymbol{\mathcal{U}}|\boldsymbol{\mathcal{Y}}=\boldsymbol{y}} \end{pmatrix} \right], \forall u \in \mathbb{R}^{p+q}. \tag{3.23}$$

And $g(\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{\theta}, \rho, \boldsymbol{\sigma})$ can then be rewritten as:

$$g(\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{\theta}, \rho, \boldsymbol{\sigma}) = \left\| Y_\Lambda - Z_{X\Lambda} \begin{pmatrix} \boldsymbol{\beta} \\ \boldsymbol{u} \end{pmatrix} + Z_{X\Lambda} \begin{pmatrix} \widehat{\boldsymbol{\beta}}_{\boldsymbol{\theta},\rho,\boldsymbol{\sigma}} \\ \mu_{\boldsymbol{\mathcal{U}}|\boldsymbol{\mathcal{Y}}=\boldsymbol{y}} \end{pmatrix} - Z_{X\Lambda} \begin{pmatrix} \widehat{\boldsymbol{\beta}}_{\boldsymbol{\theta},\rho,\boldsymbol{\sigma}} \\ \mu_{\boldsymbol{\mathcal{U}}|\boldsymbol{\mathcal{Y}}=\boldsymbol{y}} \end{pmatrix} \right\|^2 \tag{3.24}$$

$$= \left\| Y_\Lambda - Z_{X\Lambda} \begin{pmatrix} \widehat{\boldsymbol{\beta}}_{\boldsymbol{\theta},\rho,\boldsymbol{\sigma}} \\ \mu_{\boldsymbol{\mathcal{U}}|\boldsymbol{\mathcal{Y}}=\boldsymbol{y}} \end{pmatrix} \right\|^2 + \left\| Z_{X\Lambda} \begin{pmatrix} \boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_{\boldsymbol{\theta},\rho,\boldsymbol{\sigma}} \\ \boldsymbol{u} - \mu_{\boldsymbol{\mathcal{U}}|\boldsymbol{\mathcal{Y}}=\boldsymbol{y}} \end{pmatrix} \right\|^2 \tag{3.25}$$

$$= \left\| Y_\Lambda - Z_{X\Lambda} \begin{pmatrix} \widehat{\boldsymbol{\beta}}_{\boldsymbol{\theta},\rho,\boldsymbol{\sigma}} \\ \mu_{\boldsymbol{\mathcal{U}}|\boldsymbol{\mathcal{Y}}=\boldsymbol{y}} \end{pmatrix} \right\|^2 + \begin{pmatrix} \boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_{\boldsymbol{\theta},\rho,\boldsymbol{\sigma}} \\ \boldsymbol{u} - \mu_{\boldsymbol{\mathcal{U}}|\boldsymbol{\mathcal{Y}}=\boldsymbol{y}} \end{pmatrix}^\top Z_{X\Lambda}^\top Z_{X\Lambda} \begin{pmatrix} \boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_{\boldsymbol{\theta},\rho,\boldsymbol{\sigma}} \\ \boldsymbol{u} - \mu_{\boldsymbol{\mathcal{U}}|\boldsymbol{\mathcal{Y}}=\boldsymbol{y}} \end{pmatrix}. \tag{3.26}$$

$Z_{X\Lambda}^\top Z_{X\Lambda}$ can be Cholesky decomposed as

$$Z_{X\Lambda}^\top Z_{X\Lambda} = \begin{pmatrix} R_X & \mathbf{0} \\ R_{ZX} & L_{\boldsymbol{\theta},\rho,\boldsymbol{\sigma}}^\top \end{pmatrix}^\top \begin{pmatrix} R_X & \mathbf{0} \\ R_{ZX} & L_{\boldsymbol{\theta},\rho,\boldsymbol{\sigma}}^\top \end{pmatrix}, \tag{3.27}$$

where

$$L_{\boldsymbol{\theta},\rho,\boldsymbol{\sigma}} L_{\boldsymbol{\theta},\rho,\boldsymbol{\sigma}}^\top = Z_{\boldsymbol{\sigma\theta}}^\top Z_{\boldsymbol{\sigma\theta}} + \sqrt{\sigma_1^2\sigma_2^2}\Sigma_{\boldsymbol{u}}^{-1}. \tag{3.28}$$

Thereafter,

$$
\begin{aligned}
g(\boldsymbol{\beta},\boldsymbol{u},\boldsymbol{\theta},\rho,\boldsymbol{\sigma}) &= \left\|Y_{\boldsymbol{\sigma}} - X_{\boldsymbol{\sigma}}\widehat{\boldsymbol{\beta}}_{\boldsymbol{\theta},\rho,\boldsymbol{\sigma}} - Z_{\boldsymbol{\sigma}\boldsymbol{\theta}}\mu_{\boldsymbol{\mathcal{U}}|\boldsymbol{\mathcal{Y}}=\boldsymbol{y}}\right\|^2 + \sigma_1^2\sigma_2^2\mu_{\boldsymbol{\mathcal{U}}|\boldsymbol{\mathcal{Y}}=\boldsymbol{y}}^\top\Sigma_{\boldsymbol{u}}^{-1}\mu_{\boldsymbol{\mathcal{U}}|\boldsymbol{\mathcal{Y}}=\boldsymbol{y}} \\
&\quad + \left\|R_X(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}_{\boldsymbol{\theta},\rho,\boldsymbol{\sigma}})\right\|^2 + \left\|R_{ZX}(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}_{\boldsymbol{\theta},\rho,\boldsymbol{\sigma}}) + L_{\boldsymbol{\theta},\rho,\boldsymbol{\sigma}}^\top(\boldsymbol{u}-\mu_{\boldsymbol{\mathcal{U}}|\boldsymbol{\mathcal{Y}}=\boldsymbol{y}})\right\|^2.
\end{aligned} \tag{3.29}
$$

By setting

$$
r(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\theta},\rho,\boldsymbol{\sigma}},\mu_{\boldsymbol{\mathcal{U}}|\boldsymbol{\mathcal{Y}}=\boldsymbol{y}}) = \left\|Y_{\boldsymbol{\sigma}} - X_{\boldsymbol{\sigma}}\widehat{\boldsymbol{\beta}}_{\boldsymbol{\theta},\rho,\boldsymbol{\sigma}} - Z_{\boldsymbol{\sigma}\boldsymbol{\theta}}\mu_{\boldsymbol{\mathcal{U}}|\boldsymbol{\mathcal{Y}}=\boldsymbol{y}}\right\|^2 + \sigma_1^2\sigma_2^2\mu_{\boldsymbol{\mathcal{U}}|\boldsymbol{\mathcal{Y}}=\boldsymbol{y}}^\top\Sigma_{\boldsymbol{u}}^{-1}\mu_{\boldsymbol{\mathcal{U}}|\boldsymbol{\mathcal{Y}}=\boldsymbol{y}}, \tag{3.30}
$$

and returning to the calculation of $f_{\boldsymbol{\mathcal{Y}}}(\boldsymbol{y})$, it comes

$$
\begin{aligned}
f_{\boldsymbol{\mathcal{Y}}}(\boldsymbol{y}) &= \frac{\int \exp\left[-\frac{r(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\theta},\rho,\boldsymbol{\sigma}},\mu_{\boldsymbol{\mathcal{U}}|\boldsymbol{\mathcal{Y}}=\boldsymbol{y}}) + \left\|R_X(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}_{\boldsymbol{\theta},\rho,\boldsymbol{\sigma}})\right\|^2 + \left\|R_{ZX}(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}_{\boldsymbol{\theta},\rho,\boldsymbol{\sigma}}) + L_{\boldsymbol{\theta},\rho,\boldsymbol{\sigma}}^\top(\boldsymbol{u}-\mu_{\boldsymbol{\mathcal{U}}|\boldsymbol{\mathcal{Y}}=\boldsymbol{y}})\right\|^2}{2\sigma_1^2\sigma_2^2}\right] d\boldsymbol{u}}{(2\pi\sigma_1^2)^{N/2}(2\pi\sigma_2^2)^{N/2}(2\pi)^{q/2}|\Sigma_{\boldsymbol{u}}|^{1/2}} \\
&= \frac{\exp\left[-\frac{r(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\theta},\rho,\boldsymbol{\sigma}},\mu_{\boldsymbol{\mathcal{U}}|\boldsymbol{\mathcal{Y}}=\boldsymbol{y}}) + \left\|R_X(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}_{\boldsymbol{\theta},\rho,\boldsymbol{\sigma}})\right\|^2}{2\sigma_1^2\sigma_2^2}\right]}{(2\pi\sigma_1^2)^{N/2}(2\pi\sigma_2^2)^{N/2}(2\pi)^{q/2}|\Sigma_{\boldsymbol{u}}|^{1/2}} \times \\
&\quad \int \exp\left[-\frac{\left\|R_{ZX}(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}_{\boldsymbol{\theta},\rho,\boldsymbol{\sigma}}) + L_{\boldsymbol{\theta},\rho,\boldsymbol{\sigma}}^\top(\boldsymbol{u}-\mu_{\boldsymbol{\mathcal{U}}|\boldsymbol{\mathcal{Y}}=\boldsymbol{y}})\right\|^2}{2\sigma_1^2\sigma_2^2}\right] d\boldsymbol{u}.
\end{aligned} \tag{3.31}
$$

By setting $\boldsymbol{v} = R_{ZX}(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}_{\boldsymbol{\theta},\rho,\boldsymbol{\sigma}}) + L_{\boldsymbol{\theta},\rho,\boldsymbol{\sigma}}^\top(\boldsymbol{u}-\mu_{\boldsymbol{\mathcal{U}}|\boldsymbol{\mathcal{Y}}=\boldsymbol{y}})$, $d\boldsymbol{u} = \frac{1}{|L_{\boldsymbol{\theta},\rho,\boldsymbol{\sigma}}|}d\boldsymbol{v}$ and

$$
\begin{aligned}
f_{\boldsymbol{\mathcal{Y}}}(\boldsymbol{y}) &= \frac{\exp\left[-\frac{r(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\theta},\rho,\boldsymbol{\sigma}},\mu_{\boldsymbol{\mathcal{U}}|\boldsymbol{\mathcal{Y}}=\boldsymbol{y}}) + \left\|R_X(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}_{\boldsymbol{\theta},\rho,\boldsymbol{\sigma}})\right\|^2}{2\sigma_1^2\sigma_2^2}\right](\sigma_1^2\sigma_2^2)^{\frac{q}{2}}}{(2\pi\sigma_1^2)^{N/2}(2\pi\sigma_2^2)^{N/2}|\Sigma_{\boldsymbol{u}}|^{1/2}|L_{\boldsymbol{\theta},\rho,\boldsymbol{\sigma}}|} \int \frac{1}{(2\pi\sigma_1^2\sigma_2^2)^{\frac{q}{2}}}\exp\left[-\frac{\|\boldsymbol{v}\|^2}{2\sigma_1^2\sigma_2^2}\right] d\boldsymbol{v} \\
&= \frac{\exp\left[-\frac{r(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\theta},\rho,\boldsymbol{\sigma}},\mu_{\boldsymbol{\mathcal{U}}|\boldsymbol{\mathcal{Y}}=\boldsymbol{y}}) + \left\|R_X(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}_{\boldsymbol{\theta},\rho,\boldsymbol{\sigma}})\right\|^2}{2\sigma_1^2\sigma_2^2}\right](\sigma_1^2\sigma_2^2)^{\frac{q}{2}}}{(2\pi\sigma_1^2)^{N/2}(2\pi\sigma_2^2)^{N/2}|\Sigma_{\boldsymbol{u}}|^{1/2}|L_{\boldsymbol{\theta},\rho,\boldsymbol{\sigma}}|}.
\end{aligned} \tag{3.32}
$$

The log-likelihood to be maximized can therefore be expressed as,

$$
\begin{aligned}
\ell(\boldsymbol{\beta},\boldsymbol{\theta},\rho,\boldsymbol{\sigma}|\boldsymbol{y}) &= -\frac{r(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\theta},\rho,\boldsymbol{\sigma}},\mu_{\boldsymbol{\mathcal{U}}|\boldsymbol{\mathcal{Y}}=\boldsymbol{y}}) + \left\|R_X(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}_{\boldsymbol{\theta},\rho,\boldsymbol{\sigma}})\right\|^2}{2\sigma_1^2\sigma_2^2} - \frac{N-q}{2}\log(\sigma_1^2\sigma_2^2) \\
&\quad -\frac{1}{2}\log(|\Sigma_{\boldsymbol{u}}|) - \frac{1}{2}\log(|L_{\boldsymbol{\theta},\rho,\boldsymbol{\sigma}}|^2).
\end{aligned} \tag{3.33}
$$

□

By profiling out $\boldsymbol{\beta}$, the partially profiled log-likelihood is

$$\begin{aligned}\tilde{\ell}(\boldsymbol{\theta},\rho,\boldsymbol{\sigma}|\boldsymbol{y}) &= -\frac{r(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\theta},\rho,\boldsymbol{\sigma}},\mu_{\boldsymbol{\mathcal{U}}|\boldsymbol{\mathcal{Y}}=\boldsymbol{y}})}{2\sigma_1^2\sigma_2^2}-\frac{N-q}{2}\log(\sigma_1^2\sigma_2^2)\\ &\quad -\frac{1}{2}\log(|\Sigma_{\boldsymbol{u}}|)-\frac{1}{2}\log(|L_{\boldsymbol{\theta},\rho,\boldsymbol{\sigma}}|^2),\end{aligned} \tag{3.34}$$

replacing $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\theta},\rho,\boldsymbol{\sigma}}$ by $\boldsymbol{\beta}$. Then, the partially profiled deviance comes

$$\begin{aligned}-2\tilde{\ell}(\boldsymbol{\theta},\rho,\boldsymbol{\sigma}|\boldsymbol{y}) &= \frac{r(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\theta},\rho,\boldsymbol{\sigma}},\mu_{\boldsymbol{\mathcal{U}}|\boldsymbol{\mathcal{Y}}=\boldsymbol{y}})}{\sigma_1^2\sigma_2^2}+(N-q)\log(\sigma_1^2\sigma_2^2)\\ &\quad +\log(|\Sigma_{\boldsymbol{u}}|)+\log(|L_{\boldsymbol{\theta},\rho,\boldsymbol{\sigma}}|^2).\end{aligned} \tag{3.35}$$

This deviance is finally the criterion which will be minimized to obtaining the ML estimates of the parameters.

Corollary 3.3.1. *Suppose that $\boldsymbol{y}=(y_1^\top,y_2^\top)^\top$ satisfies the bivariate linear mixed-effects model expressed by Equations (3.8 and 3.9). Taking into account the notations in the Theorem 4.2.1, the ML estimators $\widehat{\boldsymbol{\beta}}$, $\widehat{\boldsymbol{\sigma}}$, $\widehat{\boldsymbol{\theta}}$, $\widehat{\rho}$ of $\boldsymbol{\beta}$, $\boldsymbol{\sigma}$, $\boldsymbol{\theta}$ and ρ satisfy*

$$\left(\widehat{\boldsymbol{\theta}},\widehat{\rho},\widehat{\boldsymbol{\sigma}}\right)=\underset{\boldsymbol{\theta},\rho,\boldsymbol{\sigma}}{\arg\max}\,\tilde{\ell}(\boldsymbol{\theta},\rho,\boldsymbol{\sigma}|\boldsymbol{y}) \quad \textit{and} \quad \widehat{\boldsymbol{\beta}}=\widehat{\boldsymbol{\beta}}_{\widehat{\boldsymbol{\theta}},\widehat{\rho},\widehat{\boldsymbol{\sigma}}}. \tag{3.36}$$

3.3.2 REML criterion

By integrating the marginal density of $\boldsymbol{\mathcal{Y}}$ with respect to the fixed effects, the REML criterion can be obtained [105]. This REML criterion is expressed through the following theorem.

Theorem 3.3.2. *Suppose that $\boldsymbol{y}=(y_1^\top,y_2^\top)^\top$ satisfies the bivariate linear mixed-effects model expressed by Equations (3.8 and 3.9). Taking into account the notations in the Theorem 4.2.1, the REML criterion of $\boldsymbol{\sigma}$, $\boldsymbol{\theta}$ and ρ given $\boldsymbol{y}$ is expressed as*

$$\mathscr{L}(\boldsymbol{\sigma},\boldsymbol{\theta},\rho|\boldsymbol{y})=\frac{\exp\left[-\frac{r(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\theta},\rho,\boldsymbol{\sigma}},\mu_{\mathcal{U}|\mathcal{Y}=\boldsymbol{y}})}{2\sigma_1^2\sigma_2^2}\right](\sigma_1^2\sigma_2^2)^{\frac{p+q-N}{2}}}{(2\pi)^{(2N-p)/2}|\Sigma_{\boldsymbol{u}}|^{1/2}|L_{\boldsymbol{\theta},\rho,\boldsymbol{\sigma}}||R_X|}. \tag{3.37}$$

Proof.

$$\mathscr{L}(\boldsymbol{\sigma},\boldsymbol{\theta},\rho|\boldsymbol{y}) = \int_{\mathbb{R}^p} f_{\boldsymbol{\mathcal{Y}}}(\boldsymbol{y})d\boldsymbol{\beta} \tag{3.38}$$

$$= \frac{\exp\left[-\frac{r(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\theta},\rho,\boldsymbol{\sigma}},\mu_{\mathcal{U}|\mathcal{Y}=\boldsymbol{y}})}{2\sigma_1^2\sigma_2^2}\right](\sigma_1^2\sigma_2^2)^{\frac{q}{2}}}{(2\pi\sigma_1^2)^{N/2}(2\pi\sigma_2^2)^{N/2}|\Sigma_{\boldsymbol{u}}|^{1/2}|L_{\boldsymbol{\theta},\rho,\boldsymbol{\sigma}}|}\int_{\mathbb{R}^p}\exp\left[-\frac{\left\|R_X(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}_{\boldsymbol{\theta},\rho,\boldsymbol{\sigma}})\right\|^2}{2\sigma_1^2\sigma_2^2}\right]d\boldsymbol{\beta}$$

$$= \frac{\exp\left[-\frac{r(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\theta},\rho,\boldsymbol{\sigma}},\mu_{\mathcal{U}|\mathcal{Y}=\boldsymbol{y}})}{2\sigma_1^2\sigma_2^2}\right](\sigma_1^2\sigma_2^2)^{\frac{p+q-N}{2}}}{(2\pi)^{(2N-p)/2}|\Sigma_{\boldsymbol{u}}|^{1/2}|L_{\boldsymbol{\theta},\rho,\boldsymbol{\sigma}}||R_X|}\int_{\mathbb{R}^p}(2\pi\sigma_1^2\sigma_2^2)^{-\frac{p}{2}}\exp\left[-\frac{\|\boldsymbol{t}\|^2}{2\sigma_1^2\sigma_2^2}\right]d\boldsymbol{t},$$

where $\boldsymbol{t}=\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}_{\boldsymbol{\theta},\rho,\boldsymbol{\sigma}} \implies d\boldsymbol{\beta}=\frac{1}{|R_X|}d\boldsymbol{t}$ and $\int_{\mathbb{R}^p}(2\pi\sigma_1^2\sigma_2^2)^{-\frac{p}{2}}\exp\left[-\frac{\|\boldsymbol{t}\|^2}{2\sigma_1^2\sigma_2^2}\right]d\boldsymbol{t}=1$. □

The REML criterion can also be expressed as

$$\begin{aligned}\log\left(\mathscr{L}(\boldsymbol{\sigma},\boldsymbol{\theta},\rho|\boldsymbol{y})\right) &= \tilde{\mathscr{L}}(\boldsymbol{\sigma},\boldsymbol{\theta},\rho|\boldsymbol{y})\\ &= -\frac{r(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\theta},\rho,\boldsymbol{\sigma}},\mu_{\mathcal{U}|\mathcal{Y}=\boldsymbol{y}})}{2\sigma_1^2\sigma_2^2}+\frac{p+q-N}{2}\log(\sigma_1^2\sigma_2^2)-\frac{1}{2}\log(|\Sigma_{\boldsymbol{u}}|)\\ &\quad -\frac{1}{2}\log(|L_{\boldsymbol{\theta},\rho,\boldsymbol{\sigma}}|^2)-\frac{1}{2}\log(|R_X|^2),\end{aligned} \tag{3.39}$$

or as

$$\begin{aligned}-2\tilde{\mathscr{L}}(\boldsymbol{\sigma},\boldsymbol{\theta},\rho|\boldsymbol{y}) &= \frac{r(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\theta},\rho,\boldsymbol{\sigma}},\mu_{\mathcal{U}|\mathcal{Y}=\boldsymbol{y}})}{\sigma_1^2\sigma_2^2}+(N-p-q)\log(\sigma_1^2\sigma_2^2)+\log(|\Sigma_{\boldsymbol{u}}|)\\ &\quad +\log(|L_{\boldsymbol{\theta},\rho,\boldsymbol{\sigma}}|^2)+\log(|R_X|^2),\end{aligned} \tag{3.40}$$

which will be minimized to obtaining the REML estimates of the parameters.

3.4 Simulation studies

In this Section, the consistency of the estimators is proven through simulation studies, and we compare the present estimation procedure with the EM algorithm. For the sake of simplicity, these simulation studies are performed using simulated bivariate longitudinal data sets. In the following paragraph, we explain how we choose the parameters that have been used to simulate the 'working' data sets.

The working data sets We suppose that we are following up a sample of subjects where the goal is to evaluate how the growth of the weight and the height of the individuals of this population are jointly explained by the sex, the score of nutrition (Nscore) and the age. We randomly choose through a uniform distribution the score of nutrition between 20 and 50, and the age between 18 and 37, using the R software. All the computations in this paper are done using the R software. The subject's sex is also randomly chosen. The model under which we simulate the data sets is the following:
n indicating the total number of subjects, for $i=1,\ldots,n$

$$\begin{aligned}\text{weight}_i &= (\mathbb{1}_{n_i},\text{sex}_i,\text{Nscore}_i,\text{age}_i)\beta_1+(\mathbb{1}_{n_i},\text{Nscore}_i)\gamma_{1i}+\varepsilon_{1i}\\ \text{height}_i &= (\mathbb{1}_{n_i},\text{sex}_i,\text{Nscore}_i,\text{age}_i)\beta_2+(\mathbb{1}_{n_i},\text{Nscore}_i)\gamma_{2i}+\varepsilon_{2i}\end{aligned} \tag{3.41}$$

with

$$\boldsymbol{\gamma}_i = \begin{pmatrix} \gamma_{1i} \\ \gamma_{2i} \end{pmatrix} \sim \mathcal{N}\left(\mathbf{0}, \bar{\boldsymbol{\Gamma}}\right), \varepsilon_{1i} \sim \mathcal{N}\left(0, \sigma_1^2 \mathrm{I}_{n_i}\right), \varepsilon_{2i} \sim \mathcal{N}\left(0, \sigma_2^2 \mathrm{I}_{n_i}\right), \gamma_i \perp \varepsilon_{1i} \perp \varepsilon_{2i} \tag{3.42}$$

The random effect related to the dependent variable 'weight' or 'height' is a vector composed by one random intercept and one random slope in the direction of the covariate 'Nscore'. The total number of observations is denoted by N.

We randomly choose $\beta_1, \beta_2, \sigma_1$ and σ_2 whose values are in the first column of Table 3.3. $\bar{\boldsymbol{\Gamma}}$ is also randomly chosen such that it is positive definite, with the following form:

$$\bar{\boldsymbol{\Gamma}} = \begin{pmatrix} \eta_1^2 & \rho_\eta \eta_1 \eta_2 & \rho \eta_1 \tau_1 & \rho \eta_1 \tau_2 \\ \rho_\eta \eta_1 \eta_2 & \eta_2^2 & \rho \eta_2 \tau_1 & \rho \eta_2 \tau_2 \\ \rho \eta_1 \tau_1 & \rho \eta_2 \tau_1 & \tau_1^2 & \rho_\tau \tau_1 \tau_2 \\ \rho \eta_1 \tau_2 & \rho \eta_2 \tau_2 & \rho_\tau \tau_1 \tau_2 & \tau_2^2 \end{pmatrix} \tag{3.43}$$

In order to have an almost strong correlation between the marginal random effects, we set $\rho = 0.8$ and randomly choose all other parameters involved in the obtaining of $\bar{\boldsymbol{\Gamma}}$. Thus, the obtained $\bar{\boldsymbol{\Gamma}}$ is

$$\bar{\boldsymbol{\Gamma}} = \begin{pmatrix} 27.77 & 18.80 & 41.70 & 4.93 \\ 18.80 & 36.00 & 47.47 & 5.62 \\ 41.70 & 47.47 & 97.81 & 8.91 \\ 4.93 & 5.62 & 8.91 & 1.37 \end{pmatrix} \tag{3.44}$$

3.4.1 Estimates' performances

One practical way to show the consistency of an estimator is by computing its Mean Square Error (MSE). If the MSE of an estimator is asymptotically null, this estimator converges in probability, and is then consistent. In this Section, we gradually simulate data sets with larger sizes , $(N, n) \in \{(600, 50), (600, 60), \ldots, (1000, 100), (1000, 300), \ldots, (15000, 1000)\}$. We simulate one hundred data sets of each size and calculate the estimators' MSE using these data sets. This yields one hundred MSE for each size of dataset. This allows to compute the 95% CI (confidence interval) along with the mean of the MSE (one hundred mse) obtained for the hundred data sets of the same size. The results are contained in Table 3.1 and Table 3.2. The Table 3.1 shows that the asymptotique in the longitudinal data requires not only $n \to \infty$ and $N \to \infty$, but also $N/n \to \infty$. This means that it requires a sufficient total number of observations, a sufficient number of levels for the grouping factor and a sufficient number of observations for each level of the grouping factor. For example, in the Table 3.1, when the total number of observations is $N = 1000$, the MSE of $\bar{\boldsymbol{\Gamma}}$ is better for $n = 100$, 0.47 $(0.03 - 1.29)$, than for $n = 300$, 0.69 $(0.06 - 1.94)$. Observing both Table 3.1 and Table 3.2 it is clear that, as the number of observations increase, the MSE descends to 0. We can conclude that the estimators constructed in this paper are consistent. The estimation procedure discussing in this paper may therefore be named Consistent estimates for the Multivariate Linear Mixed-Effects model (Cmlme). The Cmlme acronym will be used in the remainder of the paper for a question of simplicity.

Table 3.1: **Mean Square Error of estimators with** 95% **CI estimated on** 100 **replications for values of** $n \in \{50, 60, 100, 300\}$ **and** $N \in \{600, 1000, 3000\}$**.**

Parameter	n	$N = 600$	$N = 1000$	$N = 3000$
β_1	50	2.43 (0.11 - 7.11)	1.89 (0.22 - 4.89)	1.02 (0.07 - 2.44)
	60	2.57 (0.14 - 7.87)	2.13 (0.26 - 5.55)	0.77 (0.10 - 2.09)
	100	2.27 (0.16 - 5.61)	1.55 (0.14 - 5.17)	0.71 (0.04 - 1.85)
	300	3.16 (0.14 - 10.54)	1.70 (0.10 - 4.55)	0.51 (0.04 - 1.36)
β_2	50	5.50(0.09 - 16.35)	3.26 (0.02 - 11.64)	2.06 (0.09 - 5.98)
	60	5.06 (0.12 - 15.09)	3.22 (0.02 - 10.24)	1.78 (0.10 - 5.62)
	100	4.33 (0.03 - 13.17)	2.37 (0.02 - 6.89)	1.06 (0.02 - 3.60)
	300	4.58 (0.18 - 15.92)	2.43 (0.05 - 7.39)	0.90 (0.04 - 2.88)
σ_1	50	0.03 (0.00 - 0.14)	0.02 (0.00 - 0.09)	0.00 (0.00 - 0.02)
	60	0.04 (0.00 - 0.11)	0.02 (0.00 - 0.08)	0.00 (0.00 - 0.02)
	100	0.03 (0.00 - 0.13)	0.01 (0.00 - 0.06)	0.00 (0.00 - 0.01)
	300	0.05 (0.00 - 0.23)	0.03 (0.00 - 0.13)	0.00 (0.00 - 0.02)
σ_2	50	0.04 (0.00 - 0.15)	0.03 (0.00 - 0.08)	0.00 (0.00 - 0.02)
	60	0.04 (0.00 - 0.18)	0.03 (0.00 - 0.12)	0.01 (0.00 - 0.03)
	100	0.06 (0.00 - 0.18)	0.03 (0.00 - 0.11)	0.01 (0.00 - 0.03)
	300	0.13 (0.00 - 0.45)	0.04 (0.00 - 0.15)	0.01 (0.00 - 0.05)
$\bar{\mathbf{\Gamma}}$	50	0.90 (0.12 - 2.41)	0.62 (0.06 - 1.41)	0.45 (0.04 - 1.14)
	60	1.07 (0.06 - 2.64)	0.68 (0.08 - 1.98)	0.25 (0.03 - 0.66)
	100	0.90 (0.05 - 2.40)	0.47 (0.03 - 1.29)	0.21 (0.02 - 0.71)
	300	1.45 (0.19 - 4.39)	0.69 (0.06 - 1.94)	0.17 (0.02 - 0.57)

Table 3.2: **Mean Square Error of estimators with** 95% **CI estimated on** 100 **replications for values of** $(n, N) \in \{(500, 7000), (600, 8000), (800, 10000), (1000, 15000)\}$**.**

Parameter	$n = 500, N = 7000$	$n = 600, N = 8000$	$n = 800, N = 10000$	$n = 1000, N = 15000$
β_1	0.22 (0.01 - 0.62)	0.17 (0.01 - 0.48)	0.16 (0.01 - 0.47)	0.11 (0.00 - 0.32)
β_2	0.32 (0.00 - 1.01)	0.34 (0.01 - 1.16)	0.25 (0.02 - 0.67)	0.21 (0.00 - 0.69)
σ_1	0.00 (0.00 - 0.00)	0.00 (0.00 - 0.01)	0.00 (0.00 - 0.00)	0.00 (0.00 - 0.00)
σ_2	0.00 (0.00 - 0.01)	0.00 (0.00 - 0.01)	0.00 (0.00 - 0.01)	0.00 (0.00 - 0.00)
$\bar{\mathbf{\Gamma}}$	0.09 (0.00 - 0.25)	0.07 (0.00 - 0.19)	0.06 (0.00 - 0.19)	0.03 (0.00 - 0.09)

3.4.2 Comparison with EM-based estimates

In this Section, we compare the estimation procedure based on EM algorithm with the Cmlme. This comparison is performed regarding the accuracy of the estimates, whether or not the starting values of the two algorithms (EM and Cmlme) are naive or advised. We mean by naive starting values, those which are randomly chosen (without specific control), and by advised starting values those obtained by fitting separately each dimension of the bivariate model. The results of these marginal fitting are, indeed, the advised starting values for the bivariate model estimation procedure. The starting values are the same for both Cmlme and EM algorithms. The number of iteration required for convergence, for each algorithm, is also discussed. Our methodology consists in simulating thirty longitudinal data sets of size $(N = 3000, n = 300)$ and fit the model to each of these data sets using the EM algorithm and the Cmlme, respectively. This allows to compute both the 95% CI and the empirical mean of the thirty estimates in each case (naive and advised starting values). The obtained results are in Table 3.3 and Table 3.4. Table 3.3 contains the empirical means of the estimates with their 95% CI, and the minimum, the maximum and the average number of iterations. Table 3.4 contains the empirical relative error of the estimators with their 95% CI. These results show that in the case of naive initialization, the Cmlme estimators outperform the EM estimators. For example, the component of β_1 which is 14.00 is well estimated by Cmlme, $14.02\ (13.27 - 14.45)$ with an empirical relative error of $0.02\ (0.00 - 0.04)$, but poorly estimated by EM, $-2.05\ (-4.70 - -0.40)$ with an empirical relative error of $1.14\ (1.01 - 1.32)$. In the case of advised initialization, both Cmlme and EM algorithms perform well, but Cmlme converge faster (64 iterations in average) than EM (169 iterations in average). The number of iteration required by Cmlme whith advised initializations range from 48 to 89 and from 56 to 103 for naive initializations. The Cmlme with naive initialization therefore needs more iterations than Cmlme with advised initialization for converge. This is expected and may be explained by the fact that the advised initialization values contain some information from the data of interest, and the naive starting points do not.

The empirical mean of the random effects covariance matrix, $\bar{\Gamma}$, is well estimated with advised initializations:

$$\bar{\Gamma}_{\text{adv}}^{\text{Cmlme}} = \begin{pmatrix} 25.74 & 16.34 & 35.75 & 4.50 \\ 16.34 & 34.83 & 43.97 & 5.37 \\ 35.75 & 43.97 & 82.44 & 8.43 \\ 4.50 & 5.37 & 8.43 & 1.32 \end{pmatrix}, \quad \text{with} \quad \sigma_{\Gamma_{\text{adv}}}^{\text{Cmlme}} = \begin{pmatrix} 5.73 & 2.74 & 4.45 & 0.61 \\ 2.74 & 2.59 & 3.96 & 0.44 \\ 4.45 & 3.96 & 12.07 & 0.80 \\ 0.61 & 0.44 & 0.80 & 0.11 \end{pmatrix} \tag{3.45}$$

and

$$\bar{\Gamma}_{\text{adv}}^{\text{EM}} = \begin{pmatrix} 23.66 & 16.45 & 32.60 & 4.45 \\ 16.45 & 34.82 & 43.39 & 5.39 \\ 32.61 & 43.39 & 75.16 & 8.66 \\ 4.45 & 5.39 & 8.65 & 1.31 \end{pmatrix}, \quad \text{with} \quad \sigma_{\Gamma_{\text{adv}}}^{\text{EM}} = \begin{pmatrix} 7.49 & 2.71 & 5.29 & 0.76 \\ 2.71 & 2.59 & 3.71 & 0.45 \\ 5.30 & 3.71 & 13.36 & 0.80 \\ 0.76 & 0.45 & 0.80 & 0.11 \end{pmatrix} \tag{3.46}$$

Table 3.3: **EM estimates compared with Cmlme estimates on the same data sets. Empirical estimates with their 95% CI and the number of iteration required for convergence.**

		Naive initialization				Advised initialization			
		Cmlme		EM		Cmlme		EM	
Parameter	Value	Emp. Mean	95% CI	Emp. Mean	95% CI	Emp. Mean	95% CI	Emp. Mean	95% CI
β_1	50.67	50.79	$49.14 - 52.11$	13.47	$-76.70 - 43.37$	50.80	$49.15 - 52.12$	50.78	$49.09 - 52.01$
	-4.80	-5.00	$-8.39 - -3.66$	-4.79	$-8.08 - -3.42$	-5.02	$-8.39 - -3.66$	-4.98	$-8.38 - -3.65$
	14.00	14.02	$13.27 - 14.45$	-2.05	$-4.70 - -0.40$	14.02	$13.28 - 14.45$	14.02	$13.27 - 14.45$
	2.70	2.70	$2.66 - 2.72$	2.69	$2.66 - 2.72$	2.70	$2.66 - 2.72$	2.70	$2.66 - 2.72$
β_2	13.20	13.65	$11.79 - 15.06$	-84.47	$-114.28 - -50.63$	13.65	$11.79 - 15.07$	13.68	$11.81 - 15.14$
	-2.80	-2.80	$-4.74 - -0.43$	-2.75	$-4.90 - 0.21$	-2.81	$-4.79 - -0.43$	-2.85	$-4.80 - -0.51$
	27.00	27.00	$26.87 - 27.10$	0.90	$-1.62 - 2.68$	27.00	$26.87 - 27.10$	27.00	$26.87 - 27.10$
	1.70	1.68	$1.64 - 1.71$	1.68	$1.64 - 1.71$	1.68	$1.64 - 1.71$	1.68	$1.64 - 1.71$
σ_1	5.80	5.79	$5.62 - 5.92$	5.78	$5.64 - 5.98$	5.78	$5.64 - 5.92$	5.79	$5.65 - 5.94$
σ_2	7.60	7.61	$7.34 - 7.74$	7.59	$7.33 - 7.73$	7.61	$7.34 - 7.73$	7.63	$7.36 - 7.73$
Nbr. of iteration	Min	56	-	63	-	48	-	14	-
	Mean	71	-	109	-	64	-	169	-
	Max	103	-	157	-	89	-	645	-

$\sigma^{\text{Cmlme}}_{\bar{\Gamma}_{\text{adv}}}$ and $\sigma^{\text{EM}}_{\bar{\Gamma}_{\text{adv}}}$ contain the standard deviations of the entries of $\bar{\Gamma}^{\text{Cmlme}}_{\text{adv}}$ and $\bar{\Gamma}^{\text{EM}}_{\text{adv}}$, respectively. It seems that the empirical standard deviation of the higher entries of $\bar{\Gamma}$ are bigger with EM than with Cmlme. For example, the standard deviation of $\bar{\Gamma}_{11} = 27.77$ is 7.49 for EM, but 5.73 for Cmlme. Same remark about the standard deviations of $\bar{\Gamma}_{31}$, $\bar{\Gamma}_{32}$ and $\bar{\Gamma}_{33}$, comparing EM and Cmlme. This may be explained by the fact that Cmlme estimators are more consistent than EM

Table 3.4: **EM estimates compared with Cmlme estimates on the same data sets. Empirical relative error of the estimates with their 95% CI**

		Naive initialization				**Advised initialization**			
		Cmlme		**EM**		**Cmlme**		**EM**	
Parameter	**Value**	**R. Error**	**95% CI**	**R. Error**	**95% CI**	**R. Error**	**95% CI**	**R. Error**	**95% CI**
	50.67	0.01	$0.00 - 0.03$	0.73	$0.01 - 2.18$	0.01	$0.00 - 0.03$	0.01	$0.00 - 0.03$
β_1	-4.80	0.21	$0.02 - 0.32$	0.21	$0.00 - 0.32$	0.21	$0.02 - 0.32$	0.21	$0.00 - 0.33$
	14.00	0.02	$0.00 - 0.04$	1.14	$1.01 - 1.32$	0.02	$0.00 - 0.04$	0.02	$0.00 - 0.04$
	2.70	0.00	$0.00 - 0.01$	0.00	$0.00 - 0.01$	0.00	$0.00 - 0.01$	0.00	$0.00 - 0.01$
	13.20	0.07	$0.00 - 0.14$	7.39	$4.19 - 9.40$	0.07	$0.00 - 0.14$	0.07	$0.00 - 0.14$
β_2	-2.80	0.43	$0.00 - 0.84$	0.43	$0.02 - 1.07$	0.43	$0.00 - 0.84$	0.43	$0.00 - 0.81$
	27.00	0.00	$0.00 - 0.00$	0.96	$0.88 - 1.02$	0.00	$0.00 - 0.00$	0.00	$0.00 - 0.00$
	1.70	0.01	$0.00 - 0.02$	0.01	$0.00 - 0.02$	0.01	$0.00 - 0.02$	0.01	$0.00 - 0.03$
σ_1	5.80	0.01	$0.00 - 0.02$	0.01	$0.00 - 0.03$	0.01	$0.00 - 0.02$	0.01	$0.00 - 0.03$
σ_2	7.60	0.01	$0.00 - 0.02$	0.01	$0.00 - 0.04$	0.01	$0.00 - 0.02$	0.01	$0.00 - 0.02$

estimators. In the case of naive initializations, Cmlme provides a well estimated empirical mean of $\bar{\Gamma}$, when the estimated $\bar{\Gamma}$ provided by EM is very bad (we choose not to show it here).

$$\bar{\Gamma}^{\text{Cmlme}}_{\text{naiv}} = \begin{pmatrix} 24.76 & 16.18 & 34.43 & 4.47 \\ 16.18 & 34.77 & 43.91 & 5.35 \\ 34.43 & 43.91 & 80.67 & 8.40 \\ 4.47 & 5.35 & 8.40 & 1.32 \end{pmatrix}, \quad \text{with} \quad \sigma^{\text{Cmlme}}_{\Gamma_{\text{naiv}}} = \begin{pmatrix} 7.58 & 2.67 & 6.76 & 0.61 \\ 2.67 & 2.54 & 3.91 & 0.42 \\ 6.76 & 3.91 & 13.37 & 0.80 \\ 0.61 & 0.42 & 0.80 & 0.11 \end{pmatrix} \tag{3.47}$$

$\bar{\Gamma}^{\text{Cmlme}}_{\text{naiv}}$ compared to $\bar{\Gamma}^{\text{Cmlme}}_{\text{adv}}$ and $\sigma^{\text{Cmlme}}_{\Gamma_{\text{naiv}}}$ compared to $\sigma^{\text{Cmlme}}_{\Gamma_{\text{adv}}}$ show a slight difference which reveals a tiny sensibility of Cmlme to the starting values. This may be corrected by doing more than one evaluation of the model's deviance.

For all the simulation studies, we use the ML deviance criterion (Equation 3.35) and have minimized it using the nlminb function under R software. Thus, the estimates obtained are from the ML estimators. In this paper, we do not provide an application of REML estimates.

3.5 Application on malaria dataset

3.5.1 Data description

The data that we analyze here come from a study which was conducted in 9 villages (Avamé centre, Gbédjougo, Houngo, Anavié, Dohinoko, Gbétaga, Tori Cada Centre, Zébè and Zoungoudo) of Tori Bossito area (Southern Benin), where *P. falciparum* is the commonest species in the study area (95%) [47] from June 2007 to January 2010. The aim of this study was to evaluate the determinants of malaria incidence in the first months of life of child in Benin.

Mothers ($n = 620$) were enrolled at delivery and their newborns were actively followed-up during the first year of life. One questionnaire was conducted to gather information on women's characteristics (age, parity, use of Intermittent Preventive Treatment during pregnancy (IPTp) and bed net possession) and on the course of their current pregnancy. After delivery, thick and thin placental blood smears were examined to detect placental infection defined by the presence of asexual forms of *P. falciparum*. Maternal peripheral blood as well as cord blood were collected. At birth, newborn's weight and length were measured and gestational age was estimated.

During the follow-up of newborns, axillary temperature was measured weekly. In case of temperature higher than $37.5°$C, mothers were told to bring their children to the health center where a questionnaire was filled out. A rapid diagnostic test (RDT) for malaria was performed and a thick blood smear (TBS) made. Symptomatic malaria cases, defined as fever ($> 37.5°$C) with TBS and/or RDT positive, were treated with an artemisinin-based combination. Systematically, TBS were made every month to detect asymptomatic infections. Every three months, venous blood was sampled to quantify the level of antibody against malaria promised candidate vaccine antigens. Finally, the environmental risk of exposure to malaria was modeled for each child, derived from a statistical predictive model based on climatic, entomological parameters, and characteristics of children's immediate surroundings. Also every 3 months (at $3, 6, 9, 12, 15, 18$ months 130 of age), infant blood samples were collected.

Concerning the antibody quantification, two recombinant *P. falciparum* antigens where used to perform IgG subclass (IgG1 and IgG3) antibody. Recombinants antigens MSP2 (3D7 and

FC27) were from La Trobe University [7, 124]. GLURP-R0 (amino acids 25-514, F32 strain) and GLURP-R2 (amino acids 706-1178, 140 F32 strain) were also expressed. The antibodies were quantified in plasma at different times and ADAMSEL FLPb039 software (`http://www.malariaresearch.eu/content/software`) was used to analyze automatically the ELISA optical density (OD) leading to antibody concentrations in (μg/mL).

In this paper, we use some of the data and we rename the proteins used in the study, for reasons of the protection of these data. Thus, the proteins we use here, are named A1, A2, B and C, and are related to the antigens IgG1 and IgG3 as mentioned above. Information contained in the multivariate longitudinal dataset of malaria are described in the Table 3.5, where Y denotes an antigen which is one of the following:

$$\text{IgG1\_A1}, \text{IgG3\_A1}, \text{IgG1\_A2}, \text{IgG3\_A2}, \text{IgG1\_B}, \text{IgG3\_B}, \text{IgG1\_C}, \text{IgG3\_C} \tag{3.48}$$

Table 3.5: **Variables present in the empirical dataset**

N°	Variable	Description
1	id	Child ID
2	conc.Y	concentration of Y
3	conc_CO.Y	Measured concentration of Y in the umbilical cord blood
4	conc_M3.Y	Predicted concentration of Y in the child's peripheral blood at 3 months
5	ap	Placental apposition
6	hb	Hemoglobin level
7	inf_trim	Number of malaria infections in the previous 3 months
8	pred_trim	Quarterly average number of mosquitoes child is exposed to
9	nutri_trim	Quarterly average nutrition scores

3.5.2 Data analysis

The aim of the analysis of these data is to evaluate the effect of the malaria infection on the child's immune acquisition (against malaria). Since the antigens which characterize the child's immune status interact together in the human body, we analyze the characteristics of the joint distribution of these antigens, conditionally to the malaria infection and other factors of interest. The dependent variables are then provided by conc.Y (Table 3.5) which describes the level of the antigen Y in the children at 3, 6, 9, 12, 15 and 18 months. All other variables in the Table 3.5 are covariates. We then have 8 dependent variables which describe the longitudinal profile (in the child) of the proteins listed in Equation 3.48.

To illustrate the stability of our approach, we are fitting here a bivariate model to the data, with IgG1_A1 and IgG3_A2 as dependent variables:

$$\begin{aligned} \text{conc.IgG1\_A1} &= (\mathbb{1}, \text{ap}, \text{conc\_CO.IgG1\_A1}, \text{conc\_M3.IgG1\_A1}, \text{hb}, \text{inf\_trim}, \\ &\quad \text{pred\_trim}, \text{nutri\_trim})\beta_1 + (\mathbb{1}, \text{inf\_trim})\gamma_1 + \varepsilon_1 \end{aligned}$$

$$\begin{aligned}\text{conc.IgG3\_A2} &= (\mathbb{1}, \text{ap}, \text{conc\_CO.IgG3\_A2}, \text{conc\_M3.IgG3\_A2}, \text{hb}, \text{inf\_trim},\\ &\quad \text{pred\_trim}, \text{nutri\_trim})\beta_2 + (\mathbb{1}, \text{inf\_trim})\gamma_2 + \varepsilon_2, \end{aligned} \tag{3.49}$$

with

$$\boldsymbol{\gamma} = (\gamma_1^\top, \gamma_2^\top)^\top \sim \mathcal{N}\left(\mathbf{0} \quad \bar{\boldsymbol{\Gamma}}\right), \quad \boldsymbol{\varepsilon} = (\varepsilon_1^\top, \varepsilon_2^\top)^\top \sim \mathcal{N}\left(\mathbf{0}, \begin{pmatrix} \sigma_1^2\mathbf{I} & 0 \\ 0 & \sigma_2^2\mathbf{I} \end{pmatrix}\right). \tag{3.50}$$

Our strategy is to 1) fit the model to the data by running the Cmlme algorithm using 25 different naive starting points and 2) retain the estimates related to the best likelihood (the minimum of the 25 deviances) as the true parameters and compute the estimators' MSE using the 24 others estimates. This may allows to evaluate how much the Cmlme algorithm is sensitive to the starting points. The results are contained in the Table 3.6.

Table 3.6: **Empirical data analysis.**

	Response variables			
	conc.IgG1_A1		**conc.IgG3_A2**	
Covariates	**Estimate**	**MSE**	**Estimate**	**MSE**
Intercept	0.609	9.05×10^{-5}	-1.626	3.18×10^{-5}
ap	-0.093	1.06×10^{-5}	-0.337	1.04×10^{-6}
conc_CO.IgG1_A1	0.160	1.68×10^{-6}	–	–
conc_M3.IgG1_A1	0.148	9.85×10^{-6}	–	–
conc_CO.IgG3_A2	–	–	0.047	6.44×10^{-7}
conc_M3.IgG3_A2	–	–	0.155	2.22×10^{-7}
hb	-0.162	3.22×10^{-7}	-0.345	1.35×10^{-7}
inf_trim	0.369	1.89×10^{-6}	0.696	5.09×10^{-7}
pred_trim	-0.003	5.25×10^{-8}	0.017	1.49×10^{-8}
nutri_trim	0.024	5.81×10^{-6}	0.115	3.71×10^{-5}
σ_1 and σ_2	1.395	4.96×10^{-6}	1.626	2.42×10^{-5}

Based on these results, the influence of the starting points on the Cmlme algorithm is very low (see the MSE in Table 3.6). The estimated random effects covariance matrix is

$$\boldsymbol{\Gamma} = \begin{pmatrix} 0.58 & -0.13 & 0.74 & -0.36 \\ -0.13 & 0.23 & -0.39 & 0.37 \\ 0.74 & -0.39 & 0.94 & -0.24 \\ -0.36 & 0.37 & -0.24 & 0.34 \end{pmatrix} \tag{3.51}$$

with an MSE of 0.0095.

3.6 Conclusion

In the context of fitting multivariate linear mixed-effects model having homoscedastic dimensional residuals, we have suggested ML and REML estimation strategies by profiling the model's

deviance and Cholesky factorizing the random effect covariance matrix. This approach can be considered as the generalization of the approach used by [16] in the R software lme4 package. Through extensive simulation studies, we have illustrated that the present approach outperforms the traditional EM estimates and provides estimates that are consitent for both fixed effects and variance components. Another interesting characteristic is its robustness relative to the initial value of the optimization procedure which can be randomly chosen without affecting the estimation results. Furthermore, the profiled ML or REML criterion's optimization can be easily and rapidly performed using an existing optimizer in the R software. Further considerations of this approach may include heteroscedastic residuals as well as residuals correlated with random effects, where the theoretical consistency of the resulting estimators will be demonstrated.

CHAPTER 4

FIXED EFFECTS SELECTION IN THE LINEAR MIXED-EFFECTS MODEL USING ADAPTIVE RIDGE PROCEDURE FOR L0 PENALTY PERFORMANCE

Dans ce chapitre, nous faisons de la sélection d'effets fixes dans le modèle linéaire à effets mixtes (unidimensionnel). L'approche utilisée est une optimisation de la déviance pénalisée, où le terme de pénalité est de type L_2 en présence d'une matrice de poids itérativement mise à jour. C'est ce que nous avons appelé la procédure *Adaptive Ridge itérative* dont la qualité de selection approxime celle résultant d'une pénalité de type L_0. Ici, le calcul de la vraisemblance a été inspiré de celui proposé par Bates et al. (2014) avec une modification qui permet de "libérer" le vecteur des effets fixes afin de permettre une sélection plus efficace de ces effets. Les résultats obtenus sur des données simulées sont de très bonnes qualités, aussi bien en terme de sélection que d'estimation. Le présent chapitre fait aussi l'objet d'un article de recherche soumis dans une revue scientifique.

Fixed effects selection in the linear mixed-effects model using adaptive ridge procedure for L_0 penalty performance

Eric Houngla Adjakossa[1,2 *], Mahouton Norbert Hounkonnou[2], Gregory Nuel[1]

1 Laboratoire de Probabilités et Modèles Aléatoires /Université Pierre et Marie Curie, Case courrier 188 - 4, Place Jussieu 75252 Paris cedex 05 France
2 University of Abomey-Calavi, 072 B.P. 50 Cotonou, Republic of Benin

*** ericadjakossah@gmail.com**

abstract

This paper is concerned with the selection of fixed effects along with the estimation of fixed effects, random effects and variance components in the linear mixed-effects model. We introduce a selection procedure based on an adaptive ridge (AR) penalty of the profiled likelihood, where the covariance matrix of the random effects is Cholesky factorized. This selection procedure is intended to both low and high-dimensional settings where the number of fixed effects is allowed to grow exponentially with the total sample size, yielding technical difficulties due to the non-convex optimization problem induced by L_0 penalties. Through extensive simulation studies, the procedure is compared to the LASSO selection and appears to enjoy the model selection consistency as well as the estimation consistency.

4.1 Introduction

During the last two decades, selection procedures in the linear mixed-effects model have been an active research topic due to the appealing features of the model and the advent of modern technologies facilitating the collection of many variables in scientific studies. Many of these variables are typically included in the full model at the initial stage of modeling to reduce model approximation error, and due to the complexity of the mixed-effects models, inferences and interpretations of the estimated models become challenging as the dimension of fixed or random effects increases [54]. The selection of important fixed or random effects has thus become a fundamental problem in the analysis of grouped data using mixed-effects models, especially in the high-dimensional settings where the fixed or the random effects vector dimension is allowed to grow exponentially with the sample size.

Generally, model selection procedures can be viewed as covering three main approaches: the hypothesis testing procedures, the regularization procedures and other procedures which include the Bayesian selection methods. The testing procedures include the ordinary hypothesis tests and

the selection methods based on generalized information criteria. Examples of using testing hypothesis for models selection in the mixed-effects model context include Lin's works [112] who proposed a simple global variance component tests, which are locally asymptotically most precise and are robust in the sense that no assumption about the parametric form of the random effects is made. Despite their global form expressions which require only the fitting of conventional generalized linear models, the critical values of the global test statistics, which are based on large sample theory, are less accurate when the number of levels of each random effect is small, e.g. less than 15. Edwards and his co-workers [50] extended the traditional coefficient of determination R^2 for the linear mixed-effects model $y \sim \mathcal{N}(X\beta, \Sigma = Z\Gamma Z^\top + \sigma^2 I_N)$, where they introduced a statistic $R^2_\beta = (q-1)\nu^{-1}F(\widehat{\beta}, \widehat{\Sigma}) / \left[1 + (q-1)\nu^{-1}F(\widehat{\beta}, \widehat{\Sigma})\right]$, with $F(\widehat{\beta}, \widehat{\Sigma}) = (C\widehat{\beta})^\top \left[C(X^\top \widehat{\Sigma}^{-1} X)^{-1} C^\top\right]^{-1} C\widehat{\beta}/\text{rank}(C)$, $\nu = N - \text{rank}(X) = N - q$, $C = \left[\mathbf{0}_{(q-1)\times 1} I_{q-1}\right]$ of rank $q-1$, in testing $H_0 : C\beta = \mathbf{0}$. R^2_β measures the multivariate association between the repeated outcomes and the fixed effects in the context of longitudinal data analysis. This R^2_β statistic arises as a $1-1$ function of an appropriate F statistic (i.e., $F(\widehat{\beta}, \widehat{\Sigma})$) for testing all the fixed effects, except the intercept. More precisely, R^2_β compares the full model with a null model having no fixed effect except typically the intercept. R^2_β is then generalized to define a partial R^2 statistic for marginal fixed effects of all sorts. Although this testing-based selection procedure of fixed effects is very useful, one of its major drawback is that the choice of the denominator of R^2_β clearly affects the rate of convergence as $N \to \infty$, and may change the parameter being estimated. Examples of R^2-based selection of fixed effects in the mixed-effects model include [178], [202], [101] and references therein, where generally, too much restrictions are made on the random effects covariance matrix, and clearly may not be appropriate for a wide range of data analysis.

Based on testing procedures, a stepwise procedure can be constructed for selecting important fixed or random effects using generalized information criteria [54, 144], which are a generalization of Akaike's information criterion (AIC) [4] and the Bayesian information criterion (BIC) [167]. [144] extended the Generalized information criterion (GIC) proposed by [150] in order to construct a procedure for selecting fixed and random effects in the linear mixed-effects model. Here, following [173], the asymptotic behavior of the extended GIC method for selecting fixed effects is studied, and the results from simulations show that if the signal-to-noise ratio is moderate or high, the percentages of choosing the correct fixed effects by the GIC procedure are close to one for finite samples. Another examples of GIC like selection procedures include those proposed by [122], [85] and [21]. These strategies suggest a unified approach for choosing a parameters vector β that maximizes the penalized likelihood

$$n^{-1}\ell_n(\beta|y) - \sum_{j=1}^{p} p_\lambda(|\beta_j|) \tag{4.1}$$

which arose from the Kullback-Leibler (KL) divergence $-\ell_n(\widehat{\beta}|y) + \lambda\|\beta\|_0$ of the fitted model from the true model [3], where $\ell_n(.|y)$ is the log-likelihood function, $\widehat{\beta}$ is the maximum likelihood estimator of β and p_λ is the penalty function indexed by the regularization parameter $\lambda \geqslant 0$. The L_0-norm $\|\beta\|_0$ of β counts the number of non-vanishing components ($\beta_j \neq 0$) in β, and arises naturally in many classical model selection methods. Although involving a nice interpretation of

the best subset selection, and admitting good sampling properties [11], its computation is infeasible in high dimensional statistical endeavors [53]. Other penalty functions are regularly used in the literature. A natural generalization of L_0 penalty is the so-called bridge (noted L_q) penalty in [63], where $p_\lambda(|\beta|) = \lambda\|\beta\|_{L_q}^q$ for $0 < q \leqslant 2$. The L_q penalty encompasses L_0, L_1 - LASSO [186] - and L_2 (ridge) penalties. Since none of the L_q penalties satisfies all the required properties (sparsity, approximate unbiasedness and continuity, see [52] for more details) for their resulting parameter estimators, other penalties including SCAD [51, 52] and MCP [207] are introduced in the literature.

Testing-based stepwise selection procedures, where λ is fixed ($\lambda = 1$ for AIC, $\lambda = \log(n)/2$ for BIC, $\lambda = \log(\log n)/2$ for HQIC [85], $\lambda = (\log n + 1)/2$ for CAIC [21], for example) are computationally expensive for high-dimensional settings and ignore stochastic errors in the variable selection process [52]. Another severe drawback is their lack of stability [24]. Penalized likelihood approaches using data-driven choice of λ are generally preferred to handle the high-dimensional selection problem. These methods are referred to as regularization procedures. Here, the regularization parameter λ (also called the tuning parameter) is generally estimated using cross-validation methods [23, 52, 66, 186]. [95] consider the selection of both fixed and random effects in a general class of mixed effects models by optimizing the penalized likelihood criterion $Q_\lambda(\theta|\theta_{\text{old}}) = Q(\theta|\theta_{\text{old}}) - n\sum_{j=1}^{p} p_\lambda(|\beta_j|)$ using the EM algorithm [44], where θ is the vector of all the unknown parameters, $Q(.|\theta_{\text{old}})$ is the resulting function of the EM algorithm E-step, and p_λ is either the SCAD or the adaptive lasso - ALASSO - [209] penalty. They approximate the integral in $Q(\theta|\theta_{\text{old}})$ by using a Markov chain Monte Carlo and introduce the $\text{IC}_Q(\lambda) = -2Q(\widehat{\theta}_\lambda|\widehat{\theta}_0) + c_n(\widehat{\theta}_\lambda)$ statistic [94] for selecting the regularization parameter, where $\widehat{\theta}_0 = \arg\max_\theta \ell(\theta)$ is the unpenalized maximum likelihood estimate and $c_n(\theta)$ a function of the data and fitted model.

[54] introduce a class of variable selection methods for fixed effects using a penalized profiled likelihood method, where the unknown covariance matrix of the random effects is replaced with a suitable proxy matrix. Here, the general idea used is as follows. After writing the joint density $f(y, \gamma)$ of the response variable y and the random effects vector γ, they consider the penalized profiled likelihood $L_n(\beta, \widehat{\gamma}(\beta)) - n\sum_{j=1}^{d_n} p_\lambda(|\beta_j|)$, where $L_n(\beta, \widehat{\gamma}(\beta)) = f(y, \widehat{\gamma}(\beta))$ with $\widehat{\gamma}(\beta)$ the empirical Bayes estimate of γ [88] in which the covariance matrix of γ is replaced with the proxy matrix. d_n may increase with the sample size n and p_λ is a concave penalty function (SCAD and LASSO in their simulation and application section). Although the proxy matrix may be different from the true one, it may still yield correct model selection results at the cost of some additional bias [54]. [164] deal with theoretical and computational aspects for high-dimensional selection of fixed effects in the linear mixed-effects model, where the consistency of the estimator is proven along with a non-asymptotic oracle result for the adaptive LASSO estimator, under the assumption that the eigenvalues of $Z^\top Z$ are bounded. Here, Z is the random effects design matrix, and an explicit analytical expression of the regularization parameter λ is also given.

Other approches for variable selection in linear mixed-effects models include “fence” procedure and Bayesian techniques. [98] introduce a class of strategies known as a fence methods intended for variable selection in both linear and generalized linear mixed-effects models. The fence strategy is based on a measure of lack-of-fit that is a quantity $Q_M = Q_M(y, \theta_M)$, where y is the response variable, M indicating a candidate model for the selection and θ_M denotes the vector of parameters under M. Counting among the typically rare and ad-hoc selection procedures,

the fence method is computationally very demanding, particularly because it involves the estimation of the standard deviation of the difference of lack-of-fit measures. For more details on these ad-hoc selection procedures, see the nice review paper of [131] and references therein. Bayesian model selection requires to assign a prior distribution over the model parameters and compute the posterior probabilities of each of them. These computations can be difficult so are usually carried out by applying sophisticated Markov Chain Monte Carlo (MCMC) algorithms [131]. Exemples of Bayesian model selection include [33] and [160] who point out that these kinds of MCMC methods are generally time consuming to implement, requiring special software and depend on subjective choice of the hyperparameters in the priors.

In this paper, we discuss the selection of fixed effects in the linear mixed-effects model using an adaptive ridge (AR) penalty of the profiled log-likelihood, where the random effect covariance matrix is Cholesky factorized for solving a preliminary penalized least square problem. The profiled likelihood is calculated by slightly modifying the approach proposed by [14]. The weights matrix of the AR procedure introduced here is updated in such a way that the procedure converges toward selection with L_0 penalty, as have done [65]. The present selection strategy is intended to both low and high-dimensional settings.

The rest of the article is organized as follows. The profiled log-likelihood is introduced in Section 2. In Section 3, we present the weighted ridge procedure, and Section 4 presents the simulation studies.

4.2 profiled log-likelihood for the linear mixed-effects model

In this section, we consider the classical linear mixed-effect model setting where the number of observations n is larger than the number of covariates p. By slightly modifying the approach introduced by [14], we calculate the profiled likelihood function.

4.2.1 Model and notations

We consider the linear mixed-effects model in which the residual terms are homoscedastic and independent of the random effects as follows.

$$\mathcal{Y} = X\beta + Z\gamma + \varepsilon, \tag{4.2}$$

$$\gamma \sim \mathcal{N}(\mathbf{0}, \Gamma), \quad \varepsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_n) \quad \text{and } \gamma \perp \varepsilon, \tag{4.3}$$

where $\mathcal{Y}$ is the random response variable whose observed value is $y_{\text{obs}} \in \mathbb{R}^n$, $\gamma \in \mathbb{R}^q$ is the unobserved random effects vector with covariance matrix Γ, ε is the residual term, $\beta \in \mathbb{R}^p$ is the fixed effects vector, and X and Z are the fixed and random effects related design matrices of dimensions $n \times p$ and $n \times q$, respectively. $\sigma^2 I_n$ is the covariance matrix of ε with $\sigma > 0$ and I_n the $n \times n$ identity matrix.

As a variance-covariance matrix, Γ must be positive semidefinite. Conveniently, the model is expressed in terms of a relative covariance factor, Λ_θ, which is a $q \times q$ matrix, depending on the variance components vector, θ, that generate the symmetric $q \times q$ variance-covariance matrix, Γ,

according to

$$\Gamma = \sigma^2 \Lambda_\theta \Lambda_\theta^\top, \tag{4.4}$$

where σ is the same scale parameter as in Equation (4.3). This factorization of Γ yields the existence of a random vector $\mathcal{U}$ such that

$$\gamma = \Lambda_\theta \mathcal{U}, \tag{4.5}$$

with

$$\mathcal{U} \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_q) \tag{4.6}$$

which is called a *spherical random effects*[(1)] variable. The model can therefore be re-expressed as follows

$$\mathcal{Y} = X\beta + Z\Lambda_\theta \mathcal{U} + \varepsilon, \tag{4.7}$$

$$\mathcal{U} \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_q), \quad \varepsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_n) \quad \text{and} \quad \mathcal{U} \perp \varepsilon. \tag{4.8}$$

The parameters of the model are the fixed effects vector β and the variance components θ and σ^2. This formulation of the linear mixed-effects model allows, not only, the use of a singular matrix Λ_θ which arises in practice, but also for a relatively compact expression for the profiled log-likelihood of β and θ, conditional on y_{obs}.

4.2.2 profiled likelihood

The profiled likelihood considered here is expressed through the following theorem.

Theorem 4.2.1. *Suppose that y is a realization of a random vector $\mathcal{Y}$ satisfying the linear mixed-effects model expressed by Equations (4.7) and (4.8), where β, θ and σ^2 are the parameters to be estimated. Denoting by L_θ the matrix such that $L_\theta^\top L_\theta = (Z\Lambda_\theta)^\top Z\Lambda_\theta + I_q$, $\tilde{u}$ the conditional mean of $\mathcal{U}$ given that $\mathcal{Y} = y$, and $g(\tilde{u}) = \|y - X\beta - Z\Lambda_\theta \tilde{u}\|^2 + \|\tilde{u}\|^2$, the profiled log-likelihood of β and θ conditional on y is*

$$\tilde{\ell}(\beta, \theta | y) = -\frac{1}{2} \log |L_\theta|^2 - \frac{n}{2} \left[1 + \log \left(\frac{2\pi g(\tilde{u})}{n} \right) \right]. \tag{4.9}$$

Proof. Denoting by u a realization of $\mathcal{U}$, the density of y is expressed as

$$f(y) = \int_{\mathbb{R}^q} f(y, u) du = \int_{\mathbb{R}^q} f(y|u) f(u) du.$$

$$f(y|u) f(u) = (2\pi\sigma^2)^{-(n+q)/2} \exp \left[-\frac{\|y - X\beta - Z\Lambda_\theta u\|^2 + \|u\|^2}{2\sigma^2} \right], \tag{4.10}$$

where $\|u\|^2 = u^\top u$, with $^\top$ denoting the transpose operator.

$$\|y - X\beta - Z\Lambda_\theta u\|^2 + \|u\|^2 = \left\| \begin{pmatrix} y - X\beta \\ 0 \end{pmatrix} - \begin{pmatrix} Z\Lambda_\theta \\ \mathrm{I}_q \end{pmatrix} u \right\|^2 = g(u), \tag{4.11}$$

(1)[14] argued that the term "spherical" is related to the fact that the contours of the $\mathcal{N}(0, \sigma^2 I_q)$ probability density are spheres.

and solving the penalized least squares problem that is to minimize $g(u)$ over u implies

$$\tilde{u} = \underset{u\in\mathbb{R}^q}{\arg\min}\, g(u) \iff \begin{pmatrix} Z\Lambda_\theta \\ \mathrm{I}_q \end{pmatrix}^\top \begin{pmatrix} Z\Lambda_\theta \\ \mathrm{I}_q \end{pmatrix} \tilde{u} = \begin{pmatrix} Z\Lambda_\theta \\ \mathrm{I}_q \end{pmatrix}^\top \begin{pmatrix} y - X\beta \\ 0 \end{pmatrix}. \tag{4.12}$$

Viewed as a function of u, g is C^∞ and the first and the second differential of g are

$$\mathrm{d}g(u) = 2\,\mathrm{tr}\left\{-(y - X\beta - Z\Lambda_\theta u)^\top Z\Lambda_\theta \mathrm{d}u + u^\top \mathrm{d}u\right\}$$

and

$$\mathrm{d}^2 g(u) = 2\,\mathrm{tr}\left\{(Z\Lambda_\theta)^\top Z\Lambda_\theta \mathrm{d}u\mathrm{d}u^\top + \mathrm{d}u\mathrm{d}u^\top\right\}.$$

This implies that

$$\frac{\partial g}{\partial u}(u) = -2\left(y - X\beta - Z\Lambda_\theta u\right)^\top Z\Lambda_\theta + 2u^\top$$

and

$$\frac{\partial^2 g}{\partial u \partial u^\top}(u) = 2\left(Z\Lambda_\theta\right)^\top Z\Lambda_\theta + 2\mathrm{I}_q.$$

Then, $g(u)$ can be rewritten as

$$g(u) = g(\tilde{u}) + \frac{\partial g}{\partial u}(\tilde{u})(u - \tilde{u}) + \frac{1}{2}(u - \tilde{u})^\top \frac{\partial^2 g}{\partial u \partial u^\top}(\tilde{u})(u - \tilde{u}). \tag{4.13}$$

Referring to Equation (4.12),

$$\begin{pmatrix} Z\Lambda_\theta \\ \mathrm{I}_q \end{pmatrix} \left[\begin{pmatrix} y - X\beta \\ 0 \end{pmatrix} - \begin{pmatrix} Z\Lambda_\theta \\ \mathrm{I}_q \end{pmatrix} \tilde{u}\right] = 0 \implies \frac{\partial g}{\partial u}(\tilde{u}) = 0. \tag{4.14}$$

Then,

$$\begin{aligned} g(u) &= g(\tilde{u}) + \frac{1}{2}(u - \tilde{u})^\top \frac{\partial^2 g}{\partial u \partial u^\top}(\tilde{u})(u - \tilde{u}) \\ &= g(\tilde{u}) + (u - \tilde{u})^\top \left[\left(Z\Lambda_\theta\right)^\top Z\Lambda_\theta + \mathrm{I}_q\right](u - \tilde{u}) \\ &= g(\tilde{u}) + \|L_\theta(u - \tilde{u})\|^2, \text{ with } L_\theta^\top L_\theta = \left(Z\Lambda_\theta\right)^\top Z\Lambda_\theta + \mathrm{I}_q, \end{aligned} \tag{4.15}$$

and

$$\begin{aligned} f(y) &= (2\pi\sigma^2)^{-(n+q)/2} \int_{\mathbb{R}^q} \exp\left[-\frac{g(\tilde{u}) + \|L_\theta(u - \tilde{u})\|^2}{2\sigma^2}\right] du \\ &= (2\pi\sigma^2)^{-(n+q)/2} \exp\left[-\frac{g(\tilde{u})}{2\sigma^2}\right] \int_{\mathbb{R}^q} \exp\left[-\frac{\|L_\theta(u - \tilde{u})\|^2}{2\sigma^2}\right] du. \end{aligned} \tag{4.16}$$

$v = L_\theta(u - \tilde{u}) \implies du = \frac{1}{|L_\theta|} dv$, and

$$\begin{aligned} f(y) &= (2\pi\sigma^2)^{-n/2}|L_\theta|^{-1}\exp\left[-\frac{g(\tilde{u})}{2\sigma^2}\right]\int_{\mathbb{R}^q}(2\pi\sigma^2)^{-q/2}\exp\left[-\frac{\|v\|^2}{2\sigma^2}\right]dv \\ &= (2\pi\sigma^2)^{-n/2}|L_\theta|^{-1}\exp\left[-\frac{g(\tilde{u})}{2\sigma^2}\right] \end{aligned} \tag{4.17}$$

The log-likelihood of β, θ and σ^2 conditional on y is

$$\ell(\beta,\theta,\sigma^2|y) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2}\log|L_\theta|^2 - \frac{g(\tilde{u})}{2\sigma^2}. \tag{4.18}$$

$\frac{\partial\ell(\beta,\theta,\sigma^2|y)}{\partial\sigma^2} = 0 \implies \sigma^2 = \frac{g(\tilde{u})}{n}$. By profiling out σ^2, the profiled log-likelihood, $\tilde{\ell}(\beta,\theta|y)$, of β and θ conditional on y is expressed through

$$-2\tilde{\ell}(\beta,\theta|y) = \log|L_\theta|^2 + n\left[1+\log\left(\frac{2\pi g(\tilde{u})}{n}\right)\right].$$

□

4.3 L_0 estimator of β using iteratively weighted ridge procedure

The L_0 estimator of the fixed-effects vector β using iteratively weighted ridge procedure presented here fits both the low and the high-dimensional settings. Theoretically, this selection method may enjoy great stability since it uses neither inverse of X (fixed-effects design matrix) nor inverse of Z (random effects design matrix).

4.3.1 Adaptive Ridge penalty for the profiled likelihood

Let us assume that the true underlying fixed-effects vector β_{true} is sparse in the sense that many of its coefficients are zero. To enforce the sparsity of the estimator of β, we add a weighted L_2 (ridge) penalty for the fixed-effects vector β to the profiled log-likelihood function. Thus, we are considering the objective function

$$\tilde{\ell}_{\lambda,w}(\beta,\theta) = -2\tilde{\ell}(\beta,\theta|y_{\text{obs}}) + \lambda\beta^\top W\beta, \tag{4.19}$$

where $W = \text{diag}(w_1,\ldots,w_p)$ is a $p\times p$ diagonal matrix and $\lambda \geqslant 0$ is a regularization parameter. We aim at estimating β, θ and σ^2 by

$$(\tilde{\beta}_{\lambda,w},\tilde{\theta}_{\lambda,w}) = \underset{\beta,\theta}{\arg\min}\,\tilde{\ell}_{\lambda,w}(\beta,\theta) \quad \text{and} \quad \widetilde{\sigma^2}_{\lambda,w} = \frac{\|y - X\tilde{\beta}_{\lambda,w} - Z\Lambda_{\tilde{\theta}_{\lambda,w}}\tilde{u}\|^2 + \|\tilde{u}\|^2}{n}, \tag{4.20}$$

where $\tilde{u}$ is calculated by the penalized least squares algorithm that has been hinted at Equation (4.12). The criterion $\tilde{\ell}_{\lambda,w}(\beta,\theta)$ is minimized using one of the constrained optimization functions in R [146], to provide the estimators $\tilde{\beta}_{\lambda,w}$, $\tilde{\theta}_{\lambda,w}$ and $\widetilde{\sigma^2}_{\lambda,w}$.

4.3.2 Iteratively Weighted ridge procedure

The L_0 penalty for regularization arises naturally in many classical model selection since, indeed, it counts the number of non-vanishing parameters, giving a nice interpretation of the best subset selection and admits nice sampling properties [11]. However, its computation is infeasible in high dimensional settings and clearly argued to be a combinational problem with NP-complexity [53]. Some workarounds are reported in the literature, where the proposed procedure converge toward the L_0-penalty based selection. For example, [65] introduced an adaptive ridge procedure that helps to approximate L_0-penalty performances. Here, we are using the same procedure where the weight matrix (W defined in Equation (4.19)) diagonal elements $w_j, 1 \leqslant j \leqslant p$ are iteratively computed and defined as

$$w_j^{(k)} = \left[\left|\beta_j^{(k)}\right|^2 + \delta^2\right]^{-1}, \tag{4.21}$$

as have done [65], taking inspiration from [20, 28, 78, 154] and their simulation results. In Equation (4.21), k identifies the iteration and β_j is the jth component of β. The general expression of $w_j^{(k)}$ is $w_j^{(k)} = \left[\left|\beta_j^{(k)}\right|^\tau + \delta^\tau\right]^{\frac{q-2}{\tau}}$, where q precises the norm $\|\cdot\|_{L_q}$ for the penalty, δ calibrates which effect sizes are considered relevant and τ determines the quality of the approximation $w_j\beta_j^2 \approx |\beta_j|^q$. In practice, $\delta = 10^{-5}$ seems to perform well. For more details, see [65]. More precisely, the selection procedure performed here is as follows. For a fixed λ, $\tilde{\beta}_{\lambda,w}$ is initialized at $(1,\ldots,1)^\top$, the vector $selection$, say, which identifies the selected fixed effects is initialized at $(1,\ldots,1)^\top$ and W is initialized at $\text{diag}(1,\ldots,1)$. Then the steps come.

1) $\tilde{\beta}_{\lambda,\text{old}} \leftarrow \tilde{\beta}_{\lambda,w}$

2) perform the optimization $(\tilde{\beta}_{\lambda,w}, \tilde{\theta}_{\lambda,w}) = \arg\min_{\beta,\theta} \tilde{\ell}_{\lambda,w}(\beta,\theta)$, initializing β by $\tilde{\beta}_{\lambda,\text{old}}$ and θ by θ_0. Here, the components of θ which are variances are initialized by 1 and those which are not variances are initialized by 0. Thus, θ_0 components are 0 or 1.

3) $w_j \leftarrow \left(\tilde{\beta}^2_{\lambda,w,j} + \delta^2\right)^{-1}$, where $\tilde{\beta}_{\lambda,w,j}$ is the jth component of $\tilde{\beta}_{\lambda,w}$, and w_j is the jth element of W's diagonal.

4) $selection_{\text{old}} \leftarrow selection$ and $selection \leftarrow W \cdot \text{diag}(\tilde{\beta}_{\lambda,w}) \cdot \tilde{\beta}_{\lambda,w}$

5) if $|selection - selection_{\text{old}}| < \text{tol} = 10^{-5}$, then the selection can be considered as well performed for λ. Thus, we choose a new value for λ. If $|selection - selection_{\text{old}}| \geqslant \text{tol}$, the selection does not perform well and we go to the item 1) by choosing the current $\tilde{\beta}_{\lambda,w}$ as $\tilde{\beta}_{\lambda,\text{old}}$, without changing the value of λ.

For some λ values, the $selection$ vector may contain 0 or 1 as components. If it is 1, the corresponding fixed effect β_j is selected, and if it is 0 the corresponding β_j is not selected for λ.

For the choice of the regularization parameter λ, we propose to use the Bayesian Information Criterion (BIC) criterion defined as

$$c_{n,\lambda} = -2\ell(\hat{\beta}_{\lambda,\text{sel}}, \hat{\theta}_{\lambda,\text{sel}}, \widehat{\sigma^2}_{\lambda,\text{sel}} | y_{\text{obs}}) + \log(n) \cdot \hat{d}_\lambda, \tag{4.22}$$

where $\hat{\beta}_{\lambda,\text{sel}}, \hat{\theta}_{\lambda,\text{sel}}$ and $\widehat{\sigma^2}_{\lambda,\text{sel}}$ are the ML parameters estimators considering the selected variables. $\hat{d}_\lambda = \#\{\hat{\beta}_{\lambda,\text{sel},j} \neq 0 : 1 \leqslant j \leqslant p\} + \dim(\theta) + 1$ is the sum of the number of nonzero fixed-effects and the number of variance components. This form of $\hat{d}_\lambda$ has been suggested by [17] and empirically validated by [165].

4.4 Simulation studies

In this section, we assess the performance of our approach using simulated data. We compare the obtained results with those coming from the Lasso implementation in a situation where the number of noise variables are excessive.

Since we use one of the optimizers available in the R software for minimizing the $\tilde{\ell}_{\lambda,w}(\beta,\theta)$ criterion, for too bigger values of p, especially when $n \ll p$, the convergence of the used algorithm ("nlminb" for example) is hardly or no more reached. Due to this convergence problem, we restrict the simulation studies to the low-dimensional setting where $p > 40$ and will focus on this problem in another coming paper with the same theoretical approach.

We restrict ourselves to the case of longitudinal data study with N observations coming from n subjects where each subject i has n_i observations. The "working" data sets are simulated under the following model.

$$y_i = X_{1i}\beta^* + \gamma_i \mathbb{1}_{n_i} + \varepsilon_i, \quad \gamma_i \sim \mathcal{N}(0,\Gamma), \quad \varepsilon_i \sim \mathcal{N}(0,\sigma^2 I_{n_i}), \quad \text{for} \quad i = 1,\ldots,n; \quad (4.23)$$

where $\beta^* \in \mathbb{R}^{p_1}$ is the true fixed effects vector, X_1 is a $N \times p_1$ design covariates matrix, $\mathbb{1}_{n_i} = (1,\ldots,1) \in \mathbb{R}^{n_i}$. $\gamma \perp \varepsilon$ and $\gamma_i \perp \gamma_{i'}$ for $i \neq i'$. γ_i is an random intercept for the ith subject and ε is the residual term of the model.

We suppose that we are following up a sample of subjects where the goal is to evaluate how their weights are influenced by other variables including the age, the sex and the nutrition score "nscore", and which variables govern this influence. The covariates sex, nscore and age are staked in the model matrix X_1 and all other covariates are staked in another $N \times p_2$ model matrix X_2 such that $X = (X_1|X_2)$ with $\dim(X) = N \times p$ and $p = p_1 + p_2$. The components of the vector (variable) age are randomly sampled from a uniform distribution in $[18, 37]$ and the nscore variable is also uniform in $[20, 50]$. We would like to fit to the data, the model

$$y_i = X_i\beta + \gamma_i \mathbb{1}_{n_i} + \varepsilon_i, \quad \gamma_i \sim \mathcal{N}(0,\Gamma), \quad \varepsilon_i \sim \mathcal{N}(0,\sigma^2 I_{n_i}), \quad \text{for} \quad i = 1,\ldots,n; \quad (4.24)$$

where $\beta = (\beta_1^\top, \beta_2^\top)^\top$, with $\beta_1 \in \mathbb{R}^{p_1}$ and $\beta_2 = \mathbf{0} \in \mathbb{R}^{p_2}$. We are then challenging to identify which β components are zero.

For the working data sets, we choose $p_1 = 4$, $p = 54$, $N = 300$, $n = 90$, $\sigma = 1$, $\Gamma = 1$ and $\beta^* = (1,-1,-1,1)$. For the fixed-effects β, we have in fact fifty zeros components and only four components are not zeros. We simulate 100 data sets for which we perform both the lasso selection and the IWR (Iteratively Weighted Ridge) procedure. One hundred values of the regularization parameter λ are chosen in $[10^{-2}, 10^2]$. For each replication, the λ value that minimizes the BIC criterion is retained for selecting the significant fixed effects. Denoting by $\beta^{**} = (1,-1,-1,1,0,\ldots,0)$, the mean squared error MSE $= \mathbb{E}\left[\|\hat{\beta} - \beta^{**}\|_2^2\right]$ (with $\hat{\beta}$ coming

from lasso or IWR) is computed over the 100 replications. The cardinality of the estimated active set (i.e. $|S(\hat{\beta})|$, with $S(\hat{\beta}) = \{k : \hat{\beta}_k \neq 0\}$) is computed as well as the proportion TP of true positive (i.e. the selected set is exactly the true one). We also compute the proportion TPC in which the selected set contains the true one and the proportion ZP in which the true zeros are estimated.

Table 4.1: **Selection performances comparison between LASSO and IWR procedures.**

Performance criterion	LASSO	IWR
MSE	1.200	0.254
$\lvert S(\hat{\beta})\rvert$	$4(1.614)$	$5(1.250)$
TP	0%	35%
TPC	16%	90%
ZP	$98\%(0.028)$	$98\%(0.023)$

The selection results based on the 100 simulated data sets are indeed summarized through five performance criteria that are contained in Table 4.1, where the numbers between parentheses are the standard deviations related to the criteria mean values (printed just before these parentheses). Information from Table 4.1 show that the IWR procedure outperforms the LASSO one. IWR selects the correct model 35% of the time when LASSO has never found it (see TP values in Table 4.1). For information, we use the R software package lmmlasso [163] to perform LASSO selection. 90% of the time, the model selected by IWR contains the true one against 16% for LASSO (see TPC values from Table 4.1). Obviously, the estimations are of better qualities from IWR than from LASSO (MSE $= 0.254$ for IWR, and MSE $= 1.200$ for LASSO).

The true zero estimation proportions (ZP) are computed in Table 4.2 which contains also their number of occurrence over the simulated data sets. The LASSO seems to find more often than

Table 4.2: **True zero estimation proportions (ZP) with the number of occurrence over the 100 replications.**

	LASSO							**IWR**					
ZP	0.88	0.90	0.92	0.94	0.96	0.98	1	0.90	0.92	0.94	0.96	0.98	1
Number	2	2	3	6	14	24	49	2	2	8	16	32	40

IWR all the true zeros (49% for LASSO against 40% for IWR, in Table 4.2). This may explain the overfitting behavior of IWR ($|S(\hat{\beta})| = 5$ for IWR and $|S(\hat{\beta})| = 4$ for LASSO in Table 4.1). It therefore seems that LASSO has a stronger shrinkage capability than has IWS which shows in turn a somewhat overfitting behavior than LASSO. Through these simulations studies, it appears that the iteratively weighted aspect of the IWR procedure highers the shrinkage performance of the ordinary Ridge and results in a selection method having better performance than the LASSO.

Like in the case of LASSO, it is possible for IWR to take advantage of a warm start of the algorithm to obtain the full regularization path of the selection problem. For instance, Figure 4.1

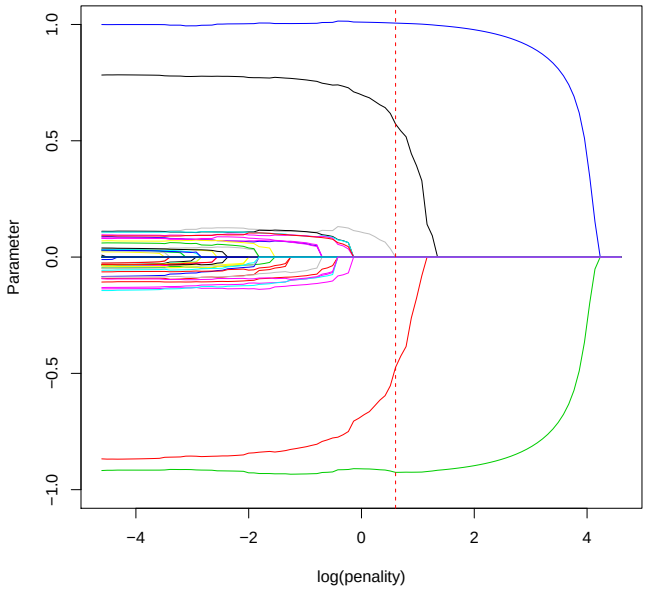

Figure 4.1: **Example of a full regularization path for the iteratively weighted ridge selection procedure with $N = 300$, $n = 90$, $p = 54$, $\sigma = 1 = \Gamma$. Only the first four components of β^* are nonzero. The dataset is simulated using the R software with set.seed(3). The vertical red dashed bar corresponds to the minimum of BIC and shows the selected variables.**

shows the full regularization path from one of the simulated data. The vertical red dashed bar corresponds to the minimum of BIC and shows the selected variables. On Figure 4.1, we clearly have $|S(\hat{\beta})| = 4$, i.e., four variables selected at the end of the procedure.

4.5 Conclusion

In this paper, we have focused on the fixed-effects selection problem in the linear mixed-effects model. We have introduced an iteratively weighted ridge procedure which enhances the shrinkage performance of the ordinary ridge in order to approximate the performance of the L_0 based penalty selection. This selection method is based on an adaptive ridge penalty of the profiled likelihood, where the covariance matrix of the random effects is Cholesky factorized. The procedure fits both the low and the high-dimensional settings and may enjoy a great numeric stability since it needs no use of the inverse of the fixed-effects design matrix or the design matrix of the random-effects. Due the problems of the lack of convergence in the high-dimensional settings using available optimizers, we restrict the simulations studies to the low-dimensional case where we find out that our selection procedure outperforms the LASSO selection. In another forcoming paper, we will focus on this convergence problem in the high-dimensional cases with the same theoretical approach.

CHAPITRE 5

APPLICATION À L'ÉTUDE DE L'ACQUISITION IMMUNITAIRE CONTRE LE PALUDISME CHEZ L'ENFANT À TORI-BOSSITO (BÉNIN)

5.1 Paludisme

Le paludisme est une maladie infectieuse à transmission vectorielle faisant intervenir trois acteurs que sont l'Homme, jouant le rôle d'hôte, infecté par un protozoaire (le parasite) du genre *Plasmodium* qui lui a été transmis par la piqûre d'un moustique (le vecteur) femelle du genre *Anopheles*. Outre les facteurs biologiques de l'hôte et du parasite, les facteurs géographiques sont essentiels pour comprendre la transmission du paludisme et expliquer la susceptibilité des individus aux infections. Le nombre de piqûres reçues par un individu dépend non seulement de l'abondance de gîtes larvaires dans l'environnement, mais aussi de la proportion d'anophèles vecteurs et d'anophèles porteurs de l'agent pathogène [43]. L'infectiosité des anophèles dépend directement du nombre de parasites circulants et par conséquent de la durée du cycle parasitaire qui elle-même dépend de la température du milieu de vie du vecteur. Par exemple, pour le *P. falciparum*, en dessus de l'optimum de $25°C$ de température, le cycle s'allonge et par conséquent la transmission diminue. Ainsi, la température, la pluviométrie et la végétation sont des facteurs essentiels à l'explication de la transmission palustre [37].

5.1.1 Vecteurs

Les vecteurs du paludisme humain appartiennent tous au genre *Anopheles* qui est de la famille des *Culicidae*, de l'ordre des *Dieptra*. Il existe plus de 450 espèces d'anophèles recensées, regroupées par régions géographiques et dont seulement 70 à 80 sont considérées comme étant des vecteurs. En Afrique tropicale, les cinq principaux vecteurs sont : *An. gambiae s.s., An. aarabiensis, An. funestus, An. nili s.l.* et *An. moucheti*. En milieu tropical, un anophèle mâle a une durée de vie d'environ 10 jours alors que la femelle quant à elle peut vivre jusqu'à deux à quatre semaines. La femelle fécondée se nourrit de jus sucré et de sang qu'elle prélève tous les deux à trois jours sur un hôte vertébré. Elle pond séparément, à la surface de l'eau, 40 à 100 œufs qui éclosent au bout de 24 à 48 heures selon la température, et part ensuite à la recherche d'un nouvel hôte pour un autre repas sanguin [130].

5.1.2 Agent pathogène

Lors de son repas sanguin, l'anophèle libère dans le sang humain les parasites responsables du paludisme chez l'Homme. Ces parasites sont des protozoaires du genre *Plasmodium*. Cinq espèces de *Plasmodium* peuvent parasiter l'Homme : *P. malariae, P. vivax, P. knowlesi, P. ovale* et *P. falciparum* qui est le plus mortel, présentant une large gamme de manifestations pathologiques. Dans les régions équatoriales, il est transmis toute l'année avec des recrudescences saisonnières et ne survient qu'en période chaude et humide dans les régions sub-tropicales. Environ trois-quarts des cas d'infection due au *P. falciparum* dans le monde sont enregistrés en Afrique [180]. Ces dernières années, il a été enregistré certains cas humains de paludisme à *P. knowlesi*, un paludisme du singe rencontré dans certaines zones de forêts d'Asie du Sud-Est.

Le cycle biologique du *Plasmodium* se décompose en une phase sexuée chez l'*Anopheles* (hôte définitif) et en une phase asexuée chez l'Homme (hôte intermédiaire). Les *Plasmodium* pénètrent dans l'organisme humain sous forme de sporozoites à la faveur d'une piqûre d'un moustique infectieux. Les sporozoites se multiplient ensuite dans les cellules du foie après y avoir été transportés par la circulation sanguine (Voir Figure 5.1). Libérés dans le sang sous forme de mérozoites, ils envahissent les globules rouges. Il en résulte des accès de fièvre. A maturité, le schizonte hépatique éclate, libérant des mérozoites, formes uninucléées qui poursuivront leur développement au cours de la phase érythrocytaire. L'invasion de l'érythrocyte débute par la liaison du mérozoite à la surface d'un érythrocyte. Les schizontes peuvent ainsi infecter d'autres globules rouges ou se transformer en gamétocytes mâles et femelles. Par la suite, un moustique se contamine par piqûre en absorbant du sang contenant des gamétocytes qui passeront à l'étape de gamètes dans le tube digestif de l'insecte. La fécondation d'un gamète femelle par un mâle produit un zygote (cel-

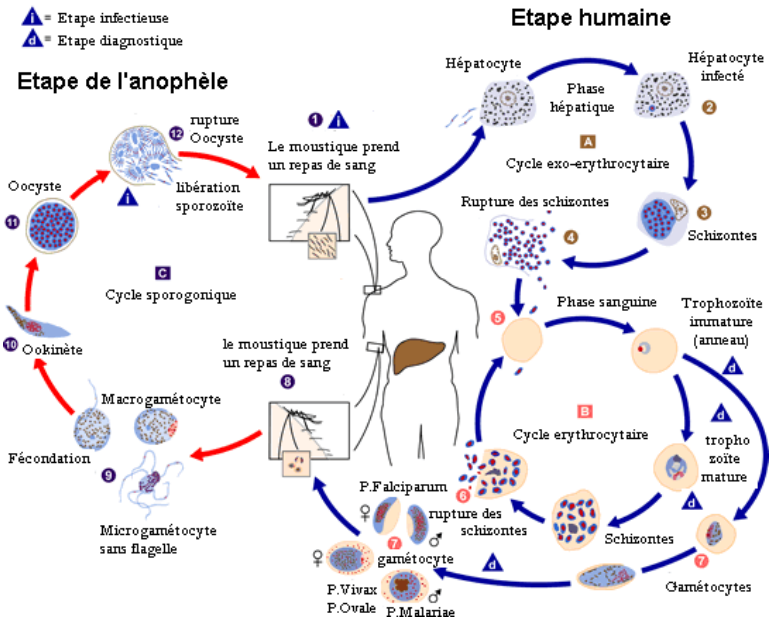

FIGURE 5.1 : **Cycle du développement et de la reproduction des *Plasmodium*.** (CDC, 2007)

lule œuf) qui se développe en sporozoite. Les sporozoites migrent ensuite dans les glandes sali-

vaires du moustique, d'où ils pourront contaminer un nouvel individu lors d'une piqûre. Ainsi, les éléments asexués (trophozoites et schizontes) sont digérés après absorption des différents stades du parasite par le moustique pendant le repas sanguin ; seuls les gamétocytes poursuivront leur développement. Rapidement, le gamète femelle se transforme en macrogamète et le gamétocyte mâle donne naissance à 8 microgamètes flagellés qui se dirigent à la rencontre du macrogamète. La fécondation donne naissance à un œuf mobile (l'ookinète) qui traverse la paroi de l'estomac, formant l'oocyte dans lequel s'individualisent les sporozoites qui, à leur tour, gagneront les glandes salivaires de l'anophèle [41].

5.1.3 Accès palustre grave et groupes à risques

P.falciparum est a l'origine des accès palustres sévères qui sévissent essentiellement chez les individus non immuns et certains groupes à risques parmi les populations vivant en zone d'endémie - femmes enceintes et enfants de moins de cinq ans - [43]. Chez un malade présentant une parasitémie avec des formes asexuées de *P.falciparum* sans aucune cause manifeste de symptômes, la présence de troubles de la conscience, de prostration, de convulsions multiples, d'ictère clinique accompagné d'autres signes de dysfonctionnement des organes vitaux,... indique qu'il souffre d'un paludisme grave [166].

Les femmes enceintes, particulièrement pendant leur première grossesse, sont susceptibles aux nouvelles infections palustres et montrent une séquestration massive d'hématies parasitées dans l'espace sanguin maternel du placenta. Or le paludisme gestationnel est généralement asymptomatique mais peut avoir de sérieuses conséquences sur le devenir de la mère et du fœtus : anémie maternelle, faible poids de naissance, mort fœtale [43, 69]. Les enfants de plus de cinq ans qui

FIGURE 5.2 : **Mère et enfant face au paludisme (IRD, 2010).**

ont été exposés régulièrement à P. falciparum développent une immunité dite "non-stérilisante" ou

immunité "clinique" qui n'apparaît qu'après un contact continu avec une large variété d'antigènes parasitaires [100]. Cette semi-immunité permet au patient d'être protégé contre la déclaration de symptômes, mais ne permet pas d'éliminer les parasites infectants [43, 199].

5.1.4 A la recherche d'un vaccin contre le paludisme

Le *P. falciparum* est connu pour son polymorphisme et sa variabilité génétique qui est relative aux protéines exprimées à la surface des érythrocytes - variant de surface antigénique (VAR) - codés par des gènes ayant plusieurs copies. Le polymorphisme quant à lui concerne les gènes représentés en une seule copie dans le génome et dépend de l'étendue des allèles disponibles et de la reproduction sexuée. D'après les études menées sur les gènes *var* - l'un des gènes les plus polymorphes - et le phénotype de cytoadhésion du parasite, le taux de mutation estimé varie entre 2 et 18 par génération [71]. Lors d'une infection multi-clonale, la majeur partie des anticorps sont synthétisés spécifiquement contre le VAR le plus représenté. Les parasites porteurs de cet antigène seront neutralisés alors que les parasites mutants seront épargnés, pouvant ainsi se développer et amener le système immunitaire à produire à nouveau des anticorps spécifiques du VAR majoritaire. Et ces parasites échappent au système immunitaire. Cette variabilité est donc à la base de toute la difficulté du système immunitaire à éliminer le parasite, et par conséquent du réel défi que représente la mise en place d'un vaccin efficace sur une longue durée [43]. Par ailleurs, le polymorphisme allélique de *P. falciparum* explique pourquoi il est nécessaire que le l'Homme soit exposé de nombreuses fois à divers antigènes avant d'acquérir une semi-immunité prévenant les accès palustres symptomatiques [100].

Trois stades dans le cycle de *P. falciparum* peuvent être la cible d'un vaccin : le stade pré-érythrocytaire, le stade érythrocytaire et le stade sexué [43, 77]. Les essais vaccinaux basés sur le stade pré-érythrocytaire ont pour objectif d'empêcher toute libération de mérozoites provenant du schizontes dans le sang. Des exemples de vaccins utilisant cette approche ont montré leur limite en phase IIb d'essai vaccinal, puisque incapables d'induire une immunité stérilisante [5, 179].

Les vaccins visant à stopper le développement des parasites asexués au stade érythrocytaire cherchent à inhiber l'invasion des hématies dans le but de contrôler la parasitémie et éviter ainsi l'évolution de l'infection vers des formes cliniques [43]. Ici, les candidats vaccins sont les antigènes portés par les mérozoites libérés par les schizontes érythrocytaires, essentiellement les Merozoite Surface Protein (MSP)-1, MSP2, MSP3, Apical Membrane Antigen (AMA)-1, Glutamate Rich Protein (GLURP). Des expériences de transfert d'immunoglobulines (Ig) purifiées d'adultes hyper-immuns à des enfants ont montré que les réponses anticorps jouaient un rôle dans l'acquisition de l'immunité adaptative dirigée contre les stades sanguins de *P. falciparum* [27, 34, 158].

Les vaccins dirigés contre le stade sexué de *P. falciparum* permettraient de bloquer la transmission vectorielle de la maladie, où les antigènes des gamètes en sont la cible. Ceci dit, il serait nécessaire de vacciner simultanément toute la population afin de stopper la transmission. De plus, cette vaccination devrait potentiellement être combinée à un vaccin des stades pré-érythrocytaire ou érythrocytaire afin d'espérer son efficacité.

La complexité du parasite, sa grande diversité antigénique et ses stratégies d'échappement immunitaire rendent le développement d'un vaccin difficile. Des efforts sont actuellement menés pour mettre au point un vaccin spécifiquement destiné à protéger la femme enceinte et son fœ-

tus aussi bien que le jeune enfant [43]. Les données que nous utilisons dans cette thèse sont issues des recherches menées en vue d'étudier la réponse anticorps spécifique d'antigènes du stade érythrocytaire, stade à l'origine des manifestations cliniques du paludisme.

5.2 Données pour l'acquisition de la réponse anticorps spécifique du paludisme

Les données que nous analysons dans le cadre de cette thèse ont été recueillies dans la commune de Tori-Bossito située dans le Nord-Ouest de Cotonou (au Bénin) à 40 km du centre ville (voir Figure 5.3). Cette commune, comptant quarante-sept villages et quartiers de ville, est caractérisée

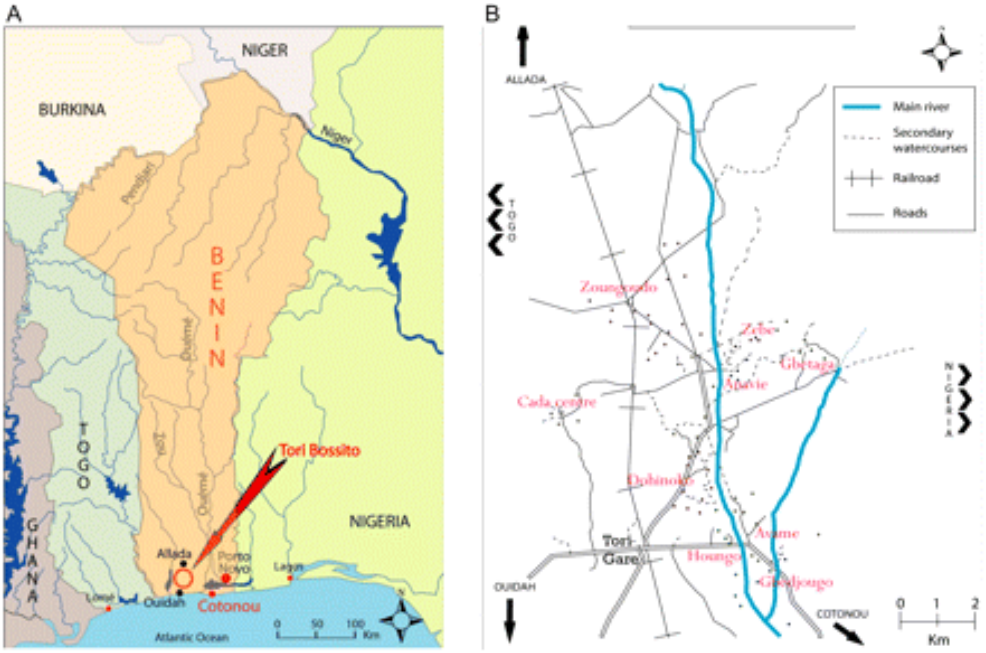

FIGURE 5.3 : **Localisation géographique de Tori-Bossito et les neuf villages d'étude au Bénin [108].**

par un climat sub-équatorial avec deux saisons de pluie et deux saisons sèches. La transmission palustre y est relativement identique chaque année (paludisme stable) fortement influencée par les pluies. Les vecteurs du paludisme rencontrés le plus souvent dans cette zone sont les *Anopheles gambiae* et les *Anopheles funestus* [35, 48].

Pour des raisons d'efficacité dans le suivi des sujets prenant part à l'étude, neuf villages parmi ceux qui sont proches des trois centres de santé de la commune de Tori-Bossito ont été sélectionnés. Il s'agit de : Avamé, Anavié, Cada-centre, Dohinoko, Gbédjougo, Gbétaga, Houngo, Zébé et Zoungoudo. Le programme avait pour objectif d'étudier les déterminants des premières infections palustres chez des nouveau-nés au cours des 18 premiers mois de vie. En outre il s'agissait également d'étudier la mise en place de la réponse immune spécifique des formes asexuées de *P. falciparum*. Pour atteindre ces deux objectifs il a été mis en place une cohorte de femmes

enceintes recrutées à l'accouchement et de leurs enfants suivis jusqu'à 18 mois. Etait susceptible d'être prise en compte par l'enquête, toute femme enceinte habitant l'un des villages retenus et accouchant dans l'une des trois maternités de la commune entre juin 2007 et juillet 2008. 620 femmes au total ont été incluses et chaque nouveau-né était suivi pendant ses dix-huit premiers mois de vie.

Les informations recueillies incluent des attributs descripteurs de la mère, de son nouveau-né aussi bien que de leur milieu de vie. A l'inclusion, était recueillies des informations épidémiologiques, comportementales et familiales, et celles nutritionnelles, environnementales, relatives à la malaria étaient enregistrées tout au long du suivi - voir [108] pour plus de détails. L'objectif de cette étude était d'évaluer les déterminants de l'incidence palustre dans les premiers mois de vie de l'enfant. Entre autres objectifs spécifiques, il s'agissait de comprendre comment est-ce que le système immunitaire de l'enfant va construire la capacité à répondre à des antigènes du stage érythrocytaire. Ainsi, les sept antigènes sélectionnés : AMA1, MSP1, MSP2 (3D7, FC27), MSP3 et GLURP (R0, R2) font partie de la phase érythrocytaire, sont immunogènes et sont des candidats vaccins prometteurs induisant une réponse anticorps protectrice démontrée [43]. Par ailleurs, ils sont localisés à la surface du mérozoite et sont donc accessibles aux anticorps lorsque les mérozoites sont libérés dans le sang. De plus, des protéines recombinantes les reproduisant sont disponibles. Pour plus de détails spécifiques à ces antigènes, voir par exemple [43].

5.3 Applications

Nous exposons ici deux illustrations relatives à l'analyse des données d'anticorps palustres mais que nous gardons sous anonymat pour des raisons de confidentialité. Ces données proviennent de l'UMR 216 et ont été codées et nous n'indiquons en conséquence aucun résultat biologique issus de nos analyses. Deux protéines recombiantes du *P. falciparum* ont été utilisées pour obtenir deux sous classes de IgG (IgG1 et IgG3) donnant ainsi des anticorps que nous avons nommés

$$\text{IgG1\_A1}, \text{IgG3\_A1}, \text{IgG1\_A2}, \text{IgG3\_A2}, \text{IgG1\_B}, \text{IgG3\_B}, \text{IgG1\_C}, \text{IgG3\_C}, \tag{5.1}$$

où A1, A2, B et C sont des antigènes du stade érythrocytaire. L'objectif de l'analyse de ces données dans le cadre de nos travaux de thèse est d'évaluer l'effet de l'infection palustre sur l'acquisition immunitaire de l'enfant pendant ces dix-huit premiers mois de vie. Puisque les antigènes qui caractérisent le statut immun de l'enfant inter-agissent entre eux dans l'organisme de l'enfant, nous analysons les caractéristiques de la distribution jointe de ces anticorps conditionnellement à l'infection palustre et à d'autres facteurs d'intérêt. Dans les modèles que nous ajustons à ces données, les variables dépendantes conc.Y (voir tableau 5.1) décrivent le niveau d'expression de l'antigène Y chez l'enfant à 3, 6, 9, 12, 15 et 18 mois, où Y désigne une protéine de la liste fournie à l'équation (5.1). Par ailleurs, toutes les autres variables (à l'exception de conc.Y) contenues dans le tableau 5.1 sont des covariables. Nous sommes ainsi en présence de huit variables dépendantes qui décrivent le profile longitudinale (chez l'enfant) des protéines listées à l'équation (5.1).

TABLE 5.1 : **Variables prises en compte par nos analyses**

Variable	Description
id	Identifiant de l'enfant
conc.Y	concentration de Y
conc_CO.Y	Concentration de Y mesurée dans le sang du cordon
conc_M3.Y	Concentration de Y prédite dans le sang périphérique de l'enfant à 3 mois
ap	Apposition placentaire
hb	Niveau d'hémoglobine
inf_trim	Nombre d'infections palustres dans les 3 derniers mois
pred_trim	Nombre moyen des prédictions entre 3-6 mois en matière d'exposition au moustique
nutri_trim	Moyenne des valeurs de nutrition mensuelle du trimestre écoulé

5.3.1 Première illustration : Classification hiérarchique de protéines palustres

Dans les modèles que nous avons ajustés aux données, nous avons mis un intercept aléatoire par enfant et une pente aléatoire par enfant suivant l'infection palustre (i.e. la variable inf_trim). L'illus-

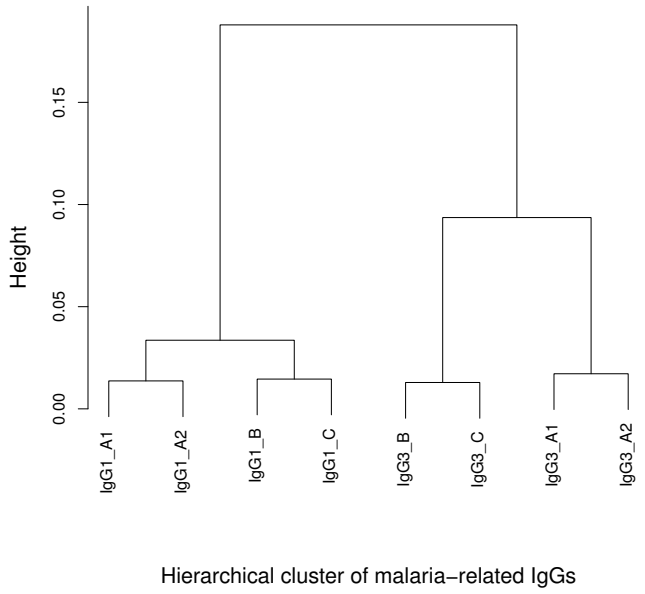

FIGURE 5.4 : **Arbre de classification hiérarchique de protéines palustres**

tration que nous faisons ici est d'analyser conjointement chacune des vingt-huit paires de protéines en vue de vérifier si certains profiles de protéines peuvent être analysés de façon indépendante, conditionnellement à la configuration du modèle ajusté. Après avoir mis en œuvre sur toutes ces paires de protéines le test de corrélation bi-varié introduit au Chapitre 2, en présence d'une correction de Bonferroni, les p-valeur obtenues vont de 0.000 à 0.932. La p-valeur 0.932 est d'ailleurs la seule qui soit en dessous du seuil de 5% et provient de la paire de protéines (IgG3_A1, IgG1_B). En

vue d'avoir une vue d'ensemble des protéines en terme de corrélation, nous avons réalisé leurs classification hiérarchique ascendante (voir Figure 5.4) en prenant $-\log(\text{p-valeur})$ comme dissimilarité. Sur la Figure 5.4, se distinguent clairement deux branches : celle de l'IgG1 et celle de l'IgG3. En d'autres mots, les protéines IgG1_A1, IgG1_A2, IgG1_B et IgG1_C sont sur une branche distincte de celle qui porte les protéines IgG3_A1, IgG3_A2, IgG3_B et IgG3_C. Par ailleurs, IgG1 et IgG3, A1 et A2, et B et C vont ensemble. Ces résultats sont biologiquement cohérents et sont compatibles avec la littérature, puisque A1 et A2 sont deux domaines de la même protéine, et B et C sont des protéines différents. Sur l'arbre (Figure 5.4), il apparait que les protéines IgG3_A1 et IgG1_B qui ne sont pas significativement corrélées d'après le test de corrélation bi-varié sont éloignées l'une de l'autre. Statistiquement, le modèle qui sera utilisé pour l'analyse jointe des huit protéines ne sera probablement pas celui qui contient vingt-sept corrélations significatives, évitant un possible sur-ajustement du modèle. En s'appuyant sur ces résultats, il semble digne d'intérêt de recourir à une procédure de régularisation dans l'ajustement du modèle conjoint à huit variables dépendantes.

5.3.2 Deuxième illustration : Estimation par optimisation de la déviance profilée du modèle

Dans cette section, nous utilisons la procédure d'estimation introduite au Chapitre 3 pour illustrer l'estimation des paramètres d'un modèle bi-varié ajusté aux données et ayant pour variables dépendantes conc.IgG1_A1 et conc.IgG3_A2 comme suit.

$$\begin{aligned}
\text{conc.IgG1_A1} &= (\mathbb{1}, \text{ap}, \text{conc_CO.IgG1_A1}, \text{conc_M3.IgG1_A1}, \text{hb}, \text{inf_trim}, \\
&\quad \text{pred_trim}, \text{nutri_trim})\beta_1 + (\mathbb{1}, \text{inf_trim})\gamma_1 + \varepsilon_1, \\
\\
\text{conc.IgG3_A2} &= (\mathbb{1}, \text{ap}, \text{conc_CO.IgG3_A2}, \text{conc_M3.IgG3_A2}, \text{hb}, \text{inf_trim}, \\
&\quad \text{pred_trim}, \text{nutri_trim})\beta_2 + (\mathbb{1}, \text{inf_trim})\gamma_2 + \varepsilon_2, \qquad (5.2)
\end{aligned}$$

avec

$$\boldsymbol{\gamma} = (\gamma_1^\top, \gamma_2^\top)^\top \sim \mathcal{N}\left(\mathbf{0} \quad \bar{\boldsymbol{\Gamma}}\right), \quad \boldsymbol{\varepsilon} = (\varepsilon_1^\top, \varepsilon_2^\top)^\top \sim \mathcal{N}\left(\mathbf{0}, \begin{pmatrix} \sigma_1^2\mathbf{I} & 0 \\ 0 & \sigma_2^2\mathbf{I} \end{pmatrix}\right). \qquad (5.3)$$

Ici, notre stratégie a consisté à 1) estimer les paramètres du modèle en démarrant la procédure à partir de vingt-cinq points de départ aléatoirement choisis, et de 2) considérer l'estimation ayant la plus grande valeur de vraisemblance comme fournissant la "vraie" valeur du vecteur de paramètres recherché. Puis nous calculons l'erreur quadratique moyenne (MSE) des vingt-quatre estimations restantes. Ceci permettrait d'évaluer la sensibilité de la procédure d'estimation au point de départ de l'algorithme, dans un cadre d'analyse de données réelles. Les résultats obtenus sont contenus dans le tableau 5.2. A partir de ces résultats, nous constatons effectivement que la sensibilité de notre procédure d'estimation au point de départ est vraiment minime, contrairement à beaucoup d'autres procédures dans la littérature. Par ailleurs, l'estimateur de la matrice de covariance des

TABLE 5.2 : **Analyse empirique de données de paludisme.**

	Variables réponse			
	conc.IgG1_A1		**conc.IgG3_A2**	
Covariables	$\hat{\beta}_1$	**MSE**	$\hat{\beta}_2$	**MSE**
Intercept	0.609	9.05×10^{-5}	-1.626	3.18×10^{-5}
ap	-0.093	1.06×10^{-5}	-0.337	1.04×10^{-6}
conc_CO.IgG1_A1	0.160	1.68×10^{-6}	$-$	$-$
conc_M3.IgG1_A1	0.148	9.85×10^{-6}	$-$	$-$
conc_CO.IgG3_A2	$-$	$-$	0.047	6.44×10^{-7}
conc_M3.IgG3_A2	$-$	$-$	0.155	2.22×10^{-7}
hb	-0.162	3.22×10^{-7}	-0.345	1.35×10^{-7}
inf_trim	0.369	1.89×10^{-6}	0.696	5.09×10^{-7}
pred_trim	-0.003	5.25×10^{-8}	0.017	1.49×10^{-8}
nutri_trim	0.024	5.81×10^{-6}	0.115	3.71×10^{-5}
	$\hat{\sigma}_1 = 1.395$	4.96×10^{-6}	$\hat{\sigma}_2 = 1.626$	2.42×10^{-5}

effets aléatoires est donné par

$$\widehat{\mathbf{\Gamma}} = \begin{pmatrix} 0.58 & -0.13 & 0.74 & -0.36 \\ -0.13 & 0.23 & -0.39 & 0.37 \\ 0.74 & -0.39 & 0.94 & -0.24 \\ -0.36 & 0.37 & -0.24 & 0.34 \end{pmatrix}, \tag{5.4}$$

avec une MSE de 0.0095.

CHAPITRE 6

CONCLUSION

6.1 Bilan

La question d'analyse jointe et longitudinale de plusieurs variables statistiques est d'actualité et suscite un intérêt croissant. Dans cette thèse, nous avons pu nous y pencher dans le cadre précis d'analyse de données longitudinales multidimensionnelles relatives à l'étude d'acquisition immunitaire contre le paludisme chez l'enfant en Afrique au sud du Sahara, plus précisément au Bénin. Notre démarche a été d'explorer dans un premier temps la littérature sur le sujet, en matière de démarches méthodologiques et algorithmiques. Ce qui nous a permis d'avoir une vue d'ensemble sur les différentes méthodes existantes afin de nous positionner en terme de contributions scientifiques.

Le chapitre 1 a donc fait l'objet d'un exposé du développement historique des méthodes d'analyse de données longitudinales unidimensionnelles où l'on analyse les caractéristiques de la distribution d'une seule variable longitudinale, conditionnellement à un ou plusieurs facteurs d'intérêt. Ce chapitre nous a aussi permis de proposer une définition plus générale du modèle linéaire à effets mixtes que celle rencontrée dans la littérature.

Le chapitre 2 expose nos travaux publiés dans PLoS One et qui introduisent des expressions plus générales (que celles rencontrées dans la littérature) d'estimateurs du maximum de vraisemblance des paramètres du modèle linéaire multidimensionnel à effets mixtes, utilisant l'algorithme EM. En effet, ces estimateurs intègrent naturellement le cadre d'analyse de données multidimensionnelles multi-niveaux dont les données longitudinales multidimensionnelles peuvent être vues comme étant un cas particulier. Nous avons également introduit un test de corrélation bi-varié permettant de tester la significativité globale des corrélations entre les effets aléatoires de deux différentes dimensions du modèle. Ce test a pour but d'aider à savoir si deux variables dépendantes peuvent être analysées séparément ou non, puisque la liaison entre elles (variables dépendantes) dépend uniquement des effets aléatoires, hypothèse faite à dessein. Ceci permettra de construire un modèle conjoint plus parcimonieux en terme de composantes de variance des effets aléatoires à travers une procédure de sélection pas-à-pas ascendante. A ce niveau, bien que la distribution asymptotique de la statistique de test soit empiriquement un χ^2 sur nos simulations, la littérature pointe plutôt du doigt un mélange pondéré de χ^2. Ce qui nous amène à envisager de considérer un approfondissement de la question dans nos prochains travaux à travers une démarche méthodologique purement théorique, ne serait-ce que dans le cadre spécifique de ce test de corrélation bi-varié.

En ce qui concerne le chapitre 3, nous y apportons une contribution méthodologique quant à l'estimation des paramètres d'un modèle linéaire multidimensionnel à effets mixtes où les résidus dimensionnels sont homoscédastiques avec la possibilité d'avoir des variances différentes dans

différentes dimensions du modèle. La procédure proposée ici est la généralisation de celle utilisée par les auteurs du package lme4 du logiciel R. Sa sensibilité au point de départ est quasi inexistante à la fois sur les données simulées que sur les données d'immunité palustre. Ce qui représente l'une des forces de la méthode qui, d'ailleurs, utilise les optimiseurs existants sous le logiciel R (où nous avons effectué tous nos travaux) pour minimiser ou maximiser le critère choisi (ML our REML) sans efforts supplémentaires. Les estimations réalisées sur des données simulées ont montré très clairement, une décroissance nette vers zéro de l'erreur quadratique moyenne lorsque le nombre d'observations augmente. cette approche semble donc fournir des estimateurs consistants, même pour les composantes de variance. Cet aspect (la consistance des estimateurs) mériterait d'être théoriquement approfondi dans nos travaux à venir. Enfin, le chapitre 4 quant à lui nous a permis de faire de la sélection d'effets fixes dans la version unidimensionnelle du modèle par pénalisation de la vraisemblance profilée des paramètres à l'exception de la variance résiduelle. La pénalité introduite ici est de type adaptive ridge itérative permettant d'approximer les performances d'une pénalité de type L_0 de la vraisemblance. Cette procédure de sélection mériterait d'être étendue à la version multidimensionnelle du modèle où on pourra sélectionner à la fois des effets fixes et des effets aléatoires en grande dimension (le nombre de covariables dépassant le nombre d'observations). L'approche utilisée dans les chapitres 3 et 4 mériterait sans doute d'intégrer ultérieurement une structure plus complexe de la matrice de covariance des résidus en présence d'une corrélation non nulle avec les effets aléatoires tout en conservant ses performances actuelles. Ce qui fera l'objet de nos réflexions dans nos travaux à venir.

6.2 Perspectives

La première perspective de notre travail concerne la loi asymptotique et à distance finie de la statistique du rapport de vraisemblance utilisée dans le chapitre 2. Si par exemple, γ_1 et γ_2 sont respectivement les vecteurs des effets aléatoires dans les dimensions 1 et 2 du modèle, le test de corrélation bi-varié du chapitre 2 a consisté à tester H_0 : Cor$(\gamma_1, \gamma_2) = 0$ contre H_1 : Cor$(\gamma_1, \gamma_2) \neq 0$, où nous avons procédé à de nombreuses simulations afin d'approcher numériquement la distribution asymptotique de la statistique de test du rapport de vraisemblance sous H_0. Nos simulations indiquent que la distribution asymptotique est celle du $\chi^2(df)$ (voir Figure 6.1). Cependant, dans la littérature, plusieurs travaux théoriques prouvent que la statistique de test du rapport de vraisemblance dans des conditions non standards (voir par exemple [31, 74, 171, 196]) où le paramètre testé est sujet à contraintes, suit asymptotiquement un mélange pondéré de χ^2 sous H_0. Or le test de corrélation bi-varié qui nous intéresse ici porte sur les corrélations entre les effets aléatoires de deux différentes dimensions ; lesquelles corrélations sont bien sûr sujettes à la contrainte d'appartenir à l'intervalle $[-1, 1]$. Dans la suite de nos travaux, nous nous pencherons donc plus théoriquement sur la distribution asymptotique de la statistique de test dans le cas précis de ce test de corrélation bi-varié.

Lorsqu'on travaille avec des jeux de données de taille très modeste, on peut recourir à une correction de type Bartlett [12] de la statistique de test S du rapport de vraisemblance afin de s'assurer, tout au moins, que la statistique utilisée (c'est à dire la statistique de test corrigée S_B) a df comme espérance mathématique, sous l'hypothèse H_0. Dans le chapitre 2, nous avons pratiqué

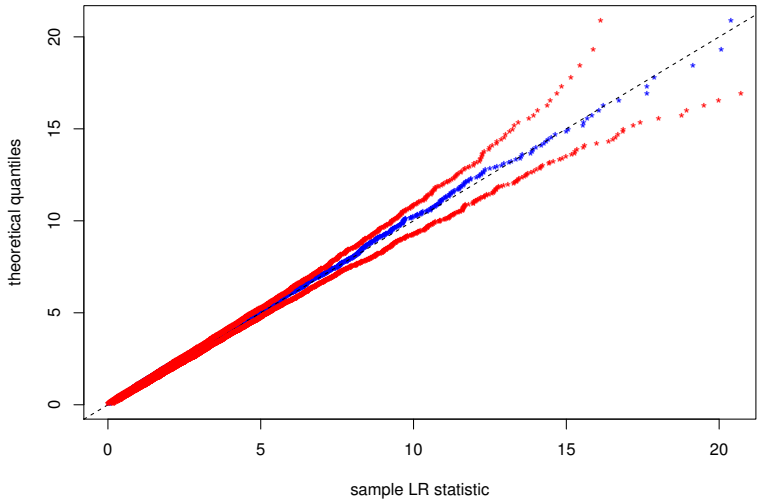

FIGURE 6.1 : Analyse empirique de la distribution asymptotique de la statistique du rapport de vraisemblance S sous H_0, utilisant 3000 replications de jeux de données longitudinales de tailles $N = 15000$ (observations) et $n = 500$ (subjects), simulées sous H_0. Les minima et les maxima de 1000×3000 réalisations simulées du $\chi^2(4)$ sont utilisés pour construire l'enveloppe rouge. La courbe bleue représente les statistiques de rapport de vraisemblance des tests de corrélation bi-varié réalisés sur les données simulées.

une correction empirique de Bartlett ($\widehat{S}_B = df \times S/\widehat{\mathbb{E}}[S|H_0]$) pour les jeux de données de petite taille. Dans la suite de nos travaux, nous entreprendrons de calculer explicitement l'espérance conditionnelle $\mathbb{E}[S|H_0]$, afin d'avoir une expression théorique générale pour la statistique corrigée dans le cas du test de corrélation bi-varié.

L'estimation des paramètres (du modèle) à la lme4 fait l'hypothèse de résidus homosédastiques. Cette hypothèse, quoique restrictive, nous permet d'obtenir de très bonnes performances en ce qui concerne la qualité des estimations. Dans la suite de nos travaux, nous pensons qu'il sera très utile de chercher à lever cette restriction en fournissant une expression générale de l'estimateur de la matrice de variance-covariance Σ_k des résidus de la dimension k du modèle.

La procédure de sélection d'effets fixes proposée au chapitre 4 fera l'objet d'une extension dans la suite de nos travaux. Cette extension prendra en compte la sélection d'effets aléatoires aussi bien que la sélection d'effets fixes dans la version multidimensionnelle du modèle.

D'autres travaux à venir incluent l'introduction d'une nouvelle méthode de sélection de SNP dans le traitement des données génome wide, où la procédure *adaptive ridge itérative* proposée au chapitre 4 sera utilisée. Par ailleurs, nous comptons aussi nous intéressés à l'analyse jointe de données longitudinales et de survie avec prise en compte ou non d'événements récurrents.

Bibliographie

[1] E. H. Adjakossa, I. Sadissou, M. N. Hounkonnou, and G. Nuel. Multivariate longitudinal analysis with bivariate correlation test. *PloS one*, 11(8) :e0159649, 2016.

[2] G. B. Airy. *On the algebraical and numerical theory of errors of observations and the combination of observations*. Macmillan&Company, 1861.

[3] H. Akaike. Information theory and an extension of the maximum likelihood principle. *Second International Symposium on Information Theory*, (Edited by B. N. Petrov and F. Csaki) :267–281, 1973.

[4] H. Akaike. A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6) :716–723, 1974.

[5] P. L. Alonso, J. Sacarlal, J. J. Aponte, A. Leach, E. Macete, J. Milman, I. Mandomando, B. Spiessens, C. Guinovart, M. Espasa, et al. Efficacy of the rts, s/as02a vaccine against plasmodium falciparum infection and disease in young african children : randomised controlled trial. *The Lancet*, 364(9443) :1411–1420, 2004.

[6] X. An, Q. Yang, and P. M. Bentler. A latent factor linear mixed model for high-dimensional longitudinal data analysis. *Statistics in medicine*, 32(24) :4229–4239, 2013.

[7] R. F. Anders, C. G. Adda, M. Foley, and R. S. Norton. Recombinant protein vaccines against the asexual blood-stages of plasmodium falciparum. *Human vaccines*, 6(1) :39–53, 2010.

[8] R. H. Baayen, D. J. Davidson, and D. M. Bates. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of memory and language*, 59(4) :390–412, 2008.

[9] J. Ballard, J. Khoury, K. Wedig, L. Wang, B. Eilers-Walsman, and R. Lipp. New ballard score, expanded to include extremely premature infants. *The Journal of pediatrics*, 119(3) : 417–423, 1991.

[10] S. Bandyopadhyay, B. Ganguli, and A. Chatterjee. A review of multivariate longitudinal data analysis. *Statistical methods in medical research*, 20(4) :299–330, 2011.

[11] A. Barron, L. Birgé, and P. Massart. Risk bounds for model selection via penalization. *Probability theory and related fields*, 113(3) :301–413, 1999.

[12] M. S. Bartlett. Properties of sufficiency and statistical tests. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, pages 268–282, 1937.

[13] D. Bates. Computational methods for mixed models. *LME4 : Mixed-Effects Modeling with R*, pages 99–118, 2014.

[14] D. Bates, D. Sarkar, M. D. Bates, and L. Matrix. The lme4 package. *R package version*, 2 (1) :74, 2007.

[15] D. Bates, M. Maechler, B. Bolker, and S. Walker. *lme4 : Linear mixed-effects models using Eigen and S4*, 2013. URL `http://CRAN.R-project.org/package=lme4`. R package version 1.0-5.

[16] D. Bates, M. Maechler, B. Bolker, and S. Walker. lme4 : Linear mixed-effects models using eigen and s4. r package version 1.1-7. *This is computer program (R package). The URL of the package is : http ://CRAN. R-project. org/package= lme4*, 2014.

[17] D. M. Bates. lme4 : Mixed-effects modeling with r. *URL http ://lme4. r-forge. r-project. org/book*, 2010.

[18] L. Beckett, D. Tancredi, and R. Wilson. Multivariate longitudinal models for complex change processes. *Statistics in medicine*, 23(2) :231–239, 2004.

[19] P. M. Bentler and D. G. Weeks. Linear structural equations with latent variables. *Psychometrika*, 45(3) :289–308, 1980.

[20] R. Biihlmann and L. Meier. Discussion of" one-step sparse estimates in nonconcave penalized likelihood models," by h. zou and r. li. *Ann. Statist*, 36 :1534–1541, 2008.

[21] H. Bozdogan. Model selection and akaike's information criterion (aic) : The general theory and its analytical extensions. *Psychometrika*, 52(3) :345–370, 1987.

[22] H. Brandsma and J. Knuver. Effects of school and classroom characteristics on pupil progress in language and arithmetic. *International Journal of Educational Research*, 13(7) : 777–788, 1989.

[23] L. Breiman. Better subset regression using the nonnegative garrote. *Technometrics*, 37(4) : 373–384, 1995.

[24] L. Breiman et al. Heuristics of instability and stabilization in model selection. *The annals of statistics*, 24(6) :2350–2383, 1996.

[25] L. F. Bringmann, N. Vissers, M. Wichers, N. Geschwind, P. Kuppens, F. Peeters, D. Borsboom, and F. Tuerlinckx. A network approach to psychopathology : new insights into clinical longitudinal data. *PloS one*, 8(4) :e60188, 2013.

[26] C. Brombin, C. Di Serio, and P. M. Rancoita. Joint modeling of hiv data in multicenter observational studies : A comparison among different approaches. *Statistical methods in medical research*, page 0962280214526192, 2014.

[27] P. C. Bull, B. S. Lowe, M. Kortok, C. S. Molyneux, C. I. Newbold, and K. Marsh. Parasite antigens on the infected red cell surface are targets for naturally acquired immunity to malaria. *Nature medicine*, 4(3) :358, 1998.

[28] E. J. Candes, M. B. Wakin, and S. P. Boyd. Enhancing sparsity by reweighted ? 1 minimization. *Journal of Fourier analysis and applications*, 14(5-6) :877–905, 2008.

[29] E. Capanna. Grassi versus ross : who solved the riddle of malaria ? *International Microbiology*, 9(1) :69–74, 2006.

[30] V. J. Carey and B. A. Rosner. Analysis of longitudinally observed irregularly timed multivariate outcomes : regression with focus on cross-component correlation. *Statistics in medicine*, 20(1) :21–31, 2001.

[31] D. Chant. On asymptotic tests of composite hypotheses in nonstandard conditions. *Biometrika*, 61(2) :291–298, 1974.

[32] R. Charnigo, R. Kryscio, M. T. Bardo, D. Lynam, and R. S. Zimmerman. Joint modeling of longitudinal data in multiple behavioral change. *Evaluation & the health professions*, 34 (2) :181–200, 2011.

[33] Z. Chen and D. B. Dunson. Random effects selection in linear mixed models. *Biometrics*, 59(4) :762–769, 2003.

[34] S. Cohen, I. McGregor, S. Carrington, et al. Gamma-globulin and acquired immunity to human malaria. *Nature*, 192 :733–7, 1961.

[35] V. Corbel, R. N ?guessan, C. Brengues, F. Chandre, L. Djogbenou, T. Martin, M. Akogbeto, J.-M. Hougard, and M. Rowland. Multiple insecticide resistance mechanisms in anopheles gambiae and culex quinquefasciatus from benin, west africa. *Acta tropica*, 101(3) :207–216, 2007.

[36] P. Corran, P. Coleman, E. Riley, and C. Drakeley. Serology : a robust indicator of malaria transmission intensity ? *Trends in parasitology*, 23(12) :575–582, 2007.

[37] G. Cottrell, B. Kouwaye, C. Pierrat, A. Le Port, A. Bouraïma, N. Fonton, M. N. Houngbonnou, A. Massougbodji, V. Corbel, and A. Garcia. Modeling the influence of local environmental factors on malaria transmission in benin and its implications for cohort study. *PLoS One*, 7(1) :e28812, 2012.

[38] D. Courtin, M. Oesterholt, H. Huismans, K. Kusi, J. Milet, C. Badaut, O. Gaye, W. Roeffen, E. J. Remarque, R. Sauerwein, et al. The quantity and quality of african children's igg responses to merozoite surface antigens reflect protection against plasmodium falciparum malaria. *PloS one*, 4(10) :e7590, 2009.

[39] R. Crouchley, D. Stott, J. Pritchard, and D. Grose. Multivariate generalised linear mixed models via sabrer (sabre in r). 2010.

[40] M. Crowder. On the use of a working correlation matrix in using generalized linear models for repeated measures. *Biometrika*, 82(2) :407–410, 1995.

[41] M. Danis and J. Mouchet. *Paludisme*. 1991.

[42] M. Davidian and D. M. Giltinan. Nonlinear models for repeated measurement data : an overview and update. *Journal of Agricultural, Biological, and Environmental Statistics*, 8 (4) :387–419, 2003.

[43] C. Dechavanne. *Construction de la réponse anticorps spécifique du paludisme chez le jeune enfant : étude combinée de l'hôte, du parasite et de leur environnement*. PhD thesis, Paris 5, 2012.

[44] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977.

[45] P. Diggle. *Analysis of longitudinal data*. Oxford University Press, 2002.

[46] P. Diggle, P. Heagerty, K.-Y. Liang, and S. Zeger. *Analysis of longitudinal data*. Number 25. Oxford University Press, 2013.

[47] A. Djènontin, S. Bio-Bangana, N. Moiroux, M.-C. Henry, O. Bousari, J. Chabi, R. Ossè, S. Koudénoukpo, V. Corbel, M. Akogbéto, et al. Culicidae diversity, malaria transmission and insecticide resistance alleles in malaria vectors in ouidah-kpomasse-tori district from benin (west africa) : A pre-intervention study. *Parasit Vectors*, 3 :83, 2010.

[48] R. Djouaka, H. Irving, Z. Tukur, and C. S. Wondji. Exploring mechanisms of multiple insecticide resistance in a population of the malaria vector anopheles funestus in benin. *PLos one*, 6(11) :e27760, 2011.

[49] S. C. Duncan and T. E. Duncan. A multivariate latent growth curve analysis of adolescent substance use. *Structural Equation Modeling : A Multidisciplinary Journal*, 3(4) :323–347, 1996.

[50] L. J. Edwards, K. E. Muller, R. D. Wolfinger, B. F. Qaqish, and O. Schabenberger. An r2 statistic for fixed effects in the linear mixed model. *Statistics in medicine*, 27(29) :6137–6157, 2008.

[51] J. Fan. Comments on wavelets in statistics : A review by a. antoniadis. *Journal of the Italian Statistical Society*, 6(2) :131–138, 1997.

[52] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456) :1348–1360, 2001.

[53] J. Fan and J. Lv. A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*, 20(1) :101, 2010.

[54] Y. Fan and R. Li. Variable selection in linear mixed effects models. *Annals of statistics*, 40 (4) :2043, 2012.

[55] J. J. Faraway. *Extending the linear model with R : generalized linear, mixed effects and nonparametric regression models*. CRC press, 2005.

[56] S. Fieuws and G. Verbeke. Joint modelling of multivariate longitudinal profiles : pitfalls of the random-effects approach. *Statistics in Medicine*, 23(20) :3093–3104, 2004.

[57] S. Fieuws and G. Verbeke. Pairwise fitting of mixed models for the joint modeling of multivariate longitudinal profiles. *Biometrics*, 62(2) :424–431, 2006.

[58] S. Fieuws and G. Verbeke. Joint models for high-dimensional longitudinal data. *Longitudinal data analysis*, pages 367–391, 2009.

[59] S. Fieuws, G. Verbeke, B. Maes, and Y. Vanrenterghem. Predicting renal graft failure using multivariate longitudinal profiles. *Biostatistics*, 9(3) :419–431, 2008.

[60] R. A. Fisher. The correlation between relatives on the supposition of mendelian inheritance. *Transactions of the royal society of Edinburgh*, 52(02) :399–433, 1919.

[61] G. Fitzmaurice, M. Davidian, G. Verbeke, and G. Molenberghs. *Longitudinal data analysis*. CRC Press, 2008.

[62] G. M. Fitzmaurice, N. M. Laird, and J. H. Ware. *Applied longitudinal analysis*, volume 998. John Wiley & Sons, 2012.

[63] L. E. Frank and J. H. Friedman. A statistical view of some chemometrics regression tools. *Technometrics*, 35(2) :109–135, 1993.

[64] E. W. Frees. *Longitudinal and panel data : analysis and applications in the social sciences*. Cambridge University Press, 2004.

[65] F. Frommlet and G. Nuel. An adaptive ridge procedure for l 0 regularization. *PloS one*, 11 (2) :e0148620, 2016.

[66] W. J. Fu. Penalized regressions : the bridge versus the lasso. *Journal of computational and graphical statistics*, 7(3) :397–416, 1998.

[67] I. Funatogawa, T. Funatogawa, and Y. Ohashi. An autoregressive linear mixed effects model for the analysis of longitudinal data which show profiles approaching asymptotes. *Statistics in medicine*, 26(9) :2113–2130, 2007.

[68] A. T. Galecki. General class of covariance structures for two or more repeated factors in longitudinal data analysis. *Communications in Statistics-Theory and Methods*, 23(11) : 3105–3119, 1994.

[69] B. Gamain, S. Gratepanche, L. H. Miller, and D. I. Baruch. Molecular basis for the dichotomy in plasmodium falciparum adhesion to cd36 and chondroitin sulfate a. *Proceedings of the National Academy of Sciences*, 99(15) :10020–10024, 2002.

[70] K. M. Gates and P. C. Molenaar. Group search algorithm recovers effective connectivity maps for individuals in homogeneous and heterogeneous samples. *Neuroimage*, 63(1) : 310–319, 2012.

[71] M. L. Gatton, J. M. Peters, E. V. Fowler, and Q. Cheng. Switching rates of plasmodium falciparum var genes : faster than we thought ? *Trends in parasitology*, 19(5) :202–208, 2003.

[72] A. Gelman and J. Hill. *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press, 2006.

[73] H. Geys, G. Molenberghs, and L. M. Ryan. Pseudolikelihood modeling of multivariate outcomes in developmental toxicology. *Journal of the American Statistical Association*, 94 (447) :734–745, 1999.

[74] V. Giampaoli and J. M. Singer. Likelihood ratio tests for variance components in linear mixed models. *Journal of Statistical Planning and Inference*, 139(4) :1435–1448, 2009.

[75] A. S. Goldberger. Best linear unbiased prediction in the generalized linear regression model. *Journal of the American Statistical Association*, 57(298) :369–375, 1962.

[76] H. Goldstein. Multilevel statistical models halsted press. *New York*, 1995.

[77] M. F. Good and D. L. Doolan. Malaria vaccine design : immunological considerations. *Immunity*, 33(4) :555–566, 2010.

[78] Y. Grandvalet. Least absolute shrinkage is equivalent to quadratic penalization. In *ICANN 98*, pages 201–206. Springer, 1998.

[79] S. M. Gray and R. Brookmeyer. Estimating a treatment effect from multidimensional longitudinal data. *Biometrics*, pages 976–988, 1998.

[80] S. M. Gray and R. Brookmeyer. Multidimensional longitudinal data : estimating a treatment effect from continuous, discrete, or time-to-event response variables. *Journal of the American Statistical Association*, 95(450) :396–406, 2000.

[81] F. Gumedze and T. Dunne. Parameter estimation and inference in the linear mixed model. *Linear Algebra and its Applications*, 435(8) :1920–1944, 2011.

[82] U. Halekoh, S. Højsgaard, and J. Yan. The r package geepack for generalized estimating equations. *Journal of Statistical Software*, 15(2) :1–11, 2006.

[83] J. D. Hamilton. State-space models. *Handbook of econometrics*, 4 :3039–3080, 1994.

[84] G. R. Hancock, W.-L. Kuo, and F. R. Lawrence. An illustration of second-order latent growth models. *Structural Equation Modeling*, 8(3) :470–489, 2001.

[85] E. J. Hannan and B. G. Quinn. The determination of the order of an autoregression. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 190–195, 1979.

[86] H. O. Hartley and J. N. Rao. Maximum-likelihood estimation for the mixed analysis of variance model. *Biometrika*, 54(1-2) :93–108, 1967.

[87] D. Harville. Extension of the gauss-markov theorem to include the estimation of random effects. *The Annals of Statistics*, pages 384–395, 1976.

[88] D. A. Harville. Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, 72(358) :320–338, 1977.

[89] S. I. Hay, C. A. Guerra, A. J. Tatem, A. M. Noor, and R. W. Snow. The global distribution and population at risk of malaria : past, present, and future. *The Lancet infectious diseases*, 4(6) :327–336, 2004.

[90] D. Hedeker and R. D. Gibbons. *Longitudinal data analysis*, volume 451. John Wiley & Sons, 2006.

[91] C. R. Henderson. Estimation of genetic parameters. In *Biometrics*, volume 6, pages 186–187. INTERNATIONAL BIOMETRIC SOC 1441 I ST, NW, SUITE 700, WASHINGTON, DC 20005-2210, 1950.

[92] C. R. Henderson. Estimation of variance and covariance components. *Biometrics*, 9(2) : 226–252, 1953.

[93] C. Horváth and J. E. Wieringa. Pooling data for the analysis of dynamic marketing systems. *Statistica Neerlandica*, 62(2) :208–229, 2008.

[94] J. G. Ibrahim, H. Zhu, and N. Tang. Model selection criteria for missing-data problems using the em algorithm. *Journal of the American Statistical Association*, 2008.

[95] J. G. Ibrahim, H. Zhu, R. I. Garcia, and R. Guo. Fixed and random effects selection in mixed effects models. *Biometrics*, 67(2) :495–503, 2011.

[96] R. I. Jennrich and M. D. Schluchter. Unbalanced repeated-measures models with structured covariance matrices. *Biometrics*, pages 805–820, 1986.

[97] K. L. Jensen, H. Spiild, and J. Toftum. Implementation of multivariate linear mixed-effects models in the analysis of indoor climate performance experiments. *International journal of biometeorology*, 56(1) :129–136, 2012.

[98] J. Jiang, J. S. Rao, Z. Gu, T. Nguyen, et al. Fence methods for mixed model selection. *The Annals of Statistics*, 36(4) :1669–1692, 2008.

[99] R. A. Johnson, D. W. Wichern, and P. Education. *Applied multivariate statistical analysis*, volume 4. Prentice hall Englewood Cliffs, NJ, 2007.

[100] R. Kobbe, R. Neuhoff, F. Marks, S. Adjei, I. Langefeld, C. Von Reden, O. Adjei, C. G. Meyer, and J. May. Seasonal variation and high multiplicity of first plasmodium falciparum infections in children from a holoendemic area in ghana, west africa. *Tropical Medicine & International Health*, 11(5) :613–619, 2006.

[101] M. Kramer. R 2 statistics for mixed models. 2005.

[102] E. Kuhn and M. Lavielle. Maximum likelihood estimation in nonlinear mixed effects models. *Computational Statistics & Data Analysis*, 49(4) :1020–1038, 2005.

[103] N. Laird, N. Lange, and D. Stram. Maximum likelihood computations with repeated measures : application of the em algorithm. *Journal of the American Statistical Association*, 82 (397) :97–105, 1987.

[104] N. M. Laird. Computation of variance components using the em algorithm. *Journal of Statistical Computation and Simulation*, 14(3-4) :295–303, 1982.

[105] N. M. Laird and J. H. Ware. Random-effects models for longitudinal data. *Biometrics*, pages 963–974, 1982.

[106] P. Lambert and F. Vandenhende. A copula-based model for multivariate non-normal longitudinal data : analysis of a dose titration safety study on a new antidepressant. *Statistics in medicine*, 21(21) :3197–3217, 2002.

[107] P. Lazarsfeld and M. Fiske. The "panel" as a new tool for measuring opinion. *The Public Opinion Quarterly*, 2(4) :596–612, 1938.

[108] A. Le Port, G. Cottrell, Y. Martin-Prevel, F. Migot-Nabias, M. Cot, and A. Garcia. First malaria infections in a cohort of infants in benin : biological, environmental and genetic determinants. description of the study site, population methods and preliminary results. *BMJ open*, 2(2) :e000342, 2012.

[109] Y. Lee and J. Nelder. Generalized linear models for the analysis of quality-improvement experiments. *Canadian Journal of Statistics*, 26(1) :95–105, 1998.

[110] K.-Y. Liang and S. L. Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1) :13–22, 1986.

[111] K.-Y. Liang, S. L. Zeger, and B. Qaqish. Multivariate regression analyses for categorical data. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 3–40, 1992.

[112] X. Lin. Variance component testing in generalised linear models with random effects. *Biometrika*, 84(2) :309–326, 1997.

[113] D. V. Lindley and A. F. Smith. Bayes estimates for the linear model. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–41, 1972.

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

[115] M. J. Lindstrom and D. M. Bates. Nonlinear mixed effects models for repeated measures data. *Biometrics*, pages 673–687, 1990.

[116] R. Littell, G. Milliken, W. Stroup, R. Wolfinger, and O. Schabenberger. Random coefficient models. *SAS system for mixed models. Cary, NC : SAS Institute Inc*, pages 253–66, 1996.

[117] R. C. Littell, G. A. Milliken, W. W. Stroup, R. D. Wolfinger, and O. Schabenberger. Sas system for mixed models cary. *Nc : sas institute*, 1996.

[118] W. Liu, Y. Li, G. H. Learn, R. S. Rudicell, J. D. Robertson, B. F. Keele, J.-B. N. Ndjango, C. M. Sanz, D. B. Morgan, S. Locatelli, et al. Origin of the human malaria parasite plasmodium falciparum in gorillas. *Nature*, 467(7314) :420–425, 2010.

[119] T. Lodewyckx, F. Tuerlinckx, P. Kuppens, N. B. Allen, and L. Sheeber. A hierarchical state space approach to affective dynamics. *Journal of mathematical psychology*, 55(1) :68–83, 2011.

[120] R. C. MacCallum, C. Kim, W. B. Malarkey, and J. K. Kiecolt-Glaser. Studying multivariate change using multilevel models and latent curve models. *Multivariate Behavioral Research*, 32(3) :215–253, 1997.

[121] T. E. MaCurdy. The use of time series processes to model the error structure of earnings in a longitudinal data analysis. *Journal of econometrics*, 18(1) :83–114, 1982.

[122] C. L. Mallows. Some comments on c p. *Technometrics*, 15(4) :661–675, 1973.

[123] J. J. McArdle. Dynamic but structural equation modeling of repeated measures data. In *Handbook of multivariate experimental psychology*, pages 561–614. Springer, 1988.

[124] J. S. McCarthy, J. Marjason, S. Elliott, P. Fahey, G. Bang, E. Malkin, E. Tierney, H. Aked-Hurditch, C. Adda, N. Cross, et al. A phase 1 trial of msp2-c1, a blood-stage malaria vaccine containing 2 isoforms of msp2 formulated with montanide (r) isa 720. *PLoS One*, 6(9) :e24413, 2011.

[125] X.-L. Meng and D. B. Rubin. Maximum likelihood estimation via the ecm algorithm : A general framework. *Biometrika*, 80(2) :267–278, 1993.

[126] J. J. Miller. Asymptotic properties and computation of maximum likelihood estimates in the mixed model of the analysis of variance. Technical report, DTIC Document, 1973.

[127] R. Miller, S. Ikram, G. Armelagos, R. Walker, W. Harer, C. Shiff, D. Baggett, M. Carrigan, and S. Maret. Diagnosis of plasmodium falciparum infections in mummies using the rapid manual parasight ?-f test. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 88(1) :31–32, 1994.

[128] G. Minois. *Histoire de l'avenir : des prophètes à la prospective*. Fayard, 1996.

[129] G. Molenberghs and G. Verbeke. Models for discrete longitudinal data. 2005.

[130] J. Mouchet. *Biodiversité du paludisme dans le monde*. John Libbey Eurotext, 2004.

[131] S. Müller, J. L. Scealy, A. H. Welsh, et al. Model selection in linear mixed models. *Statistical Science*, 28(2) :135–167, 2013.

[132] R. B. Nelsen. *An introduction to copulas*. Springer, 1999.

[133] M. Nerlove. An essay on the history of panel data econometrics. In *Proceedings of Ninth International Conference on Panel Data, Geneva, Switzerland*, page 13, 2000.

[134] P. N. Oak. World vedic heritage : a history of histories : presenting a unique unified field theory of history that from the beginning of time the world practised vedic culture and spoke sanskrit. 1984.

[135] L. M. O'Brien and G. M. Fitzmaurice. Analysis of longitudinal multiple-source binary data using generalized estimating equations. *Journal of the Royal Statistical Society : Series C (Applied Statistics)*, 53(1) :177–193, 2004.

[136] P. E. Okello, W. Van Bortel, A. M. Byaruhanga, A. Correwyn, P. Roelants, A. Talisuna, U. d ?Alessandro, and M. Coosemans. Variation in malaria transmission intensity in seven sites throughout uganda. *The American journal of tropical medicine and hygiene*, 75(2) : 219–225, 2006.

[137] F. J. Oort. Three-mode models for multivariate longitudinal data. *British Journal of Mathematical and Statistical Psychology*, 54(1) :49–78, 2001.

[138] W. H. Organization et al. World malaria report 2012. 2012. *Geneva : World Health Organization Google Scholar*, 2014.

[139] H. D. Patterson and R. Thompson. Recovery of inter-block information when block sizes are unequal. *Biometrika*, 58(3) :545–554, 1971.

[140] J. Pinheiro and D. Bates. *Mixed-effects models in S and S-PLUS*. Springer Science & Business Media, 2006.

[141] J. Pinheiro, D. Bates, S. DebRoy, D. Sarkar, et al. Linear and nonlinear mixed effects models. *R package version*, 3 :57, 2007.

[142] J. Pinheiro, D. Bates, S. DebRoy, and D. Sarkar. R core team (2014). nlme : linear and nonlinear mixed effects models. r package version 3.1–117. *URL : http ://cran. r-project. org/web/packages/nlme/index. html*, 2014.

[143] R. L. Prentice and L. P. Zhao. Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses. *Biometrics*, pages 825–839, 1991.

[144] W. Pu and X.-F. Niu. Selecting mixed-effects models based on a generalized information criterion. *Journal of multivariate analysis*, 97(3) :733–758, 2006.

[145] R Core Team. *R : A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014. URL `http://www.R-project.org/`.

[146] R Core Team. *R : A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015. URL `http://www.R-project.org/`.

[147] R Core Team. *R : A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015. URL `http://www.R-project.org/`.

[148] J. Ramsay and B. Silverman. Functional data analysis. 1997, 1997.

[149] C. R. Rao. Estimation of variance and covariance components ?minque theory. *Journal of multivariate analysis*, 1(3) :257–275, 1971.

[150] R. Rao and Y. Wu. A strongly consistent procedure for model selection in a regression problem. *Biometrika*, 76(2) :369–374, 1989.

[151] G. Reinsel. Estimation and prediction in a multivariate random effects generalized linear model. *Journal of the American Statistical Association*, 79(386) :406–414, 1984.

[152] H. Ribaudo and S. Thompson. The analysis of repeated multivariate binary quality of life data : a hierarchical model approach. *Statistical methods in medical research*, 11(1) :69–83, 2002.

[153] J. A. Rice and C. O. Wu. Nonparametric mixed effects models for unequally sampled noisy curves. *Biometrics*, 57(1) :253–259, 2001.

[154] R. C. Rippe, J. J. Meulman, and P. H. Eilers. Visualization of genomic changes by segmented smoothing using an l 0 penalty. *PloS one*, 7(6) :e38230, 2012.

[155] G. K. Robinson. That blup is a good thing : the estimation of random effects. *Statistical science*, pages 15–32, 1991.

[156] J. Rochon. Analyzing bivariate repeated measures for discrete and continuous outcome variables. *Biometrics*, pages 740–750, 1996.

[157] C. Rogier, E. Orlandi-Pradines, T. Fusai, B. Pradines, S. Briolant, and L. Almeras. [malaria vaccines : prospects and reality]. *Medecine et maladies infectieuses*, 36(8) :414–422, 2006.

[158] A. Sabchareon, T. Burnouf, D. Ouattara, P. Attanath, H. Bouharoun-Tayoun, P. Chantavanich, C. Foucault, T. Chongsuphajaisiddhi, and P. Druilhe. Parasitologic and clinical human response to immunoglobulin administration in falciparum malaria. *The American journal of tropical medicine and hygiene*, 45(3) :297–308, 1991.

[159] M. Sammel, X. Lin, and L. Ryan. Multivariate linear mixed models for multiple outcomes. *Statistics in medicine*, 18(17-18) :2479–2492, 1999.

[160] B. R. Saville and A. H. Herring. Testing random effects in the linear mixed model using approximate bayes factors. *Biometrics*, 65(2) :369–376, 2009.

[161] J. L. Schafer and R. M. Yucel. Computational strategies for multivariate linear mixed-effects models with missing values. *Journal of computational and Graphical Statistics*, 11 (2) :437–457, 2002.

[162] H. Scheffe. Alternative models for the analysis of variance. *The Annals of Mathematical Statistics*, pages 251–271, 1956.

[163] J. Schelldorfer. lmmlasso : Linear mixed-effects models with lasso. *R package version 0.1-2*, 2011.

[164] J. Schelldorfer, P. Bühlmann, G. DE, and S. VAN. Estimation for high-dimensional linear mixed-effects models using l1-penalization. *Scandinavian Journal of Statistics*, 38(2) :197–214, 2011.

[165] J. Schelldorfer, L. Meier, and P. Bühlmann. Glmmlasso : an algorithm for high-dimensional generalized linear mixed models using ?1-penalization. *Journal of Computational and Graphical Statistics*, 23(2) :460–477, 2014.

[166] P. Schlagenhauf and C. Hatz. Traitement de secours antipaludique : Actualisation 1997. *Médecine et hygiène*, 55(2165) :1126–1127, 1997.

[167] G. Schwarz et al. Estimating the dimension of a model. *The annals of statistics*, 6(2) : 461–464, 1978.

[168] S. Searle, G. Casella, and C. McCulloch. Variance components john wiley and sons. *New York, New York, USA*, 1992.

[169] S. R. Searle. An overview of variance component estimation. *Metrika*, 42(1) :215–230, 1995.

[170] S. R. S. R. Searle. Linear models. Technical report, 1971.

[171] S. G. Self and K.-Y. Liang. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, 82(398) :605–610, 1987.

[172] A. Shah, N. Laird, and D. Schoenfeld. A random-effects model for multiple characteristics with possibly missing data. *Journal of the American Statistical Association*, 92(438) :775–779, 1997.

[173] J. Shao. An asymptotic theory for linear model selection. *Statistica Sinica*, pages 221–242, 1997.

[174] N. W. Shock, r. C. Greulich, P. T. Costa, R. Andres, E. G. Lakatta, D. Arenberg, and J. D. Tobin. Normal human aging : The baltimore longitudinal study of aging. 1984.

[175] M. Sklar. *Fonctions de répartition à n dimensions et leurs marges*. Université Paris 8, 1959.

[176] T. A. Smith, R. Leuenberger, and C. Lengeler. Child mortality and malaria transmission intensity in africa. *Trends in parasitology*, 17(3) :145–149, 2001.

[177] T. A. Snijders. *Multilevel analysis*. Springer, 2011.

[178] T. A. Snijders and R. J. Bosker. Modeled variance in two-level models. *Sociological methods & research*, 22(3) :342–363, 1994.

[179] G. Snounou, A. C. Grüner, C. D. Müller-Graf, D. Mazier, and L. Rénia. The plasmodium sporozoite survives rts, s vaccination. *Trends in parasitology*, 21(10) :456–461, 2005.

[180] R. W. Snow, C. A. Guerra, A. M. Noor, H. Y. Myint, and S. I. Hay. The global distribution of clinical episodes of plasmodium falciparum malaria. *Nature*, 434(7030) :214–217, 2005.

[181] S. Sturtz, U. Ligges, and A. Gelman. R2winbugs : A package for running winbugs from r. *Journal of Statistical Software*, 12(3) :1–16, 2005. URL `http://www.jstatsoft.org`.

[182] S. Subramanian, D. Kim, and I. Kawachi. Covariation in the socioeconomic determinants of self rated health and happiness : a multivariate multilevel analysis of individuals and communities in the usa. *Journal of Epidemiology and Community Health*, 59(8) :664–669, 2005.

[183] J. Sy, J. Taylor, and W. Cumberland. A stochastic model for the analysis of bivariate longitudinal aids data. *Biometrics*, pages 542–555, 1997.

[184] H. Theil and A. S. Goldberger. On pure and mixed statistical estimation in economics. *International Economic Review*, 2(1) :65–78, 1961.

[185] R. Thiébaut, H. Jacqmin-Gadda, G. Chêne, C. Leport, and D. Commenges. Bivariate linear mixed models using sas proc mixed. *Computer methods and programs in biomedicine*, 69 (3) :249–256, 2002.

[186] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[187] R. S. Tsay. *Multivariate Time Series Analysis : With R and Financial Applications*. John Wiley & Sons, 2013.

[188] W. Tschacher and F. Ramseyer. Modeling psychotherapy process by time-series panel analysis (tspa). *Psychotherapy Research*, 19(4-5) :469–481, 2009.

[189] W. Tschacher, P. Zorn, and F. Ramseyer. Change mechanisms of schema-centered group psychotherapy with personality disorder patients. *PloS one*, 7(6) :e39687, 2012.

[190] A. Tseloni and C. Zarafonitou. Fear of crime and victimization a multivariate multilevel analysis of competing measurements. *European Journal of Criminology*, 5(4) :387–409, 2008.

[191] F. Vaida and S. Blanchard. Conditional akaike information for mixed-effects models. *Biometrika*, 92(2) :351–370, 2005.

[192] G. Verbeke and G. Molenberghs. Linear mixed models in practice : a sas oriented approach. 1997.

[193] G. Verbeke and G. Molenberghs. *Linear mixed models for longitudinal data*. Springer Science & Business Media, 2009.

[194] G. Verbeke, S. Fieuws, G. Molenberghs, and M. Davidian. The analysis of multivariate longitudinal data : A review. *Statistical methods in medical research*, 23(1) :42–59, 2014.

[195] E. F. Vonesh. A note on the use of laplace's approximation for nonlinear mixed-effects models. *Biometrika*, 83(2) :447–452, 1996.

[196] H. Vu, S. Zhou, et al. Generalization of likelihood ratio tests under nonstandard conditions. *The Annals of Statistics*, 25(2) :897–916, 1997.

[197] W.-L. Wang and T.-H. Fan. Ecm-based maximum likelihood inference for multivariate linear mixed models with autoregressive errors. *Computational Statistics & Data Analysis*, 54(5) :1328–1341, 2010.

[198] X.-F. Wang. Joint generalized models for multidimensional outcomes : A case study of neuroscience data from multimodalities. *Biometrical Journal*, 54(2) :264–280, 2012.

[199] G. E. Weiss, B. Traore, K. Kayentao, A. Ongoiba, S. Doumbo, D. Doumtabe, Y. Kone, S. Dia, A. Guindo, A. Traore, et al. The plasmodium falciparum-specific human memory b cell compartment expands gradually with repeated malaria infections. *PLoS Pathog*, 6(5) : e1000912, 2010.

[200] S. S. Wilks. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, 9(1) :60–62, 1938.

[201] H. Wu and J.-T. Zhang. *Nonparametric regression methods for longitudinal data analysis : mixed-effects modeling approaches*, volume 515. John Wiley & Sons, 2006.

[202] R. Xu. Measuring explained variation in linear mixed effects models. *Statistics in medicine*, 22(22) :3527–3541, 2003.

[203] R. Yucel. *mlmmm : ML estimation under multivariate linear mixed models with missing values*, 2010. URL `https://CRAN.R-project.org/package=mlmmm`. R package version 0.3-1.2.

[204] S. Zeger, K. Liang, and P. Albert. Models for longitudinal data : a generalized estimating equation approach. *Biometrics*, pages 1049–1060, 1988.

[205] S. L. Zeger and K.-Y. Liang. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, pages 121–130, 1986.

[206] A. Zellner. An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *Journal of the American statistical Association*, 57(298) :348–368, 1962.

[207] C. H. Zhang. Penalized linear unbiased selection. *Department of Statistics and Bioinformatics, Rutgers University*, pages 2007–003, 2007.

[208] M. Zhang, A. A. Tsiatis, M. Davidian, K. S. Pieper, and K. W. Mahaffey. Inference on treatment effects from a randomized clinical trial in the presence of premature treatment discontinuation : the synergy trial. *Biostatistics*, 12(2) :258–269, 2011.

[209] H. Zou. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476) :1418–1429, 2006.

[210] A. Zuur, E. N. Ieno, N. Walker, A. A. Saveliev, and G. M. Smith. *Mixed effects models and extensions in ecology with R*. Springer Science & Business Media, 2009.

Résumé

Dans cette thèse, nous nous sommes focalisés sur le modèle statistique linéaire à effets mixtes. Nous nous sommes d'abord intéressés à l'estimation consistante des paramètres du modèle dans sa version multidimensionnelle, puis à de la sélection d'effets fixes en dimension un. En ce qui concerne l'estimation des paramètres du modèle linéaire à effets mixtes multidimensionnel, nous avons proposé des estimateurs du maximum de vraisemblance par utilisation de l'algorithme EM, mais avec des expressions plus générales que celles de la littérature classique, permettant d'analyser non seulement des données longitudinales multivariées mais aussi des données multidimensionnelles multi-niveaux. Ici, en s'appuyant sur ces EM-estimateurs, nous avons introduit un test de rapport de vraisemblance permettant de tester la significativité globale des corrélations entre les effets aléatoires de deux dimensions du modèle. Ce qui permettrait de construire un modèle multidimensionnel plus parcimonieux en terme de paramètres de variance des effets aléatoires, par une procédure de selection pas-à-pas ascendante. Cette démarche a été suscitée par le fait que la dimension du vecteur de tous les effets aléatoires du modèle peut très rapidement croitre avec le nombre de variables à analyser, entrainant facilement des problèmes numériques dans l'optimisation du critère choisi (ML ou REML). Nous avons ensuite proposé une procédure d'estimation consistante des paramètres du modèle qui passe par la résolution d'un problème de moindres carrés pénalisés pour fournir une expression explicite de la déviance à minimiser. La procédure de sélection d'effets fixes proposée ici est de type adaptive ridge itérative et permet d'approximer les performances de sélection d'une pénalité de type L_0 de la vraisemblance des paramètres du modèle. Nos résultats ont été appuyés par des études de simulation à plusieurs niveaux, mais aussi par l'analyse de plusieurs jeux de données réelles.

Abstract

This thesis focuses on the statistical linear mixed-effects model, where we have been interested in its multivariate version's parameters estimation but also in the unidimensional selection of fixed effects. Concerning the parameters estimation of the multivariate linear mixed-effects model, we have first introduced more general expressions of the EM algorithm-based estimators which fit the multivariate longitudinal data analysis framework but also the framework of the multivariate multilevel data analysis. Since the dimensionality of the total vector of random effects in the multivariate model can grow with the number of the outcome variables leading often to computational problems in the likelihood optimization, we introduced a likelihood ratio test for testing the global effect of the correlations between the random effects of two dimensions of the model. This bivariate correlation test is intended to help in constructing a more parsimonious model regarding the variance components of the random effects, using a stepwise procedure. Secondly, we have introduced another estimation procedure that yields to consistent estimates for all the model parameters. This procedure is based on the Cholesky factorization of the random effects covariance matrix and the resolution of a preliminary penalized means square problem, and leads to an explicite expression of the profiled deviance of the model. For selecting fixed effects in the one dimensional mixed-effects model, we introduce an iterative adaptive ridge procedure for approximating

L_0 penalty selection performances. All the results in this manuscript have been accompanied by extensive simulation studies along with real data analysis examples.