Université de Montréal

Modèle de mélange gaussien à effets superposés pour l'identification de sous-types de schizophrénie

par

Samy Nefkha-Bahri

Département de mathématiques et de statistique

Faculté des arts et des sciences

Mémoire présenté à la Faculté des études supérieures et postdoctorales

en vue de l'obtention du grade de

Maître ès sciences (M.Sc.)

en statistique

Mars 2020

Sommaire

Ce travail s'inscrit dans l'effort de recherche ayant pour but d'identifier des sous-types de schizophrénie à travers des données de connectivité cérébrale tirées de l'imagerie par résonance magnétique fonctionelle. Des techniques de regroupement en grappes, dont l'algorithme Espérance-Maximisation (EM) pour l'estimation des paramètres de modèles de mélange gaussien, ont été utilisées sur des données de ce type dans des recherches précédentes. Cette approche capture des effets de processus cérébraux normaux qui sont sans intérêt pour l'identification de sous-types maladifs. Dans le présent travail, les données de la population des individus témoins (non-atteints de la maladie) sont modélisées par un mélange fini de densités gaussiennes. Chaque densité représente un sous-type supposé de fonctionnement cérébral normal. Une nouvelle modélisation est proposée pour les données de la population des individus atteints : un mélange de densités gaussiennes où chaque densité a une moyenne correspondant à la somme d'un état normal et d'un état maladif. Il s'agit donc d'un modèle de mélange gaussien dans lequel se superposent des sous-types de fonctionnement cérébral normal et des sous-types de maladie. On présuppose que les processus normaux et maladifs sont additifs et l'objectif est d'isoler et d'estimer les effets maladifs. Un algorithme de type EM spécifiquement conçu pour ce modèle est développé. Nous disposons en outre de données de connectivité cérébrale de 242 individus témoins et 242 patients diagnostiqués schizophrènes. Des résultats de l'utilisation de cet algorithme sur ces données sont rapportés.

Mots-clés : mélange gaussien, effets superposés, regroupement en grappes, schizophrénie.

Summary

This work is part of the research effort to identify subtypes of schizophrenia through brain connectivity data from functional magnetic resonance imaging. Clustering techniques, including the Esperance-Maximization algorithm (EM) for estimating parameters of Gaussian mixture models, have been used on such data in previous research. This approach captures the effects of normal brain processes that are irrelevant to the identification of disease subtypes. In this work, the population data of control (non-disease) individuals are modeled by a finite mixture of Gaussian densities. Each density represents an assumed subtype of normal brain function. A new model is proposed for the population data of affected individuals : a mixture of Gaussian densities where each density has an mean corresponding to the sum of a normal state and a disease state. Therefore, it is a mixture in which subtypes of normal brain function and subtypes of disease are superimposed. It is assumed that normal and unhealthy processes are additive and the goal is to isolate and estimate the unhealthy effects. An EM algorithm specifically designed for this model is developed. Data were obtained from functional magnetic resonance imaging of 242 control individuals and 242 patients diagnosed with schizophrenia. Results obtained using this algorithm on this data set are reported.

Key words : gaussian mixture, superimposed effects, clustering, schizophrenia.

Table des matières

Liste des tableaux

Table des figures

Introduction

«La psychiatrie est un domaine en crise». C'est cela qu'avancent certains scientifiques [voir Wiecki et collab., 2015, p. 2]. Les nombreux et grands développements technologiques et scientifiques des dernières décennies ont mené à une révolution en médecine de telle sorte que beaucoup des décisions importantes dans le cadre d'une pratique moderne de la médecine sont basées sur des signes observables et quantifiables ainsi que sur des tests objectifs. L'efficacité et la fiabilité de ces tests dépendent en grande partie du récent développement des connaissances en biologie et en biologie moléculaire en particulier qui ont mené à une compréhension poussée et fondamentale des phénomènes observés en clinique. Une des branches de la médecine semble toutefois se démarquer des autres : la psychiatrie se pratique encore largement sur la base d'une symptomatologie car l'esprit et le cerveau humain persistent à apparaître relativement mystérieux [Tamminga et collab., 2014, p. 1].

Parmi les syndromes diagnostiqués sur la base de signes et symptômes, il y a la schizophrénie. À l'époque du DSM-IV [2001], la schizophrénie était divisée en plusieurs sous-types. La version la plus récente du DSM, le DSM-5 [2013], propose un seul diagnostic et remplace la division en sous-types par une échelle de sévérité. Il existe toujours une nécessité de catégoriser le syndrome selon la diversité des symptômes et de leur sévérité [voir DSM-5, p. 99 et DSM-IV-TR, p. 349].

D'aucuns pourraient dire que, fondamentalement, l'approche diagnostique demeure la même par le fait qu'elle reste axée sur la symptomatologie et qu'elle reflète implicitement une incapacité à différencier ce qui pourrait être des syndromes se distinguant les uns des autres de manière essentielle dans leurs processus pathologiques.

L'objectif de la présente recherche n'est pas de se prononcer sur le fait que la psychiatrie soit ou non en crise ni de juger de la pertinence des amendements effectués au DSM.

En revanche, les avancées des dernières décennies dans le domaine de la classification statistique ainsi que dans le domaine de l'imagerie médicale nous permettent d'entrevoir de nouvelles avenues de recherche qui peuvent s'appliquer à des problématiques diverses et autrefois jugées irréductibles. En cela, la psychiatrie ne se distingue pas des autres branches de la médecine et la schizophrénie ne se distingue pas des autres syndromes médicaux. De nombreuses recherches ayant une approche quantitative aux problèmes psychiatriques ont déjà été effectuées et de nombreuses techniques de regroupement en grappes (*clustering*) ont été utilisées sur des données quantifiant la connectivité cérébrale ou les niveaux de symptômes de gens atteints. Néanmoins, l'identification de sous-types de schizophrénie se heurte à des difficultés. Une hypothèse expliquant cela est que les techniques de regroupement en grappes utilisées jusqu'à ce jour ne considèrent pas le fait que les individus atteints de troubles psychiatriques varient dans leurs processus cérébraux maladifs mais aussi dans leurs processus cérébraux normaux. Ainsi, le regroupement en grappes effectué sur des données représentant la connectivité cérébrale s'est effectué, jusqu'à date, en capturant des processus normaux mélangés à des processus pathologiques, ce qui empêche l'isolation et l'estimation correcte des effets des processus pathologiques. Nous proposons une approche nouvelle qui tente de pallier à ce défaut méthodologique.

Nous disposons de données provenant de 242 patients diagnostiqués schizophrènes ainsi que de 242 individus non-schizophrènes pour un total de 484 individus. Les données propres à un individu représentent des niveaux de connectivité entre différentes zones de son cerveau et ont été obtenues grace à l'imagerie par résonance magnétique fonctionnelle (IRMf).

Évidemment, chaque individu est unique et il y a une variabilité dans ces données. Nous formulons l'hypothèse que les données obtenues pourraient être caractérisées comme étant la manifestation d'un mélange de densités gaussiennes. Paramétriser individuellement les densités gaussiennes supposées qui composent le mélange dont seraient tirées les données observées nous permettrait d'identifier des sous-types biologiques. L'algorithme Espérance-Maximisation (EM) proposé par Dempster et collab. [1977] permet, en théorie du moins, de proposer une solution à cette problématique et d'identifier des densités gaussiennes et d'y attribuer des observations. Une de ces densités et les observations qui y sont attribuées forment une grappe (*cluster*).

Dans notre cas, nous rajoutons l'hypothèse que les données observées pour un individu schizophrène sont le résultat d'un effet attribuable à une condition maladive parmi des conditions maladives et que cet effet se superpose de manière additive à un état normal parmi des états normaux. Les données observées de la population malade étudiée sont donc modélisables comme un mélange de densités gaussiennes où chaque densité a une moyenne correspondant à la somme d'un état normal et d'un état maladif. Nous allons utiliser, avec les adaptations requises, l'algorithme EM dans l'objectif d'identifier et de paramétriser ces densités gaussiennes. Nous commencerons toutefois par une présentation de la schizophrénie, de l'IRM, des efforts antérieurs de regroupement en grappes sur des données relatives à la schizophrénie ainsi qu'une justification des hypothèses de recherche. Ceci constituera le contenu du premier chapitre. Celui-ci sera suivi d'un chapitre où les données seront décrites en plus amples détails et où le résultat de l'utilisation de méthodes de regroupement en grappes déjà existantes sur nos données sera exposé. Au chapitre 3, les fondements théoriques du modèle de mélange gaussien et de l'algorithme EM seront exposés et serviront de point de départ pour le développement théorique du nouveau modèle de mélange gaussien et de l'algorithme adapté, objet du chapitre 4. Les deux derniers chapitres concerneront des résultats de simulations servant à qualifier le comportement de l'algorithme adapté et les résultats de l'utilisation du nouvel algorithme sur nos données.

Chapitre 1

Schizophrénie, IRM et données de travail

1.1. Présentation de la schizophrénie

La schizophrénie est une maladie à l'étiologie inconnue pour laquelle il n'existe aucun test diagnostic [DSM-5, p.101]. C'est la présence de certains signes et symptômes tels que la présence de délusions (croyances fortes et totalement déraisonnables), d'hallucinations, d'un dérèglement grave de la pensée et/ou du discours ou de comportements grossièrement désorganisés qui permet le diagnostic d'un trouble dans le spectrum schizophrénique. Ces derniers symptômes sont parfois qualifiés de symptômes positifs. Il y aussi des symptômes dits négatifs qui peuvent se manifester tels que l'anhédonie (perte de la sensation du plaisir), une asociabilité ou une avolition (perte de volonté et de capacité à réaliser des activités dirigées vers un but).

Malgré le grand nombre et la variété de signes et symptômes possibles, aucun n'est pathognomonique à la schizophrénie et donc aucun ne permet automatiquement le diagnostic. En plus des symptômes majeurs mentionnés et qui sont utilisés pour diagnostiquer la maladie, de très nombreuses et variées caractéristiques symptomatiques peuvent se présenter chez un individu atteint de la maladie [DSM-5, p.101]. La schizophrénie peut donc se manifester de plusieurs manières très différentes et de fait, «les individus atteints du désordre varieront [entre eux] de manière substantielle sur la majorité des caractéristiques, puisque la schizophrénie est un syndrome clinique hétérogène» [DSM-5, p.100].

Ces éléments, la nature chronique de la maladie et l'absence de médication pouvant la guérir contribuent au fait que la schizophrénie est un phénomène complexe et aux conséquences potentiellement graves pour les gens qui en sont atteints. Il est pertinent de mentionner les

termes de Joyce et collab. [2017, p. 37] qui décrivent la recherche pour trouver des sous-types de schizophrénie comme «un domaine de recherche particulièrement difficile et pertinent [...] et [cette] maladie résistante au traitement continue d'affliger les vies des patients et de défier les scientifiques à tel point qu'elle est considérée comme une entité clinique distincte.» Ces dernières années ont été marquées par des appels à un changement paradigmatique de la recherche concernant les syndromes psychiatriques dans le but de pouvoir finalement identifier des différences cliniquement significatives basées sur des marqueurs biologiques. La méthode classique consistant à comparer des individus témoins à des individus malades ne suffit plus (voir par exemple Kapur et collab. 2012).

1.2. Imagerie par résonance magnétique fonctionnelle (IRMf)

1.2.1. Principe général de l'IRM

L'imagerie par résonance magnétique existe depuis des décennies et a été développée à l'origine pour analyser la structure interne du corps d'un patient, notamment pour identifier la présence de tumeurs. Le principe général de cette technique d'imagerie médicale est de créer un puissant champ magnétique. Un standard couramment utilisé est un champ de 3 Tesla. La présence de ce champ influe sur l'alignement des protons des atomes d'hydrogène du corps. Ceux-ci sont alors perturbés par l'émission temporaire d'ondes radios. Quand l'émission cesse, les électrons se réalignent sur le champ magnétique. D'un tissu organique à l'autre, le temps nécessaire pour que les protons se réalignent (appelé temps de relaxation) est différent et leur densité en protons d'hydrogènes varie. Cela permet donc de distinguer les tissus et donc, ultimement, d'en créer une image précise [Weishaupt et collab., 2006, p. 1-4, 11].

1.2.2. Signal BOLD et IRMf

La technique de l'IRM a été adaptée pour permettre la visualisation de l'activité de zones du cerveau. On parle alors d'imagerie par résonance magnétique fonctionnelle (IRMf). L'acquisition de l'image est basée sur le fait qu'une activation d'une zone cérébrale mène à une augmentation de l'irrigation sanguine vers cette zone ainsi qu'à une consommation d'oxygène et d'énergie par les tissus activés. Dans les secondes suivant une période d'activité,

l'augmentation de l'irrigation est telle que la concentration en oxyhémoglobine (hémoglobine transportant une molécule d'oxygène) augmente localement tandis que la concentration en désoxyhémoglobine (hémoglobine ne transportant pas de molécule d'oxygène) diminue. Cette dernière molécule a des propriétés magnétiques et ceci permet donc de traduire cette variation en signal BOLD (acronyme de *blood oxygenation level-dependant*) ou en un signal combinant le signal BOLD avec d'autres mesures de changements physiologiques ou métaboliques mesurables telles que la variation du volume sanguin, du flot sanguin ou du taux métabolique d'oxygène. L'image générée par l'interprétation de ce signal est donc un reflet indirect à basse fréquence de l'activité neuronale. Le lien entre l'activité et le signal BOLD est néanmoins fort [Bijsterbosch et collab., 2017, p. 2] et l'interprétation du signal BOLD est la technique la plus couramment utilisée pour produire une image en IRMf [Soares et collab., 2016, p. 9]. Plusieurs paramètres peuvent influer de manière importante sur le signal recueilli. Parmi ceux-ci, il y a la puissance du champ magnétique, celle-ci augmentant généralement la sensibilité et la précision du signal recueilli au prix de la présence d'un artefact d'image plus prononcé. L'intervalle de temps entre l'excitation de la couche et l'acquisition du signal est un paramètre de «l'importance la plus critique» [Soares et collab., 2016, p. 10]. La fréquence d'acquisition du signal à travers le temps, l'augmentation de la fréquence devant se faire au coût d'une certaine précision spatiale du signal et vice-versa, la direction du vecteur d'acquisition de l'information, son angle et la méthode de découpage du cerveau sont d'autres paramètres. Le signal est généralement pris sur une série de fines couches successives du haut vers le bas du cerveau. Typiquement, la taille d'un voxel (élément de volume) est d'environ 3mm^3 et l'acquisition de l'information du cerveau entier se fait à chaque 2 ou 3 secondes. Ainsi, une séance d'IRMf sur un sujet mène à l'acquisition d'une série chronologique pour chaque voxel de son cerveau.

1.2.3. Pré-traitement des données

L'interprétation du signal BOLD ne peut pas se faire directement. Il faut procéder à plusieurs étapes de prétraitement des données avant de pouvoir les analyser avec des méthodes statistiques ou autres. Il est important de procéder à la minimisation des artefacts résultant des distortions spatiales et des irrégularités de signal. Les fonctions physiologiques de l'individu subissant la routine d'IRMf telles que le fonctionnement de son coeur, le mouvement de

sa tête, sa respiration ou encore les délais associés à l'augmentation de l'irrigation sanguine dans les secondes après l'activation neuronale qui mène à la génération du signal BOLD ont aussi un effet sur le signal recueilli et cet effet doit en être éliminé autant que possible [Soares et collab., 2016, p. 12].

Le mouvement de tête du sujet est une des sources de bruit sur les données d'IRMf parmi les plus «plus problématiques» et une correction de l'effet de ces mouvements est indispensable [Bijsterbosch et collab., 2017, p. 27-29]. En effet, un mouvement de tête mène souvent à un déplacement spatial des voxels du cerveau ce qui peut mener à la disparition de ces voxels et leur substitution par des voxels mesurant une activité différente de par le fait que la proportion de chaque tissu présent dans le voxel est modifiée suite au mouvement. Celui-ci peut aussi faire en sorte que certaines parties du cerveau ne soient plus dans la même couche subissant l'effet de l'appareil, ce qui peut avoir pour conséquence que la mesure subséquente peut-être influencée par la mesure précédente ou que la régularité des prises de mesures soit perturbée. Tel que mentionné ci-haut, les fonctions physiologiques telles que la respiration et l'activité cardiaque normale peuvent elles aussi contribuer à créer des artefacts dans le signal mais sont généralement plus régulières et donc plus simples à corriger.

Plusieurs approches existent pour minimiser les phénomènes qui viennent d'être décrits [Soares et collab., 2016, p.11-12] telles qu'éliminer certains signaux ou images présentant un écart trop grand par rapport à une valeur de référence ou utiliser des procédés statistiques permettant d'estimer et d'isoler l'effet de ces phénomènes sur le signal [Bijsterbosch et collab., 2017, p. 35-43]. Il est certain qu'il faut procéder au moins au réalignement des images d'un sujet car des mouvements de tête, si petit soient-ils, vont se produire.

Une autre procédure très importante est de transformer l'image obtenue du cerveau de l'individu pour qu'elle s'ajuste à celle d'un cerveau standard de référence, tel que l'un de ceux de l'Institut et hôpital neurologique de Montréal, qui servira donc d'espace stéréotaxique commun aux individus. En outre, certaines structures cérébrales peuvent être ramenées à une représentation sur un plan en deux dimensions [Bijsterbosch et collab., 2017, p. 32]. Cette procédure est appelée «normalisation» dans certains écrits scientifiques [Soares et collab., 2016, p. 13] mais ne doit pas être confondue avec l'utilisation de la transformation de Fisher (laquelle sera mentionnée en plus amples détails au chapitre 2). Ainsi, le réalignement des

images d'un sujet est une procédure servant à corriger les données au niveau intraindividuel et la normalisation est une étape servant à régulariser les données au niveau interindividuel.

Mentionnons finalement qu'on peut procéder au lissage de signal pour éliminer les anomalies à très hautes fréquences. Ceci consiste généralement à évaluer une moyenne pondérée des signaux de voxels voisins pour convoluer le signal vers une densité normale, ce qui est sensé être approximativement la densité moyenne du signal BOLD. Néanmoins, ceci doit être fait avec précaution pour éviter d'éliminer de petits signaux locaux qui représentent des activités cérébrales significatives [Soares et collab., 2016, p. 14] [Bijsterbosch et collab., 2017, p. 30-31].

1.2.4. Analyse de la connectivité fonctionnelle entre régions cérébrales

L'acquisition répétée du signal BOLD, même à l'état de repos, suivi de son prétraitement tel que décrit ci-haut mène à la création de séries temporelles pour chaque voxel du cerveau. Une possibilité est de s'intéresser à évaluer la connectivité entre différentes parties du cerveau. Ceci se fait en comparant la similarité du signal d'un voxel (ou d'un groupe de voxels) à l'autre. La présence d'une telle similarité entre l'activité d'une région et d'une autre ne prouve pas la présence d'un lien physiologique direct mais indique probablement qu'une région envoie de l'information à une autre région. Ainsi, ce qui est appelé la connectivité fonctionnelle n'est autre qu'une corrélation temporelle observée et sans directionnalité entre les séries temporelles du signal provenant de deux parties du cerveau [Bijsterbosch et collab., 2017, p. 2,3].

Il n'est pas pratique de s'intéresser à la connectivité entre chaque paire de voxel pour plus d'une raison : d'une part, le nombre de paires est extrêmement grand, ce qui rend les manipulations et calculs complexes et mène à devoir faire beaucoup plus d'estimations que le nombre d'observations disponibles (ceci ayant des conséquences importantes sur la qualité des estimations), et d'autre part, nous savons que le cerveau est constitué d'unités fonctionnelles généralement faites d'un nombre de neurones considérablement plus grand que le nombre moyen couvert par un voxel [Bijsterbosch et collab., 2017, p. 82-84]. Il est donc naturel de réunir des voxels en zones ou groupes fonctionnels qui ne sont pas nécessairement contigus au niveau spatial. Cette parcellisation du cerveau peut-être facilitée par les connaissances découlant de la cartographie physiologique mentionnée ci-haut. Une mauvaise parcellisation

peut affecter de manière très négative les analyses et cela même si l'erreur de parcellisation n'est pas grossière, pour la simple raison que l'activité très différente et faiblement corrélée d'une structure fonctionnelle physiquement proche d'une autre peut biaiser la corrélation de l'activité de cette dernière structure avec l'activité d'une tierce structure. Il existe encore de l'incertitude sur la manière qui serait optimale de parcelliser le cerveau. Cette incertitude découle d'un certain manque de connaissances et du fait qu'elle dépend de la résolution spatiale voulue ainsi que de la partie ou de la fonction du cerveau qui est d'intérêt [Bijsterbosch et collab., 2017, p. 85].

Plusieurs méthodes sont utilisées pour procéder à la parcellisation. Celle-ci peut être réalisée à l'aide de méthodes numériques et/ou mathématiques telles que du regroupement en grappes, des techniques de décomposition linéaire ou d'évaluation du gradient spatial. Elle peut aussi être réalisée à l'aide d'atlas anatomiques. Ainsi, on peut identifier des régions d'intérêt constituées de voxels choisis (parfois appelées *seed*) qui sont définies selon des caractéristiques structurelles et anatomiques ou encore fonctionnelles. La parcellisation peut donc mener à des parcelles constituées de zones non-contiguës du cerveau [Poldrack, 2007].

Une fois les séries temporelles enregistrées pour chaque voxel et la parcellisation choisie, il y a plusieurs méthodes pour créer les séries temporelles de chaque parcelle à partir des séries temporelles des voxels composant la dite parcelle. Une méthode est d'évaluer la moyenne du signal BOLD pour tous les voxels de la parcelle. Une autre consiste à utiliser des techniques de régression. Une fois ces séries créées, on peut évaluer la corrélation de Pearson ou une autre mesure entre chaque paire de séries : c'est la mesure de connectivité entre chaque paire de parcelles du cerveau [Bijsterbosch et collab., 2017, p. 86-94]. Une méthode de recherche couramment utilisée consiste à évaluer la mesure de connectivité entre une région d'intérêt et le reste du cerveau, constituant ainsi un connectome. Cela donne une perspective restreinte de la fonction du cerveau mais répéter cette procédure pour toutes les parcelles du cerveau mène à une carte de connectivité de l'ensemble du cerveau [Bijsterbosch et collab., 2017, p. 52-54].

L'IRMf a le potentiel de dresser une bonne carte de toutes les connections du cerveau (le connectome) et, puisque cette corrélation pourrait être un biomarqueur pour différentes maladies cérébrales, peut possiblement nous aider à trouver des étiologies différentes à des

problèmes d'ordre psychiatrique [Bijsterbosch et collab., 2017, p. 7]. Dans tous les cas, l'utilisation de l'IRMf a mené jusqu'à date à de nombreux développements des connaissances sur diverses fonctions ou dysfonctions liées au cerveau. Les avantages de l'IRMf sont nombreux : la technique n'est pas invasive, les résultats sont reproductibles et la résolution de l'image est potentiellement très grande [Soares et collab., 2016, p. 1,2].

1.3. Lien entre IRMf et schizophrénie et hypothèses de recherche

Il a déjà été mentionné qu'une grande portion de la recherche effectuée jusqu'à présent dans le domaine de la psychiatrie a été réalisée en employant la méthode qui consiste à comparer, à l'aide de statistiques inférentielles classiques, un groupe constitué d'individus malades à un groupe témoin constitué d'individus sains. Ceci a mené à l'identification de plusieurs anomalies biologiques chez les gens atteints de divers troubles mentaux mais cette approche ne permet pas d'identifier des sous-groupes et donc ne permet pas d'identifier des syndromes étiologiquement distincts [Kapur et collab., 2012, p. 1]. Il est utile de citer les propos de Gates et collab. [2014, p. 1] :

> «*les études portant sur l'étiologie d'un quelconque désordre ignorent, typiquement, l'hétérogénéité potentielle qui existe dans les classifications actuelles basées sur les symptômes. Selon ces standards, les analyses sont faites comme si le groupe [d'individus diagnostiqués] et le groupe auquel il est comparé représentaient chacun [une] population homogène. [...] Une revue de [la] littérature [scientifique] révèle que ces suppositions sont rarement correctes. (...)*»

Il n'y a donc pas une bonne adéquation entre ces différences découvertes entre les groupes malades et témoins et la variété d'observations cliniques car, entre autres, il existe une hétérogénéité tant chez les individus non-malades que chez les individus malades. Il est vrai que «les méthodes de regroupement en grappes ont été utilisées de manière extensives pour stratifier tous les désordres psychiatriques» mais Marquand et collab. [2016] nous informent par contre qu'une petite variété d'algorithmes a été utilisée et que le regroupement en grappes a été fait de manière «prédominante» sur des mesures cliniques. Une des conclusions de leur revue est que le regroupement en grappes a un problème de reproductibilité s'il est effectué sur la base de symptômes. Les mesures quantitatives biologiques telles que l'IRMf ont été moins utilisées et ceci est probablement dû au fait qu'elles sont généralement complexes et

multidimensionnelles ce qui rend considérablement difficile l'utilisation de méthodes de regroupement en grappes pour les analyser (voir *Personalized Psychiatry* 2019, p. 124). Jusqu'à ce jour, il y a une faible convergence des résultats, particulièrement pour la schizophrénie et l'autisme [Marquand et collab., 2016, p. 435]. En bref, on est encore loin d'avoir identifié des sous-types clairement distincts qui reflèteraient des différences cliniquement pertinentes.

Les données résultantes de l'IRMf pourraient servir à identifier des sous-groupes. En effet, les résultats aux imageries neurologiques des individus atteints de schizophrénie présentent des différences «évidentes» par rapport aux gens non-atteints [DSM-5, p.101] [Davatzikos et collab., 2005]. Il existe des arguments probants de l'existence de liens entre la dysfonction de plusieurs réseaux cérébraux et certains symptômes schizophréniques. Plus spécifiquement, les résultats de nombreuses études sont compatibles avec une augmentation de la connectivité cérébrale entre certaines régions anatomiquement distinctes chez les individus atteints, mais cette maladie est généralement associée à une diminution de connectivité fonctionnelle dans la plupart des réseaux cérébraux [Pettersson-Yeo et collab., 2011, p. 1116-1118, 1120] [Brandl et collab., 2019] [Dong et collab., 2017], à tel point que la schizophrénie est considérée comme un syndrome de dysconnectivité cérébrale fonctionnelle [Friston et Frith, 1995]. Les symptômes psychotiques semblent corrélés avec une augmentation de la connectivité entre différentes dimensions du connectome, c'est-à-dire une augmentation de la connectivité entre différents réseaux du cerveau tandis que certains symptômes sont associés à une diminution de connectivité pour certaines dimensions du connectome [Wang et collab., 2018, p. 8]. Des études de connectivité fonctionnelle montrent des distinctions importantes entre les circuits corticaux responsables de différents symptômes de diverses maladies mentales (notamment chez des individus atteints de schizophrénie) et montrent que le fonctionnement de ces circuits pourrait varier selon l'étiologie de la maladie [Wang et collab., 2018, p .1]. Il y a de plus des associations identifiées entre certains gènes de susceptibilité et des manifestations symptomatiques de la schizophrénie [Clementz et collab., 2015, p. 1]. Des études ont aussi montré que les patients schizophréniques produisent des signaux BOLD anormaux du cortex préfrontal ainsi qu'une plus grande variabilité dans l'activation de plusieurs régions [Huettel et collab., 2004, p. 456]. En résumé, la schizophrénie est liée à de très diverses anomalies de la structure du cerveau et de la connectivité de certaines régions cérébrales et ces anomalies sont parfois liées à certaines catégories de symptômes. Chez les individus atteints

de la maladie, plusieurs anomalies structurelles sont généralement plus prononcées chez les hommes que chez les femmes [Voineskos et collab., 2019]. Tout ceci contribue à l'hétérogénité des patients atteints de schizophrénie et/ou de troubles psychiatriques. L'existence de cette hétérogénité est un fait établi dans une panoplie d'études qui ont de plus indiqué l'existence de sous-types du syndrome. Ces études renforcent donc l'hypothèse qu'un unique diagnostic clinique peut résulter d'étiologies différentes et ouvrent la voie à une meilleure classification médicale des troubles dans différents spectrums de maladie [1]. Des stratégies visant à mettre à profit cette hétérogénéité peuvent aider à découvrir des différences neurobiologiques et à identifier des sous-types de patients [Voineskos et collab., 2019].

L'IRMf offre de nombreuses possibilités quant à l'étude du cerveau mais il n'existe pas un standard rigoureux pour définir ce qui est normal en terme de connectivité ou d'activation cérébrale. Les individus sains sont tous uniques et beaucoup d'hétérogénéité existe chez eux [2]. D'un individu sain à l'autre, l'architecture cérébrale varie. Cette variabilité est plus prononcée d'une partie du cortex à l'autre et l'ampleur de l'hétérogénéité pour une partie du cerveau est liée à la complexité des fonctions remplies par cette partie. Cette variabilité reflète de grandes différences génétiques ainsi que la grande diversité des environnements influant sur le développement du cerveau de chaque individu. Tout cela résulte en une grande variabilité entre individus sains dans les données d'IRMf [Mueller et collab., 2013]. Il y a donc une superposition de processus cérébraux normaux divers ainsi que de processus maladifs divers chez les individus schizophrènes.

Notre recherche est fondée sur quatre points fondamentaux établis avec certitude par les récents développements scientifiques. D'abord, une nouvelle classification des problèmes psychiatriques, et notamment de la schizophrénie, est nécessaire ou du moins souhaitable pour prendre en considération les différences étiologiques et potentiellement mener à de meilleurs traitements. D'autre part, les individus atteints de troubles psychiatriques dont la schizophrénie présentent une hétérogénéité entre eux. Le même phénomène est établi pour les individus sains. Finalement, les données d'IRMf pourraient aider à identifier des patrons de connectivité associés à des sous-types normaux et maladifs.

1. Voir par exemple Dollfus et collab. 1996, Clementz et collab. 2015, Drysdale et collab. 2017, Gates et collab. 2014, Sponheim et collab. 2001.

2. Voir Gates et collab. 2014, p. 1 et Feczko et collab., 2019, Van Dam et collab. [2017] et Miller et Van Horn [2002] [2007].

Nous remarquons que toutes les tentatives de regroupement en grappes sur des données relatives à des individus atteints de troubles psychiatriques réalisées à ce jour ne tiennent pas compte du fait qu'il y a une variabilité chez les individus normaux et il est raisonnable de penser que cette variabilité pourrait influer sur les données relatives aux gens atteints d'un trouble. Il est donc amplement justifié d'explorer un modèle qui suppose l'existence d'effets maladifs divers en plus de distinctions dans les types d'activité cérébrale normale et qui considère que les données résultantes de l'IRMf sur un individu malade peuvent être modélisées comme la superposition d'un des effets maladifs sur un type d'activité cérébrale normale. Notre approche prend donc en considération les deux variabilités (normales et maladives) et le regroupement en grappes sur les données des individus malades sera donc effectué sur la base d'une ressemblance au niveau de l'effet de maladie. Mentionnons le travail qui semble unique de Dong et collab. [2016] qui ont noté les problèmes méthodologiques qui résultent du fait d'ignorer l'hétérogénéité chez les gens atteints de certains syndromes tels que la schizophrénie, la maladie d'Alzheimer ou l'autisme et/ou de procéder à un regroupement en grappes directement sur les images obtenues de l'IRMf. Ceux-ci ont proposé un algorithme qui considère l'effet de maladie comme un effet transitif ou une transformation à partir du stade normal. Il existe toutefois des différences entre cette dernière approche et ce que nous proposons ici et à notre connaissance, notre approche est nouvelle et n'a pas été été essayée jusqu'à ce jour.

Chapitre 2

Étude exploratoire des données et utilisation de méthodes de regroupement en grappes déjà existantes

2.1. Défis du regroupement en grappes

Il a été mentionné au chapitre précédent que les méthodes de regroupement en grappes ont été utilisées dans de nombreuses recherches visant à identifier des sous-groupes de divers troubles psychiatriques. Toutefois, ces algorithmes ont souvent été utilisés sur des mesures psychométriques et non sur des données décrivant une activité biologique. Nos données de travail sont des données résultant de sessions d'IRMf sur divers individus. Il est donc intéressant de tenter de regrouper en grappes les données dont nous disposons avec des algorithmes existant déjà. De nombreuses difficultés peuvent être rencontrées lorsqu'il y a tentative de regrouper en grappes des données biologiques en groupes significativement différents[1]. Il est important d'être conscient de ces difficultés pour analyser justement les résultats de ces regroupements.

2.1.1. Bruit

Il est possible que la variabilité attribuable au bruit de mesure soit bien supérieure à la variabilité attribuable aux sous-types se trouvant dans les données [Marquand et collab., 2016, p. 437]. Dans ces cas-ci, plusieurs méthodes de regroupement en grappes pourraient être incapables de livrer une solution représentant les différences significatives pour la simple raison que l'expression de ces différences serait masquée par un bruit aléatoire au point de fausser les estimations nécessaires au regroupement en grappes.

1. Dans ce contexte, la significativité peut être statistique et/ou biologique et/ou clinique.

Il a déjà été mentionné que les mouvements de tête peuvent causer des nuisances difficiles à éliminer sur les données résultant de l'IRMf. C'est un problème qui peut se trouver dans nos données d'autant plus que celles-ci proviennent de sites de recherche différents. On peut donc s'attendre à ce que l'effet des mouvements de tête soit différent selon le protocole suivi (donc selon le site) et selon le fait que le sujet soit atteint de schizophrénie ou non.

2.1.2. Validation des résultats et reproductibilité

Il peut être difficile de savoir si les grappes obtenues représentent des sous-types biologiquement ou cliniquement significatifs. L'objectif pratique est souvent d'améliorer la connaissance des phénomènes biologiques causant la maladie et de produire de nouvelles approches thérapeutiques adaptées à ces phénomènes. Des méthodes de validation des résultats ont été proposées. Il peut être utile d'appliquer le modèle à un autre ensemble de données ou de mesurer la reproductibilité des résultats de l'algorithme sur les mêmes données. L'identification de correspondances avec d'autres mesures biologiques ou cliniques a aussi été proposé comme méthode de validation des résultats [Marquand et collab., 2016, p. 438].

2.1.3. Identification du nombre de grappes

Plusieurs algorithmes de regroupement en grappes diviseront les données qu'elles doivent diviser même si les différences entre les groupes ne sont pas significatives. Plusieurs algorithmes fonctionnent avec la condition préalable au fonctionnement qu'un nombre de grappes soit fixé. Il en résulte qu'il est souvent nécessaire d'identifier un nombre optimal ou adéquat de grappes réellement contenues dans les données, mais c'est une identification qui peut être impossible à réaliser avec certitude.

2.1.4. Grande dimensionnalité

Les données biologiques sont souvent représentables avec des données numériques de très grandes dimensions. Une règle générale proposée auparavant est de considérer 10 dimensions comme un nombre élevé [Ronan et collab., 2016, p. 2]. Ceci mène au fait que des résultats très différents mais tout aussi valides peuvent être obtenus selon la méthode de regroupement en grappes utilisée ou selon les variables considérées pour le regroupement. Ceci est d'autant plus problématique que les populations humaines peuvent exhiber énormément de variabilité sur de très nombreux aspects, ce qui se reflète souvent comme une grande variabilité dans des

données à grande dimensionnalité. Si le nombre de dimensions est élevé, il est probable que deux individus similaires diffèrent largement sur un petit nombre de dimensions, ce qui peut mener un algorithme basé sur une mesure de distance à considérer des individus qui sont en réalité relativement proches comme éloignés les uns des autres et vice-versa. La méthode K-moyennes et d'autres sont particulièrement mal adaptées aux données de dimensionnalité élevée [Ronan et collab., 2016, p. 2]. De manière plus générale, certaines dimensions peuvent être disproportionnellement influentes dans la détermination de la solution de l'algorithme de regroupement en grappes. Dans certains cas, il peut être justifié de considérer des méthodes de réduction de la dimensionnalité, notamment l'analyse en composantes principales. L'IRMf permet en théorie d'obtenir une image de l'entièreté du cerveau avec une résolution de l'ordre de 1mm^3, mais le cerveau est probablement la structure biologique la plus complexe qui soit connue. Le problème de dimensionnalité peut donc être très important.

2.1.5. Approche possible : regroupement en grappes sur les résultats d'un ensemble de regroupements en grappes

Il a été proposé d'utiliser singulièrement différentes méthodes de regroupement en grappes sur les mêmes données puis de procéder à un autre regroupement sur un score reflétant la similarité entre individus selon les différentes méthodes de regroupement. L'hypothèse justifiant cette approche est que puisque les différents algorithmes de regroupement fonctionnent de manières fondamentalement distinctes et peuvent donc tous mener à des résultats erronés de manières différentes, il est probable qu'une erreur ponctuelle de regroupement ne se reproduise pas d'un algorithme à l'autre et donc qu'elle ne soit pas corrélée avec les erreurs résultantes des autres algorithmes. Diviser en grappes sur le résultat d'un ensemble de méthode aurait donc pour effet d'atténuer les erreurs ponctuelles [Ronan et collab., 2016, p. 8-9].

2.2. Données de travail

Nous disposons des données de $N_c = 242$ sujets dans le groupe témoin (X) et de $N_s = 242$ sujets dans le groupe maladie (Y). Chaque individu dans ce dernier groupe est diagnostiqué comme atteint de schizophrénie selon les critères du DSM-IV ou du DSM-5. Les données des individus proviennent de 10 sites de recherche différents. Les symptômes de chaque individu

ont été évalués selon l'échelle PANSS ou l'échelle SAPS/SANS. Les résultats des évaluations réalisées selon la première échelle ont finalement été convertis vers l'échelle SAPS/SANS. [2] Le nombre total d'invididus provenant de chaque site varie de 18 à 84. De ce nombre, il y a toujours une proportion équivalente d'individus non-malades et d'individus malades de manière à minimiser les différences entre les malades et les non-malades en ce qui concerne la distribution du sexe, la distribution de l'âge et la mesure de déplacement (fd) lors des séances d'IRMf. Les protocoles suivis pour recueillir les données varient de manière substantielle entre les sites. En effet, la récolte des données d'IRMf s'est faite, selon le site, au repos ou pendant la réalisation de tâches très variées et le nombre obtenu de volumes d'IRMf varie de 96 à 273 pour des séances ayant duré de 3 à 13 minutes. Le tableau 2.1 contient des statistiques descriptives des sujets dont sont tirées nos données de travail selon le site de provenance.

Le connectome est composé de b corrélations entre les q parcelles du cerveau, c'est-à-dire les corrélations entre les séries temporelles moyennes (pour chaque parcelle) résultantes de l'IRMf. L'utilisation des corrélations est justifiée par le fait que c'est une mesure populaire et assez précise [Smith et collab., 2011]. Puisque nous ne considérons pas la corrélation entre une parcelle et elle-même, $b = (q^2 - q)/2 = q(q-1)/2$. L'ensemble des données peuvent être schématiquement représentées comme l'agrégation de X et Y :

$$X = \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,b} \\ x_{2,1} & x_{2,2} & \dots & x_{2,b} \\ \vdots & \ddots & \ddots & \vdots \\ x_{N_c,1} & \dots & \dots & x_{N_c,b} \end{bmatrix}$$

$$Y = \begin{bmatrix} y_{1,1} & y_{1,2} & \dots & y_{1,b} \\ y_{2,1} & y_{2,2} & \dots & y_{2,b} \\ \vdots & \ddots & \ddots & \vdots \\ y_{N_s,1} & \dots & \dots & y_{N_s,b} \end{bmatrix}$$

Chaque ligne de ces matrices est donc la vectorisation de la matrice du connectome d'un individu.

2. Voir Chang et collab. [2019] et Orban [2019]

Site	Nombres de sujet (proportion du total %)	Pourcentage d'hommes (%)	Âge moyen en années (écart-type)	Valeur fd moyenne (écart-type)	Nombre moyen de volumes (écart-type)
1	84 (17,36)	90,48	33,57 (11,89)	0,22 (0,05)	115,39 (27,65)
2	82 (16,94)	86,59	33,35 (8,89)	0,14 (0,052)	187,76 (28,69)
3	62 (12,81)	61,29	32,82 (7,74)	0,15 (0,05)	118,61 (20,06)
4	50 (10,33)	66,00	31,7 (10,23)	0,15 (0,04)	144,24 (31,98)
5	34 (7,02)	76,47	36,74 (9,98)	0,14 (0,029)	89,35 (7,86)
6	70 (14,46)	55,71	30,37 (7,81)	0,15 (0,044)	221,36 (44,42)
7	28 (5,79)	71,43	37 (8,33)	0,13 (0,045)	184,04 (23,85)
8	28 (5,79)	64,29	31,07 (7,04)	0,12 (0,039)	178,04 (8,83)
9	28 (5,79)	85,71	31,61 (9,04)	0,09 (0,038)	208,57 (4,19)
10	18 (3,72)	83,33	34,06 (9,89)	0,11 (0,04)	157,5 (9,43)
Tous les sites	484 (100)	74,38	32,96 (9,45)	0,15 (0,056)	159,09 (50,52)

Tableau 2.1. Statistiques décrivant les sujets dont sont tirées les données selon le site de provenance

2.2.1. Prétraitement des données

2.2.1.1. *Parcellisation*

La parcellisation du cerveau est réalisée selon l'atlas *Multiresolution Intrinsic Segmentation Template* (MIST) [Urchs et collab., 2017] qui offre une possibilité de parcellisation du cerveau sur une base fonctionnelle pour $q = 7$ ce qui fait donc que $b = 21$.

2.2.1.2. *Transformée de Fisher*

Les données sont des corrélations. Pour faciliter les analyses et manipulations, la tranformée de Fisher leur est appliquée. Cette transformée est définie ainsi :

$$r \to f(r) = \tfrac{1}{2}\log(\tfrac{1+r}{1-r}).$$

Les corrélations transformées peuvent être modélisées comme suivant une distribution approximativement normale même si les données à partir desquelles sont obtenues les corrélations ne le sont pas tout à fait [Fisher, 1921].

2.2.1.3. *Régression linéaire pour estimer certains effets et les éliminer*

Comme mentionné auparavant, nos données proviennent de 10 sites différents et les protocoles pour les prises des mesures d'IRMf varient de manière importante entre les sites.

Les individus sont aussi d'âges variables et des deux sexes. Ces variables pourraient avoir un effet sur les données. Nous procédons donc à l'établissement d'un modèle linéaire général expliquant le vecteur des données d'un individu x_i comme étant

$$x_i = \beta_0 + \beta_1 sz_i + \beta_2 \hat{a}ge_i + \beta_3 sexe_i + \beta_4 fd_i + \beta_{site_i} + \epsilon_i$$

où β_0, β_1, β_2 et β_3 sont respectivement une constante, les coefficients de l'effet d'un diagnostic de schizophrénie, de l'âge et du fait d'être de sexe masculin. La mesure fd_i (pour *frame displacement*) est une mesure du mouvement de l'individu x_i pendant la session d'IRMf sur les 3 axes possibles de déplacement de la tête soit les axes horizontal, vertical et rotationnel et β_4 est le coefficient qui y est associé. Il est à noter que dans les données à notre disposition, si la mesure fd dépassait 0,5 sur un volume, alors ce volume a été retiré ainsi que le volume le précédant et deux volumes le suivant. La variable ϵ_i désigne une erreur. La variable β_{site_i} désigne l'effet de site dont est tiré l'individu x_i. Le dixième site a été arbitrairement choisi comme site de référence ; l'effet de chacun des autres sites sera estimé par rapport à ce site. Les estimations des effets des variables $\hat{a}ge$, $sexe$, fd et $site$ sont soustraites aux données de chaque individu. L'objectif est de ne réaliser le regroupement en grappes et les manipulations subséquentes que sur des données représentant l'effet de maladie et la variabilité des individus, donc de ne laisser dans les données que ce qui est sensé être représenté par sz et ϵ. Si cette étape n'est pas réalisée, il y a une possibilité que le regroupement en grappes soit guidé par ces effets car ceux-ci peuvent être les plus grands contributeurs à la variabilité des données [Dong et collab., 2016]. Dans notre cas, nous avons remarqué que si nous effectuons un regroupement sur les données sans tenter d'éliminer les effets attribuables au site de provenance, alors c'est principalement ces effets qui guideront la division en sous-groupes (c'est-à-dire que la distribution des sites de provenance des individus constituant une grappe sera très différente d'une grappe à l'autre). Ceci est compatible avec le fait que nos données proviennent de différents sites et que certains sites n'ont pas récolté les mesures d'IRMf lors du repos des sujets mais plutôt lors de l'accomplissement de certaines tâches ce qui a donc des effets potentiellement drastiques sur les données.

La figure 2.1 illustre les 21×14 valeurs des coefficients résultant de cette régression multiple pour les 21 dimensions et les 14 coefficients du modèle. On note que les variables $\hat{a}ge$ et $sexe$ ne semblent avoir aucun impact ou un très faible impact et cela sur l'entièreté des dimensions des données, contrairement aux effets attribuables au site de provenance et

fd qui sont parfois significativement différents de 0 sur certaines dimensions. Un diagnostic de schizophrénie a un faible effet statistiquement significatif (valeur-p $< 0{,}05$) sur la majorité des dimensions du connectome.

La figure 2.2 illustre les valeurs des 484×21 résidus de la régression selon le site. Ceux-ci semblent distribués symétriquement autour de 0 pour tous les sites. De plus, il ne semble y avoir que peu de résidus qui sont des données aberrantes.

La figure 2.3 illustre les distances de Mahalanobis des données comparativement aux quantiles d'une distribution χ^2. Il est évident que l'ensemble des données n'est pas distribué selon une unique densité multivariée normale et ceci est donc compatible avec la possibilité que les données soient modélisables comme un mélange de densités gaussiennes multivariées.

Une autre approche possible pour tenter d'éliminer les effets pouvant influencer le regroupement en grappes est de séparer les données en 10 groupes selon le site de provenance puis de faire une régression linéaire pour les données de chaque groupe de manière séparée. Le but serait d'estimer l'effet des variables $\hat{a}ge$, $sexe$, fd et le coefficient β_0 qui serait sensé contenir l'effet attribuable au site puis de soustraire ces effets estimés aux données d'origine.

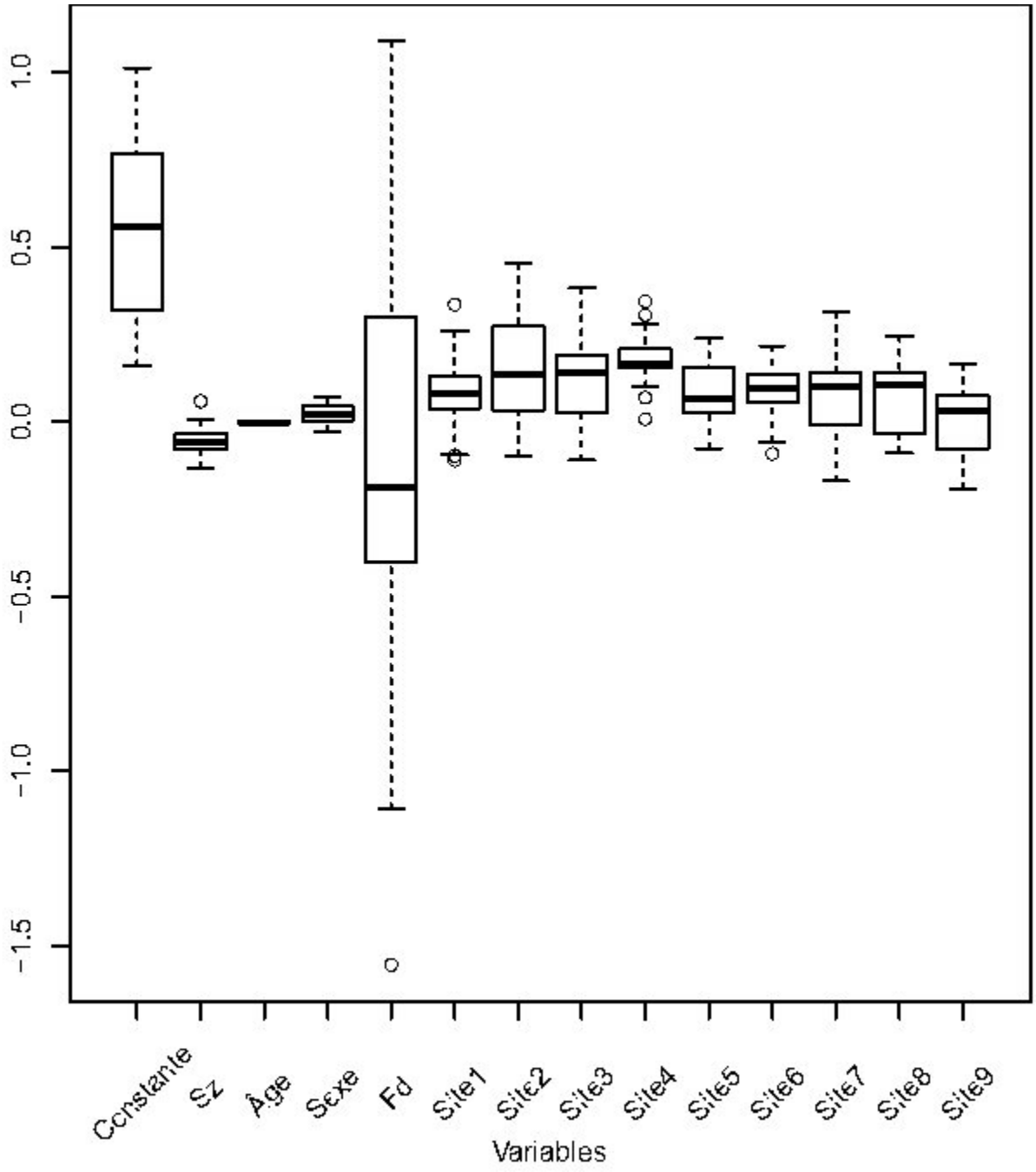
Toutes les manipulations relatives aux régressions ont été effectuées avec le logiciel R 3.6 [R Core Team, 2017].

2.3. Méthodes de regroupement en grappes utilisées

2.3.1. Algorithme K-moyennes

L'algorithme K-moyennes est un algorithme simple qui consiste à initialiser K points, chaque point étant sensé représenter la moyenne d'une grappe (parfois appelée «centroïde»). Chaque point des données à diviser est alors assigné au centroïde le plus proche selon une mesure de distance (euclidienne par exemple). La moyenne des points sert alors à ré-évaluer la moyenne de chaque grappe et le processus est ainsi répété jusqu'à ce qu'un critère de convergence soit vérifié. Les résultats de l'algorithme dépendent évidemment de la distance évaluée entre la moyenne de la grappe et les points assignés, de l'initialisation et du nombre K de grappes choisi par l'utilisateur [MacQueen et collab., 1967].

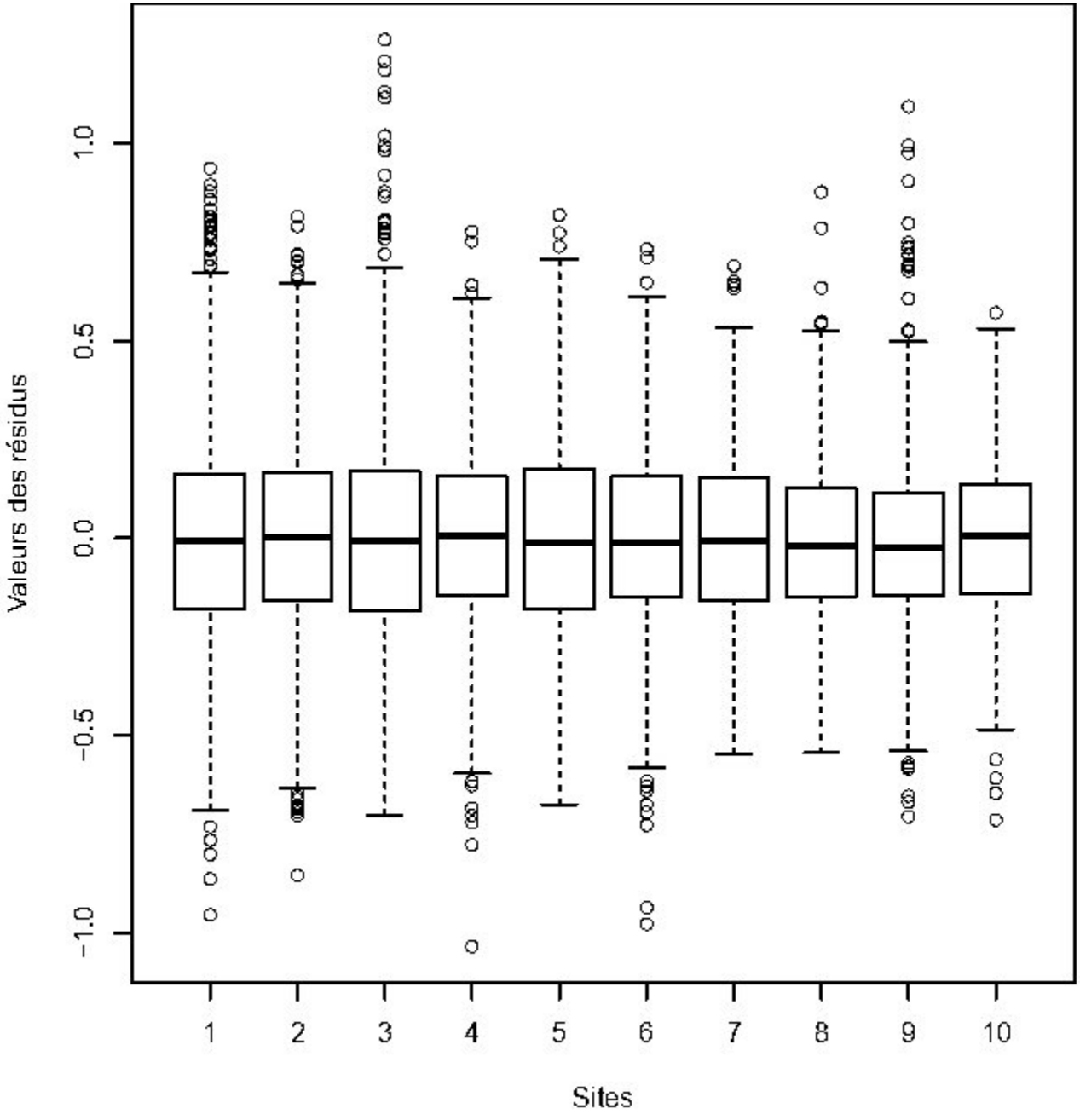
Figure 2.1. Coefficients de régression pour chaque variable explicative en incluant les résultats pour tous les sites



2.3.2. Regroupement en grappes hiérarchique

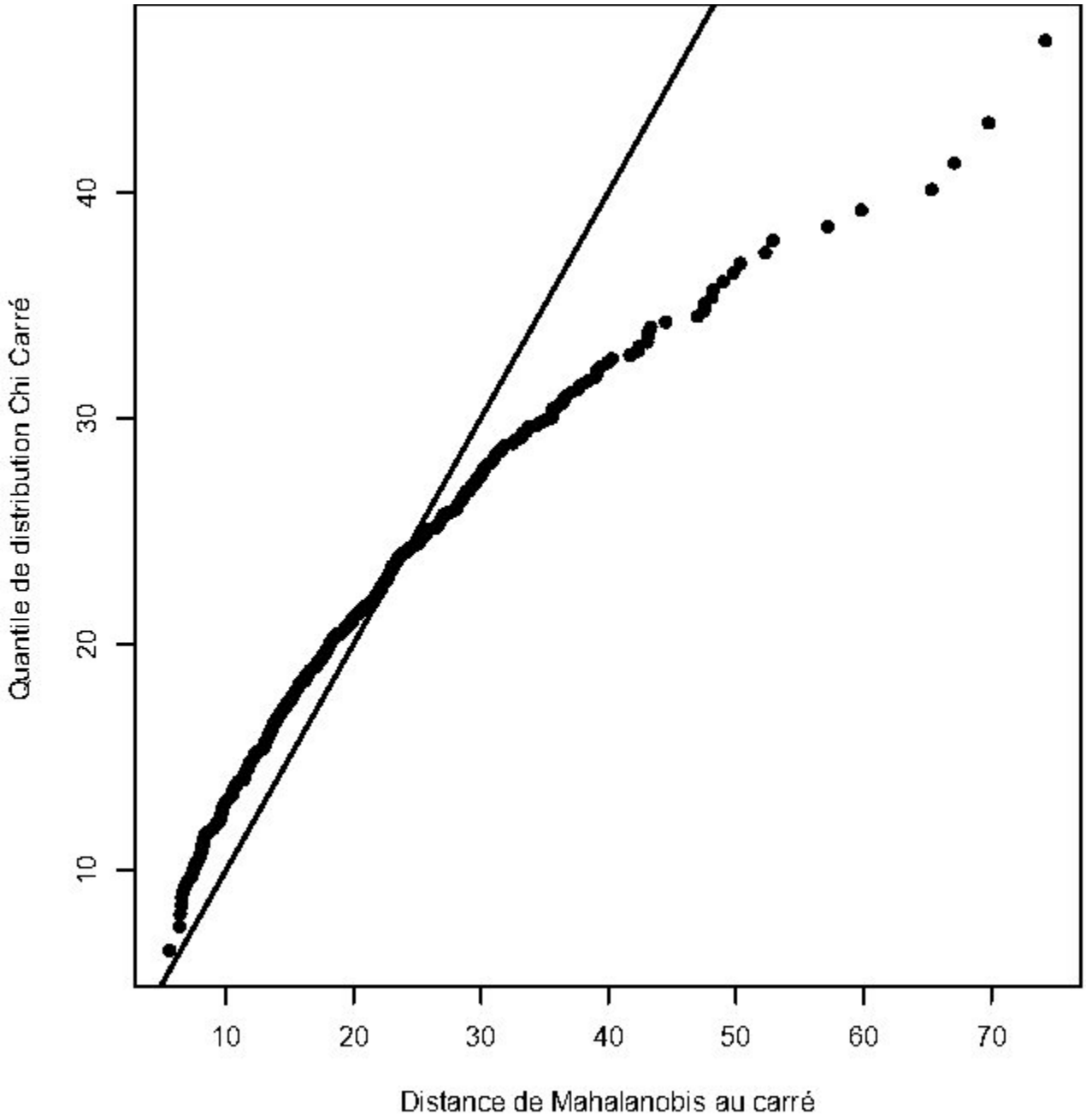
Il s'agit d'un algorithme itératif conceptuellement simple. À chaque itération, les deux individus ou grappes ($C1$ et $C2$ dont les moyennes sont respectivement m_{C_1} et m_{C_2}) les plus

FIGURE 2.2. Résidus de la régression pour chaque site



proches sont joints pour former une plus grande grappe (notée $C_1 \cup C_2$ dont la moyenne est $m_{C_1 \cup C_2}$). L'algorithme se poursuit jusqu'à ce qu'un critère d'arrêt soit vérifié ou jusqu'à ce que tous les individus se trouvent dans la même grappe. Nous avons utilisé la distance de Ward comme critère de proximité entre les grappes. Cette méthode vise à minimiser l'augmentation des carrés des distances. Soit :

Figure 2.3. Distances de Mahalanobis des données comparées aux quantiles d'une distribution χ^2 à 21 degrés de liberté



$$\Delta(C_1,C_2) = \sum_{i\in C_1\cup C_2} \|x_i - m_{C_1\cup C_2}\|^2 - \sum_{i\in C_1} \|x_i - m_{C_1}\|^2 - \sum_{i\in C_2} \|x_i - m_{C_2}\|^2. \qquad (2.3.1)$$

À chaque itération, l'algorithme va joindre dans une seule grappe la paire $C1$ et $C2$ qui va minimiser $\Delta(C1,C2)$ [Ward Jr, 1963].

2.3.3. Algorithme Espérance-Maximisation pour estimation des paramètres d'un mélange de densités gaussiennes

Une description en détails des fondements mathématiques et statistiques de ce modèle et de l'algorithme est fourni au prochain chapitre et sert de base pour les développements subséquents.

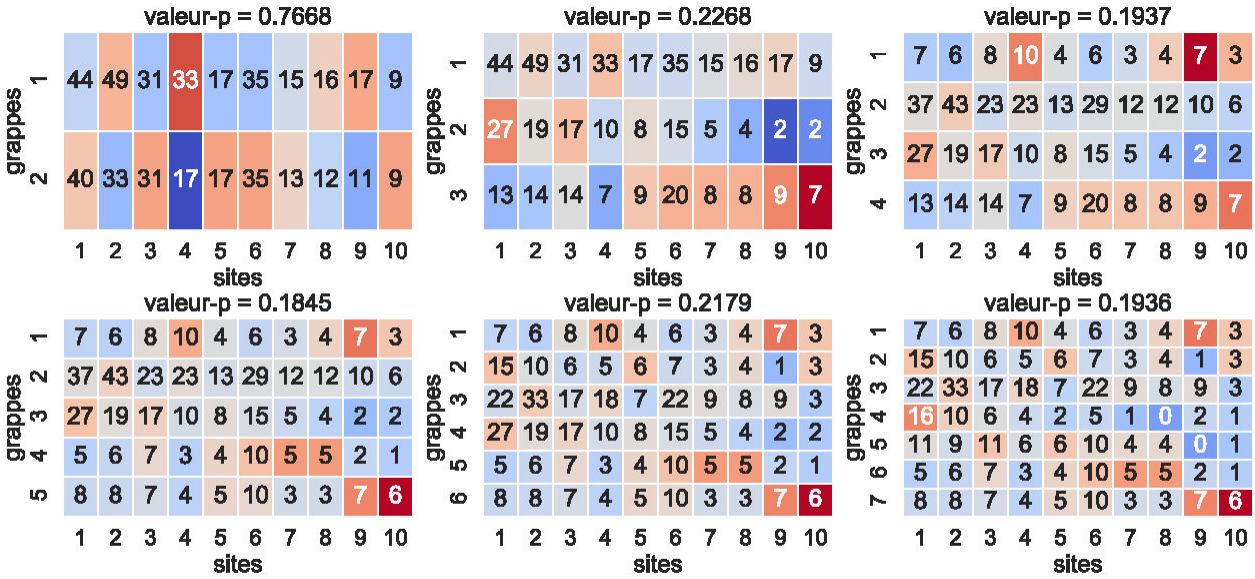
2.4. Résultats

Nous avons effectué des regroupements en grappes hiérarchiques en utilisant la distance de Ward, avec l'algorithme EM pour l'estimation des paramètres des densités composant un mélange gaussien et avec la méthode K-moyennes. Toutes les données à notre disposition ont été incluses pour réaliser ces regroupements. Chaque méthode de regroupement en grappes a été utilisée pour obtenir tous les nombres de grappes (K) compris inclusivement entre $K = 2$ et $K = 7$. Il y a donc eu un total de 18 regroupements en grappes effectués sur les données, ce qui correspond au produit du nombre de méthodes employées (3) avec le nombre de valeurs possibles pour K (6). Nous avons calculé un score de cooccurence allant de 0 à 18 pour chaque paire d'individus, 1 point étant alloué à chaque fois que deux individus se retrouvaient dans la même grappe suite à un regroupement en grappes parmi les 18 regroupements réalisés. Ce score de cooccurence a été utilisé pour faire un regroupement en grappes hiérarchique en utilisant la distance de Ward. Toutes ces étapes ont été réalisées avec scikit-learn v. 0.20 [Pedregosa et collab., 2011] et SciPy 1.2.1 [Jones et collab., 2001–2019] sur le language de programmation Python 3.6.

2.4.1. Distribution des sites de provenance dans les grappes

Nous avons effectué des tests du χ^2 pour tester l'indépendance du site de provenance d'un individu et le fait qu'il soit attribué à une grappe. Au seuil de 5% et pour toutes les valeurs possibles de K, il n'y a pas de raisons suffisantes de conclure qu'il y a une dépendance entre ces deux variables (voir la figure 2.4). Il semble donc que les effets attribuables au site de provenance n'ont pas d'impact significatif sur les résultats des regroupements en grappes.

FIGURE 2.4. Tests du χ^2 pour tester l'indépendance du site de provenance et l'attribution à une grappe, $K = 2$ jusqu'à $K = 7$



Note : Le bleu et le rouge représentent respectivement la négativité et la positivité d'une différence relative entre l'observation et la valeur attendue.
Une couleur plus intense représente une valeur absolue plus élevée de cette différence relative.

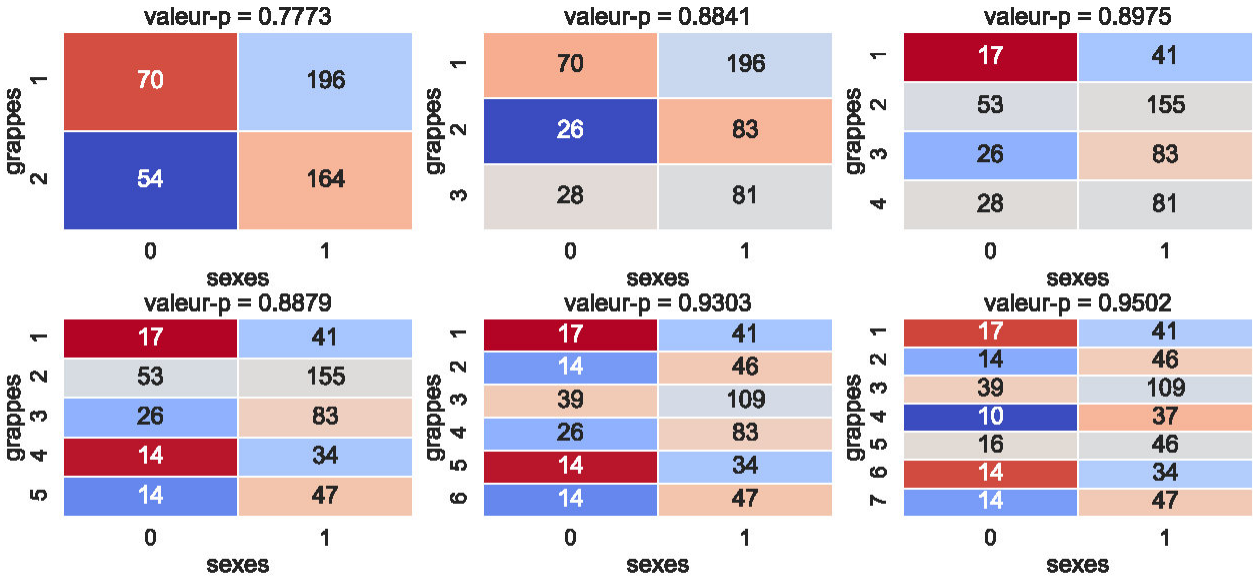
2.4.2. Distribution du sexe dans les grappes

Nous avons effectué des tests du χ^2 pour tester l'indépendance du sexe d'un individu et le fait qu'il soit attribué à une grappe. Au seuil de 5% et pour toutes les valeurs possibles de K, il n'y a pas de raisons suffisantes de conclure qu'il y a une dépendance entre ces deux variables (voir figure 2.5). Il semble donc que les effets attribuables à la variable *sexe* n'ont pas d'impact significatif sur les résultats des regroupements en grappes.

2.4.3. Mesure fd dans les grappes

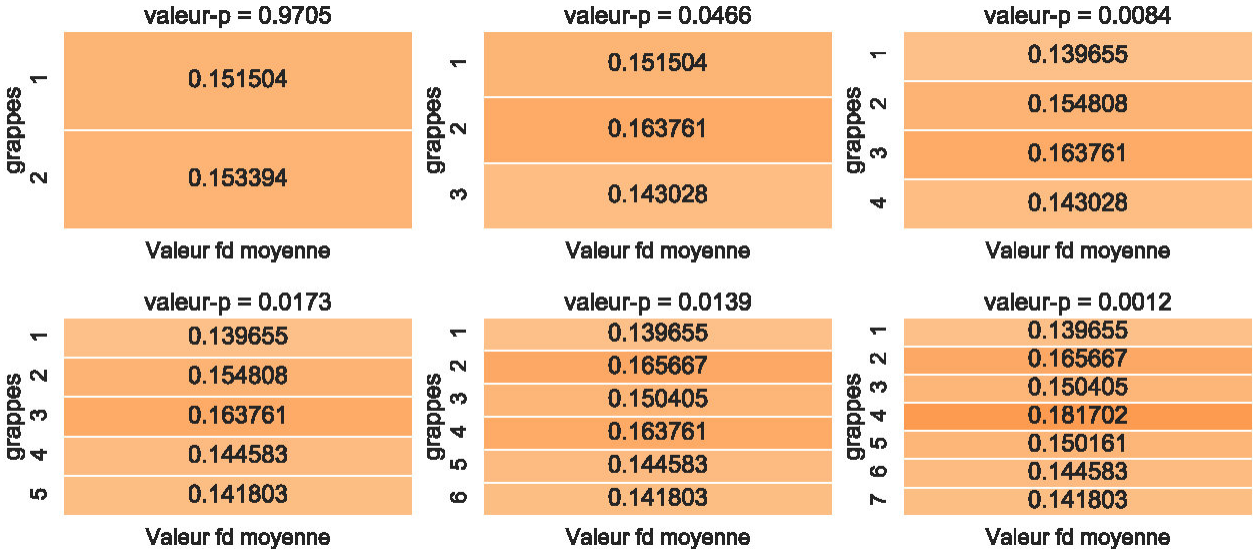
Pour tester l'égalité de la mesure fd moyenne à travers les grappes, nous avons pensé à utiliser une analyse de la variance (ANOVA). Le test de Levene nous permet de travailler avec l'hypothèse que les variances sont homogènes entre les grappes. Toutefois, le test de Shapiro-Wilk montre, au seuil de 5%, qu'il faudrait rejeter l'hypothèse de normalité des résidus. Nous avons donc décidé d'utiliser le test non-paramétrique de Kruskal-Wallis pour tester l'égalité de la médiane de la mesure fd à travers les grappes. Au niveau de 5%, nous pouvons conclure que la mesure de mouvement fd est inégale à travers les grappes (avec une exception lorsque $K = 2$). Il semble que la soustraction de l'effet estimé par la régression

FIGURE 2.5. Tests du χ^2 pour tester l'indépendance du sexe et l'attribution à une grappe, $K = 2$ jusqu'à $K = 7$



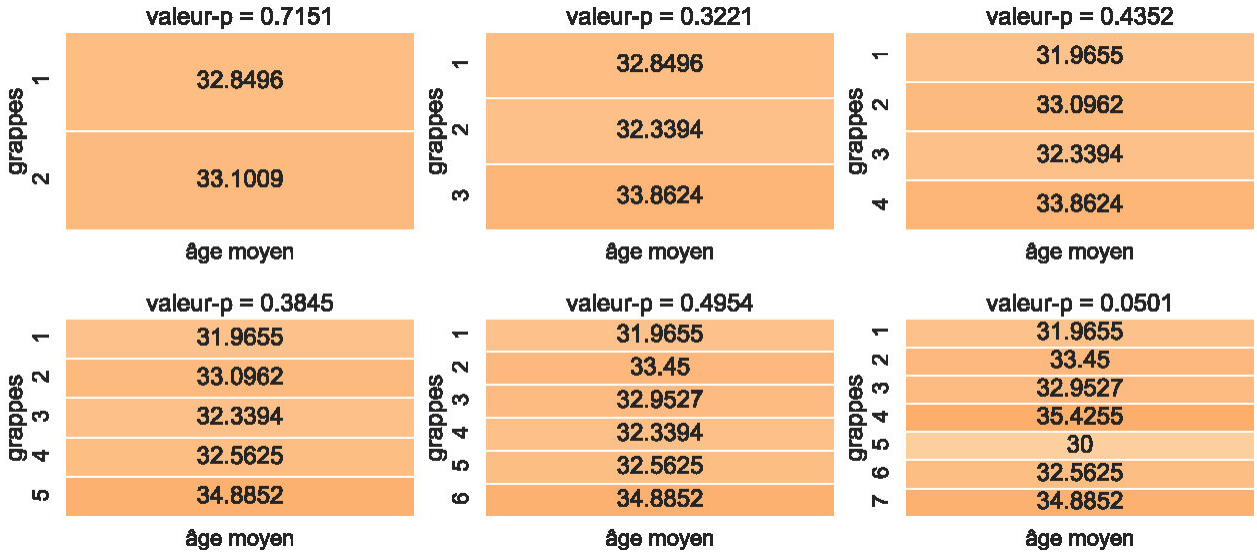
Note : Le bleu et le rouge représentent respectivement la négativité et la positivité d'une différence relative entre l'observation et la valeur attendue.
Une couleur plus intense représente une valeur absolue plus élevée de cette différence relative.

FIGURE 2.6. Tests de Kruskal-Wallis pour tester l'égalité de la médiane de la mesure fd à travers les grappes, $K = 2$ jusqu'à $K = 7$



linéaire de la variable fd n'ait pas été suffisante pour éliminer tous les effets attribuables à cette variable (voir figure 2.6).

Figure 2.7. Tests de Kruskal-Wallis pour tester l'égalité de l'âge médian à travers les grappes, $K = 2$ jusqu'à $K = 7$



2.4.4. Âges dans les grappes

Pour les mêmes raisons que pour la mesure fd moyenne, nous avons décidé d'utiliser le test non-paramétrique de Kruskal-Wallis pour comparer l'âge médian entre les grappes. Au seuil de 5% et pour toutes les valeurs possibles de K, il n'y a pas de raisons suffisantes de conclure que l'âge des individus est différent d'une grappe à l'autre (voir figure 2.7).

2.4.5. Connectivité moyenne à travers les grappes

Pour les mêmes raisons que pour la mesure fd moyenne ainsi que pour l'âge moyen, nous avons opté pour le test non-paramétrique de Kruskal-Wallis pour comparer, entre les grappes, la médiane de la connectivité moyenne des individus. La connectivité moyenne d'un individu correspond à la moyenne des données pour cet individu (donc la moyenne d'une ligne de X ou Y). Au niveau de 5% et pour toutes les valeurs possibles de K, nous pouvons conclure que la connectivité moyenne est inégale à travers les grappes (voir figure 2.8).

2.4.6. Nombres d'individus malades dans les grappes

Nous avons effectué des tests du χ^2 pour tester l'indépendance du fait qu'un individu soit atteint de la maladie et le fait qu'il soit attribué à une grappe. Au seuil de 5%, nous pouvons conclure qu'il y a une dépendance entre ces deux variables (voir figure 2.9). Il semble

FIGURE 2.8. Tests de Kruskal-Wallis pour tester l'égalité de la médiane de la connectivité moyenne à travers les grappes, $K = 2$ jusqu'à $K = 7$

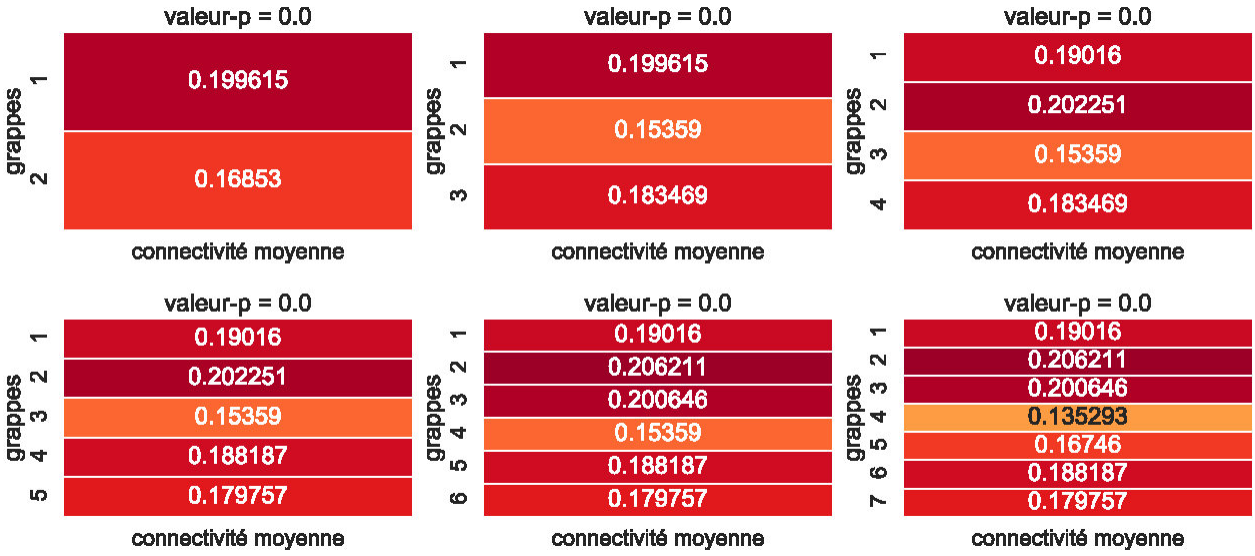
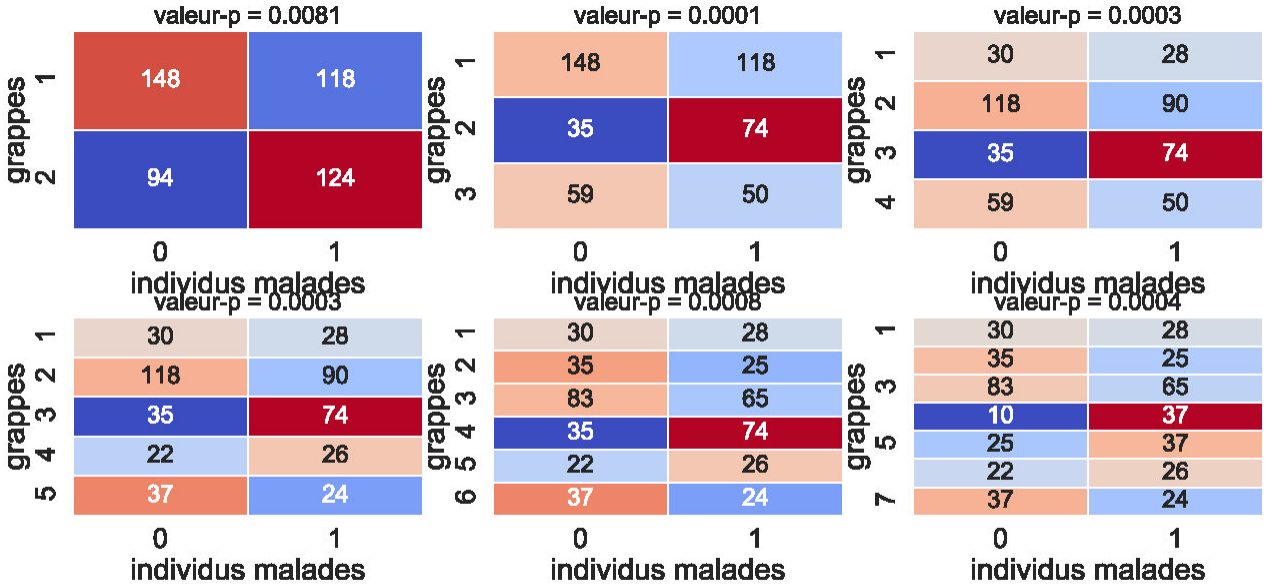


FIGURE 2.9. Tests du χ^2 pour tester l'indépendance d'un diagnostic de schizophrénie et l'attribution à une grappe, $K = 2$ jusqu'à $K = 7$



Note : Le bleu et le rouge représentent respectivement la négativité et la positivité d'une différence relative entre l'observation et la valeur attendue.
Une couleur plus intense représente une valeur absolue plus élevée de cette différence relative.

donc que les effets attribuables à la maladie ont des impacts significatifs sur les données et influencent le regroupement en grappes.

2.4.7. Comparaison des symptômes

Comme mentionné ci-haut, nous disposons de l'évaluation de *six* types de symptômes sur l'échelle SAPS/SANS pour tous les individus malades. Ces évaluations portent sur les hallucinations, les délusions, la désorganisation, l'amotivation, l'expressivité et la cognition des individus atteints. La figure 2.10 montre des niveaux moyens de symptômes évidemment similaires d'une grappe à l'autre. À titre indicatif, une MANOVA a été réalisée pour tester les moyennes des valeurs des symptômes d'une grappe à l'autre. Cette analyse n'a pas pu déceler une quelconque différence significative d'une grappe à l'autre (valeur-p $> 0{,}41$). [3] De la même manière, si on utilise le test de Kruskal-Wallis pour comparer, entre les grappes, la médiane du niveau d'un symptôme, alors la valeur-p obtenue (après correction de Bonferroni pour comparaisons multiples) est bien au-dessus du seuil de 5% et cela, pour tous les symptômes. Il sera intéressant de comparer ces résultats à ceux obtenus après l'utilisation du nouvel algorithme développé dans les chapitres subséquents.

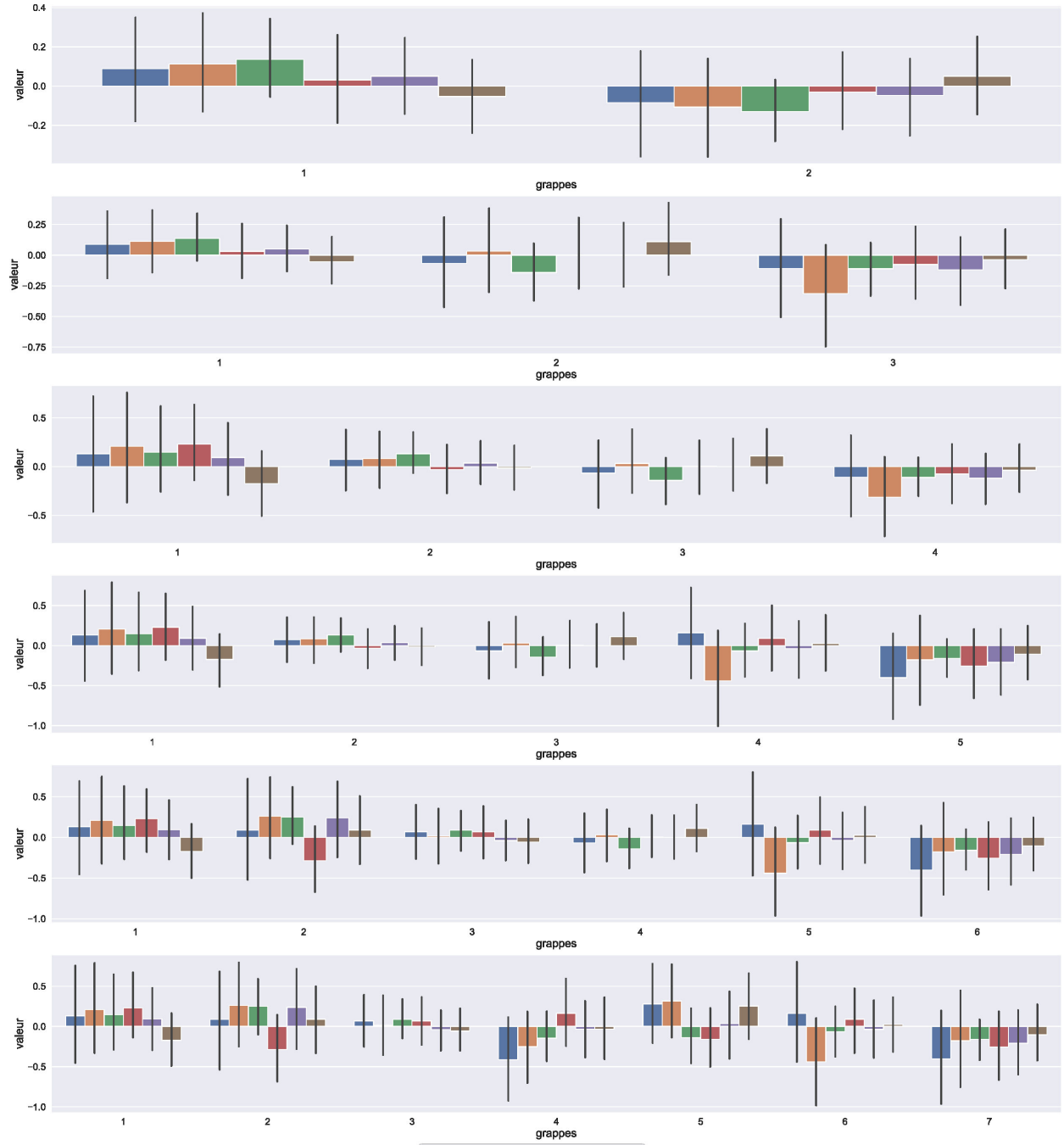
2.4.8. Observations

Comme attendu, l'effet des mouvements de tête sont réels et sont difficiles à éliminer. Il peut paraître étonnant qu'une régression linéaire ait été suffisante pour estimer (et donc éliminer) un effet statistiquement significatif du site de provenance. On remarque aussi que la proportion d'individus atteints de la maladie varie de manière statistiquement significative entre les grappes. L'effet de la maladie a donc une réelle répercussion sur les données d'IRMf à notre disposition et il est donc raisonnable d'utiliser ces données pour la suite de notre recherche.

Cette partie de notre étude tend à supporter l'affirmation que la schizophrénie est associée à une diminution de la connectivité fonctionnelle. On remarque qu'à mesure que K augmente, les grappes ont tendances à se créer en divisant une grappe déjà existante en deux sous-grappes, une exhibant une connectivité moyenne plus élevée et l'autre, une connectivité moyenne plus basse. Cette dernière grappe contient invariablement un ratio élevé du nombre d'individus malades par rapport au nombre d'individus non-malades, comme on peut le constater en observant les figures précédentes. Le niveau moyen de connectivité semble donc guider le regroupement en grappes (voir figures 2.9 et 2.8) tandis que les symptômes sont

3. Ce test a été réalisé avec la librairie Python *statsmodels*.

Figure 2.10. Différence entre les niveaux moyens des symptômes dans les grappes et le niveau moyen des symptômes de tous les individus, $K = 2$ jusqu'à $K = 7$



Note : Les barres verticales désignent des intervalles de confiance de 95%

très similaires d'une grappe à l'autre. La motivation pour trouver d'autres grappes ou une nouvelle manière de diviser les données est donc intacte.

Chapitre 3

Modèle de mélange gaussien fini et algorithme Espérance-Maximisation

Tel que mentionné au chapitre précédent, X est la matrice où chaque ligne de b dimensions représente le connectome vectorisé d'un individu sain (une observation). Par contre, dans un but de compatibilité avec certaines références utilisées (notamment McLachlan et Peel 2000, p. 6, et autres), tous les développements mathématiques sont effectués avec l'idée qu'un vecteur représentant les données d'un individu est vertical de dimensions $b \times 1$; $\boldsymbol{x}_i = \left[\; x_{i,1}, x_{i,2}, \ldots x_{i,b} \; \right]^T$. Il existe une méthode très flexible qui peut nous permettre de modéliser ces vecteurs de données et d'autres vecteurs de données distribués de manières diverses : l'utilisation d'un modèle probabiliste de mélange de densités (voir par exemple la figure 3.1). Le principe de ce modèle est de considérer que la population étudiée consiste en K sous-populations qui sont chacune modélisable par une densité. L'entièreté de la population n'est donc, sous ce modèle, qu'un mélange de K densités distinctes. Ainsi, la matrice X est supposée être une réalisation aléatoire (ou encore un échantillon aléatoire) de densité $f(\boldsymbol{x} \mid \boldsymbol{\Theta}) = \sum_{g=1}^{K} \pi_g f_g(\boldsymbol{x} \mid \theta_g)$, où $\sum_{g}^{K} \pi_g = 1$, π_g étant la proportion (strictement positive) qu'occupe la densité f_g dans le mélange de K densités, K étant fixé et θ_g désignant l'ensemble des paramètres de la densité f_g [McLachlan et Peel, 2000, p.6, p.29]. Les densités f_g peuvent être ou ne pas être gaussiennes, mais l'utilisation d'un mélange de densités gaussiennes est beaucoup plus populaire que l'utilisation d'alternatives [Melnykov et collab., 2010, p. 82]. De plus, une modélisation standard implique généralement de considérer les densités f_g comme étant toutes du même type ($f_g(\boldsymbol{x} \mid \theta_g) = f(\boldsymbol{x} \mid \theta_g)$ pour tout g) [McNicholas, 2016, p.335]. Dans le cas où le modèle fait appel à un mélange de densités gaussiennes, les matrices de

covariances doivent être déterminées comme uniques ou non à chaque densité du mélange et/ou être déterminées comme étant diagonales, complètes ou même encore décomposées de manière à rendre le modèle plus parcimonieux.

Rappelons que l'un de nos objectifs est en fait d'isoler, à travers les données à notre disposition, des sous-groupes et de décrire à travers certaines statistiques les caractéristiques de chacun de ces sous-groupes. Ainsi, une fois le modèle établi (donc l'accomplissement de l'estimation de tous les paramètres θ_g pertinents) à partir de l'échantillion à notre disposition, la constitution de grappes se fait en attribuant chaque observation de l'échantillon à la densité qui a la probabilité *a posteriori* la plus élevée d'être à l'origine de cette observation [Melnykov et collab., 2010, p. 82].

L'algorithme Espérance-Maximisation (EM) est une procédure itérative typique pour arriver à l'estimation des paramètres des densités composant le mélange gaussien ou autre sur la base du fait que les données sont incomplètes ou supposées être incomplètes. [1] Le principe de l'algorithme, comme son nom l'indique, est d'alterner de manière cyclique entre l'étape Espérance (E) et l'étape Maximisation (M). Une réalisation des étapes E et M constitue une itération. L'étape E consiste à calculer la probabilité qu'un individu ait été généré par une densité gaussienne dont les paramètres de moyenne et de variance ont été estimés à l'étape M précédente (ou à l'étape d'initialisation de l'algorithme si c'est la première fois que les probabilités sont évaluées). L'étape M tire son nom du fait que les paramètres sont estimés par la maximisation d'une log-vraisemblance conditionnelle. La solution est l'attribution déterminée à une grappe pour chaque individu (s'il y a un besoin de regroupement en grappes) et l'estimation finale des paramètres quand un critère de convergence adéquat est vérifié.

L'algorithme nécessite donc, avant d'être mis en marche, que soit décidé le nombre de densités (donc de grappes) dont les paramètres doivent être estimés. Ces densités composent le mélange dont les données à la disposition de l'utilisateur sont une manifestation, sous le modèle. Fixer le nombre de grappes de manière correcte et de bons points de départ lors de l'étape d'initialisation sont deux défis associés à l'emploi de l'algorithme EM [McLachlan et Peel, 2000, p. 4-5].

Ici est présenté le développement menant à des estimateurs lorsque les matrices de covariances sont distinctes et complètes pour chaque densité gaussienne. Il s'agit du choix naturel

1. Plus de détails sur la signification du caractère incomplet des données seront mentionnés ci-bas.

dans notre contexte puisque des corrélations existent (du moins, *a priori*) entre les différentes dimensions d'un connectome.

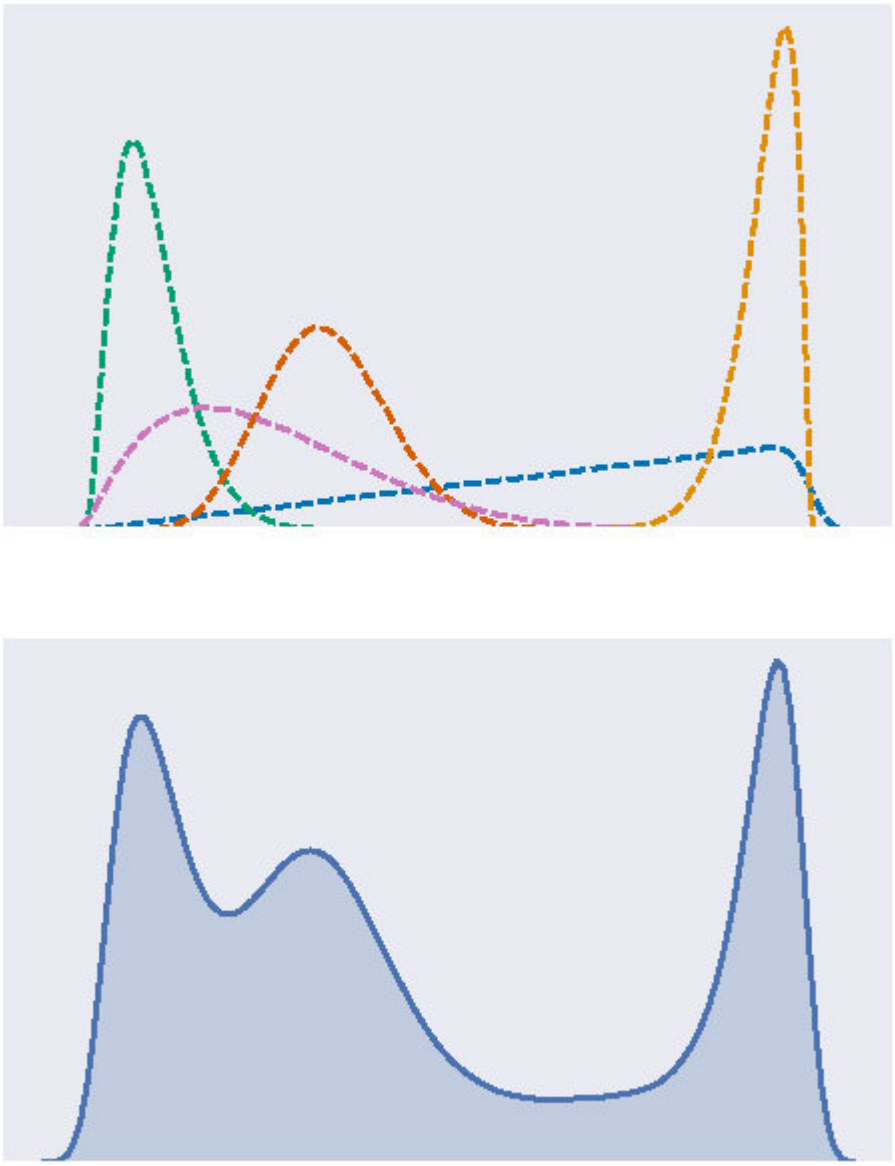
Il est donc supposé que nous avons un mélange de K densités gaussiennes comme modèle pour N individus et que chaque densité a une matrice de variance-covariance propre à elle. Tel que mentionné ci-haut, K doit être choisi avant de commencer l'algorithme. Soit π_g la proportion qu'occupe la densité f_g dans le mélange de densités gaussiennes, μ_g son vecteur du paramètre de moyenne de dimensions $b \times 1$ et Σ_g sa matrice de variance-covariance de dimensions $b \times b$. Le mélange s'écrit alors $f(\boldsymbol{x} \mid \boldsymbol{\Theta}) = \sum_{g=1}^{K} \pi_g f_g(\boldsymbol{x} \mid \theta_g) = \sum_{g=1}^{K} \pi_g \mathcal{N}(x\ ; \mu_g; \Sigma_g)$.

3.1. Algorithme Espérance-Maximisation

3.1.1. Étape Espérance

Pour pouvoir obtenir les probabilités que les données de chaque individu soient générées par les densités f_g du mélange, il faut estimer les paramètres π_g, μ_g, Σ_g pour chaque densité. Supposons l'existence d'une matrice Z (disons de dimensions $N \times K$) contenant des labels notés z tels que $z_{ig} = 1$ si l'individu i fait parti de la grappe g, c'est-à-dire que ses données sont attribuables à la densité f_g parmi les K densités constituant le mélange et $z_{ig} = 0$ dans le cas contraire. Les données seraient dites complètes si on pouvait observer l'entièreté de X et Z mais seul X nous est disponible, c'est-à-dire que tous les z_{ig} sont des labels inconnus, et donc les données observées sont dites incomplètes. La log-vraisemblance des données complètes s'écrit $\log(f(X,Z \mid \Theta)) = \sum_{i=1}^{N}\sum_{g=1}^{K} z_{ig} \log(\pi_g f_g(x_i \mid \theta_g))$. Soit $\bar{\Theta}$ les estimations des paramètres de l'étape actuelle. Il s'agit donc d'une constante connue, tout comme X est constant et connu. L'estimation à venir des paramètres et l'ensemble des labels (Z) sont des inconnus et peuvent donc être considérés comme deux variables aléatoires. Comme mentionné auparavant, l'algorithme EM est une procédure itérative utilisant $\bar{\Theta}$ et le fait que Z est inconnu pour trouver les probabilités qu'une observation soit générée par une ou l'autre des densités composant le modèle. On pose $Q(\Theta \mid \bar{\Theta}) = E[\log(f(X,Z \mid \Theta)) \mid X, \bar{\Theta}]$. Pour chaque g fixé, z_{ig} peut-être considéré comme une variable distribuée selon une loi de Bernoulli. Ainsi, $Q(\Theta \mid \bar{\Theta}) = \sum_{g=1}^{K}\sum_{i=1}^{N} \log(\pi_g) p(z_{ig} = 1 \mid x_i, \bar{\Theta}) + \sum_{g=1}^{K}\sum_{i=1}^{N} \log(f_g(x_i \mid \theta_g)) p(z_{ig} = 1 \mid x_i, \bar{\Theta})$. Dans ce contexte, $\bar{\Theta}_g = (\bar{\pi}_g, \bar{\mu}_g, \bar{\Sigma}_g)$. De plus, la probabilité conditionnelle pour l'individu i d'avoir

Figure 3.1. 5 densités bêta et densité de probabilité résultant de leur mélange



été généré par la densité f_g est :

$$p_{ig} = p(z_{ig} = 1 \mid x_i, \bar{\Theta}) = \frac{\bar{\pi}_g |\bar{\Sigma}_g|^{-\frac{1}{2}} \exp\{-\frac{1}{2}(x_i - \bar{\mu}_g)^T \bar{\Sigma}_g^{-1}(x_i - \bar{\mu}_g)\}}{\sum_{g'=1}^{K} \bar{\pi}_{g'} |\bar{\Sigma}_{g'}|^{-\frac{1}{2}} \exp\{-\frac{1}{2}(x_i - \bar{\mu}_{g'})^T \bar{\Sigma}_{g'}^{-1}(x_i - \bar{\mu}_{g'})\}}.$$

La barre au-dessus des paramètres désigne l'estimation actuelle des paramètres. Évidemment, sous ce modèle et pour chaque individu i, la somme des probabilités que i soit généré par une densité parmi celles constituant le mélange gaussien est de 1, ce que l'on peut écrire : $\sum_{g=1}^{K} p_{ig} = 1$. Si on fait du regroupement en grappes, chaque individu est attribué à la densité gaussienne (parmi K densités) qui a la plus grande probabilité de l'avoir généré. À partir de la fonction Q, nous pouvons établir des estimateurs selon le principe du maximum de vraisemblance (voir McNicholas 2016, p. 335).

3.1.2. Étape Maximisation : estimateur de π_g

Pour obtenir l'estimateur de π_g, il y a la contrainte $\sum_{g=1}^{K} \pi_g = 1$ à respecter. Introduisons le multiplicateur de Lagrange λ et posons $p_{ig} = p(z_{ig} = 1 \mid x_i, \bar{\Theta})$ pour alléger le texte. Il faut donc résoudre l'équation :

$$\frac{\partial}{\partial \pi_g}\Big[Q(\Theta \mid \bar{\Theta}) + \lambda\Big(\sum_{g=1}^{K} \pi_g - 1\Big)\Big] = 0.$$

Pour un g donné, $\frac{\partial}{\partial \pi_g} \sum_{i=1}^{N} \log(\pi_g) p_{ig} = \sum_{i=1}^{N} \frac{1}{\pi_g} p_{ig}$ et $\frac{\partial}{\partial \pi_g} \pi_g = 1$. Ceci nous permet d'écrire $-\lambda = \sum_{i=1}^{N} \frac{1}{\pi_g} p_{ig}$. Il est donc justifié d'écrire que $-\pi_g \lambda = \sum_{i=1}^{N} p_{ig}$. En appliquant la sommation $\sum_{g=1}^{K}$ sur les deux côtés de la précédente égalité, on se retrouve avec le fait que $\lambda = -N$ et donc avec l'estimateur suivant :

$$\hat{\pi}_g = \frac{1}{N} \sum_{i=1}^{N} p(z_{ig} = 1 \mid x_i, \bar{\Theta}). \tag{3.1.1}$$

3.1.3. Étape Maximisation : estimateur de μ_g

Dans le contexte d'un mélange gaussien, $\theta_g = (\mu_g, \Sigma_g)$, c'est-à-dire que les paramètres de la densité f_g sont un vecteur de moyenne μ_g et une matrice symétrique de covariances Σ_g. Ainsi, $\log(f_g(x_i \mid \theta_g)) = C - \frac{1}{2}\log(|\Sigma_g|) - \frac{1}{2}(x_i - \mu_g)^T \Sigma_g^{-1}(x_i - \mu_g)$ où C désigne une

constante n'ayant pas d'impact sur les développements subséquents. La symétrie de Σ_g nous permet d'écrire, pour un g donné : $\frac{\partial}{\partial \mu_g} \log(f_g(x_i \mid \theta_g)) = -\frac{1}{2}\frac{\partial}{\partial \mu_g}[(x_i-\mu_g)^T\Sigma_g^{-1}(x_i-\mu_g)] = -\frac{1}{2}[(x_i-\mu_g)^T(\Sigma_g^{-1})^T(-1) + (x_i-\mu_g)^T(\Sigma_g^{-1})(-1)] = (\Sigma_g^{-1})(x_i-\mu_g)$.

Ainsi, $\frac{\partial}{\partial \mu_g}(Q) = 0 = \sum_{i=1}^{N}(x_i-\mu_g)^T(\Sigma_g^{-1})p_{ig}$ et cela permet d'isoler l'estimateur :

$$\hat{\mu}_g = \frac{\sum_{i=1}^{N} x_i p(z_{ig}=1 \mid x_i, \bar{\Theta})}{\sum_{i=1}^{N} p(z_{ig}=1 \mid x_i, \bar{\Theta})}. \tag{3.1.2}$$

3.1.4. Étape Maximisation : estimateur de Σ_g

Notons d'abord que $|\Sigma^{-1}| = \frac{1}{|\Sigma|} = |\Sigma|^{-1}$. La trace et la diagonale de la matrice A seront notées respectivement $\mathrm{tr}(A)$ et $\mathrm{diag}(A)$. Les deux résultats suivants sont établis [2] pour une matrice A symétrique et une matrice B : $\frac{\partial \log|A|}{\partial A} = 2A^{-1} - \mathrm{diag}(A^{-1})$; $\frac{\partial \mathrm{tr}(AB)}{\partial A} = B + B^T - \mathrm{diag}(B)$. Or, $\sum_{g=1}^{K}\sum_{i=1}^{N}\log(f_g(x_i \mid \theta_g))p_{ig} = \sum_{g=1}^{K}\frac{1}{2}\log(|\Sigma_g^{-1}|)\sum_{i=1}^{N}p_{ig} - \frac{1}{2}\sum_{g=1}^{K}\sum_{i=1}^{N}p_{ig}\cdot tr[\Sigma_g^{-1}(x_i-\mu_g)(x_i-\mu_g)^T]$ et Σ_g est une matrice symétrique. Ainsi, pour un g fixé, $\frac{\partial}{\partial \Sigma_g^{-1}}(Q) = 0 = \frac{1}{2}\sum_{i=1}^{N}p_{ig}\Big(2\Sigma_g - \mathrm{diag}(\Sigma_g)\Big) - \frac{1}{2}\sum_{i=1}^{N}p_{ig}\Big(2(x_i-\mu_g)(x_i-\mu_g)^T - \mathrm{diag}[(x_i-\mu_g)(x_i-\mu_g)^T]\Big) =$

$$\frac{1}{2}\sum_{i=1}^{N}p_{ig}\Big(2[\Sigma_g - (x_i-\mu_g)(x_i-\mu_g)^T] - \mathrm{diag}[\Sigma_g - (x_i-\mu_g)(x_i-\mu_g)^T]\Big).$$

Ceci est une expression matricielle de la forme $2D - \mathrm{diag}(D) = 0$. Cela implique que $2D = \mathrm{diag}(D)$ et donc que $D = 0$. On peut donc affirmer que $\sum_{i=1}^{N}p_{ig}\Big(2[\Sigma_g - (x_i-\mu_g)(x_i-\mu_g)^T]\Big) = 0$ et ceci mène à l'estimateur suivant :

$$\hat{\Sigma}_g = \frac{\sum_{i=1}^{N} p(z_{ig}=1 \mid x_i, \bar{\Theta})(x_i-\mu_g)(x_i-\mu_g)^T}{\sum_{i=1}^{N} p(z_{ig}=1 \mid x_i, \bar{\Theta})}. \tag{3.1.3}$$

3.1.5. Simultanéité des estimations

Remarquons que le seul estimateur qui dépend d'un autre est l'estimateur de Σ_g qui dépend de l'estimateur de μ_g. Ainsi, un ordre naturel pour effectuer les estimations lors de l'étape M est d'estimer π_g, puis estimer μ_g, puis estimer Σ_g (en utilisant l'estimation

2. Petersen et collab., p. 15.

de μ_g). Une fois l'initialisation effectuée et K déterminé, l'algorithme EM peut être utilisé tel que décrit par le pseudo-code (voir Algorithme 1). La figure 3.2 montre une densité de probabilité inusuelle résultant du mélange de plusieurs densités gaussiennes et l'estimation des paramètres de ces densités résultant de l'utilisation de l'algorithme EM.

```
1  tant que critère de convergence ou d'arrêt non vérifié faire
2      pour chaque observation dans X faire
3          pour chaque g compris entre 1 et K faire
4              Évaluer la probabilité conditionnelle que l'observation soit générée par
                densité paramétrisée par Θ̄_g
5      pour  chaque g compris entre 1 et K faire
6          Évaluer π̄_g (pour chaque g);
7          Évaluer μ̄_g (pour chaque g);
8          Évaluer Σ̄_g (pour chaque g);
9          Θ̄_g ← (π̄_g, μ̄_g, Σ̄_g)
10     Vérifier critère de convergence ou d'arrêt;
```

Algorithme 1 : Espérance-Maximisation

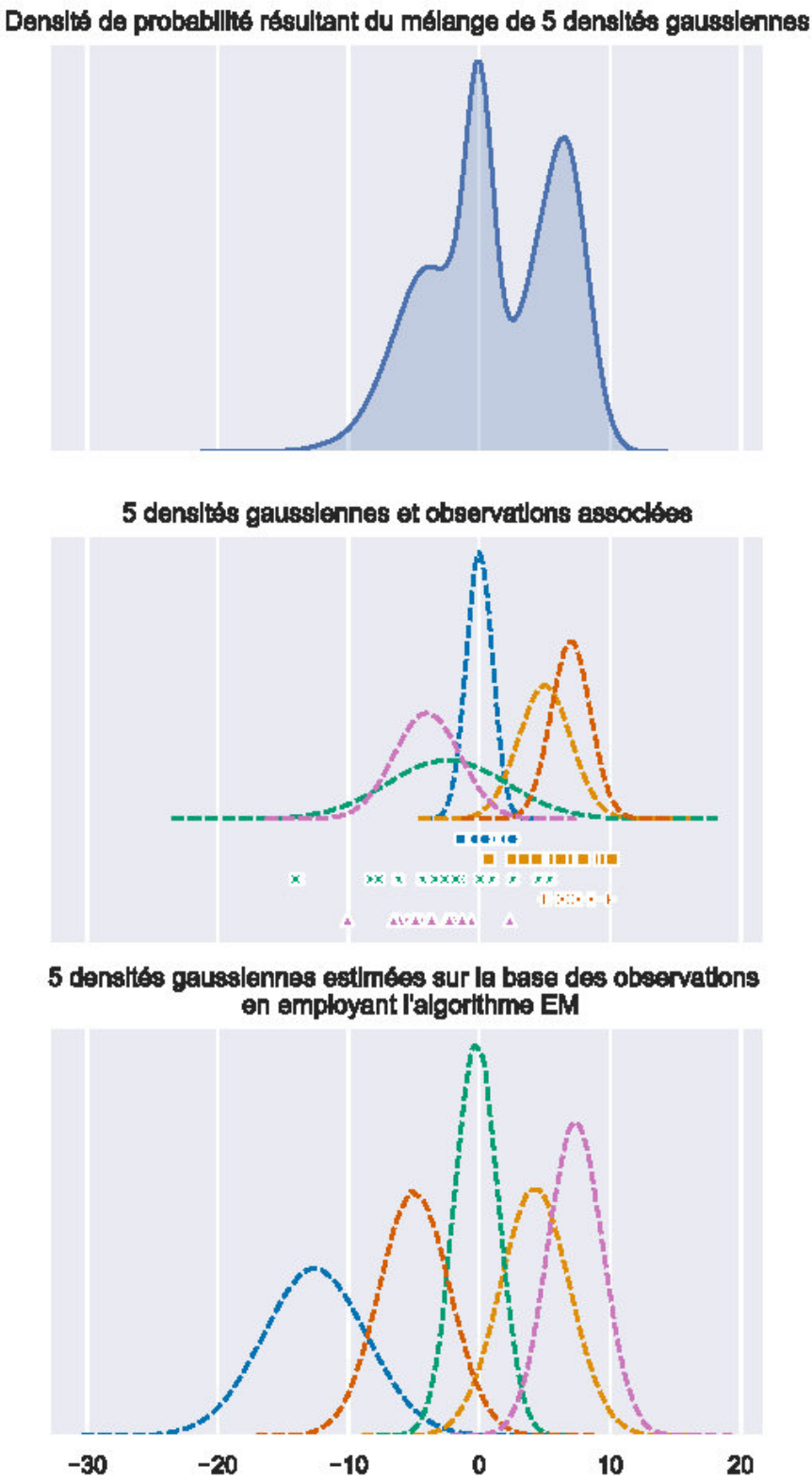
3.1.6. Choix de K

Il a déjà été mentionné qu'une problématique associée avec l'utilisation d'un modèle probabiliste de mélange de densités est de déterminer le nombre de densités sans nécessairement avoir d'informations sur le nombre réel ou optimal de sous-groupes constituant les données à l'étude. Une méthode qui peut être utilisée consiste à répéter l'utilisation de l'algorithme EM avec des valeurs de K différentes et de sélectionner le modèle maximisant la valeur du critère d'information bayésien (communément désigné par l'acronyme anglais BIC) qui est défini de la manière suivante :

$$2\log_{\mathcal{M}}(f(x,\hat{\theta})) - m_{\mathcal{M}}\log(N).$$

où $\log_{\mathcal{M}}(f(x,\hat{\theta})) = \sum_{i=1}^{N}\sum_{g=1}^{K} \mathbb{1}_{z_{ig=1}} \log(\pi_g \mathcal{N}(x_i\ ;\bar{\Theta}_g))$, c'est-à-dire la log-vraisemblance du modèle $\mathcal{M}$ et où $m_{\mathcal{M}} = Kb(b+1)/2 + bK + K - 1$, c'est-à-dire le nombre de paramètres uniques et indépendants à estimer dans le cas où chaque densité se voit attribuer sa propre matrice complète de covariances [Schwarz et collab., 1978] [Fraley et Raftery, 1998, p. 582].

Figure 3.2. Illustration sur l'algorithme EM et les estimations en résultant



3.1.7. Remarques sur l'initialisation et la convergence de l'algorithme

Un des désavantages de l'algorithme EM est qu'une mauvaise initialisation (par exemple : trop éloignée de la vraie solution ou trop proche d'un maximum local) peut mener à de mauvais résultats, c'est-à-dire à la convergence vers une solution non-optimale ou à un échec de l'algorithme. Au niveau numérique, la convergence de l'algorithme est lente. Il n'est pas certain que la convergence sera vers le point qui maximisera la vraisemblance. Il est en effet possible que la convergence soit vers un maximum local. Pour cette raison, certains auteurs recommandent que l'algorithme EM soit essayé à plusieurs reprises avec des points de départ différents [Wu et collab., 1983]. De plus, une mauvaise initialisation peut favoriser l'émergence de matrices singulières et/ou de dégénérescence des estimateurs, ce qui est généralement fatal pour la bonne implémentation de l'algorithme. La qualité de l'initialisation est donc «cruciale» pour que de bons estimés soient obtenus avec l'algorithme EM [Melnykov et collab., 2010, p. 86]. Celle-ci se réalise généralement en utilisant un autre algorithme de regroupement en grappes et le résultat de cette méthode sera alors le point de départ du regroupement en grappes EM. Dans notre cas, la fonction du module *sklearn* utilisée permet d'initialiser en faisant appel à différents algorithmes de regroupement en grappes [3]. Nous avons utilisé la méthode d'initialisation par défaut qui est l'algorithme K-moyennes. C'est une méthode qui donne généralement de bons points d'initialisation pour l'algorithme EM si le nombre de grappes n'est pas élevé [Hu, 2015]. L'initialisation peut être une tâche complexe et il n'existe aucune méthode qui permette d'initialiser de manière optimale peu importe la situation [Blömer et Bujna, 2013].

3.2. Emploi d'une régularisation bayésienne

Il vient d'être mentionné que l'algorithme EM peut échouer. Pratiquement, ceci est souvent dû à l'émergence de matrices de covariances singulières et ceci arrive surtout lorsque les matrices sont uniques à chaque densité gaussienne ou lorsque le nombre de densités est élevé. L'emploi d'une régularisation bayésienne peut éliminer les échecs dus à l'émergence de singularité de matrice sans avoir d'effets majeurs sur les résultats [Fraley et Raftery, 2007].

3. Pedregosa et collab. [2011]

Supposons les densités *a priori* suivantes pour les moyennes et les matrices de covariances :

$$\mu_g \mid \Sigma_g \sim \mathcal{N}(\mu_{pr}, \Sigma_g/\kappa_{pr}) \propto |\Sigma_g|^{-1/2} \exp\{-\frac{\kappa_{pr}}{2} tr[(\mu_g - \mu_{pr})^T \Sigma^{-1} (\mu_g - \mu_{pr})]\};$$
$$\Sigma_g \sim \mathcal{W}^{-1}(\Lambda_{pr}, \nu_{pr}) \propto |\Sigma_g|^{-\frac{\nu_{pr}+b+1}{2}} \exp\{-\frac{1}{2} tr[{\Sigma_g}^{-1} \Lambda_{pr}^{-1}]\},$$

où $\mathcal{N}(\mu, \Sigma)$ et $\mathcal{W}^{-1}(\Lambda, \nu)$ représentent respectivement une loi normale multivariée et une loi de Wishart inverse. Les hyperparamètres sont supposés égaux pour toutes les densités. Tel qu'à la section 3.1.1 :

$$Q(\Theta \mid \bar{\Theta}) = \sum_{g=1}^{K} \sum_{i=1}^{N} \log(\pi_g) p_{ig} + \sum_{g=1}^{K} \sum_{i=1}^{N} \log(f_g(x_i \mid \theta_g)) p_{ig}.$$

Si certains termes constants sont négligés, l'expression Q peut être développée en l'expression simplifiée (Q_s) ainsi :

$$Q_s = \sum_{g=1}^{K} \Bigg\{ -\frac{1}{2} \log(|\Sigma_g|) \sum_{i=1}^{N} p_{ig} - \frac{1}{2} \sum_{i=1}^{N} p_{ig} \cdot tr(\Sigma_g^{-1}(x_i - \mu_g)(x_i - \mu_g)^T) + \sum_{i=1}^{N} p_{ig} \log(\pi_g) \Bigg\}.$$

Nous introduisons la fonction *a priori* conjuguée Normale-Inverse-Wishart, dorénavant désignée par NIW, pour tous les groupes. Cette fonction est proportionnelle à :

$$\prod_{g=1}^{K} |\Sigma_g|^{-\frac{\nu_{pr}+b+2}{2}} \exp\Bigg\{ -\frac{1}{2} tr\big[{\Sigma_g}^{-1} \Lambda_{pr}^{-1}\big] \Bigg\} \exp\Bigg\{ -\frac{\kappa_{pr}}{2} tr\big[(\mu_g - \mu_{pr})^T \Sigma_g^{-1} (\mu_g - \mu_{pr})\big] \Bigg\}.$$

La fonction $\log(NIW) + Q_s$ est à maximiser pour obtenir les estimateurs voulus et correspond, à quelques constantes près, à :

$$\sum_{g=1}^{K} \Bigg\{ \left(\frac{\nu_{pr} + b + 2}{2} \right) \log|\Sigma_g^{-1}| - \frac{1}{2} tr\big[{\Sigma_g}^{-1} \Lambda_{pr}^{-1}\big] - \frac{\kappa_{pr}}{2} tr\big[(\mu_g - \mu_{pr})^T \Sigma_g^{-1} (\mu_g - \mu_{pr})\big] \Bigg\}$$
$$+ \sum_{g=1}^{K} \Bigg\{ \frac{1}{2} \log|\Sigma_g^{-1}| \sum_{i=1}^{N} p_{ig} - \frac{1}{2} \sum_{i=1}^{N} p_{ig} \cdot tr\big[\Sigma_g^{-1}(x_i - \mu_g)(x_i - \mu_g)^T\big] + \sum_{i=1}^{N} p_{ig} \log(\pi_g) \Bigg\}.$$

3.2.1. Étape Maximisation : estimateur de μ_g

Pour un g donné, $\dfrac{\partial}{\partial \mu_g} \Big(\log(NIW) + Q_s \Big) = 0$ mène directement à la relation $\kappa_{pr}(\mu_g - \mu_{pr}) = \sum_{i=1}^{N} p_{ig}(x_i - \mu_g)$ et donc à l'estimateur suivant :

$$\hat{\mu}_g = \frac{\sum_{i=1}^{N} p_{ig} x_i + \kappa_{pr} \mu_{pr}}{\kappa_{pr} + \sum_{i=1}^{N} p_{ig}}. \tag{3.2.1}$$

3.2.2. Étape Maximisation : estimateur de Σ_g

Pour un g donné, avec Λ_{pr}^{-1} symétrique (voir la section 3.2.3) :

$$\frac{\partial}{\partial \Sigma_g^{-1}}\Big(\log(NIW) + Q_s\Big) = 0 =$$
$$\frac{\nu_{pr}+b+2}{2}\Big(2\Sigma_g - \mathrm{diag}(\Sigma_g)\Big) + \frac{-1}{2}\Big(2\Lambda_{pr}^{-1} - \mathrm{diag}(\Lambda_{pr}^{-1})\Big) + \frac{\sum_{i=1}^N p_{ig}}{2}\Big(2\Sigma_g - \mathrm{diag}(\Sigma_g)\Big)$$
$$+ \frac{-\kappa}{2}\Big(2(\mu_g-\mu_{pr})(\mu_g-\mu_{pr})^T - \mathrm{diag}((\mu_g-\mu_{pr})(\mu_g-\mu_{pr})^T)\Big)$$
$$+ \frac{-\sum_{i=1}^N p_{ig}}{2}\Big(2(x_i-\mu_g)(x_i-\mu_g)^T - \mathrm{diag}((x_i-\mu_g)(x_i-\mu_g)^T)\Big).$$

Ceci est une expression matricielle de la forme $2D - \mathrm{diag}(D) = 0$. Ceci implique que $D = 0$. L'estimateur est donc :

$$\hat{\Sigma}_g = \frac{\Lambda_{pr}^{-1} + \kappa(\mu_g-\mu_{pr})(\mu_g-\mu_{pr})^T + \sum_{i=1}^N p_{ig}(x_i-\mu_g)(x_i-\mu_g)^T}{\nu+b+2+\sum_{i=1}^N p_{ig}}. \tag{3.2.2}$$

3.2.3. Valeurs des hyperparamètres

Suivant les indications de Fraley et Raftery [2005, p. 10-11], les valeurs suivantes pour les hyperparamètres constituent une possibilité qui fonctionne de manière satisfaisante :

(1) $\mu_{pr} = \frac{1}{N}\sum_{i=1}^N x_i$;

(2) $\nu_{pr} = b+2$;

(3) $\kappa_{pr} = 0.01$;

(4) $\Lambda_{pr}^{-1} = \frac{S}{K^{2/b}}$;

(5) $S = \frac{1}{N-1}\sum_{i=1}^N (x_i-\mu_{pr})(x_i-\mu_{pr})^T$.

Chapitre 4

Modèle de mélange gaussien à effets superposés et algorithme Espérance-Maximisation adapté

Nous utilisons ici la même approche qu'au chapitre précédent, c'est-à-dire un modèle probabiliste de mélange de densités. Celui-ci diffère par contre dans le fait qu'il faut modéliser, en plus de la matrice de données X, la matrice Y dont chaque ligne représente le connectome vectorisé de chaque individu malade (une observation) tel que décrit au chapitre 2. Tous les développements mathématiques sont effectués pour le moment avec l'idée qu'un vecteur représentant les données d'un individu est vertical de dimensions $b \times 1$ tel qu'illustré et mentionné au chapitre 3. Il est supposé que nous avons un mélange de K_c densités gaussiennes comme modèle pour les N_c individus sains et que chaque densité a une matrice de variance-covariance propre à elle. Chaque individu malade parmi les N_s individus malades est modélisé comme ayant un état normal de base, donc qu'il fait à l'origine parti d'une grappe g parmi K_c grappes distinctes, puis qu'ensuite un état maladif j parmi K_s états maladifs distincts s'est superposé à son état normal faisant en sorte qu'il fait finalement partie d'une grappe gj parmi $K_c \cdot K_s$ grappes.

Soit $\pi_{m,gj}$ la proportion qu'occupe la densité $f_{m,gj}$ dans le mélange de $K_c \cdot K_s$ densités gaussiennes, $\mu_{m,j}$ le vecteur du paramètre de moyenne de la densité j parmi les K_s densités représentant les effets de maladie et $\Sigma_{m,j}$ la matrice de variance-covariance respectivement représentés par des matrices de dimensions $K_c \times K_s$, $b \times 1$ et $b \times b$.

Si nous établissions un modèle analogue à celui du chapitre précédent, alors il serait possible que le fait que des effets de maladie s'additionnent à l'effet témoin mène l'algorithme EM à un point où sont confondus une partie d'un ou des effets de maladie avec un ou des effets

de groupe témoin. Ceci peut provoquer une situation où plus d'une solution serait possible pour les mêmes données X, Y. Notamment, le modèle ne permettrait pas toujours d'identifier l'effet témoin pour les individus malades. Comme mentionné par McLachlan et Peel [2000, p. 27], une manière de prévenir l'apparition d'un tel phénomène dû à l'interchangeabilité de certains effets est d'imposer des contraintes sur certains paramètres. Les contraintes que nous imposerons seront explicitées ci-bas. D'abord, nous introduisons μ_c qui est la moyenne de la catégorie contrôle et μ_m qui est la moyenne de la catégorie malade. Les vecteurs μ_c et μ_m sont de dimensions $b \times 1$; ils sont respectivement communs à tous les individus témoins et tous les individus malades. Ainsi, pour les individus sains et les individus malades, nous avons respectivement les mélanges gaussiens suivants en guise de modèle :

$$f(\boldsymbol{x} \mid \boldsymbol{\Theta}) = \sum_{g=1}^{K_c} \pi_{c,g} f_{c,g}(\boldsymbol{x} \mid \theta_g) = \sum_{g=1}^{K_c} \pi_{c,g} \mathcal{N}(\ \boldsymbol{x}\ ; \mu_c + \mu_{c,g}; \Sigma_{c,g});$$

$$f(\boldsymbol{y} \mid \boldsymbol{\Theta}) = \sum_{j=1}^{K_s} \sum_{g=1}^{K_c} \pi_{m,gj} f_{m,gj}(\boldsymbol{y} \mid \theta_{gj}) = \sum_{j=1}^{K_s} \sum_{g=1}^{K_c} \pi_{m,gj} \mathcal{N}(\ \boldsymbol{y}\ ; \mu_m + \mu_{c,g} + \mu_{m,j}; \Sigma_{m,j}).$$

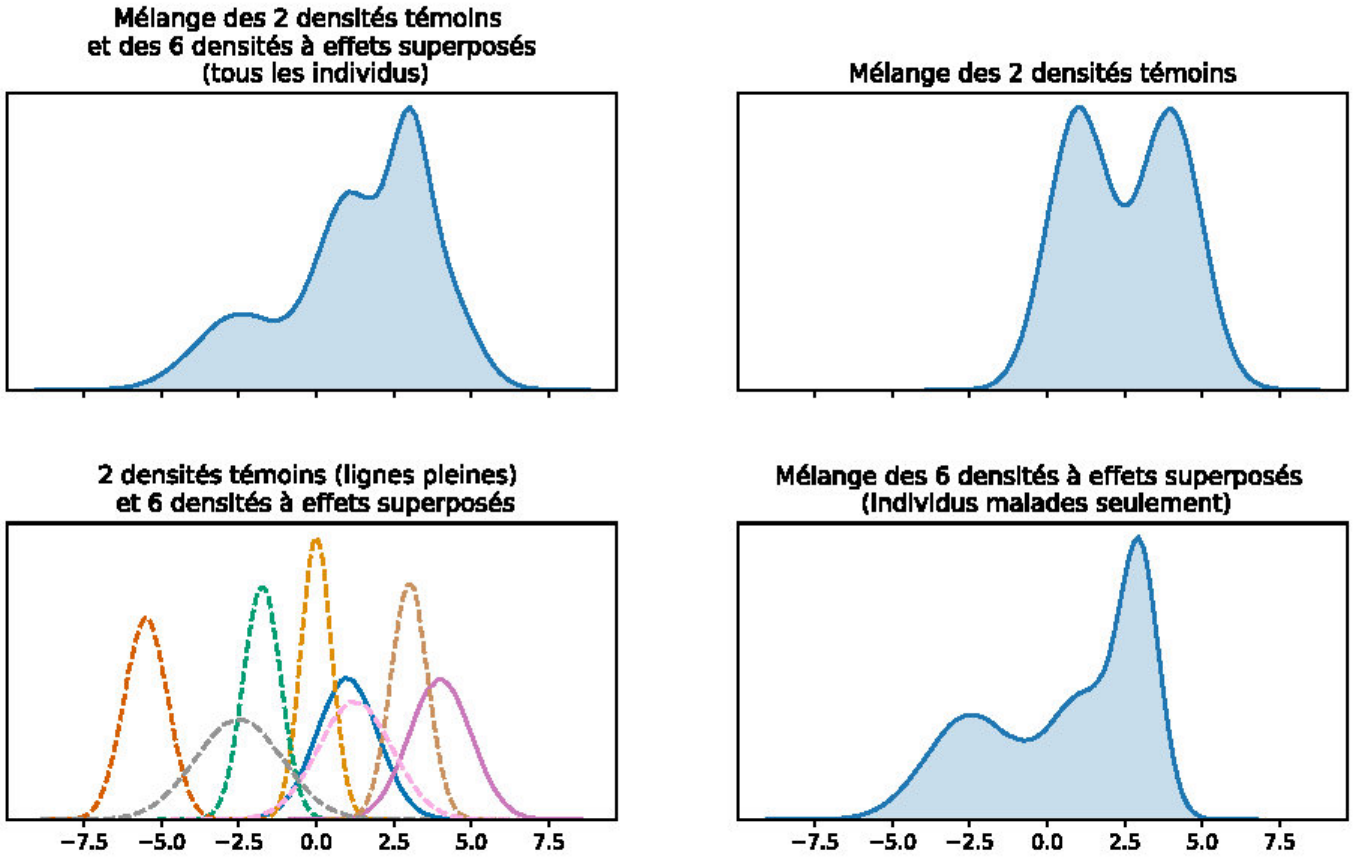
Les paramètres indicés par c,g sont tels que ceux indicés par g au chapitre précédent.[1] Ainsi, la matrice X et la matrice Y ne sont qu'une réalisation aléatoire suivant respectivement une de ces deux densités. De manière analogue à ce qui a été mentionné auparant, $\sum_{g}^{K_c} \sum_{j}^{K_s} \pi_{m,gj} = 1$. Le paramètre θ_{gj} désigne l'ensemble des paramètres de la densité $f_{m,gj}$. Dans notre modèle, les densités $f_{c,g}$ et $f_{m,gj}$ sont toutes gaussiennes, bien qu'il soit probablement possible d'employer un modèle faisant appel à des densités autres que gaussiennes. Encore une fois, nous considérons les matrices de covariances comme distinctes et complètes pour chaque densité car des corrélations existent (du moins, *a priori*) entre les différentes dimensions d'un connectome tant chez les sujets malades que les sujets témoins. La figure 4.1 illustre un exemple de ces mélanges gaussiens sans contraintes.

4.1. Algorithme Espérance-Maximisation pour mélange gaussien à effets superposés (EMS)

Le principe général de l'algorithme EMS utilisé pour arriver à une estimation des paramètres du mélange gaussien reste le même : l'algorithme alterne entre l'étape E et l'étape M

1. L'indice c a été rajouté pour désambiguïser les paramètres du modèles qui concernent les effets des groupes témoins des paramètres qui concernent les effets de maladie qui eux ont été notés avec un m.

FIGURE 4.1. Mélange de densités sous le modèle proposé



et ces deux étapes ne diffèrent pas, fondamentalement, de ce qui a été décrit au chapitre 3. K_c et K_s doivent aussi être fixés avant de faire fonctionner l'algorithme. Par contre, il y a plus de paramètres à estimer et il faut modifier certains estimateurs.

4.1.1. Étape Espérance

Pour pouvoir obtenir les probabilités que les données de chaque individu soient générées par les densités $f_{c,g}$ et $f_{m,j}$ composant le mélange, il faut estimer les paramètres $\pi_{c,g}, \pi_{m,gj}, \mu_{c,g}, \mu_{m,j}, \Sigma_{c,g}, \Sigma_{m,j}$ relatifs à chaque densité. Supposons l'existence de matrices Z_1

et Z_2 contenant respectivement des labels z_1 et z_2 binaires prenant la valeur 0 ou 1 tels que $z_{ig,1} = 1$ si et seulement si l'individu i du groupe témoin est dans la grappe g, c'est-à-dire que ses données sont attribuables à la densité f_g parmi les K_c densités constituant le mélange, et tels que $z_{igj,2} = 1$ si et seulement si l'individu i du groupe maladie est dans la grappe de maladie g,j, c'est-à-dire que son état est modélisable par la densité $f_{m,gj}$ où la moyenne de l'effet normal est $\mu_{c,g}$ et la moyenne de l'effet de maladie est $\mu_{m,j}$. Il y a donc $K_c \cdot K_s$ grappes possibles pour les individus malades. Les labels $z_{ig,1}$ et $z_{igj,2}$ sont inconnus et font donc en sorte que nos données sont incomplètes. L'estimation à venir des paramètres et l'ensemble des labels ($Z1$ et $Z2$) sont des inconnus et peuvent donc être considérés comme deux variables aléatoires. La log-vraisemblance des données complètes s'écrit :

$$\begin{aligned}\log\big(f(X,Y,Z_1,Z_2 \mid \Theta)\big) = &\sum_{i=1}^{N_c}\sum_{g=1}^{K_c} z_{ig,1}\log(\pi_{c,g} f_{c,g}(x_i \mid \theta_g)) \\ &+ \sum_{i=1}^{N_s}\sum_{g=1}^{K_c}\sum_{j=1}^{K_s} z_{igj,2}\log\big(\pi_{m,gj} f_{m,gj}(y_i \mid \theta_{gj})\big).\end{aligned}$$

Notons que $E[z_{ig,1}|X,Y,\bar{\Theta}] = p_{c,ig} = p(z_{ig,1} = 1 \mid x_i, \bar{\Theta})$ et $E[z_{igj,2}|X,Y,\bar{\Theta}] = p_{m,igj} = p(z_{igj,2} = 1 \mid y_i, \bar{\Theta})$. Dans ce contexte, $\bar{\Theta}_{g,j} = (\bar{\pi}_{m,gj}, \bar{\mu}_{c,g}, \bar{\mu}_{m,j}, \bar{\Sigma}_{m,j})$. L'espérance (E) de ceci nous mène à l'expression :

$$Q(\Theta|\bar{\Theta}) = \sum_{i=1}^{N_c}\sum_{g=1}^{K_c} p_{c,ig}\log(\pi_{c,g} f_{c,g}(x_i \mid \theta_g)) + \sum_{i=1}^{N_s}\sum_{g=1}^{K_c}\sum_{j=1}^{K_s} p_{m,igj}\log\big(\pi_{m,gj} f_{m,gj}(y_i \mid \theta_{gj})\big).$$

Notons que ces probabilités conditionnelles sont :

$$p_{c,ig} = \frac{\bar{\pi}_{c,g}|\bar{\Sigma}_{c,g}|^{-\frac{1}{2}} \exp\{-\frac{1}{2}(x_i - \bar{\mu}_c - \bar{\mu}_{c,g})^T \bar{\Sigma}_{c,g}^{-1}(x_i - \bar{\mu}_c - \bar{\mu}_{c,g})\}}{\sum_{g'} \bar{\pi}_{c,g'}|\bar{\Sigma}_{cg'}|^{-\frac{1}{2}} \exp\{-\frac{1}{2}(x_i - \bar{\mu}_c - \bar{\mu}_{c,g'})^T \bar{\Sigma}_{cg'}^{-1}(x_i - \bar{\mu}_c - \bar{\mu}_{c,g'})\}};$$

$$p_{m,igj} = \frac{\bar{\pi}_{m,gj}|\bar{\Sigma}_{m,j}|^{-\frac{1}{2}} \exp\{-\frac{1}{2}(y_i - \bar{\mu}_m - \bar{\mu}_{c,g} - \bar{\mu}_{m,j})^T \bar{\Sigma}_{m,j}^{-1}(y_i - \bar{\mu}_m - \bar{\mu}_{c,g} - \bar{\mu}_{m,j})\}}{\sum_{g'}\sum_{j'} \bar{\pi}_{m,g'j'}|\Sigma_{m,j'}|^{-\frac{1}{2}} \exp\{-\frac{1}{2}(y_i - \bar{\mu}_m - \bar{\mu}_{c,g'} - \bar{\mu}_{m,j'})^T \bar{\Sigma}_{m,j'}^{-1}(y_i - \bar{\mu}_m - \bar{\mu}_{c,g'} - \bar{\mu}_{m,j'})\}}.$$

4.1.2. Régularisation bayésienne

Comme mentionné à la section 3.2, l'algorithme EM peut échouer si des matrices singulières se présentent durant la progression de l'algorithme. Il n'y a aucun doute que le même problème peut mener à l'échec de l'algorithme EMS. Nous proposons donc l'emploi d'une

régularisation bayésienne inspirée des travaux de Fraley et Raftery [2007]. Les hyperparamètres sont supposés égaux pour toutes les densités *a priori*. Supposons les densités *a priori* suivantes pour les moyennes et les matrices de covariances :

$$\mu_{c,g} \mid \Sigma_{c,g} \sim \mathcal{N}(\eta_c, \Sigma_{c,g}/\kappa_1) \propto |\Sigma_{c,g}|^{-1/2} \exp\Big\{ -\frac{\kappa_1}{2} tr\big[(\mu_{c,g}-\eta_c)^T \Sigma_{c,g}^{-1}(\mu_{c,g}-\eta_c)\big]\Big\};$$

$$\mu_{m,j} \mid \Sigma_{m,j} \sim \mathcal{N}(\eta_m, \Sigma_{m,j}/\kappa_2) \propto |\Sigma_{m,j}|^{-1/2} \exp\Big\{ -\frac{\kappa_2}{2} tr\big[(\mu_{m,j}-\eta_m)^T \Sigma_{m,j}^{-1}(\mu_{m,j}-\eta_m)\big]\Big\};$$

$$\Sigma_{c,g} \sim \mathcal{W}^{-1}(\Lambda_1,\nu_1) \propto |\Sigma_{c,g}|^{-\frac{\nu_1+b+1}{2}} \exp\Big\{ -\frac{1}{2} tr\big[{\Sigma_{c,g}}^{-1}\Lambda_1^{-1}\big]\Big\};$$

$$\Sigma_{m,j} \sim \mathcal{W}^{-1}(\Lambda_2,\nu_2) \propto |\Sigma_{m,j}|^{-\frac{\nu_2+b+1}{2}} \exp\Big\{ -\frac{1}{2} tr\big[{\Sigma_{m,j}}^{-1}\Lambda_2^{-1}\big]\Big\}.$$

Posons $N_1 = \nu_1 + b + 2$; $N_2 = \nu_2 + b + 2$; $\Psi_{c,ig} = (x_i - \mu_{c,g} - \mu_c)$ et $\Psi_{m,igj} = (y_i - \mu_{c,g} - \mu_m - \mu_{m,j})$. Ces densités permettent d'introduire la fonction *a priori* conjuguée Normale-Inverse-Wishart (qui sera notée NIW) pour tous les groupes. Notons Q_s une fonction équivalente à Q mais dont certains termes constants ont été négligés. Nous pouvons écrire la fonction $\log(NIW) + Q_s$ de la manière suivante :

$$\begin{aligned}
&\sum_{g=1}^{K_c}\Big(\frac{N_1}{2}\Big)\log|\Sigma_{c,g}^{-1}| + \sum_{j=1}^{K_s}\Big(\frac{N_2}{2}\Big)\log|\Sigma_{m,j}^{-1}| - \sum_{g=1}^{K_c}\frac{1}{2}tr\big[{\Sigma_{c,g}}^{-1}\Lambda_1^{-1}\big] - \sum_{j=1}^{K_s}\frac{1}{2}tr\big[{\Sigma_{m,j}}^{-1}\Lambda_2^{-1}\big] \\
&- \sum_{g=1}^{K_c}\frac{\kappa_1}{2}tr\big[(\mu_{c,g}-\eta_c)^T\Sigma_{c,g}^{-1}(\mu_{c,g}-\eta_c)\big] - \sum_{j=1}^{K_s}\frac{\kappa_2}{2}tr\big[(\mu_{m,j}-\eta_m)^T\Sigma_{m,j}^{-1}(\mu_{m,j}-\eta_m)\big] \\
&+ \sum_{j=1}^{K_s}\left\{\frac{1}{2}\log|\Sigma_{m,j}^{-1}|\sum_{i=1}^{N_s}\sum_{g=1}^{K_c}p_{m,igj} - \frac{1}{2}\sum_{i=1}^{N_s}\sum_{g=1}^{K_c}p_{m,igj}\cdot tr[\Sigma_{m,j}^{-1}\Psi_{m,igj}\Psi_{m,igj}^T]\right\} \\
&+ \sum_{g=1}^{K_c}\left\{\frac{1}{2}\log|\Sigma_{c,g}^{-1}|\sum_{i=1}^{N_c}p_{c,ig} - \frac{1}{2}\sum_{i=1}^{N_c}p_{c,ig}\cdot tr[\Sigma_{c,g}^{-1}\Psi_{c,ig}\Psi_{c,ig}^T]\right\}
\end{aligned}$$

C'est cette fonction qui est à maximiser pour obtenir les estimateurs. Il est à noter que nous n'employons pas de densité *a priori* pour μ_c et μ_m car le modèle proposé est sensé s'utiliser sur un nombre conséquent d'observations, celui-ci rendant non-nécessaire l'imposition d'une loi *a priori* pour ces paramètres. De plus, l'utilisation de densités *a priori* pour ces paramètres pourrait complexifier outre mesure la dérivation des estimateurs. Ne pas imposer de loi *a*

priori est l'équivalent d'imposer une densité *a priori* impropre. Ceci peut être obtenu aevc une loi *a priori* de la forme $\mathcal{N}(\eta, I/\kappa)$ avec $\kappa \to 0$, ce qui est une forme de densité non-informative [Gelman et collab., 2013, p. 90, 392].

4.1.3. Étape maximisation : estimateur de $\pi_{c,g}$ et $\pi_{m,gj}$

Suivant une démarche identique à celle se trouvant à la section 3.1.2 :

$$\hat{\pi}_{c,g} = \frac{1}{N_c} \sum_{i=1}^{N_c} p_{c,ig}. \tag{4.1.1}$$

La contrainte pour $\pi_{m,gj}$ est $\sum_{g=1}^{K_c} \sum_{j=1}^{K_s} \pi_{m,gj} = 1$. Introduisons le multiplicateur de Lagrange λ. Il faut donc résoudre l'équation :

$$\frac{\partial}{\partial \pi_{m,gj}} \Big[Q + \lambda \Big(\sum_{g=1}^{K_c} \sum_{j=1}^{K_s} \pi_{m,gj} - 1 \Big) \Big] = 0.$$

Pour un g et j donnés, $\dfrac{\partial}{\partial \pi_{m,gj}} \displaystyle\sum_{i=1}^{N_s} \log(\pi_{m,gj}) p_{m,igj} = \sum_{i=1}^{N_s} \frac{1}{\pi_{m,gj}} p_{m,igj}$ et $\dfrac{\partial}{\partial \pi_{m,gj}} \pi_{m,gj} = 1$. Ceci nous permet d'écrire $-\lambda = \sum_{i=1}^{N_s} \dfrac{1}{\pi_{m,gj}} p_{m,igj}$. Il est donc justifié d'écrire que $-\pi_{m,gj}\lambda = \sum_{i=1}^{N_s} p_{m,igj}$. En appliquant les sommations $\sum_{g=1}^{K_c} \sum_{j=1}^{K_s}$ sur les deux côtés de la précédente égalité, on se retrouve avec le fait que $\lambda = -N_s$ et donc avec l'estimateur suivant :

$$\hat{\pi}_{m,gj} = \frac{1}{N_s} \sum_{i=1}^{N_s} p_{m,igj}. \tag{4.1.2}$$

4.1.4. Contraintes

Il a déjà été mentionné que la superposition des effets pouvait potentiellement mener à des situations où plus d'une solution serait possible et donc nuire à l'identifiabilité du modèle pour certaines données. Imposer des contraintes symétriques sur les paramètres de moyenne représentant les effets de maladie et les effets témoins permet d'éviter ce problème. Notons :

(1) $\pi_{m,g+} = \sum_{j=1}^{K_s} \pi_{m,gj}$;

(2) $\pi_{m,+j} = \sum_{g=1}^{K_c} \pi_{m,gj}$.

Il résulte de l'équation (4.1.2) que $\sum_{i=1}^{N_s} \sum_{j=1}^{K_s} p_{m,igj} = N_s \sum_{j=1}^{K_s} \hat{\pi}_{m,gj} = N_s \hat{\pi}_{m,g+}$ et $\sum_{i=1}^{N_s} \sum_{g=1}^{K_c} p_{m,igj} = N_s \sum_{g=1}^{K_c} \hat{\pi}_{m,gj} = N_s \hat{\pi}_{m,+j}$. Les deux contraintes sur les moyennes des effets sont :

(1) $\sum_{j=1}^{K_s} \hat{\pi}_{m,+j}\mu_{m,j} = 0$;

(2) $\sum_{g=1}^{K_c}(N_c\hat{\pi}_{c,g} + N_s\hat{\pi}_{m,g+})\mu_{c,g} = 0$.

4.1.5. Étape maximisation : estimateur de μ_c

$\frac{\partial}{\partial\mu_c}(\log(NIW) + Q_s) = 0$ mène à la relation :
$\sum_{i=1}^{N_s}\sum_{g=1}^{K_c} p_{c,ig}\Sigma_{c,g}^{-1}(x_i - \mu_{c,g}) = \sum_{i=1}^{N_s}\sum_{g=1}^{K_c} p_{c,ig}\Sigma_{c,g}^{-1}\mu_c$. Il en résulte donc l'estimateur suivant :

$$\hat{\mu}_c = \left[\sum_{i=1}^{N_s}\sum_{g=1}^{K_c} p_{c,ig}\Sigma_{c,g}^{-1}\right]^{-1}\left[\sum_{i=1}^{N_s}\sum_{g=1}^{K_c} p_{c,ig}\Sigma_{c,g}^{-1}(x_i - \mu_{c,g})\right]. \tag{4.1.3}$$

4.1.6. Étape maximisation : estimateur de μ_m

$\frac{\partial}{\partial\mu_m}(\log(NIW) + Q_s) = 0$ mène à la relation :
$\sum_{j=1}^{K_s}\sum_{i=1}^{N_s}\sum_{g=1}^{K_c} p_{m,igj}\Sigma_{m,j}^{-1}(y_i - \mu_{c,g} - \mu_{m,j}) = \sum_{j=1}^{K_s}\sum_{i=1}^{N_s}\sum_{g=1}^{K_c} p_{m,igj}\Sigma_{m,j}^{-1}\mu_m$. L'estimateur est donc :

$$\hat{\mu}_m = \left[\sum_{j=1}^{K_s}\sum_{i=1}^{N_s}\sum_{g=1}^{K_c} p_{m,igj}\Sigma_{m,j}^{-1}\right]^{-1}\left[\sum_{j=1}^{K_s}\sum_{i=1}^{N_s}\sum_{g=1}^{K_c} p_{m,igj}\Sigma_{m,j}^{-1}(y_i - \mu_{c,g} - \mu_{m,j})\right]. \tag{4.1.4}$$

4.1.7. Étape maximisation : estimateur de $\Sigma_{c,g}$ et $\Sigma_{m,j}$

Suivant une démarche fondamentalement identique à celle qui se trouve à la section 3.2.2, nous obtenons :

$$\hat{\Sigma}_{c,g} = \frac{\Lambda_1^{-1} + \kappa_1(\mu_{c,g} - \eta_c)(\mu_{c,g} - \eta_c)^T + \sum_{i=1}^{N_c} p_{c,ig}\Psi_{c,ig}\Psi_{c,ig}^T}{N_1 + \sum_{i=1}^{N_c} p_{c,ig}}; \tag{4.1.5}$$

$$\hat{\Sigma}_{m,j} = \frac{\Lambda_2^{-1} + \kappa_2(\mu_{m,j} - \eta_m)(\mu_{m,j} - \eta_m)^T + \sum_{i=1}^{N_s}\sum_{g=1}^{K_c} p_{m,igj}\Psi_{m,igj}\Psi_{m,igj}^T}{N_2 + \sum_{i=1}^{N_s}\sum_{g=1}^{K_c} p_{m,igj}}. \tag{4.1.6}$$

4.1.8. Étape maximisation : estimateur de $\mu_{m,j}$ sous contrainte

La contrainte est $\sum_{j=1}^{K_s}\hat{\pi}_{m,+j}\mu_{m,j} = 0$, avec $\sum_{i=1}^{N_s}\sum_{g=1}^{K_c} p_{m,igj} = N_s\sum_{g=1}^{K_c}\hat{\pi}_{m,gj} = N_s\hat{\pi}_{m,+j}$, tel que résultant de l'équation (4.1.2). Pour un j donné et puisque $\Sigma_{m,j}$ est symétrique, suivant la démarche de la section 3.1.3, $\frac{\partial}{\partial\mu_{m,j}}(\log(NIW) + Q_s) = 0$ mène à la relation $\sum_{i=1}^{N_s}\sum_{g=1}^{K_c} p_{m,igj}(y_i - \mu_m - \mu_{c,g} - \mu_{m,j}) = \kappa_2(\mu_{m,j} - \eta_m)$. Nous pouvons donc noter l'estimateur sans contrainte pour un j donné :

$$\tilde{\mu}_{m,j} = \frac{\kappa_2 \eta_m + \sum_{i=1}^{N_s} \sum_{g=1}^{K_c} p_{m,igj}(y_i - \mu_m - \mu_{c,g})}{\sum_{i=1}^{N_s} \sum_{g=1}^{K_c} p_{m,igj} + \kappa_2}. \tag{4.1.7}$$

Si on pose $\tilde{\mu}_m = \sum_{j=1}^{K_s} \hat{\pi}_{m,+j} \tilde{\mu}_{m,j}$, alors l'estimateur :

$$\hat{\mu}_{m,j} = \tilde{\mu}_{m,j} - \tilde{\mu}_m \tag{4.1.8}$$

satisfait la contrainte. En effet, il suffit de multiplier l'équation par $\hat{\pi}_{m,+j}$ et d'appliquer la sommation $\sum_{j=1}^{K_s}$ pour le vérifier.

4.1.9. Étape maximisation : estimateur de $\mu_{c,g}$ sous contrainte

La contrainte est $\sum_{g=1}^{N_c}(N_c\hat{\pi}_{c,g} + N_s\hat{\pi}_{m,g+})\mu_{c,g} = 0$ avec $\sum_{i=1}^{N_s}\sum_{j=1}^{K_s} p_{m,igj} = N_s \sum_{j=1}^{K_s} \hat{\pi}_{m,gj} = N_s\hat{\pi}_{m,g+}$, tel que résultant de l'équation (4.1.2). Pour un g donné, et puisque $\Sigma_{c,g}$ est symétrique, $\dfrac{\partial}{\partial \mu_{c,g}}(\log(NIW) + Q_s) = 0$ mène à la relation
$\kappa_1\Sigma_{c,g}^{-1}\eta_c + \sum_{i=1}^{N_c} p_{c,ig}\Sigma_{c,g}^{-1}(x_i - \mu_c) + \sum_{i=1}^{N_s}\sum_{j=1}^{K_s} p_{m,igj}\Sigma_{m,j}^{-1}(y_i - \mu_m - \mu_{m,j}) = \kappa_1\Sigma_{c,g}^{-1}\mu_{c,g} + \sum_{i=1}^{N_c} p_{c,ig}\Sigma_{c,g}^{-1}\mu_{c,g} + \sum_{i=1}^{N_s}\sum_{j=1}^{K_s} p_{m,igj}\Sigma_{m,j}^{-1}\mu_{c,g}$. Nous pouvons donc noter l'estimateur sans contrainte pour un g donné :

$$\begin{aligned}\tilde{\mu}_{c,g} = &\left[\kappa_1\Sigma_{c,g}^{-1} + \Sigma_{c,g}^{-1}\sum_{i=1}^{N_c} p_{c,ig} + \sum_{j=1}^{K_s}\Sigma_{m,j}^{-1}\sum_{i=1}^{N_s} p_{m,igj}\right]^{-1} \times \\ &\left[\kappa_1\Sigma_{c,g}^{-1}\eta_c + \Sigma_{c,g}^{-1}\sum_{i=1}^{N_c} p_{c,ig}(x_i - \mu_c) + \sum_{j=1}^{K_s}\Sigma_{m,j}^{-1}\sum_{i=1}^{N_s} p_{m,igj}(y_i - \mu_m - \mu_{m,j})\right].\end{aligned} \tag{4.1.9}$$

Si on pose $\tilde{\mu}_c = \sum_{g=1}^{K_c}(N_c\hat{\pi}_{c,g} + N_s\hat{\pi}_{m,g+})\tilde{\mu}_{c,g}/(N_c + N_s)$, alors l'estimateur :

$$\hat{\mu}_{c,g} = \tilde{\mu}_{c,g} - \tilde{\mu}_c \tag{4.1.10}$$

satisfait la contrainte. En effet, il suffit de multiplier l'équation par $(N_c\hat{\pi}_{c,g} + N_s\hat{\pi}_{m,g+})$ et d'appliquer la sommation $\sum_{g=1}^{N_c}$. La vérification du résultat découle de la relation $\sum_{g=1}^{N_c}(N_c\hat{\pi}_{c,g} + N_s\hat{\pi}_{m,g+}) = N_c + N_s$.

4.1.10. Remarque sur la simultanéité des estimations

Nous remarquons que plusieurs estimateurs dépendent de la valeur d'autres estimations effectuées, théoriquement, à la même étape. Pour éviter que ceci biaise les estimateurs, nous

proposons que les estimations soient réalisées de manière séquentielles les unes après les autres mais dans un ordre aléatoirement établi à chaque itération. Une fois cet ordre établi, chaque paramètre estimé à une itération t sera estimé à l'aide des estimations obtenues à l'itération t si cela est nécessaire et possible. Si cela est nécessaire mais que les estimations de l'itération t ne sont pas encore disponibles, ce seront les estimations obtenues à l'itération $t-1$ qui seront utilisées.

4.1.11. Valeurs des hyperparamètres

Nous proposons comme valeur pour les hyperparamètres des valeurs analogues à celles utilisées par Fraley et Raftery [2007] :

(1) $\mu_4 = \dfrac{1}{N_c}\sum_{i=1}^{N_c} x_i\,;$

(2) $\mu_3 = \dfrac{1}{N_s}\sum_{i=1}^{N_s} y_i\,;$

(3) $\eta_c = \dfrac{1}{N_c}\sum_{i=1}^{N_c}(x_i - \mu_4)\,;$

(4) $\eta_m = \dfrac{1}{N_s}\sum_{i=1}^{N_s}(y_i - \mu_3)\,;$

(5) $\nu_1 = \nu_2 = b + 2\,;$

(6) $\kappa_i = 0.01 \;\forall\; i \in \{1{,}2\}\,;$

(7) $\Lambda_1^{-1} = \dfrac{S_1}{K_c^{2/b}}\,;$

(8) $\Lambda_2^{-1} = \dfrac{S_2}{K_s^{2/b}}\,;$

(9) $S_1 = \dfrac{1}{N_c - 1}\sum_{i=1}^{N_c}(x_i - \mu_4)(x_i - \mu_4)^T\,;$

(10) $S_2 = \dfrac{1}{N_s - 1}\sum_{i=1}^{N_s}(y_i - \mu_3)(y_i - \mu_3)^T.$

4.1.12. Choix de K_c et K_s

Il a déjà été mentionné que l'utilisation d'un modèle probabiliste de mélange de densités nécessite de déterminer le nombre de densités le composant. Dans ce cas-ci, il faut déterminer K_c et K_s. Tout comme dans le cas simple, on peut répéter l'utilisation de l'algorithme EMS et comparer les valeurs du BIC. La définition du BIC ne change pas (voir la section 3.1.6). La vraisemblance du modèle devient $\log_{\mathcal{M}}(f(x,\hat{\theta})) = \sum_{i=1}^{N_c}\sum_{g=1}^{K_c} \mathbb{1}_{z_{ig}=1}\log(\pi_{c,g}\mathcal{N}(x_i\,;\bar{\Theta}_g)) +$

$\sum_{i=1}^{N_s} \sum_{g=1}^{K_c} \sum_{j=1}^{K_s} \mathbb{1}_{z_{igj}=1} \log(\pi_{m,gj} \mathcal{N}(y_i \ ; \bar{\Theta}_{g,j}))$. Le nombre de paramètres à estimer dans le cas où chaque densité se voit attribuer sa propre matrice de covariances est $m_{\mathcal{M}} = (K_c + K_s)b(b+1)/2 + b(K_c + K_s) + (K_c K_s - 1) + (K_c - 1)$.

4.2. Initialisation des paramètres

Il a déjà été mentionné qu'une initialisation correcte était importante pour minimiser les probabilités d'échec ou de mauvaises performances de l'algorithme. Contrairement au cas où on utilise l'algorithme EM simple et où on peut souvent et raisonnablement se fier à un algorithme de regroupement en grappes tel que l'algorithme K-Moyennes pour trouver un point initial, il n'y a pas une méthode qui se présente de manière évidente pour initialiser les estimateurs de $\mu_{m,j}$. Cela est dû à la nature de la présente modélisation des données où un effet se superpose à un autre effet : si une partition était réalisée sur les données avec un algorithme usuel, disons en K_s ou $K_s \cdot K_c$ groupes, le résultat pour $\hat{\mu}_{m,j}$ serait nécessairement un mélange d'un ou de plusieurs effets de maladie avec un ou des effets de groupe témoin ($\mu_{c,g}$). Il est aussi possible que l'effet de groupe témoin soit décisif au point de masquer l'effet de maladie pour le critère de décision de l'algorithme employé lors de l'initialisation, empêchant ainsi l'obtention de points d'initialisation de qualité.

Il est donc nécessaire de proposer une méthode d'initialisation et d'évaluer sa performance. L'objectif n'est évidemment pas d'arriver à une estimation parfaite des estimateurs ; c'est une tâche réservée à l'algorithme EMS. La méthode suivante, développée de manière heuristique, permet parfois de débuter avec des estimateurs relativement proches de la solution.

Les algorithmes décrits par les pseudo-codes inclus ci-dessous (voir les algorithmes 2 et 3) fonctionnent si l'on dispose de X qui est une matrice de dimensions $N_c \times b$ contenant les données des N_c individus contrôles et de Y qui est une matrice de dimensions $N_s \times b$ contenant les données des N_s individus malades, chaque ligne représentant les données d'un seul individu (une observation). Ceci implique que le fonctionnement de ces algorithmes n'est possible qu'en présence de $\hat{\pi}_{c,g}$ de dimensions $K_c \times 1$, de $\hat{\pi}_{m,gj}$ de dimensions $K_c \times K_s$, de $\hat{\mu}_m$ et $\hat{\mu}_c$ chacun de dimensions $1 \times b$. Il faut aussi $\hat{\mu}_{c,g}$ de dimensions $K_c \times b$ et $\hat{\mu}_{m,j}$ de dimensions $K_s \times b$, c'est-à-dire que l'effet d'appartenance à une grappe parmi les g grappes témoins ou les j effets de maladie est une ligne de b dimensions dans $\hat{\mu}_{c,g}$ ou $\hat{\mu}_{m,j}$ respectivement. De plus, pour

chaque g compris dans $\{1,2\ldots,K_c\}$ et pour chaque j compris dans $\{1,2\ldots,K_s\}$, il faut une matrice de dimensions $b \times b$ pour les estimateurs $\hat{\Sigma}_{c,g}$ et $\hat{\Sigma}_{m,j}$ des paramètres de covariances. Dans les pseudo-codes, les valeurs $g^{(1)}$ et $g^{(2)}$ représentent deux indices d'itération sur les valeurs possibles de g et g_{max} désigne une ou des valeurs possibles de g.

L'algorithme d'initialisation proposé implique aussi l'utilisation d'un tableau S de dimensions $(K_c \cdot K_s) \times (K_c \cdot K_s)$, de vecteurs de dimensons $1 \times b$ (notés m dans le pseudo-code), d'un vecteur π_j de dimensions $1 \times K_s$ et de matrices $X^{(1)}$ et $Y^{(1)}$, $Y^{(2)}$, $Y^{(3)}$, $Y^{(4)}$ qui sont de mêmes dimensions que X et Y respectivement. Dans le pseudo-code, $S_{h^{(1)},*}$ désigne la ligne $h^{(1)}$ du tableau S. La valeur maximale d'une ligne de S est désignée par $s_{h^{(1)},max}$. Pour chaque valeur de $s_{h^{(1)},max}$ et chaque valeur de g (correspondant à des valeurs de $g^{(1)}$ et $g^{(2)}$), il faut calculer un score désigné par $q_{s_{h^{(1)},max},g}$.

Le principe général de cette méthode d'initialisation est d'obtenir de bons estimateurs $\hat{\pi}_{c,g}, \hat{\mu}_{c,g}, \hat{\Sigma}_{c,g}$ en utilisant l'algorithme EM sur les sujets des groupes contrôles. Ensuite, il s'agit de regrouper en $K_c \cdot K_s$ grappes les sujets malades en utilisant l'algorithme EM. Ceci devrait être possible car il n'y a que $K_c \cdot K_s$ combinaisons différentes d'effets $\mu_{c,g}$ et $\mu_{m,j}$ et ces combinaisons caractérisent et différentient, sous notre modèle, tous les sujets malades. Une fois les sujets malades séparés en $K_c \cdot K_s$ grappes, un effet estimé de groupe témoin (disons $\hat{\mu}_{c,1}$) est soustrait aux données des sujets malades composant une grappe (parmi $K_c \cdot K_s$ grappes) et un autre effet (disons $\hat{\mu}_{c,2}$) est soustrait aux données des sujets d'une autre grappe parmi ces grappes. L'objectif de cette étape est d'éliminer (à peu près) le réel effet $\mu_{c,g}$ pour certaines grappes, ne laissant apparaître à peu près que l'effet $\mu_{m,j}$. Si deux grappes de sujets malades se ressemblent (selon le critère de l'algorithme K-Moyennes) après l'élimination de différents effets $\mu_{c,g}$, une conclusion possible est qu'elles sont caractérisées par le même effet de maladie $\mu_{m,j}$. Ceci mène à augmenter le score de ressemblance (contenu dans S) entre les deux grappes. Si l'effet témoin estimé soustrait aux données des deux grappes est le même, alors il convient de pénaliser le score de ressemblance. Le raisonnement justifiant cette pénalisation est que s'il y a proximité (selon le critère de l'algorithme K-Moyennes) entre ces individus alors que les effets témoins estimés qui leur ont été retirés sont les mêmes et cela malgré le fait que la combinaison d'effets témoins et de maladie sont sensés différer entre ces observations, alors c'est possiblement que ces observations se ressemblent pour d'autres raisons que le fait d'avoir le même effet de maladie.

Le processus est répété avec toutes les grappes et toutes les combinaisons possibles de $\hat{\mu}_{c,1}$ et $\hat{\mu}_{c,2}$ ce qui mène, dans les meilleurs des cas, à identifier l'effet $\mu_{c,g}$ correspondant à chaque grappe de sujets malades. Une fois identifié, celui-ci est soustrait à toutes les données des grappes malades et l'algorithme EM peut être utilisé pour isoler les estimateurs $\hat{\pi}_{m,gj}$, $\hat{\mu}_{m,j}$ et $\hat{\Sigma}_{m,j}$. Un exemple simple de l'utilisation de l'algorithme avec des illustrations est présenté à l'annexe A.

```
1 Attribuer à μ̂_c le vecteur (1 × b) qui est la moyenne des observations de X;
2 pour chaque observation dans X faire
3     X^(1)_obs ← X_obs − μ̂_c;
4 Séparer X^(1) en K_c grappes avec l'algorithme EM;
5 pour chaque g compris entre 1 et K_c inclusivement faire
6     Attribuer à π̂_{c,g} la proportion de la grappe g;
7     Attribuer à μ̂_{c,g} la moyenne de la grappe g;
8     Attribuer à Σ̂_{c,g} la matrice de covariances de la grappe g;
```

Algorithme 2 : Initialisation pour les sujets contrôles

Dans le chapitre suivant, des données relatives à la performance de cet algorithme d'initialisation sont rapportées. Une fois l'initialisation effectuée et K_c et K_s déterminés, l'algorithme EMS peut être utilisé tel que décrit par le pseudo-code (voir l'algorithme 4).

Attribuer à $\hat{\mu}_m$ le vecteur $(1 \times b)$ qui est la moyenne des observations de Y;
pour *chaque observation dans* Y **faire**
    $Y_{obs}^{(1)} \leftarrow Y_{obs} - \hat{\mu}_m$;
Séparer $Y^{(1)}$ en $K_c \cdot K_s$ grappes à l'aide de l'algorithme EM;
pour *chaque grappe* $h^{(1)}$ *parmi les* $K_c \cdot K_s$ *grappes* **faire**
    Attribuer à $n_{h^{(1)}}$ le nombre d'observations de Y qui se trouvent dans la grappe $h^{(1)}$;
    pour *chaque observation de* $Y^{(1)}$ **faire**
        pour *chaque* $g^{(1)}$ *compris entre* 1 *et* K_c *inclusivement* **faire**
            Attribuer à $Y_{obs}^{(2)}$ la valeur $(Y_{obs}^{(1)} - \mu_{c,g^{(1)}})$;
            pour *Pour chaque* $g^{(2)}$ *compris entre* 1 *et* K_c *inclusivement* **faire**
                Attribuer à $Y_{obs}^{(3)}$ la valeur $(Y_{obs}^{(1)} - \mu_{c,g^{(2)}})$;
            pour *chaque grappe* $h^{(2)}$ *parmi les* $K_c \cdot K_s$ *grappes* **faire**
                si $h^{(1)}$ *n'est pas égal à* $h^{(2)}$ **alors**
                    Attribuer à $m_{h^{(2)}}$ le vecteur $(1 \times b)$ qui est la moyenne dans $Y^{(3)}$ des observations comprises dans la grappe $h^{(2)}$
                Séparer en $(K_c \cdot K_s - 1)$ grappes les $(K_c \cdot K_s - 1)$ moyennes $m_{h^{(2)}}$ avec l'algorithme K-Moyennes;
                pour *chaque observation de* $Y^{(2)}$ *qui est dans la grappe* $h^{(1)}$ **faire**
                    Avec l'algorithme K-Moyennes, attribuer $Y_{obs}^{(2)}$ à la grappe la plus proche;
                    si $g^{(1)}$ *n'est pas égal à* $g^{(2)}$ **alors**
                        Rajouter $1/n_{h^{(1)}}$ à $S_{h^{(1)},h^{(2)}}$ (sauvegarder information sur $g^{(1)}$ et $g^{(2)}$ pour prochaine étape);
                    sinon
                        Soustraire $1/n_{h^{(1)}}$ à $S_{h^{(1)},h^{(2)}}$ (sauvegarder information sur $g^{(1)}$ et $g^{(2)}$ pour prochaine étape);
pour *chaque* $h^{(1)}$ *parmi les* $K_c \cdot K_s$ *grappes* **faire**
    Attribuer à $s_{h^{(1)},max}$ la valeur maximale contenue dans $S_{h^{(1)},*}$
    pour *chaque* g *qui a contribué à faire augmenter* $s_{h^{(1)},max}$ **faire**
        incrémenter $q_{s_{h^{(1)},max},g}$ de 1
    Attribuer à g_{max} le ou les g maximisant tous les $q_{s_{h^{(1)},max},g}$;
    si g_{max} *est unique* **alors**
        pour *chaque observation de* $Y^{(1)}$ *qui se trouve dans la grappe* $h^{(1)}$ **faire**
            $Y_{obs}^{(4)} \leftarrow Y_{obs}^{(1)} - \mu_{c,g_{max}}$;
    sinon
        pour *chaque observation de* $Y^{(1)}$ *qui se trouve dans la grappe* $h^{(1)}$ **faire**
            choisir g au hasard parmi g_{max} possibles;
            $Y_{obs}^{(4)} \leftarrow Y_{obs}^{(1)} - \mu_{c,g}$;
    Séparer $Y^{(4)}$ en K_s grappes avec l'algorithme EM;
    pour *chaque* j *compris entre 1 et* K_s*inclusivement* **faire**
        Attribuer à $\hat{\pi}_j$ *la proportion de la grappe* j*;*
        Attribuer à $\hat{\mu}_{m,j}$ *la moyenne de la grappe* j*;*
        Attribuer à $\hat{\Sigma}_{m,j}$ *la matrice de covariances de la grappe* j
    $\hat{\pi}_{m,gj} \leftarrow \hat{\pi}_{c,g} \times \hat{\pi}_j$

Algorithme 3 : Initialisation pour les sujets malades

tant que *critère de convergence ou d'arrêt non vérifié* **faire**
  pour *chaque observation dans* X **faire**
    pour *chaque* g *compris entre* 1 *et* K_c **faire**
      Évaluer la probabilité conditionnelle que l'observation soit générée par la densité paramétrisée par $\bar{\Theta}_g$
  pour *chaque observation dans* Y **faire**
    pour *chaque* g *compris entre* 1 *et* K_c **faire**
      pour *chaque* j *compris entre* 1 *et* K_s **faire**
        Évaluer la probabilité conditionnelle que l'observation soit générée par la densité paramétrisée par $\bar{\Theta}_{g,j}$
  Établir un ordre aléatoire pour estimation de $\pi_{c,g}, \pi_{m,gj}, \mu_m, \mu_c, \mu_{c,g}, \mu_{m,j}, \Sigma_{m,j}, \Sigma_{c,g}$;
  Suivant l'ordre aléatoire établi, utiliser les probabilités pour établir de nouveaux $\hat{\pi}_{c,g}, \hat{\pi}_{m,gj}, \hat{\mu}_m, \hat{\mu}_c, \hat{\mu}_{c,g}, \hat{\mu}_{m,j}, \hat{\Sigma}_{m,j}, \hat{\Sigma}_{c,g}$ (pour chaque g et chaque j);
  Vérifier critère de convergence ou d'arrêt

Algorithme 4 : Espérance-Maximisation double

Chapitre 5

Simulations

Il est nécessaire de procéder à des simulations pour évaluer les performances de l'algorithme EMS et de l'algorithme d'initialisation proposé. Les étapes de ces simulations consistent d'abord à générer de manière aléatoire des données fictives satisfaisant les contraintes du modèle et selon des paramètres variables explicités ci-bas (voir le résumé dans le tableau 5.1). Par la suite, nous appliquerons l'algorithme d'initialisation décrit au chapitre précédent, puis l'algorithme EMS, et enfin nous évaluerons à quel point les individus sont correctement assignés à leur grappe respective. Un total de 10 échantillons sont générés pour chaque cas de figure (scénario).

Pour évaluer la difficulté du regroupement en grappes pour chaque scénario, nous utilisons un ratio de similarité proposé par Chen et collab. [2002]. Dans ce contexte, w_i désigne une quelconque observation (d'un groupe témoin ou malade), μ_{w_i} désigne la moyenne exacte de la grappe à laquelle fait partie la dite observation, N désigne le nombre total d'observations (donc $N = N_c + N_s$), K désigne le nombre total de grappes différentes (donc $K = K_c + K_c \cdot K_s$), n_k désigne le nombre d'observations dans la grappe k et μ_k désigne la moyenne exacte de la grappe k. Le ratio de similarité utilisé est défini comme $1 - \left(\frac{N}{N-1}\right)\left(\frac{H}{H+S}\right)$ où $H = \frac{1}{N}\sum_{i=1}^{N} \|w_i - \mu_{w_i}\|_2$ et où $S = \frac{1}{\sum_{k^{(1)}\neq k^{(2)}}^{K} n_{k^{(1)}} n_{k^{(2)}}} \sum_{k^{(1)}\neq k^{(2)}}^{K} n_{k^{(1)}} n_{k^{(2)}} \|\mu_{k^{(1)}} - \mu_{k^{(2)}}\|_2$. Le ratio se calcule donc avec une mesure d'homogénéité (H) des grappes basée sur la distance euclidienne entre un élément d'une grappe et le centroïde de la grappe et une mesure de séparation (S) entre les grappes basée sur la distance euclidienne entre les centroïdes des grappes.

Pour évaluer l'assignation correcte des individus à leur grappe, nous utilisons l'indice de Rand ajusté (dorénavant désigné par IRA) [Hubert et Arabie, 1985] [Pedregosa et collab.,

2011]. Cette mesure peut prendre une valeur comprise entre 0 et 1 (approximativement). Cette dernière valeur signifie une correspondance parfaite entre l'assignation aux grappes après l'estimation des paramètres résultant de l'algorithme EMS et les grappes réelles de provenance des individus. La valeur de 0 a une signification inverse, c'est-à-dire une absence de correspondance.

L'algorithme EMS peut, en théorie, fonctionner sans arrêt. Nous limitons le fonctionnement de l'algorithme EMS par les 3 évènements suivants : un échec de l'algorithme dû à l'apparition d'une matrice de covariances singulière, la réalisation de 2500 itérations ou la réalisation d'un critère de convergence défini ci-bas. Le nombre de 2500 a été choisi car l'algorithme EM est connu pour être de convergence lente et nous avons parfois observé une convergence en moins de 100 itérations, donc ce nombre maximal d'itérations semble être raisonnable pour permettre la convergence.

Finalement, soit l'itération t et $\rho^{(t)}$ une composante d'un des vecteurs constituant les estimateurs des paramètres de moyenne ou d'une matrice des covariances à l'itération t, disons $\hat{\mu}_{c,g}^{(t)}$, $\hat{\mu}_{c}^{(t)}$, $\hat{\mu}_{m,j}^{(t)}$, $\hat{\mu}_{m}^{(t)}$, $\hat{\Sigma}_{c,g}^{(t)}$ ou $\hat{\Sigma}_{m,j}^{(t)}$, alors l'algorithme est arrêté si le critère de convergence suivant se réalise :

$$\max_{\forall \rho} \left(\left| \frac{\rho^{(t)} - \rho^{(t-1)}}{\rho^{(t-1)}} \right| \right) < Seuil; \forall t \in \{t-2, t-1, t\}$$

c'est-à-dire que pour chaque composante des estimateurs, la valeur absolue de la différence relative entre la valeur de la composante lors de l'itération (t) et sa valeur lors de l'itération précédente $(t-1)$ est inférieure à un seuil (fixé à 10^{-5}) et cela pour 3 itérations consécutives.

5.1. Description des données générées

5.1.1. Scénario 1

Dans ce scénario, il y a 3 effets de maladie et 3 effets de groupe témoin répartis sur 486 individus. Une grappe malade contient 27 individus et une grappe témoin contient 81 individus, faisant en sorte qu'il y a $N_c = 243$ individus malades et $N_s = 243$ individus témoins. Les données ont $b = 20$ dimensions. Le paramètre μ_c est un effet fixé à 3 sur toutes les dimensions. De manière analogue pour les individus malades, μ_m est un effet fixé à -3.

Les effets, fixés à 3, de la première et seconde grappes témoins ($\mu_{c,g}, g \in \{1,2\}$) sont communs sur 4 dimensions et propre à chaque grappe sur 1 dimension. De plus, les deux

Tableau 5.1. Résumé des paramètres des simulations

Scénario	1	2	3	4	5	6	7	8
N_c	243	216	256	243	243	243	243	243
N_s	243	216	256	243	243	243	243	243
K_c	3	2	4	3	3	3	3	3
K_s	3	3	2	3	3	3	3	3
b	20	20	20	20	20	20	20	20
μ_c	3	1.5	5	0.27125	0.27125	5	1	0.27125
$\mu_{c,g}$	3	1.5	5	0.5425	0.5425	10	2	0.5425
μ_m	-3	-1.5	-5	-0.27125	-0.27125	-5	-1	-0.27125
$\mu_{m,j}$	-3	-1.5	-5	-0.5425	-0.5425	-10	-2	-0.5425
Variance	1	1	1	0.0401 à 0.0984	0.0401 à 0.0984	1	1	Données réelles
H	4.416	4.418	4.415	1.099	1.048	4.415	4.415	1.046
S	17.267	8.865	26.536	2.721	1.971	44.621	8.922	1.971
Ratio de similarité	0.796	0.667	0.857	0.712	0.652	0.91	0.668	0.653

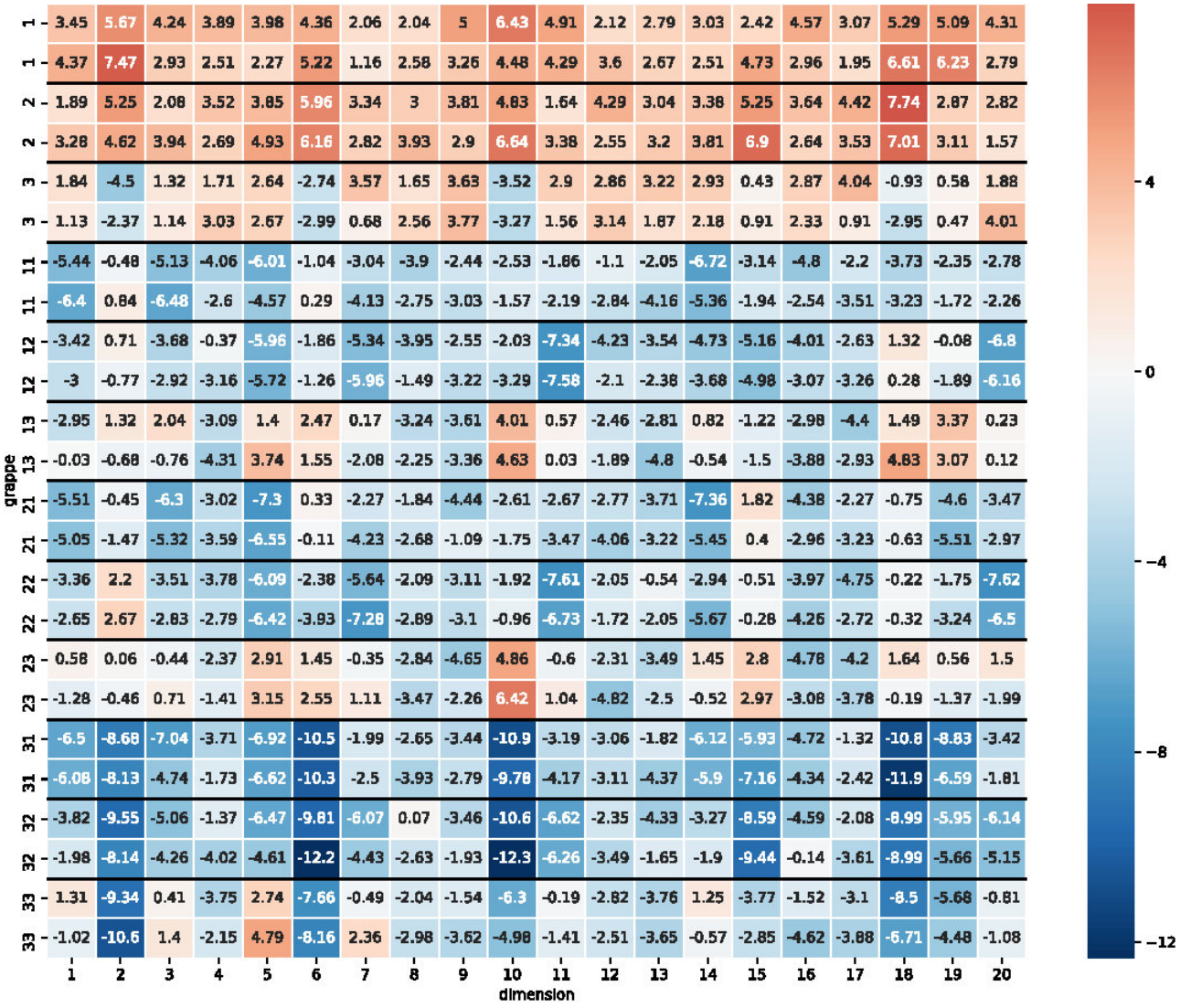
premiers effets de maladie ($\mu_{m,j}, j \in \{1,2\}$) sont fixés à -3 et se trouvent chacun sur 7 dimensions parmi les 20 dimensions possibles. Trois de ces 7 dimensions sont partagées avec un effet de groupe témoin. Seulement une des 7 dimensions est occupée par plus d'un effet de maladie. L'effet $\mu_{m,j}$ est fixé à $0{,}7 \cdot -3 = -2{,}1$ sur les dimensions où il se superpose à un effet de groupe témoin (c'est-à-dire les dimensions où $\mu_{c,g}$ a un effet différent de 0).

Pour chaque dimension, la variance («bruit») est de 1 et la matrice de covariances est diagonale. La matrice de covariances est donc la matrice identité. Le troisième effet témoin et le troisième effet de maladie sont générés de manière à ce que les contraintes du modèle soient respectées. Le résultat final est que les grappes résultantes sont assez bien délimitées les unes par rapport aux autres mais ont quand même des similarités. La figure 5.1 illustre les données de 24 de ces individus et permet de voir à quel point 2 individus qui sont sensés faire partie de la même grappe sont similaires.

5.1.2. Scénario 2

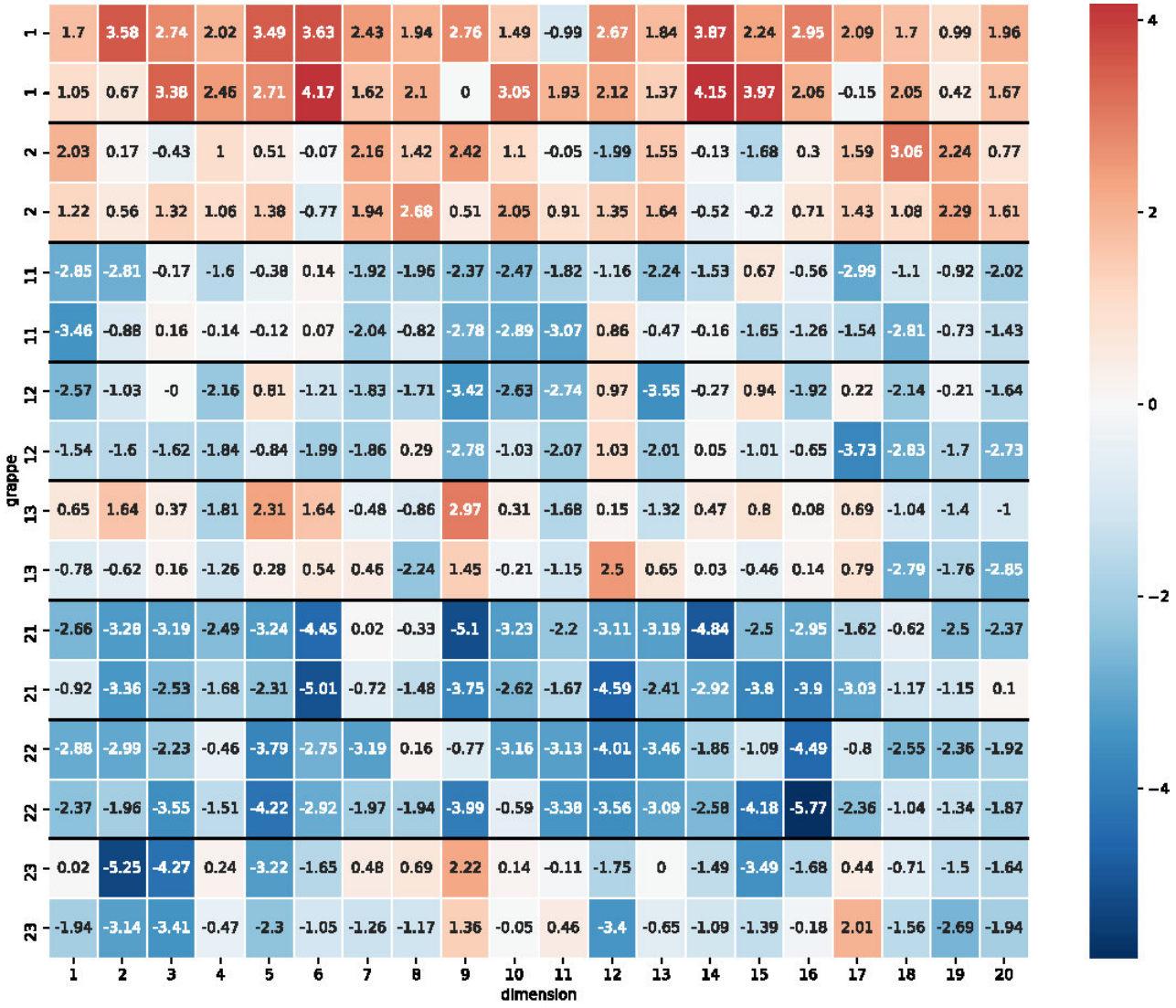
Dans ce scénario, il y a 3 effets de maladie et 2 effets de groupe témoin répartis sur 432 individus. Une grappe malade contient 36 individus et une grappe témoin contient 108 individus, faisant en sorte qu'il y a $N_c = 216$ individus malades et $N_s = 216$ individus témoins. Le paramètre μ_c est un effet fixé à 1,5 sur toutes les dimensions des données des

FIGURE 5.1. Scénario 1 : données de 24 individus, 2 individus pour chaque grappe générée



individus témoins. De manière analogue pour les individus malades, μ_m est un effet fixé à $-1{,}5$. L'effet $\mu_{c,g}$ de la première grappe témoin g est de 3 et se trouve sur 8 des 20 dimensions. L'effet de la deuxième grappe témoin est généré de manière à ce que les contraintes du modèle soient respectées. De plus, les deux premiers effets de maladie ($\mu_{m,j}, j \in \{1{,}2\}$) sont fixés à $-1{,}5$ et se trouvent chacun sur 7 dimensions parmi les 20 dimensions possibles. 3 de ces 7 dimensions sont partagées avec un effet de groupe témoin. Seulement une des 7 dimensions est occupée par plus d'un effet de maladie. L'effet $\mu_{m,j}$ est fixé à $0{,}7 \cdot -1{,}5 = -1{,}05$ sur les dimensions où il se superpose à un effet $\mu_{c,g}$.

FIGURE 5.2. Scénario 2 : données de 16 individus, 2 individus pour chaque grappe générée



Le reste des paramètres est identique aux paramètres du précédent scénario. Ainsi, il y a moins de grappes à estimer mais les données sont plus difficiles à distinguer les unes des autres. La figure 5.2 illustre les données de 16 de ces individus et permet de voir à quel point 2 individus qui sont sensés faire partie de la même grappe sont similaires.

5.1.3. Scénario 3

Dans ce scénario, il y a 2 effets de maladie et 4 effets de groupe témoin répartis sur 512 individus. Une grappe malade contient 32 individus et une grappe témoin contient 64 individus, faisant en sorte qu'il y a $N_c = 256$ individus malades et $N_s = 256$ individus

témoins. Le paramètre μ_c est un effet fixé à 5 sur toutes les dimensions des données des individus témoins. De manière analogue pour les individus malades, μ_m est un effet fixé à -5.

Les effets des trois premières grappes témoins ($\mu_{c,g}, g \in \{1,2,3\}$) sont sur 7 ou 8 dimensions avec 1 ou 2 dimensions en commun entre chaque paire d'effets. Cet effet est fixé à 5. L'effet de la quatrième grappe témoin est généré de manière à ce que les contraintes du modèle soient respectées. Le premier effet de maladie se trouve sur 3 des 20 dimensions mais sur 2 de ces dimensions se trouve un effet de groupe témoin. Quand un effet de maladie se superpose sur une dimension à un effet de groupe témoin, il est fixé à $-3{,}5$. Sinon, il est fixé à -5. L'effet du deuxième effet de maladie est fixé de manière à ce que soient respectées les contraintes du modèle.

Le reste des paramètres est identique aux paramètres du précédent scénario. Dans ce cas-ci, les deux effets de maladie devraient être plus difficiles à estimer car ils se trouvent sur un nombre restreint de dimensions et se superposent largement à des effets témoins.

La figure 5.3 illustre les données de 16 de ces individus et permet de voir à quel point 2 individus qui sont sensés faire partie de la même grappe sont similaires.

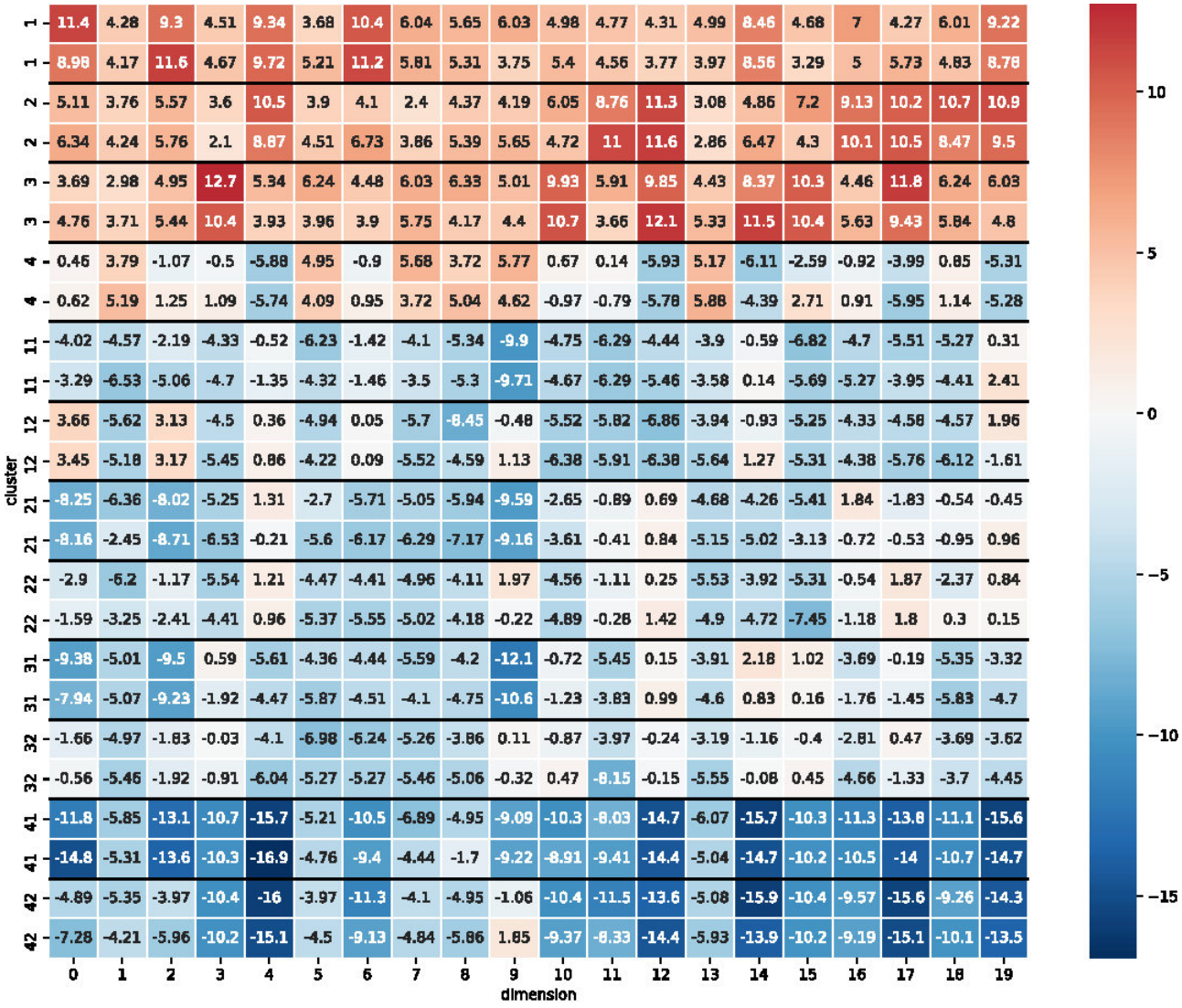
5.1.4. Scénario 4

Ce scénario est conceptuellement identique au scénario 1 sauf que la magnitude des effets est très différente et est inspirée des données à notre disposition après la réalisation des étapes de prétraitement des données (voir le chapitre 2). Les effets μ_c, μ_m, $\mu_{c,g}$ et $\mu_{m,j}$ sont fixés respectivement à 0,271258, $-0{,}271258$, 0,542515 et $-0{,}542515$. La valeur de l'effet $\mu_{c,g}$ correspond en fait à la moyenne des données. L'effet de maladie est fixé à $-0{,}379761$ dans l'éventualité où il se trouve sur la même dimension qu'un effet de groupe témoin. Les matrices de covariances sont toujours diagonales mais les variances sont comprises entre 0,04 et 0,099. Cela aussi est inspiré des données à notre disposition. La figure 5.4 illustre les données de 24 de ces individus et permet de voir à quel point 2 individus qui sont sensés faire partie de la même grappe sont similaires.

5.1.5. Scénario 5

Ce scénario est identique au précédent sauf pour certains paramètres. D'abord, les effets de groupes témoins sont chacuns sur 3 des 20 dimensions et ne se superposent jamais. Les

FIGURE 5.3. Scénario 3 : données de 24 individus, 2 individus pour chaque grappe générée

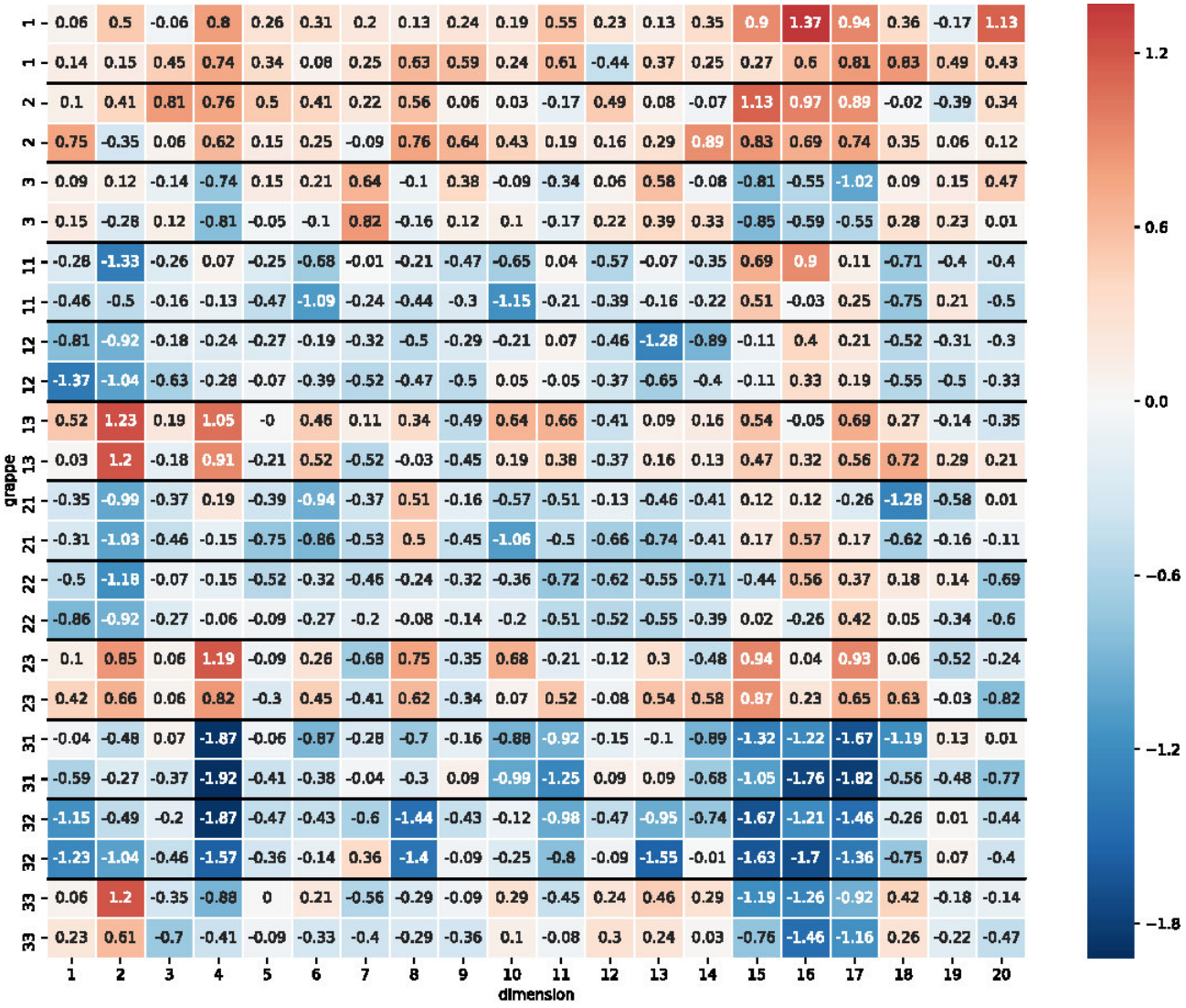


effets de maladie se trouvent sur 3 des 20 dimensions et se superposent chacun sur une seule dimension où agit un effet de groupe témoin. Il s'agit donc d'un cas où séparer les grappes les unes des autres est relativement simple. La figure 5.5 illustre les données de 24 de ces individus et permet de voir à quel point 2 individus qui sont sensés faire partie de la même grappe sont similaires.

5.1.6. Scénario 6

Ce scénario est identique au précédent en ce qui concerne la distribution des effets mais ceux-ci sont de magnitude différente. Les effets μ_c, μ_m, $\mu_{c,g}$ et $\mu_{m,j}$ sont fixés respectivement

FIGURE 5.4. Scénario 4 : données de 24 individus, 2 individus pour chaque grappe générée

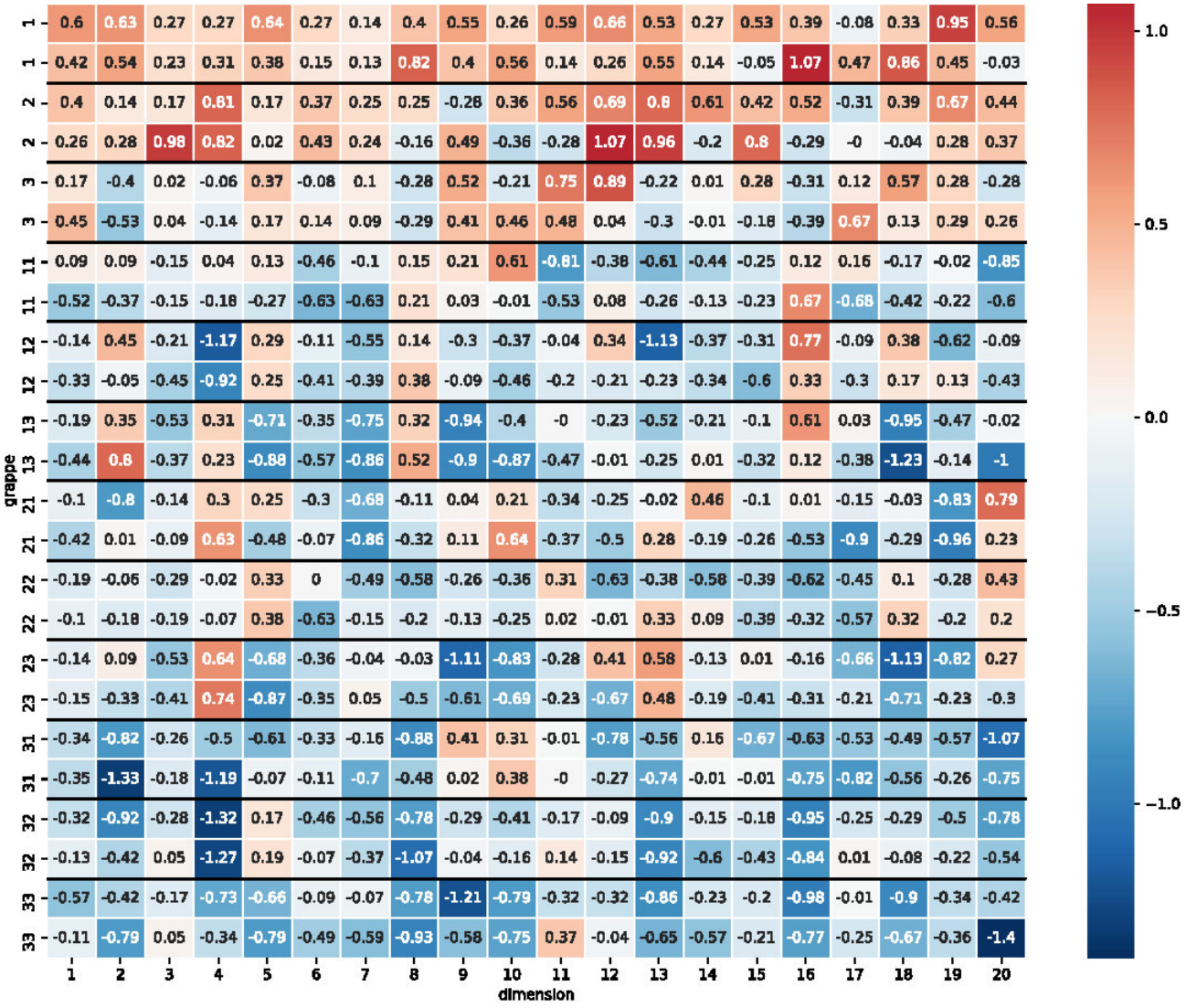


à 10, −10, 10 et 10. L'effet de maladie est fixé à −7 dans l'éventualité où il se trouve sur la même dimension qu'un effet de groupe témoin. Les matrices de covariances sont identiques à la matrice identité. Il en résulte des données très facilement distinguables les unes des autres, tel que la figure 5.6 le montre.

5.1.7. Scénario 7

Ce scénario est identique au précédent en ce qui concerne la distribution des effets mais ceux-ci sont de magnitude différente. Les effets μ_c, μ_m, $\mu_{c,g}$ et $\mu_{m,j}$ sont fixés respectivement à 2, −2, 2 et 2. L'effet de maladie est fixé à −1,4 dans l'éventualité où il se trouve sur la

FIGURE 5.5. Scénario 5 : données de 24 individus, 2 individus pour chaque grappe générée

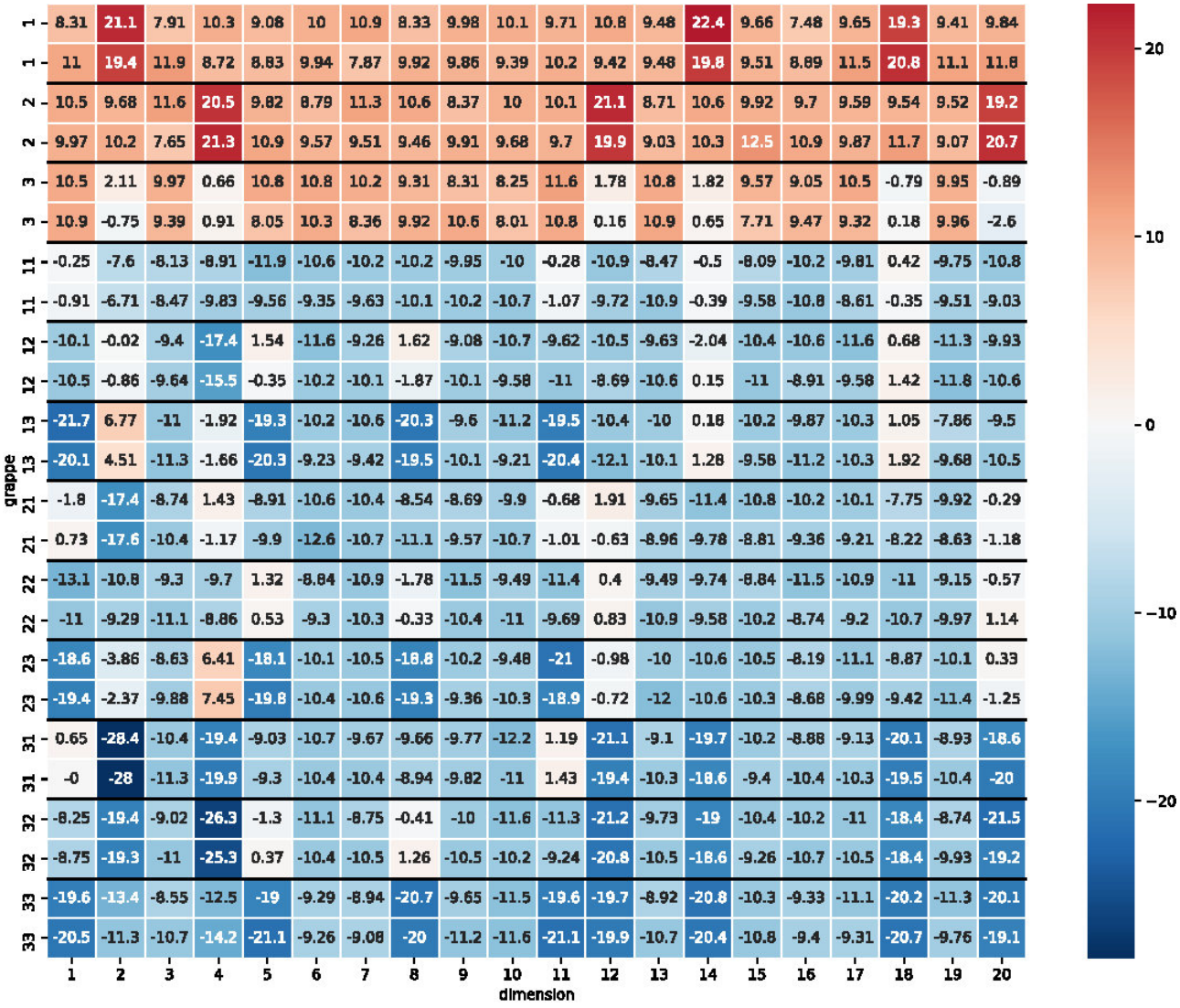


même dimension qu'un effet de groupe témoin. Les matrices de covariances sont identiques à la matrice identité. Il en résulte des données facilement distinguables les unes des autres, tel que la figure 5.7 le montre. En revanche, le fait que les effets ne soient pas aussi grands par rapport à leur variance respective pourrait rendre le regroupement en grappes et les estimations plus difficiles.

5.1.8. Scénario 8

Ce scénario est identique au scénario 4 sauf en ce qui concerne les matrices de covariances. Dans ce cas-ci, on prend le produit d'un nombre généré de manière aléatoire suivant une

FIGURE 5.6. Scénario 6 : données de 24 individus, 2 individus pour chaque grappe générée



distribution uniforme comprise entre 0,8 et 1,2 avec les 20 premières dimensions de la matrice de covariances de l'ensemble des données réelles à notre disposition. Il en résulte des données un peu plus difficiles à distinguer que celles du scénario 4 tel que la figure 5.8 l'illustre.

5.2. Résultats et commentaires

D'abord, il est important de mentionner qu'il n'y a eu aucun cas où l'algorithme a échoué. Cela signifie donc que la régularisation bayésienne employée permet bien d'éviter l'apparition de matrices singulières et cela malgré le fait qu'il y a eu des situations où l'algorithme

FIGURE 5.7. Scénario 7 : données de 24 individus, 2 individus pour chaque grappe générée

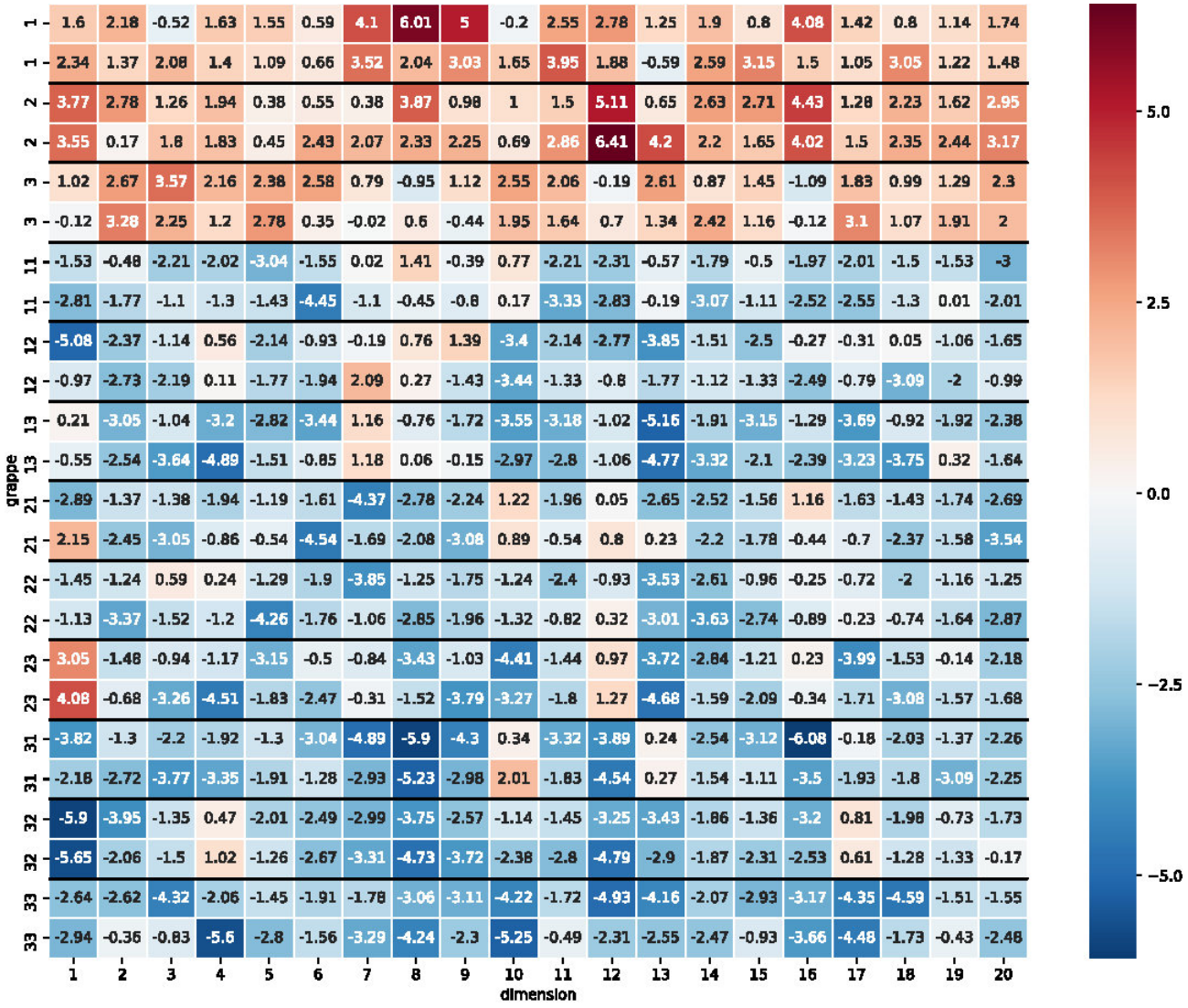
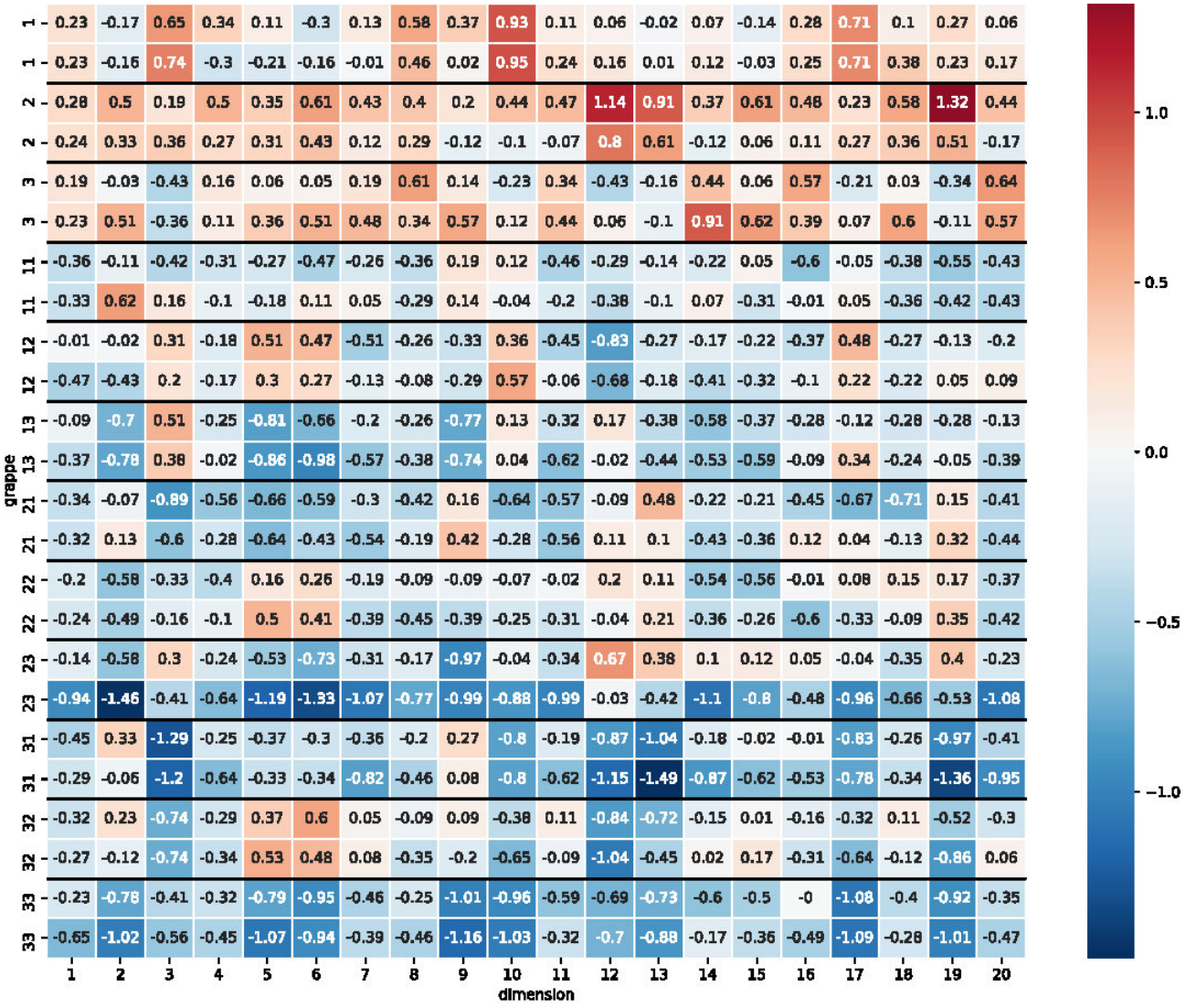


Tableau 5.2. IRA moyens pour les individus témoins

	IRA	
Scénario	Initial (écart-type)	Post-EMS (écart-type)
1	0.8 (0.23)	0.8 (0.23)
2	1.0 (0.0)	1.0 (0.0)
3	1.0 (0.0)	1.0 (0.0)
4	0.74 (0.2)	0.73 (0.19)
5	0.99 (0.0)	0.99 (0.01)
6	1.0 (0.0)	1.0 (0.0)
7	0.87 (0.21)	0.88 (0.2)
8	0.65 (0.23)	0.76 (0.26)

FIGURE 5.8. Scénario 8 : données de 24 individus, 2 individus pour chaque grappe générée



d'initialisation a très mal performé, ce qui mène potentiellement à des situations favorables à l'apparition de telles matrices.

Pour le scénario 1, où les effets sont assez prononcés mais où il y a beaucoup de superposition de ces effets, l'ensemble des procédures mènent à de bons résultats. Pour le scénario 2, les résultats sont un peu moins bons mais cela est probablement attribuable au fait que les effets de maladie sont moins distinguables. On remarque aussi que les écarts-types sont très importants, notamment pour les IRA des effets témoins des individus malades (voir le tableau 5.3). Ceci veut en fait dire que des fois, l'ensemble de la procédure d'initialisation suivi de l'utilisation de l'algorithme EMS permet d'améliorer de manière conséquente les

Tableau 5.3. IRA moyens pour les individus malades

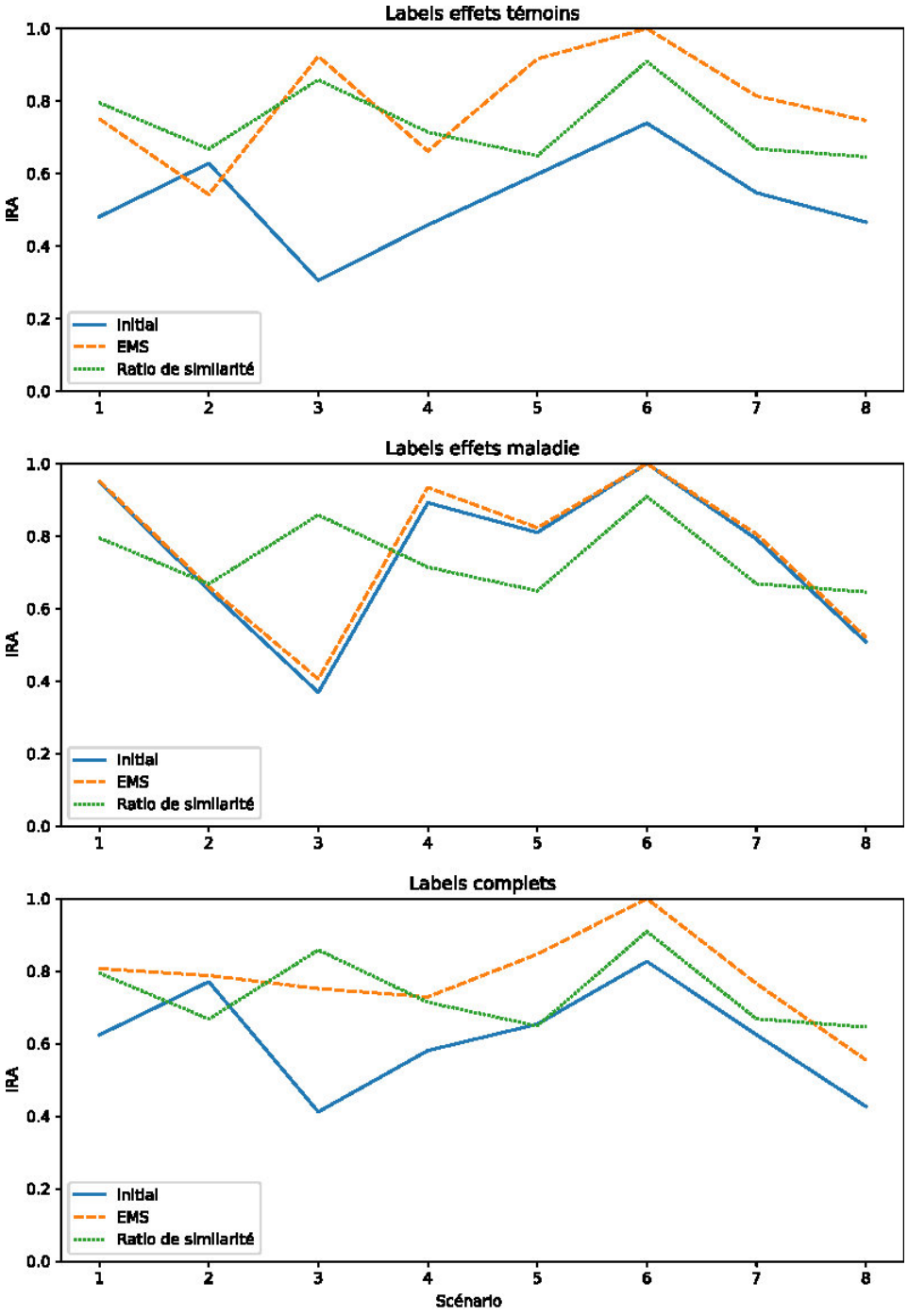
	IRA labels effets témoins seulement (écart-type)		IRA labels effets de maladie seulement (écart-type)		IRA labels complets (écart-type)	
Scénario	Initial	Post-EMS	Initial	Post-EMS	Initial	Post-EMS
1	0.48 (0.06)	0.75 (0.23)	0.95 (0.15)	0.95 (0.14)	0.62 (0.09)	0.81 (0.19)
2	0.63 (0.37)	0.54 (0.46)	0.65 (0.25)	0.66 (0.32)	0.77 (0.16)	0.79 (0.18)
3	0.31 (0.05)	0.92 (0.07)	0.37 (0.43)	0.41 (0.43)	0.41 (0.09)	0.75 (0.21)
4	0.46 (0.06)	0.66 (0.14)	0.89 (0.19)	0.93 (0.16)	0.58 (0.08)	0.73 (0.11)
5	0.6 (0.06)	0.92 (0.07)	0.81 (0.21)	0.82 (0.2)	0.65 (0.1)	0.85 (0.13)
6	0.74 (0.1)	1.0 (0.0)	1.0 (0.0)	1.0 (0.0)	0.83 (0.08)	1.0 (0.0)
7	0.55 (0.05)	0.81 (0.17)	0.79 (0.2)	0.81 (0.19)	0.63 (0.08)	0.77 (0.12)
8	0.47 (0.07)	0.75 (0.24)	0.51 (0.03)	0.52 (0.03)	0.43 (0.04)	0.56 (0.11)

estimateurs et le regroupement en grappes (et donc d'améliorer le score des IRA) mais que l'inverse se produit aussi parfois. La même remarque s'applique aux résultats du scénario 3, particulièrement pour les labels d'effets de maladie seuls. Pour le scénario 4 qui est conceptuellement identique au premier sauf en ce qui concerne la magnitude des effets, les résultats sont égaux ou meilleurs que les résultats du scénario 2. Tout naturellement, les résultats des scénarios 5, 6, 7 et 8 s'améliorent ou deviennent moins bons en fonction de l'augmentation de la difficulté relativement aux scénarios précédents.

On remarque qu'en général, le ratio de similarité semble s'accorder assez bien avec les IRA moyens sauf pour le scénario 3 qui est sensé être un cas où le regroupement en grappes est plus simple (voir figure 5.9). Dans ce dernier scénario, les procédures ont performé en moyenne moins bien mais aussi avec plus de variabilité. Les IRA pour les labels d'effets témoins se sont toutefois grandement améliorés même si les IRA initiaux sont assez médiocres. Ceci est compatible avec le fait qu'il n'y avait que 2 effets de maladie et 4 groupes témoins : il est possible que le fait qu'il y avait plus de choix pour l'assignation initiale à un groupe témoin des individus malades contribue à augmenter les possibilités de mauvaises décisions lors de l'étape d'initialisation, ce qui peut avoir de grands impacts sur les résultats pour les IRA des labels d'effets de maladie. De plus, il y a une forte superposition des effets dans ce scénario et ceci peut expliquer les résultats pour les estimateurs initiaux des effets de maladie.

La plus évidente conclusion générale de ces simulations est que les résultats sont extrêmement dépendants de la qualité de l'initialisation. En effet, on voit une très évidente corrélation

Figure 5.9. IRA moyens pour les individus malades selon le scénario



entre l'IRA initial et celui obtenu après avoir utilisé l'algorithme EMS (voir figure 5.9). De plus, l'algorithme EMS n'a amélioré que de manière très marginale les estimateurs et le regroupement en grappes relatifs aux individus des groupes témoins (voir tableau 5.2) et c'est surtout sur les estimateurs et le regroupement en grappes relatifs aux individus malades

que l'effet de l'algorithme EMS est substantiel. En d'autres termes, l'algorithme EM est utilisé pour l'initialisation des individus des groupes témoins et l'algorithme EMS n'améliore que peu ou pas ces résultats de l'algorithme EM ou de la procédure d'initialisation pour les groupes témoins. On remarque aussi que la procédure d'initialisation mène à des IRA élevés pour les labels d'effets de maladie et cela pour tous les scénarios sauf le scénario 3 où les écart-types sont élevés. En effet, dans certains cas, la procédure d'initialisation donne un IRA de 1 et parfois un IRA proche de 0. Nous pouvons donc conclure que la procédure d'initialisation a le mérite de pouvoir généralement isoler de bons estimateurs initiaux pour les effets de maladie.

Le fait que l'algorithme EMS est sensé améliorer la vraisemblance des estimations, couplé au fait que l'augmentation des IRA est attribuable au fait que les effets contrôles des individus malades sont mieux identifiés, permet de conclure que l'algorithme EMS a l'avantage, si les effets ne sont pas trop superposés ou difficiles à délier, d'améliorer l'exactitude des estimateurs des effets de maladie. Une autre observation importante est que l'amélioration des IRA pour le regroupement en grappes des individus malades est presqu'entièrement attribuable à l'amélioration de leur assignation aux bons groupes témoins. Cette assignation est généralement correcte pour la majorité des individus malades à travers les scénarios. Ceci suggère donc une nouvelle stratégie constituée de deux étapes.

5.3. Procédure en deux étapes

Les simulations pour la procédure en deux étapes consistent à générer aléatoirement des données selon les mêmes paramètres que ceux décrits dans la section précédente. L'algorithme EMS est alors employé pour tenter d'améliorer les estimateurs. Ensuite, nous avons soustrait à chaque individu malade l'effet témoin estimé de la grappe à laquelle il a été attribué par l'algorithme EMS. Les individus malades ont alors été regroupés en $K_s + 1$ (plutôt que K_s) grappes avec l'algorithme EM. Par la suite, nous avons ignoré les individus de la grappe contenant le moins d'individus, c'est-à-dire celle dont la proportion estimée $\hat{\pi}$ est la plus petite. Nous avons conservé les paramètres estimés des K_s grappes restantes comme nouveaux points de départ pour utiliser l'algorithme EMS une deuxième fois. La justification du fait d'ignorer les individus de la plus petite grappe dans notre contexte est que les résultats des simulations de la section précédente montrent qu'habituellement, une minorité d'individus se

Tableau 5.4. IRA moyens (écart-types) pour les individus malades, labels effets témoins seulement, procédure à deux étapes

Scénario	Initialisation	Post-EMS	Initialisation 2	Post-EMS 2
1	0.49 (0.04)	0.83 (0.14)	0.83 (0.14)	0.86 (0.14)
2	0.69 (0.37)	0.69 (0.37)	0.69 (0.37)	0.89 (0.16)
3	0.32 (0.06)	0.88 (0.05)	0.88 (0.05)	1.0 (0.0)
4	0.48 (0.03)	0.67 (0.12)	0.67 (0.12)	0.68 (0.13)
5	0.56 (0.08)	0.9 (0.13)	0.9 (0.13)	0.93 (0.12)
6	0.78 (0.14)	0.98 (0.05)	0.98 (0.05)	1.0 (0.0)
7	0.56 (0.04)	0.84 (0.15)	0.84 (0.15)	0.89 (0.15)
8	0.39 (0.13)	0.73 (0.16)	0.73 (0.16)	0.9 (0.16)

Tableau 5.5. IRA moyens (écart-types) pour les individus malades, labels effets de maladie seulement, procédure à deux étapes

Scénario	Initialisation	Post-EMS	Initialisation 2	Post-EMS 2
1	0.88 (0.19)	0.9 (0.2)	0.88 (0.02)	1.0 (0.01)
2	0.66 (0.25)	0.66 (0.24)	0.83 (0.03)	0.77 (0.2)
3	0.19 (0.28)	0.11 (0.3)	0.89 (0.04)	1.0 (0.0)
4	0.92 (0.17)	0.94 (0.17)	0.87 (0.01)	0.97 (0.05)
5	0.89 (0.17)	0.89 (0.17)	0.87 (0.02)	0.93 (0.07)
6	0.91 (0.19)	0.91 (0.18)	0.88 (0.02)	1.0 (0.0)
7	0.74 (0.19)	0.79 (0.19)	0.83 (0.04)	0.92 (0.04)
8	0.47 (0.1)	0.49 (0.09)	0.63 (0.15)	0.81 (0.21)

voient attribuer le mauvais label pour l'effet témoin après avoir utilisé l'algorithme EMS. Il est raisonnable de penser qu'il soit possible de regrouper plusieurs de ces individus dans une grappe ayant des estimateurs plus éloignés des paramètres réels. Les tableaux 5.4, 5.5 et 5.6 contiennent les IRA moyens pour les individus malades résultant des simulations pour les 8 scénarios. Les résultats contenus dans ces tableaux sont illustrés dans la figure 5.10. Les IRA résultant de la deuxième initialisation et de la deuxième utilisation de l'algorithme EMS sont respectivement contenus dans les colonnes intitulées «initialisation 2» et «post-EMS 2».

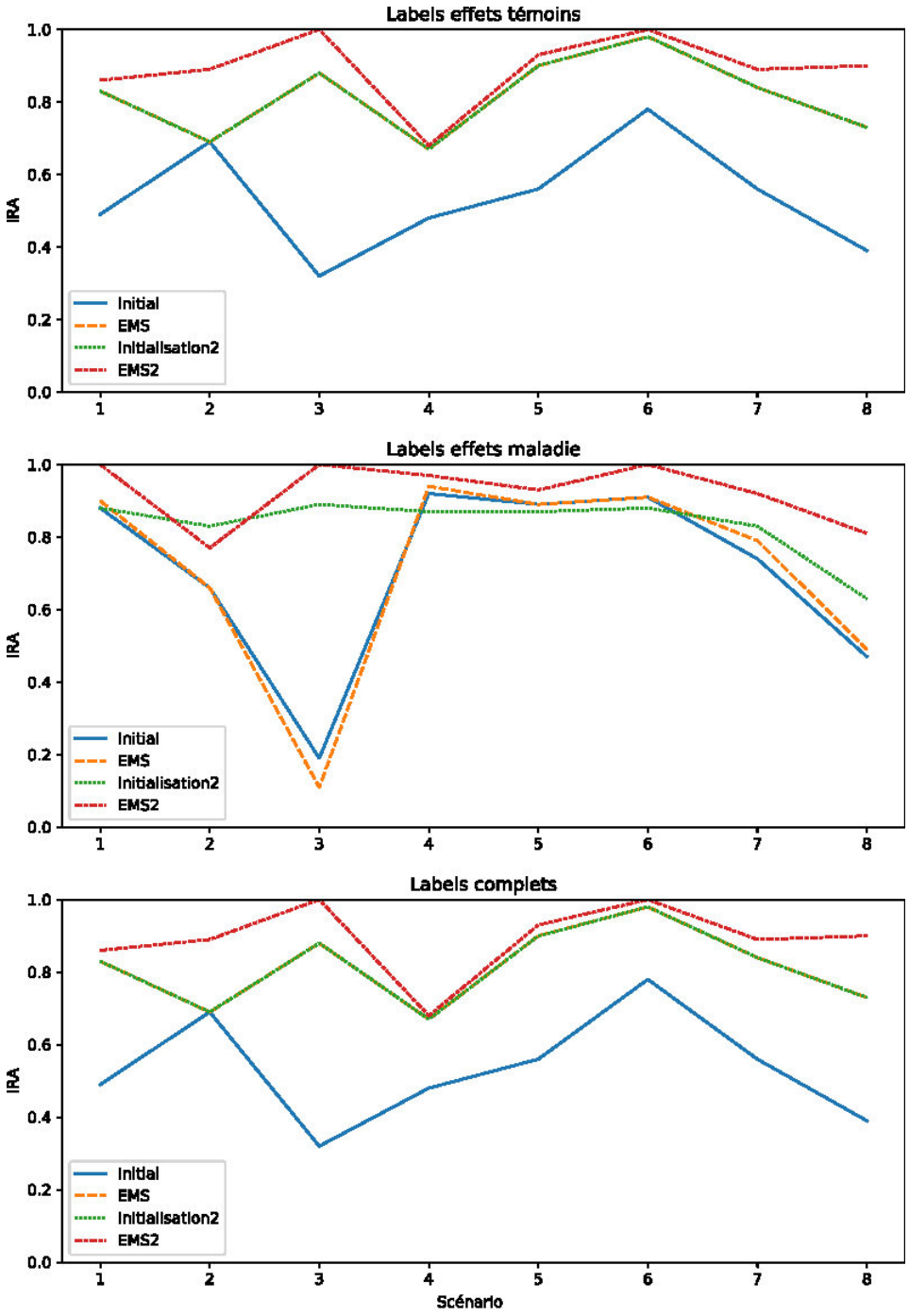
Les résultats finaux sont évidemment meilleurs que ceux obtenus après une procédure consistant seulement à initialiser les estimateurs puis à utiliser l'algorithme EMS. Il est évident que les IRA moyens à chacune des sous-étapes de la présente procédure sont fortement corrélées avec les IRA moyens de la sous-étape la précédant. On peut donc conclure qu'il est souhaitable de tenter une procédure de regroupement en grappes constituée d'au moins deux étapes. Cette procédure devrait contenir des sous-étapes d'initialisation subséquentes à la première tentative d'initialisation et qui visent à améliorer les estimateurs après avoir

Tableau 5.6. IRA moyens (écart-types) pour les individus malades, labels complets, procédure à deux étapes

Scénario	Initialisation	Post-EMS	Initialisation 2	Post-EMS 2
1	0.63 (0.07)	0.83 (0.14)	0.82 (0.06)	0.9 (0.1)
2	0.77 (0.16)	0.77 (0.16)	0.87 (0.06)	0.82 (0.15)
3	0.42 (0.05)	0.63 (0.12)	0.94 (0.03)	1.0 (0.0)
4	0.6 (0.07)	0.73 (0.1)	0.71 (0.04)	0.75 (0.08)
5	0.67 (0.11)	0.88 (0.12)	0.87 (0.06)	0.9 (0.11)
6	0.83 (0.14)	0.94 (0.13)	0.9 (0.03)	1.0 (0.0)
7	0.62 (0.11)	0.79 (0.17)	0.8 (0.07)	0.87 (0.09)
8	0.41 (0.05)	0.56 (0.13)	0.6 (0.11)	0.79 (0.18)

tenté d'identifier les individus qui ont été assignés à des groupes de manière erronée. Il est raisonnable de penser que différentes stratégies pourraient être utilisées. On peut par exemple tenter d'initialiser les estimateurs en ignorant les individus se regroupant dans une grappe marginale au regard de la proportion qu'elle occupe dans le mélange (comme réalisé ici) ou ayant une dispersion anormalement élevée ou encore ayant une moyenne estimée qui semble incompatible avec les données à l'étude. La prudence est nécessaire car il est possible que des données contiennent réellement une grappe constituée d'un petit nombre d'individus ou étant paramétrisée par des effets superposés (effet de maladie) particulièrement distincts ou encore avec une très grande variance.

Figure 5.10. IRA moyens pour les individus malades selon le scénario, procédure en 2 étapes



Chapitre 6

Résultats de l'utilisation de l'algorithme EMS sur les données à l'étude

On suppose que les données réelles à notre disposition sont modélisables selon le modèle décrit au chapitre 4. Il reste toutefois que le nombre réel de grappes pour les observations témoins (K_c) et pour les observations malades (K_s) est inconnu. L'algorithme EMS a donc été utilisé avec différentes valeurs de K_c et K_s et la valeur du BIC a été calculée pour aider à l'identification du nombre de paramètres le plus compatible avec les données (voir le tableau 6.1). La procédure d'initialisation en deux étapes a été employée telle que décrite au chapitre précédent. Il est pertinent de rapporter que l'algorithme EMS a convergé (à la deuxième étape) vers une solution en moins de 500 itérations dans la majorité des cas.

On remarque que $K_c = 2$ est la valeur qui maximise le BIC pour chaque valeur de K_s. Le BIC est maximisé pour $K_c = 2$ et $K_s = 2$ suivi du cas où $K_c = 2$ et $K_s = 3$. Les figures décrivant la valeur moyenne des symptômes (à laquelle a été soustraite la valeur moyenne pour tous les individus, chaque symptôme séparément) des individus regroupés selon l'effet de maladie pour $K_c = 2$ et pour K_s allant de 2 à 5 sont rapportées (voir figure 6.1). Il est possible de suspecter que la valeur moyenne de certains symptômes, notamment

Tableau 6.1. Valeur du BIC en fonction de K_c et K_s

K_c \ K_s	1	2	3	4	5
1	9697.801	9781.247	9631.119		
2		9901.142	9790.574	9680.669	9534.40
3		9730.328	9737.963	9438.42	
4		9596.842	9781.777	9591.47	

la désorganisation, soit inégale d'une grappe à l'autre. Pour $K_c = 2$ et $K_s = 2$, le test de Kruskal-Wallis comparant la médiane du niveau d'un seul symptôme entre les grappes retourne une valeur-p, après correction de Bonferroni pour comparaisons multiples, de 0.062 pour la désorganisation. Cette valeur-p est toutefois bien supérieure à 5% pour les autres symptômes. À titre indicatif, si nous réalisons une MANOVA comparant l'ensemble de tous les symptômes, la valeur-p obtenue est de 0,128. Pour $K_c = 2$ et $K_s = 3$, la valeur-p obtenue après réalisation d'une MANOVA est de 0,048.

La figure 6.2 illustre les mêmes différences de valeurs moyennes mais sans regrouper les malades selon l'effet de maladie. On remarque une plus grande dispersion des résultats. Le test de Kruskal-Wallis, après correction de Bonferroni pour comparaisons multiples, retourne une valeur-p de 0,035 pour les hallucinations quand $K_s = 3$ et de 0,014 pour la désorganisation quand $K_s = 4$.

La figure 6.3 illustre les mêmes différences de valeurs moyennes mais pour les résultats d'un regroupement en grappes basé sur trois modèles différents. Le premier modèle est un modèle de mélange gaussien régulier avec 2 grappes où le regroupement a été réalisé sur les individus malades seulement. Les deux autres modèles sont des modèles de mélange gaussien à effets superposés à 2 grappes pour les individus malades et 1 ou 2 grappes pour les individus sains. Dans ces trois cas, les individus ont été regroupés selon l'effet de maladie.

Selon le critère de maximisation du BIC, le modèle où $K_c = 2$ et $K_s = 2$ devrait être choisi. Il est cependant à noter qu'un modèle ne peut être choisi sans considérer sa pertinence dans un tel contexte clinique.

Figure 6.1. Différence entre les niveaux moyens des symptômes dans les grappes de malades et le niveau moyen des symptômes de tous les individus, $K_c = 2$ et K_s compris entre 2 et 5. Note : les individus ont été regroupés selon l'effet de maladie. Les nombres entre parenthèses représentent le nombre d'individus dans les grappes.

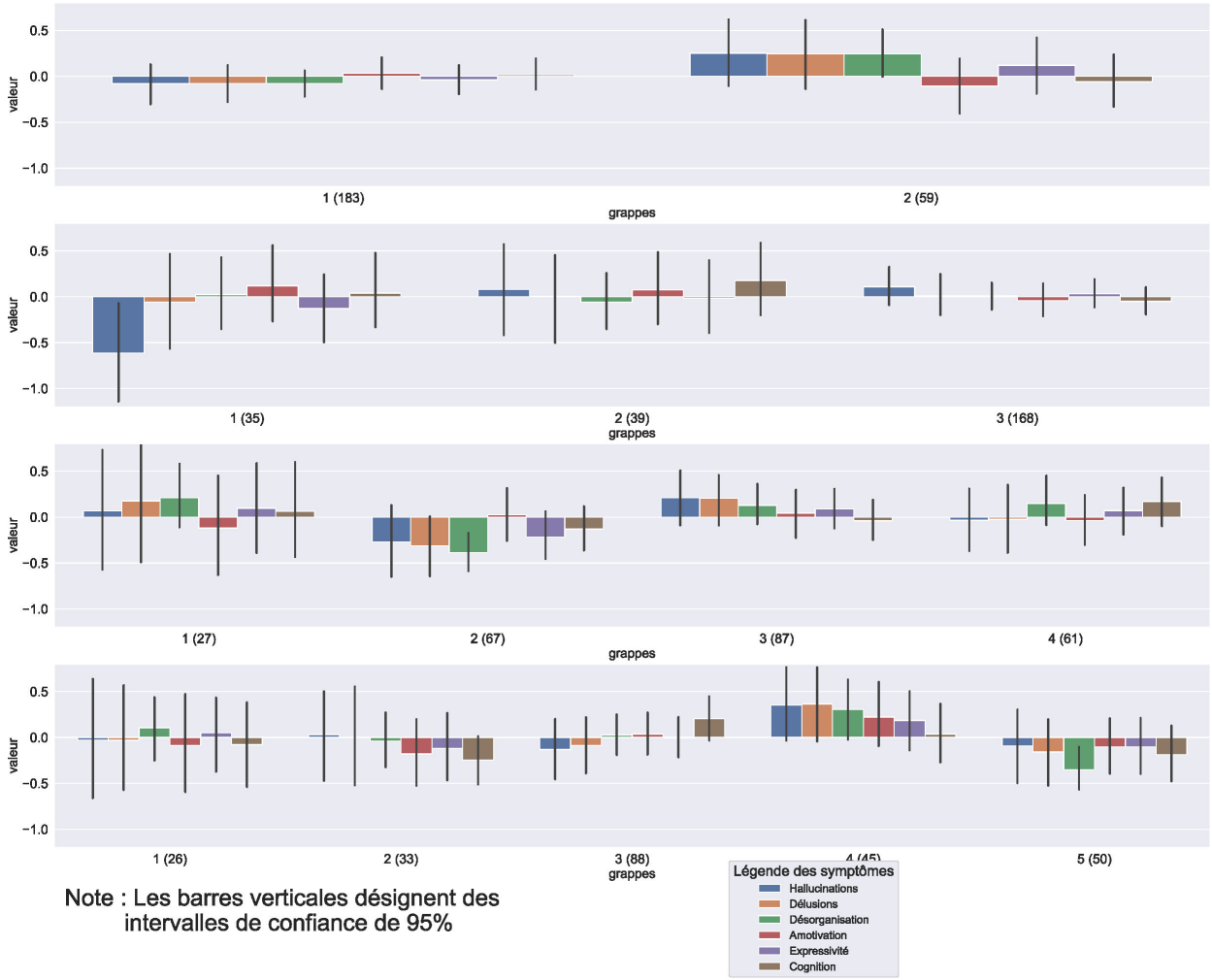
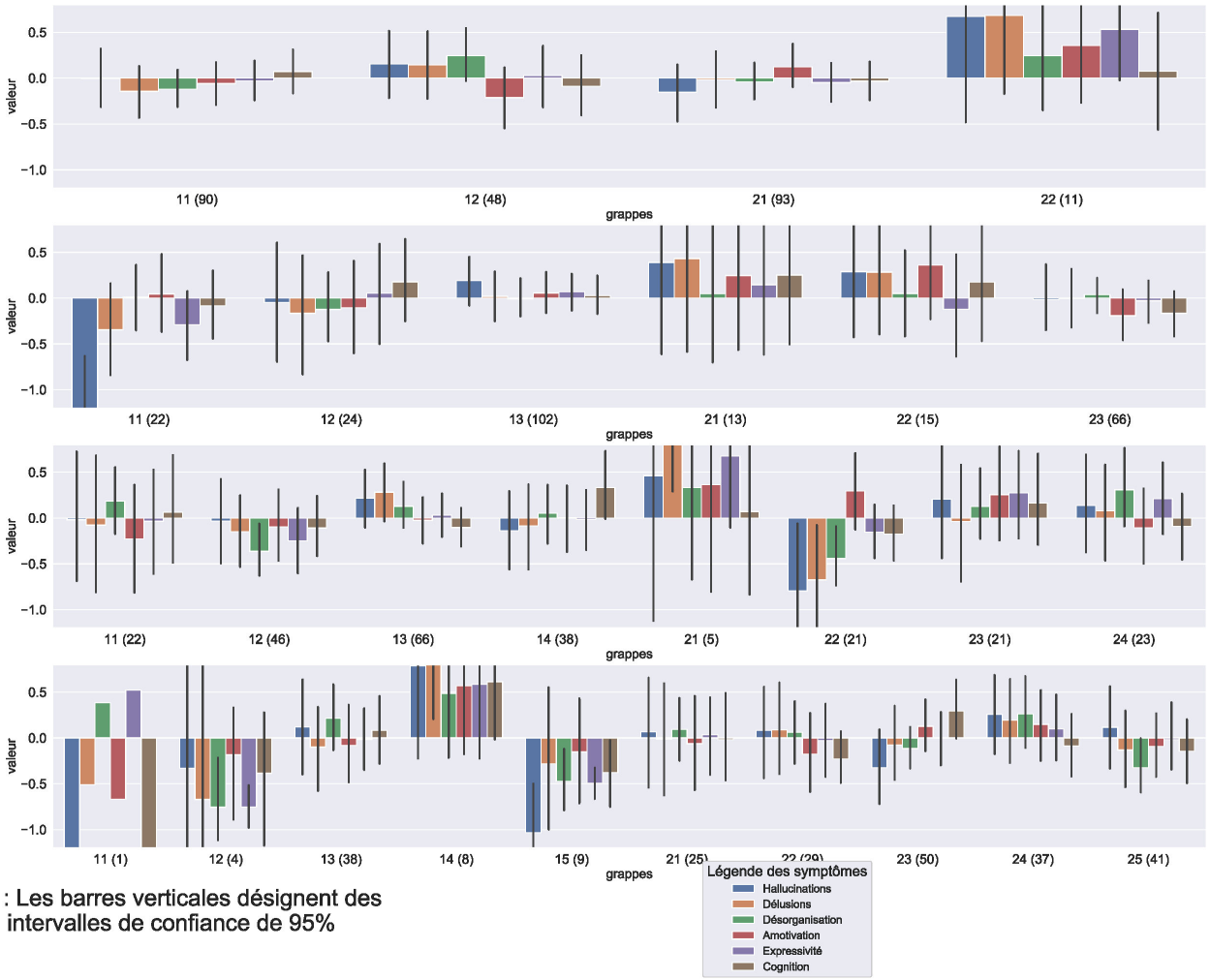
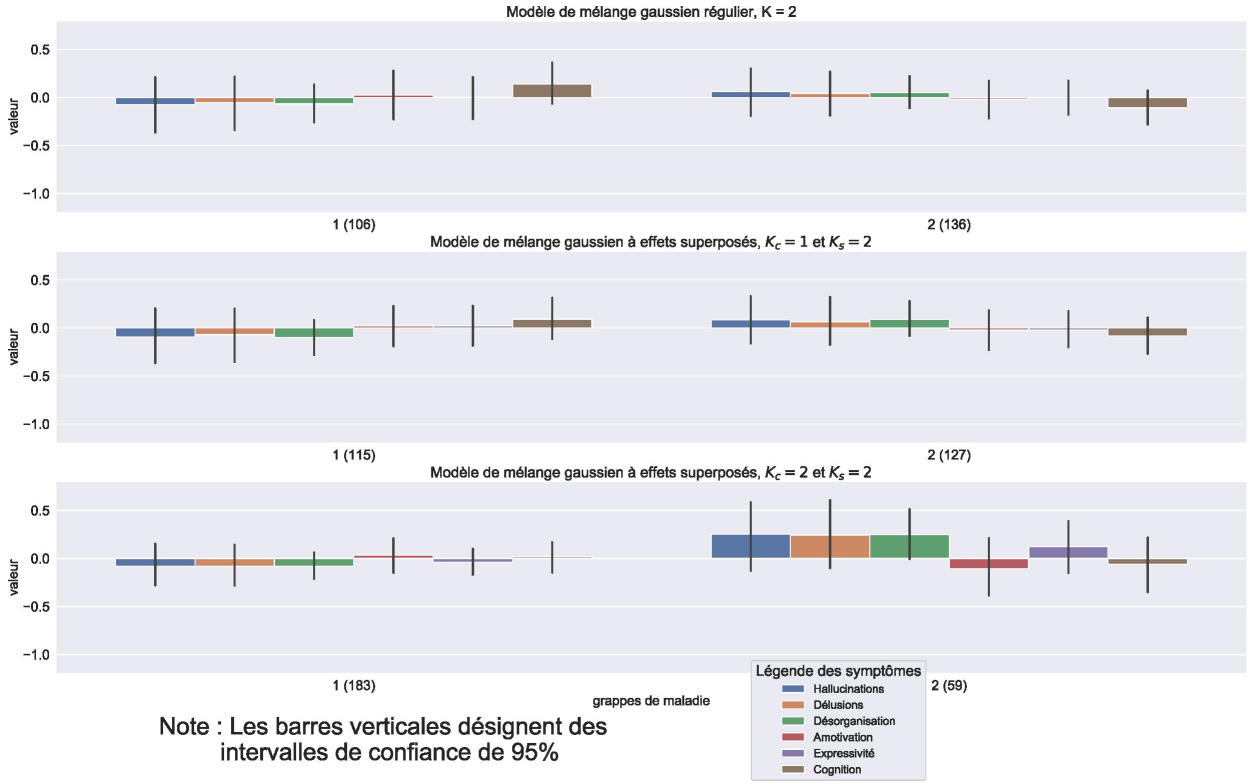


Figure 6.2. Différence entre les niveaux moyens des symptômes dans les grappes de malades et le niveau moyen des symptômes de tous les individus, $K_c = 2$ et K_s compris entre 2 et 5. Note : les individus ont été regroupés selon les $K_c \cdot K_s$ grappes du modèle. Les nombres entre parenthèses représentent le nombre d'individus dans les grappes.



Note : Les barres verticales désignent des intervalles de confiance de 95%

Figure 6.3. Différence entre les niveaux moyens des symptômes dans les grappes de malades et le niveau moyen des symptômes de tous les individus pour le modèle de mélange gaussien régulier avec 2 grappes, le modèle de mélange gaussien à effets superposés $K_c = 1$ et $K_s = 2$ et le même modèle avec $K_c = 2$ et $K_s = 2$. Les nombres entre parenthèses représentent le nombre d'individus dans les grappes.



Discussion et conclusion

La présente recherche visait à identifier des sous-groupes significativement différents de schizophrénie. Cette maladie est une entité clinique diverse, complexe et qui résiste dans une certaine mesure aux efforts des scientifiques visant à mieux la caractériser et mieux la traiter.

Notre approche était fondée sur un modèle où s'additionnaient les effets témoins sensés représenter l'hétérogénéité qui existe dans les populations humaines non-atteintes de schizophrénie aux divers effets de maladie dont nous suspectons l'existence. Ceci a mené à l'établissement d'un modèle de mélange gaussien pour des effets superposés. L'algorithme EMS fonctionne sur des données modélisables par un tel mélange et peut permettre d'identifier correctement les effets témoins et de maladie des densités constituant le mélange.

Il peut sembler que nous ne trouvions pas de grappes clairement distinguables les unes des autres avec cette méthode dans les données réelles tirées de l'IRMf à notre disposition et cela peut être dû à plusieurs facteurs. Selon certains, il n'est pas raisonnable de chercher à établir des sous-types bien distincts de problèmes psychiatriques car ceux-ci montrent une grande variété de symptômes souvent non-spécifiques à un syndrome particulier, ces symptômes étant liés à des polymorphismes génétiques divers ayant chacun un petit effet [Marquand et collab., 2019, p.128]. La présente recherche ne nous permet pas d'aller jusqu'à confirmer ce scepticisme quant au bien-fondé de la recherche visant à identifier des sous-types bien distincts. En fait, il est possible de considérer les résultats obtenus ici comme encourageants. D'une part, il semble que les données à notre disposition contiennent plus d'un groupe. Les données s'ajustent en effet mieux à un modèle contenant deux effets témoins et plusieurs effets de maladie. D'autre part, les résultats de la validation externe avec les profils de symptômes ne sont pas incompatibles avec la présence de différences significatives. Il est aussi important de souligner qu'une limitation de notre étude est que la validation externe des résultats n'a

été tentée qu'à travers la comparaison des profils de symptômes mais ceci n'est pas la seule ni nécessairement la meilleure manière de valider les résultats d'un regroupement en grappes.

Dans notre cas, un problème notable est que nos données sont une agrégation de données provenant de différents sites ayant suivi des protocles bien divers et qu'il persiste des difficultés à éliminer les effets attribuables au site de provenance et relatifs aux mouvements des sujets. Il est aussi possible que la parcellisation choisie du cerveau ou une autre étape de préparation ou de pré-traitement des données contribue à atténuer de réelles différences entre des sous-groupes.

L'initialisation peut être un problème complexe dans le cas de l'algorithme EM, mais le problème semble encore plus ardu dans le cas de l'algorithme EMS. Les simulations (voir chapitre 5) montrent que la méthode qui a été développée ici peut mener à de bons résultats dans certains cas mais qu'elle n'est pas bien adaptée à d'autres situations. Il est possible qu'une meilleure initialisation permette d'obtenir de meilleurs résultats. Toutefois, Marquand et collab. [2019, p.128] affirment que les variations dans les désordres psychiatriques se trouvent souvent dans des données à très grande dimensionnalité. Il semble évident que l'algorithme employé pourrait difficilement performer sur de telles données car la procédure d'initialisation serait potentiellement impossible à réaliser suffisamment bien pour des données trop complexes. Il pourrait être pertinent de comparer la procédure en deux étapes ici proposée (avec l'algorithme EMS) à une procédure en deux étapes similaires (par exemple en utilisant l'algorithme EM) ou encore à une procédure constituée de plusieurs étapes.

Si les connectomes étudiés avaient plus de dimensions, il est possible que certaines différences significatives soient ressorties plus clairement, mais cela aurait probablement nécessité beaucoup plus de données. La littérature scientifique récente montre par ailleurs que les données tirées de l'IRMf manquent de précision et ont une mauvaise réplicabilité. Une solution maintenant proposée est d'accumuler beaucoup plus d'observations sur chaque individu, c'est-à-dire d'augmenter drastiquement la durée des scéances d'IRMf. [1] Nous disposions de 159 volumes, en moyenne, par individu (voir chapitre 2). Il semble donc qu'une limitation sérieuse de notre recherche soit la quantité de données recueillies pour chaque individu.

1. Voir par exemple Gratton et collab. [2019].

Il y a de nombreuses voies à explorer pour trouver de meilleurs résultats. La modélisation proposée et l'algorithme développé dans la présente recherche se montreront donc peut-être plus utile sur d'autres données et pourront peut-être contribuer davantage à l'effort de recherche dans le futur.

Bibliographie

American Psychiatric Association et collab. *DSM-IV-TR, Manuel diagnostique et statistique des troubles mentaux (4 ème éd. texte révisé)*. Masson, Paris, 2001.

American Psychiatric Association et collab. *Diagnostic and statistical manual of mental disorders (DSM-5®)*. American Psychiatric Pub, 2013.

Janine Bijsterbosch, Stephen M Smith, et Christian F Beckmann. *Introduction to resting state FMRI functional connectivity*. Oxford University Press, 2017.

Johannes Blömer et Kathrin Bujna. Simple methods for initializing the em algorithm for gaussian mixture models. *CoRR*, 2013.

Felix Brandl, Mihai Avram, Benedikt Weise, Jing Shang, Beatriz Simões, Teresa Bertram, Daniel Hoffmann Ayala, Nora Penzel, Deniz A Gürsel, Josef Bäuml, et collab. Specific substantial dysconnectivity in schizophrenia : A transdiagnostic multimodal meta-analysis of resting-state functional and structural magnetic resonance imaging studies. *Biological psychiatry*, 85(7) :573–583, 2019.

Winston Chang, Joe Cheng, JJ Allaire, Yihui Xie, et Jonathan McPherson. *shiny : Web Application Framework for R*, 2019. URL `https://CRAN.R-project.org/package=shiny`. R package version 1.4.0.

Gengxin Chen, Saied A Jaradat, Nila Banerjee, Tetsuya S Tanaka, Minoru SH Ko, et Michael Q Zhang. Evaluation and comparison of clustering algorithms in analyzing es cell gene expression data. *Statistica Sinica*, pages 241–262, 2002.

Brett A Clementz, John A Sweeney, Jordan P Hamm, Elena I Ivleva, Lauren E Ethridge, Godfrey D Pearlson, Matcheri S Keshavan, et Carol A Tamminga. Identification of distinct psychosis biotypes using brain-based biomarkers. *American Journal of Psychiatry*, 173(4) : 373–384, 2015.

Christos Davatzikos, Dinggang Shen, Ruben C Gur, Xiaoying Wu, Dengfeng Liu, Yong Fan, Paul Hughett, Bruce I Turetsky, et Raquel E Gur. Whole-brain morphometric study of schizophrenia revealing a spatially complex set of focal abnormalities. *Archives of general psychiatry*, 62(11) :1218–1227, 2005.

Arthur P Dempster, Nan M Laird, et Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society : Series B (Methodological)*, 39(1) :1–22, 1977.

Sonia Dollfus, Brian Everitt, Jean Marie Ribeyre, Francoise Assouly-Besse, Charlie Sharp, et Michel Petit. Identifying subtypes of schizophrenia by cluster analyses. *Schizophrenia Bulletin*, 22(3) :545–555, 1996.

Aoyan Dong, Nicolas Honnorat, Bilwaj Gaonkar, et Christos Davatzikos. Chimera : Clustering of heterogeneous disease effects via distribution matching of imaging patterns. *IEEE transactions on medical imaging*, 35(2) :612–621, 2016.

Debo Dong, Yulin Wang, Xuebin Chang, Cheng Luo, et Dezhong Yao. Dysfunction of large-scale brain networks in schizophrenia : a meta-analysis of resting-state functional connectivity. *Schizophrenia bulletin*, 44(1) :168–181, 2017.

Andrew T Drysdale, Logan Grosenick, Jonathan Downar, Katharine Dunlop, Farrokh Mansouri, Yue Meng, Robert N Fetcho, Benjamin Zebley, Desmond J Oathes, Amit Etkin, et collab. Resting-state connectivity biomarkers define neurophysiological subtypes of depression. *Nature medicine*, 23(1) :28, 2017.

Eric Feczko, Oscar Miranda-Dominguez, Mollie Marr, Alice M. Graham, Joel T. Nigg, et Damien A. Fair. The heterogeneity problem : Approaches to identify psychiatric subtypes. *Trends in Cognitive Sciences*, xx(xx) :1–18, 2019.

Ronald A Fisher. On the probable error of a coefficient of correlation deduced from a small sample. *Metron*, 1 :3–32, 1921.

Chris Fraley et Adrian E Raftery. How many clusters ? which clustering method ? answers via model-based cluster analysis. *The computer journal*, 41(8) :578–588, 1998.

Chris Fraley et Adrian E Raftery. Bayesian regularization for normal mixture estimation and model-based clustering. Technical report, Washington Univ Seattle Dept of Statistics, 2005.

Chris Fraley et Adrian E Raftery. Bayesian regularization for normal mixture estimation and model-based clustering. *Journal of classification*, 24(2) :155–181, 2007.

Karl J Friston et Christopher D Frith. Schizophrenia : a disconnection syndrome. *Clin Neurosci*, 3(2) :89–97, 1995.

Kathleen M Gates, Peter CM Molenaar, Swathi P Iyer, Joel T Nigg, et Damien A Fair. Organizing heterogeneous samples using community detection of gimme-derived resting state functional networks. *PloS one*, 9(3) :e91322, 2014.

Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, et Donald B Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 2013.

Caterina Gratton, Brian T Kraus, Deanna J Greene, Evan M Gordon, Timothy O Laumann, Steven M Nelson, Nico UF Dosenbach, et Steven E Petersen. Defining individual-specific functional neuroanatomy for precision psychiatry. *Biological Psychiatry*, 2019.

Zhengyu Hu. *Initializing the EM algorithm for data clustering and sub-population detection*. PhD thesis, The Ohio State University, 2015.

Lawrence Hubert et Phipps Arabie. Comparing partitions. *Journal of classification*, 2(1) : 193–218, 1985.

Scott A Huettel, Allen W Song, Gregory McCarthy, et collab. *Functional magnetic resonance imaging*, volume 1. Sinauer Associates Sunderland, MA, 2004.

Eric Jones, Travis Oliphant, Pearu Peterson, et collab. SciPy : Open source scientific tools for Python, 2001–2019. URL `http://www.scipy.org/`.

Dan W Joyce, Angie A Kehagia, Derek K Tracy, Jessica Proctor, et Sukhwinder S Shergill. Realising stratified psychiatry using multidimensional signatures and trajectories. *Journal of translational medicine*, 15(1) :15, 2017.

Shitij Kapur, Anthony G Phillips, et Thomas R Insel. Why has it taken so long for biological psychiatry to develop clinical tests and what to do about it ? *Molecular psychiatry*, 17(12) : 1174, 2012.

James MacQueen et collab. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA, 1967.

Andre F Marquand, Thomas Wolfers, Maarten Mennes, Jan Buitelaar, et Christian F Beckmann. Beyond lumping and splitting : a review of computational approaches for stratifying

psychiatric disorders. *Biological psychiatry : cognitive neuroscience and neuroimaging*, 1 (5) :433–447, 2016.

Andre F Marquand, Thomas Wolfers, et Richard Dinga. Phenomapping : Methods and measures for deconstructing diagnosis in psychiatry. In *Personalized Psychiatry*, pages 119–134. Springer, 2019.

GJ McLachlan et D Peel. Finite mixture models', wiley series in probability and statistics, a wiley-interscience publication. 2000.

Paul D McNicholas. Model-based clustering. *Journal of Classification*, 33(3) :331–373, 2016.

Volodymyr Melnykov, Ranjan Maitra, et collab. Finite mixture models and model-based clustering. *Statistics Surveys*, 4 :80–116, 2010.

Michael B Miller et John Darrell Van Horn. Individual variability in brain activations associated with episodic retrieval : a role for large-scale databases. *International journal of psychophysiology*, 63(2) :205–213, 2007.

Michael B Miller, John Darrell Van Horn, George L Wolford, Todd C Handy, Monica Valsangkar-Smyth, Souheil Inati, Scott Grafton, et Michael S Gazzaniga. Extensive individual differences in brain activations associated with episodic retrieval are reliable over time. *Journal of Cognitive Neuroscience*, 14(8) :1200–1214, 2002.

Sophia Mueller, Danhong Wang, Michael D Fox, BT Thomas Yeo, Jorge Sepulcre, Mert R Sabuncu, Rebecca Shafee, Jie Lu, et Hesheng Liu. Individual variability in functional connectivity architecture of the human brain. *Neuron*, 77(3) :586–595, 2013.

Pierre Orban. *Shiny app to convert scores between the PANSS ans SAPS/SANS scales*, 2019. URL `https://github.com/pnplab/convert_app`.

Ives Cavalcante Passos, Benson Mwangi, et Flavio Kapczinski. *Personalized Psychiatry*. Springer International Publishing, 2019.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, et E. Duchesnay. Scikit-learn : Machine Learning in Python . *Journal of Machine Learning Research*, 12 :2825–2830, 2011.

Kaare Brandt Petersen, Michael Syskind Pedersen, et collab. The matrix cookbook. *Technical University of Denmark*, 2012.

William Pettersson-Yeo, Paul Allen, Stefania Benetti, Philip McGuire, et Andrea Mechelli. Dysconnectivity in schizophrenia : where are we now ? *Neuroscience & Biobehavioral Reviews*, 35(5) :1110–1124, 2011.

Russell A Poldrack. Region of interest analysis for fmri. *Social cognitive and affective neuroscience*, 2(1) :67–70, 2007.

R Core Team. *R : A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017. URL `https://www.R-project.org/`.

Tom Ronan, Zhijie Qi, et Kristen M Naegle. Avoiding common pitfalls when clustering biological data. *Sci. Signal.*, 9(432) :re6–re6, 2016.

Gideon Schwarz et collab. Estimating the dimension of a model. *The annals of statistics*, 6 (2) :461–464, 1978.

Stephen M Smith, Karla L Miller, Gholamreza Salimi-Khorshidi, Matthew Webster, Christian F Beckmann, Thomas E Nichols, Joseph D Ramsey, et Mark W Woolrich. Network modelling methods for fmri. *Neuroimage*, 54(2) :875–891, 2011.

José M Soares, Ricardo Magalhães, Pedro S Moreira, Alexandre Sousa, Edward Ganz, Adriana Sampaio, Victor Alves, Paulo Marques, et Nuno Sousa. A hitchhiker's guide to functional magnetic resonance imaging. *Frontiers in neuroscience*, 10 :515, 2016.

SR Sponheim, WG Iacono, PD Thuras, et M Beiser. Using biological indices to classify schizophrenia and other psychotic patients. *Schizophrenia Research*, 50(3) :139–150, 2001.

Carol A Tamminga, Godfrey Pearlson, Matcheri Keshavan, John Sweeney, Brett Clementz, et Gunvant Thaker. Bipolar and schizophrenia network for intermediate phenotypes : outcomes across the psychosis continuum. *Schizophrenia bulletin*, 40(Suppl_2) :S131–S137, 2014.

Sebastian Urchs, Jonathan Armoza, Yassine Benhajali, Jolène St-Aubin, Pierre Orban, et Pierre Bellec. Mist : A multi-resolution parcellation of functional brain networks. *MNI Open Research*, 1, 2017.

Nicholas T Van Dam, David O'Connor, Enitan T Marcelle, Erica J Ho, R Cameron Craddock, Russell H Tobe, Vilma Gabbay, James J Hudziak, F Xavier Castellanos, Bennett L Leventhal, et collab. Data-driven phenotypic categorization for neurobiological analyses : beyond dsm-5 labels. *Biological psychiatry*, 81(6) :484–494, 2017.

Aristotle N Voineskos, Grace R Jacobs, et Stephanie H Ameis. Neuroimaging heterogeneity in psychosis : Neurobiological underpinnings and opportunities for prognostic and therapeutic innovation. *Biological psychiatry*, 2019.

Danhong Wang, Meiling Li, Meiyun Wang, Franziska Schoeppe, Jianxun Ren, Huafu Chen, Dost Öngür, Roscoe O Brady, Justin T Baker, et Hesheng Liu. Individual-specific functional connectivity markers track dimensional and categorical features of psychotic illness. *Molecular psychiatry*, page 1, 2018.

Joe H Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301) :236–244, 1963.

D Weishaupt, B Marincek, et VD Koechli. *How does MRI work ? An introduction to the physics and function of magnetic resonance imaging. 2.* 2006.

Thomas V Wiecki, Jeffrey Poland, et Michael J Frank. Model-based cognitive neuroscience approaches to computational psychiatry : clustering and classification. *Clinical Psychological Science*, 3(3) :378–399, 2015.

CF Jeff Wu et collab. On the convergence properties of the em algorithm. *The Annals of statistics*, 11(1) :95–103, 1983.

Annexe A

Exemple d'initialisation

À des fins d'illustation de la procédure d'initialisation, nous générons des données simples (voir la figure A.5). Dans ce contexte, X désigne l'ensemble des sujets témoins et Y désigne l'ensemble des sujets malades. Notons $X_{i,*}$ la i^{me} ligne de X (donc l'ensemble des données d'un seul individu) et $X_{*,j}$ la j^{me} colonne de X. Une notation analogue est utilisée pour Y. Les matrices de covariances de chaque grappe sont identiques à la matrice identité. Les figures A.1, A.2, A.3 et A.4 représentent les paramètres de moyennes associés à chaque grappe ou à l'ensemble des individus témoins ou des individus malades.

A.1. Initialisation pour sujets témoins

En premier lieu, nous calculons $\hat{\mu}_c = \frac{1}{N_c} \sum_{i=1}^{N_c} X_{i,*}$ (voir la figure A.6). Puis, nous calculons $X^{(1)}_{i,*} = X_{i,*} - \hat{\mu}_c$ pour chaque ligne i (voir la figure A.7). L'algorithme EM est alors utilisé pour séparer $X^{(1)}$ en K_c grappes et mène à l'obtention de $\hat{\pi}_{c,g}$, $\hat{\Sigma}_{c,g}$ et $\hat{\mu}_{c,g}$ pour chaque g compris entre 1 et K_c inclusivement (voir figure A.8). Il est possible que $\hat{\mu}_{c,g}$ exhibe des labels interchangés, mais ceci n'a pas d'impacts sur les résultats finaux.

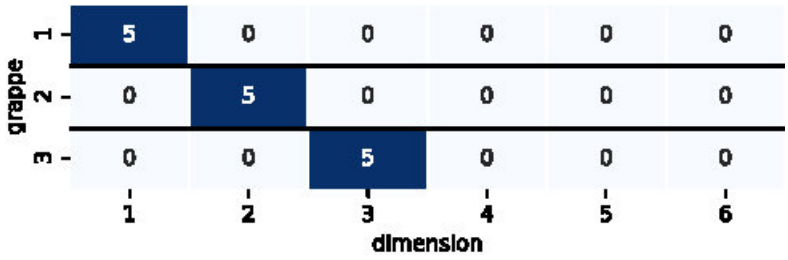


FIGURE A.1. $\mu_{c,g}$

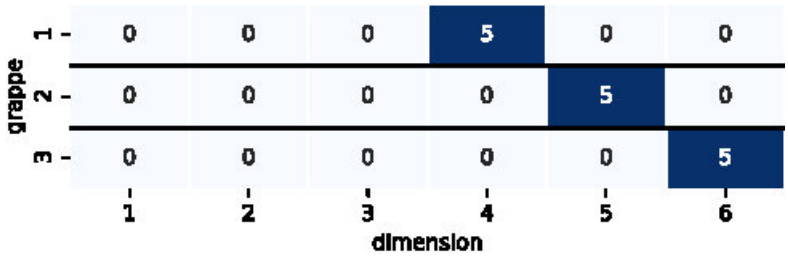


FIGURE A.2. $\mu_{m,j}$

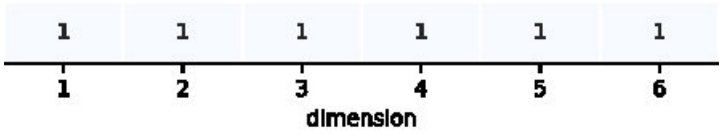


FIGURE A.3. μ_c

FIGURE A.4. μ_m

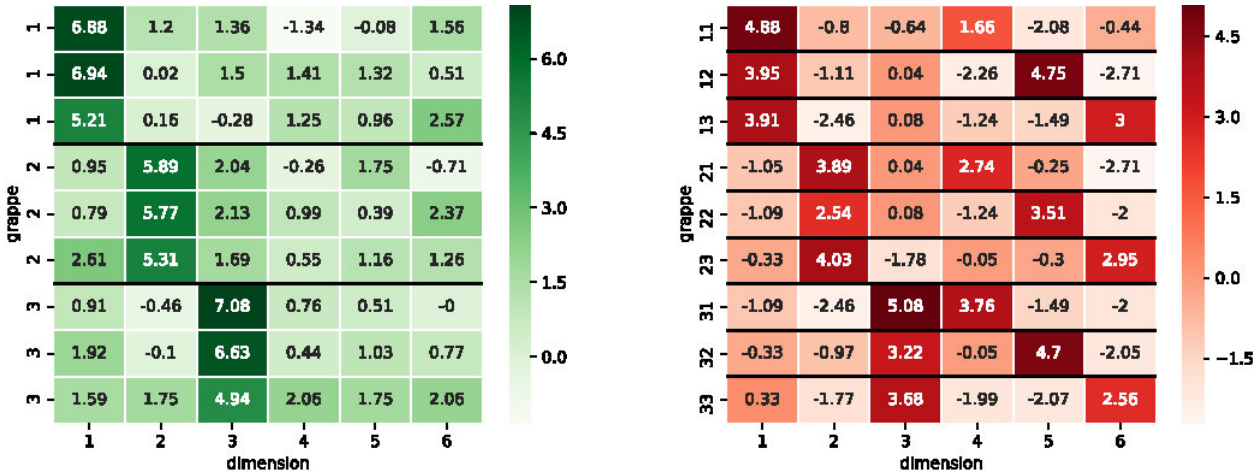


FIGURE A.5. Sujets témoins (gauche) et sujets malades (droite)

A.2. Initialisation pour sujets malades

En premier lieu, nous calculons $\hat{\mu}_m = \frac{1}{N_s} \sum_{i=1}^{N_s} Y_{i,*}$ (voir figure A.9). Puis, nous calculons $Y_{i,*}^{(1)} = Y_{i,*} - \hat{\mu}_m$ pour chaque ligne i (voir la figure A.10). L'algorithme EM est alors utilisé pour séparer $Y^{(1)}$ en $K_c \cdot K_s$ grappes et mène à l'obtention de $\hat{\Sigma}_{m,j}$. La figure A.8 représente $Y^{(1)}$ et les labels de grappes obtenus suite à l'utilisation de l'algorithme EM. Ces labels sont des valeurs que $h^{(1)}$ et $h^{(2)}$ peuvent prendre. Dans ce présent contexte, il n'y a qu'une

FIGURE A.6. $\hat{\mu}_c$

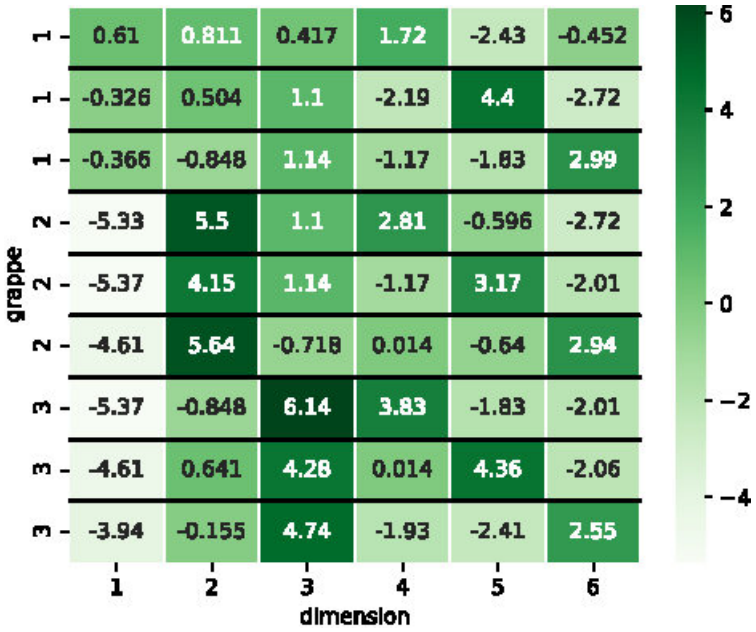


FIGURE A.7. $X^{(1)}$

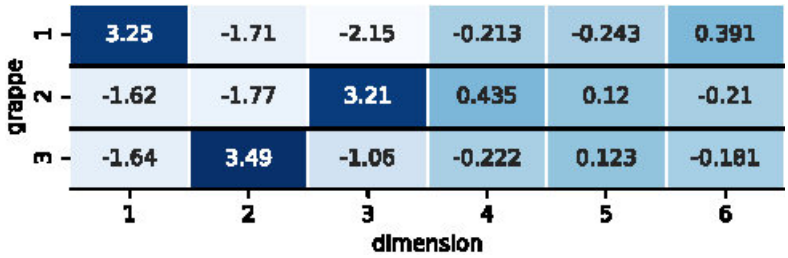


FIGURE A.8. $\hat{\mu}_{c,g}$

observation dans chacune de ces grappes et donc $n_{h^{(1)}}$ (désignant le nombre d'observations dans une grappe) vaut 1 pour toutes les valeurs possibles de $h^{(1)}$. Pour la même raison, $m_{h^{(2)}}$, qui désigne la moyenne des observations dans la grappe $h^{(2)}$, est équivalent à la seule observation se trouvant dans la grappe.

La prochaine étape consiste à choisir une valeur de g (entre 1 et K_c), disons $g^{(1)}$ et à calculer $Y^{(2)}_{i,*} = Y_{i,*} - \hat{\mu}_{c,g^{(1)}}$ pour chaque ligne i. Comme le montre la figure A.11, cette opération fait que les 3 premières lignes ne semblent différer les unes des autres que par ce qui ressemble à des effets de maladie différents. Autrement dit, $\mu_{c,g}$ et μ_c ont été presque

totalement retirés des données de ces individus et il ne reste approximativement qu'une partie des différents effets de maladie. Ensuite, on choisit une valeur de g, disons $g^{(2)}$, et on calcule $Y^{(3)}_{i,*} = Y_{i,*} - \hat{\mu}_{c,g^{(2)}}$ pour chaque ligne i. Le résultat est analogue à $Y^{(2)}$ (voir la figure A.12). Dans cette présente illustration, $g^{(1)}$ est différent de $g^{(2)}$. Ces manipulations ont pour but de faire ressortir des effets de maladie similaires après l'élimination des effets témoins. Remarquons par exemple que la ligne de la grappe 1 (figure A.12) et 6 (figure A.11) se ressemblent alors qu'elles diffèrent plus dans la figure A.10.

On sélectionne alors une grappe $h^{(1)}$, disons $h^{(1)} = 6$. Toutes les observations dans $Y^{(3)}$ (dans ce cas-ci, une seule observation) dont le label est $h^{(1)}$ sont exclues. Les observations restantes (celles dont le label n'est pas équivalent au label de la grappe $h^{(1)}$) sont divisées en $K_c \cdot K_s - 1$ grappes avec l'algorithme EM. Ces grappes sont désignées par des valeurs différentes de $h^{(2)}$. Cette étape de regroupement en grappes ne sert qu'à faire en sorte que les moyennes des observations dans ces grappes ($m_{h^{(2)}}$) sont considérées comme les centroïdes de grappes utilisés par l'algorithme K-Moyennes. Par la suite, les observations dans $Y^{(2)}$ (dans ce cas-ci, une seule observation) dont le label est $h^{(1)} = 6$ sont attribuées à la grappe $h^{(2)}$ dont le centroïde est $m_{h^{(2)}}$. Par exemple, l'algorithme K-Moyennes attribue l'observation (dans $Y^{(2)}$) dont le label est $h^{(1)} = 6$ à la grappe dont le label est $h^{(1)} = 1$.

On rajoute alors $1/n_{h^{(1)}}$ à $S_{h^{(1)},h^{(2)}} = S_{0,6}$ pour chaque observation dans $h^{(1)}$ attribuée à $h^{(2)}$ à cette étape. Il est aussi important de sauvegarder les informations pour savoir quelle combinaison de $g^{(1)}$ et $g^{(2)}$ ont mené à ce résultat. Il faut répéter toutes les étapes précédentes depuis $Y^{(1)}$ avec toutes les combinaisons possibles de $g^{(1)}$ et $g^{(2)}$. Si $g^{(1)} = g^{(2)}$, il faut pénaliser et donc soustraire $1/n_{h^{(1)}}$ à $S_{h^{(1)},h^{(2)}}$.

La figure A.13 montre la matrice S des scores de ressemblance. On peut y voir par exemple que l'observation dont le label $h^{(1)} = 8$ a été jugé similaire aux observations dont le label était $h^{(1)} = 3$ et $h^{(1)} = 4$ lorsque des effets témoins estimés différents leur ont été soustraits. L'observation dont le label $h^{(1)} = 8$ a été jugé similaire à l'observation dont le label était $h^{(1)} = 0$ quand le même effet témoin estimé leur a été soustrait. Si on se réfère aux figures A.10 et A.11, on voit que c'est un résultat cohérent avec la structure des données présentes. En effet, l'observation dont le label est $h^{(1)} = 8$ a été généré avec le même effet de maladie que les observations dont le label était $h^{(1)} = 3$ et $h^{(1)} = 4$ mais avec un effet témoin différent.

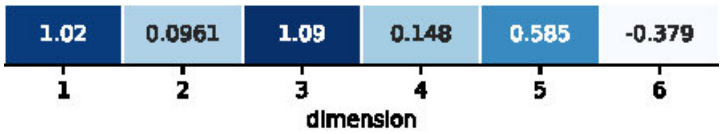


FIGURE A.9. $\hat{\mu}_m$

Donc, en retirant un effet témoin estimé différent à ces individus, la ressemblance (c'est-à-dire approximativement le même effet de maladie) entre eux devrait apparaître, menant ainsi l'algorithme K-Moyennes à attribuer une de ces observations à l'autre. En retirant le même effet témoin estimé aux observations dont le label est $h^{(1)} = 8$ et $h^{(1)} = 0$, ce qui reste est sensé être un effet de maladie différent, et donc il convient de pénaliser cette ressemblance car elle semble attribuable à autre chose qu'un effet de maladie similaire.

Une fois qu'un score de ressemblance entre les grappes est établi, c'est-à-dire que toutes les étapes ont été réalisées pour toutes les combinaisons possibles d'effets témoins et que les scores dans S ont été calculés, il faut identifier les effets témoins soustraits qui ont contribué à faire augmenter le score de ressemblance et à faire apparaître des ressemblances entre ce qui est sensé être le reste des effets de maladie. Une fois l'effet témoin estimé le plus probable identifié pour chaque observation, disons $\hat{\mu}_{c,g_{max},i}$, on calcule $Y^{(4)}_{i,*} = Y_{i,*} - \hat{\mu}_{c,g_{max},i}$ pour chaque ligne i (voir figure A.14). Dans le meilleur des cas, on a identifié correctement l'effet témoin pour chaque observation et le regroupement en grappes avec l'algorithme EM des données contenues dans $Y^{(4)}$ permet d'isoler de bons estimateurs des effets de maladie (voir figure A.15).

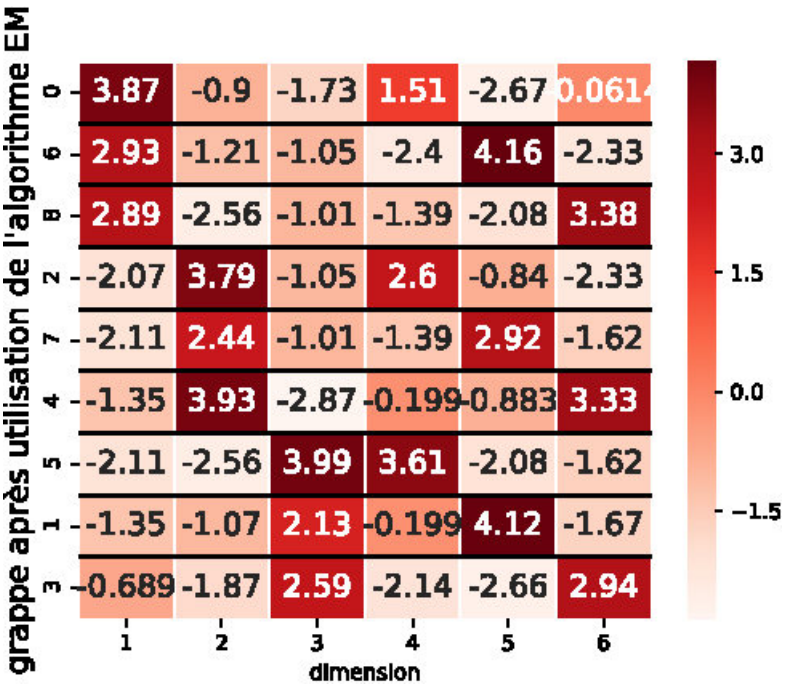


FIGURE A.10. $Y^{(1)}$

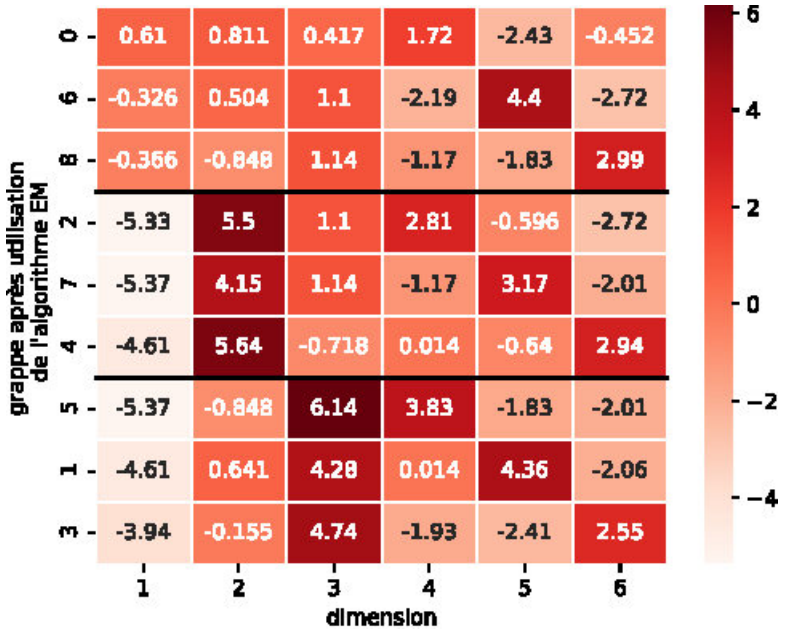


FIGURE A.11. $Y^{(2)}$

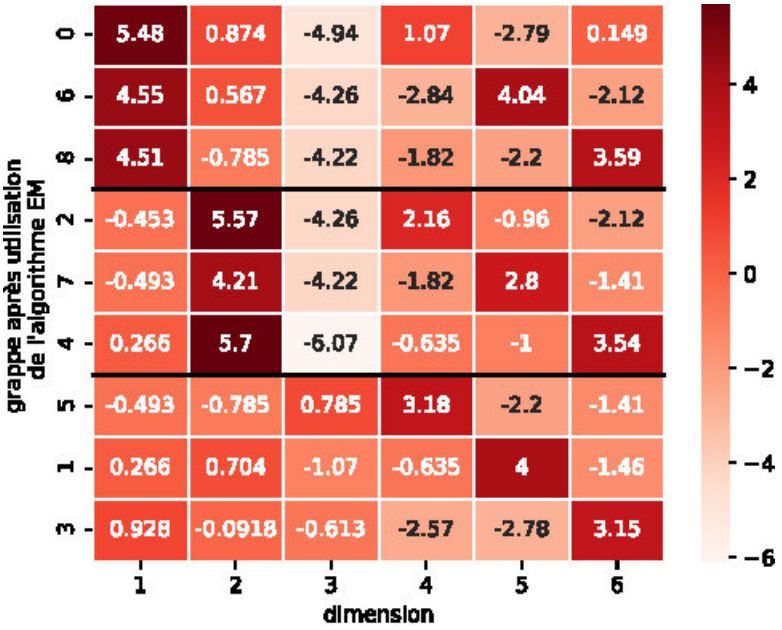


FIGURE A.12. $Y^{(3)}$

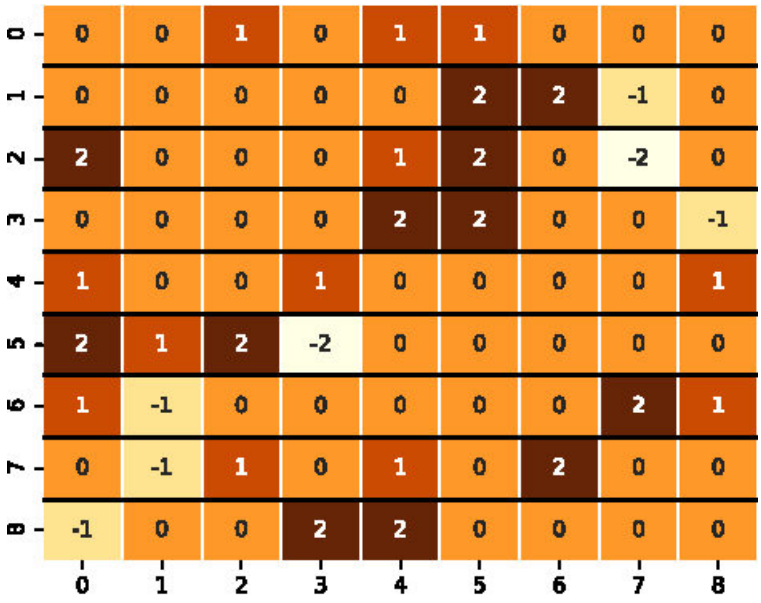


FIGURE A.13. Matrice S des scores de ressemblance

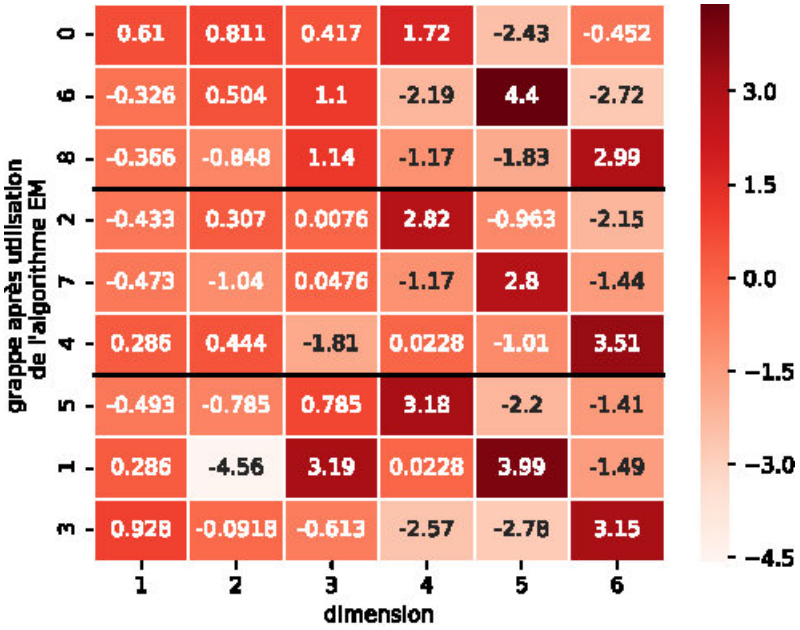


FIGURE A.14. $Y^{(4)}$

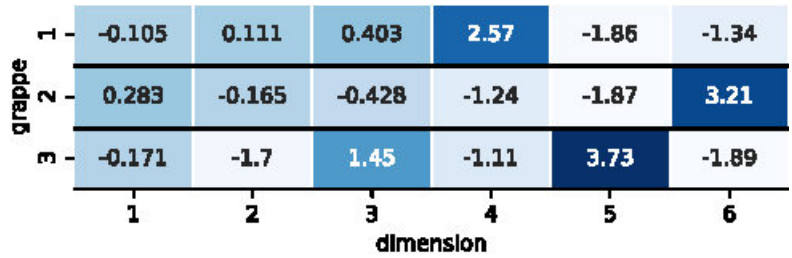


FIGURE A.15. $\hat{\mu}_{m,j}$