AIX-MARSEILLE UNIVERSITE
ECOLE DOCTORALE DE MATHEMATIQUES ET INFORMATIQUE (E.D. 184)

Thèse présentée pour obtenir le grade universitaire de docteur

Discipline : Mathématiques appliquées

Modèles de covariance pour l'analyse et la classification de signaux électroencéphalogrammes

Juliette SPINNATO

Thèse effectuée à l'I2M Sous la direction de Bruno TORRESANI et au LNC sous la direction de Boris BURLE

Florence FORBES	INRIA Grenoble Rhone-Alpes (France)	Examinatrice
Maureen CLERC	INRIA Sophia-Antipolis (France)	Rapporteure
Jean-Marc LINA	Centre de recherches mathématiques, Montréal (Canada)	Rapporteur

Résumé

Cette thèse s'inscrit dans le contexte de l'analyse et de la classification de signaux électroencéphalogrammes (EEG) par des méthodes d'analyse discriminante. Ces signaux multi-capteurs qui sont, par nature, très fortement corrélés spatialement et temporellement sont considérés dans ce travail le plan temps-fréquence. En particulier, nous nous intéressons à des signaux de type *potentiels évoqués* basse fréquence qui sont bien représentés dans l'espace des ondelettes. Par la suite, nous considérons donc les signaux représentés par des coefficients multi-échelles et qui ont une structure matricielle électrodes $\times$ coefficients.
Les signaux EEG sont de plus, considérés comme un mélange entre l'activité d'intérêt que l'on souhaite extraire et l'activité spontanée (couramment appelée "bruit de fond"), qui est largement prépondérante.

La problématique principale est ici de distinguer des signaux issus de différentes conditions expérimentales (classes). Dans le cas binaire, nous nous focalisons sur l'approche probabiliste de l'analyse discriminante et des modèles de mélange gaussien sont considérés, décrivant dans chaque classe les signaux en termes de composantes fixes (moyenne) et aléatoires. Cette dernière, caractérisée par sa matrice de covariance, permet de modéliser différentes sources de variabilité. Essentielle à la mise en œuvre de l'analyse discriminante, l'estimation de cette matrice (et de son inverse) peut être dégradée dans le cas de grandes dimensions et/ou de faibles échantillons d'apprentissage, cadre applicatif de cette thèse. Nous nous intéressons aux alternatives qui se basent sur la définition de modèle(s) de covariance(s) particulier(s) et qui permettent de réduire le nombre de paramètres à estimer.

Premièrement, une extension de l'analyse discriminante au cas matriciel permettant d'exploiter la structure matricielle des signaux est proposée. Dans ce travail, un signal EEG est modélisé à partir d'une distribution gaussienne dont la matrice de covariance spatio-temporelle, supposée séparable, se factorise sous la forme d'un produit de Kronecker. Ce modèle permet de diminuer le nombre de paramètres à estimer et de dissocier les analyses spatiales et temporelles pour l'extraction de caractéristiques discriminantes. Une formalisation du problème de l'analyse discriminante binaire sous forme d'une décomposition en valeurs singulières de la différence des moyennes des deux classes est introduite. La méthode est appliquée à deux jeux de données réels. L'analyse des composantes discriminantes extraites et leur utilisation pour un problème de classification démontre la pertinence de la méthode.

Deuxièmement, nous proposons une approche qui consiste à modéliser plus finement la variabilité inter-essais au travers d'un modèle linéaire mixte gaussien. Le signal observé est dans ce cas décomposé en une somme de deux composantes : une composante commune à tous les signaux issus d'une même condition expérimentale et une composante aléatoire modélisant la variabilité propre au signal observé dans une classe donnée. De plus, la variabilité du signal observé se décompose en deux sources : la variabilité du bruit de fond et la variabilité du signal d'intérêt. Cette modélisation de la matrice de covariance permet de réduire de façon significative le nombre de paramètres à estimer et donc, dans un contexte de classification, produit d'excellent résultats avec un ensemble d'apprentissage de taille limitée et déséquilibré. Une estimation du modèle par l'intégration de méthodes de type lasso est également étudiée. La méthode est illustrée et validée sur les deux jeux de données.

Mots-clés. Analyse discriminante, données matricielles, matrice de covariance séparable, modèle de mélange gaussien, modèle linéaire mixte, décomposition en valeurs singulières, transformation en ondelettes discrète, signaux électroencéphalogrammes.

Abstract

The present thesis finds itself within the framework of analyzing and classifying electroencephalogram signals (EEG) using discriminant analysis. Those multi-sensor signals which are, by nature, highly correlated spatially and temporally are considered, in this work, in the time-frequency domain. In particular, we focus on low-frequency *evoked-related potential*-type signals (ERPs) that are well described in the wavelet domain. Thereafter, we will consider signals represented by multi-scale coefficients and that have a matrix structure electrodes $\times$ coefficients. Moreover, EEG signals are seen as a mixture between the signal of interest that we want to extract and spontaneous activity (also called "background noise") which is overriding.

The main problematic is here to distinguish signals from different experimental conditions (class). In the binary case, we focus on the probabilistic approach of the discriminant analysis and Gaussian mixtures are used, describing in each class the signals in terms of fixed (mean) and random components. The latter, characterized by its covariance matrix, allow to model different variability sources. The estimation of this matrix (and of its inverse) is essential for the implementation of the discriminant analysis and can be deteriorated by high-dimensional data and/or by small learning samples, which is the application framework of this thesis. We are interested in alternatives that are based on specific covariance model(s) and that allow to decrease the number of parameters to estimate.

First, an extension of the discriminant analysis that allows to use the matrix structure is proposed. In this work, an EEG signal is modeled from a Gaussian distribution for which the spatio-temporal covariance matrix, assumed separable, can be decomposed as a Kronecker product. This model allows to decrease the number of parameters to estimate and dissociate spatial and temporal analysis for the extraction of discriminating features. A formulation of the binary discriminant analysis problem as a singular value decomposition of the means of the two classes is introduced. The method is applied on two real data sets. The analysis of the extracted discriminating features and their use for a classification problem shows the relevance of the model.

Second, an approach that finely models between-trial variability through a Gaussian linear mixed model is proposed. The observed signal is here decomposed as a sum of two components : a fixed part common to every signal from a same experimental condition and a random part that models the variability of the signal observed in a given class. Moreover, the variability of the observed signal can be decomposed into two sources : the background noise variability and the variability of the signal of interest. This covariance matrix modeling allows to significantly reduce the number of parameters to estimate which produces excellent results in a classification framework with small and unbalanced training set. An estimate of the model by introducing lasso-type methods is also studied. The method is illustrated and validated on the same two data sets.

Key words. Discriminant analysis, matrix-based data, separable covariance matrix, Gaussian mixtures, linear mixed model, singular value decomposition, discrete wavelet transform, electroencephalogramm signals

Table des matières

Première partie

Introduction et état de l'art : techniques pour la modélisation et la classification de signaux EEG

Chapitre 1

Introduction générale

Positionnement et contributions

L'objectif de cette thèse est de développer des modèles décrivant la variabilité de signaux neurophysiologiques électro- ou magnéto- encéphalogrammes (EEG et MEG), et des procédures et/ou algorithmes permettant d'exploiter ces modèles à des fins d'analyse de ces signaux (estimation de paramètres, décompositions en composantes associées à des phénomènes spécifiques,...). Dans le contexte applicatif considéré, les signaux sont généralement associés à des classes, et les problématiques de classification ou décision sont centrales. Ces problématiques sont donc centrales, tant dans les modèles que nous proposons que dans leur validation.

Sur l'importance de l'analyse en essai unique

Classiquement, le traitement de ces signaux fortement bruités nécessite une étape de moyennage de plusieurs répétitions (essais) produites dans des conditions identiques (par exemple, réponses consécutives à un même stimulus au cours d'une expérience, pour un ou plusieurs sujets). Cette moyenne présente l'avantage de faire ressortir les composantes pertinentes générées en réponse à des stimuli (*potentiels évoqués*) qui, dans une analyse en essai unique, sont difficilement identifiables car "noyées" dans un bruit de bien plus grande importance. Le moyennage n'est évidemment pas sans contre-partie. Il ne permet clairement pas de rendre compte de variabilités intrinsèques au signal, qui peuvent être riches en information. De plus, il a été montré que l'absence de prise en compte de ces variabilités (comme par exemple la variabilité de latence) peut affecter la représentativité et donc la pertinence du signal moyenné (Callaway et al., 1984; Knuth et al., 2006; Burle et al., 2008).

Par ailleurs, la nécessité de développer des analyses en essai unique est aussi devenue importante avec l'émergence de nouvelles applications en EEG. Nous nous intéressons dans cette thèse aux problématiques de classification, très largement exploitées dans le cadre des interfaces cerveau-machine (ICM) dont le but est de "décoder" les signaux cérébraux acquis et de les retranscrire en commandes pour un système informatisé. Le but premier de ces dispositifs est de s'adresser à des personnes lourdement handicapées ou bien atteintes de maladies graves affectant leur motricité. Dans ce cas, des contraintes supplémentaires sont introduites dans l'analyse des signaux EEG, en particulier sur la phase d'apprentissage du classifieur pour laquelle le nombre d'essais considérés doit être réduit autant que possible tout en garantissant de bons résultats en phase de test.

L'analyse en essai unique permet de prendre en compte des informations complémentaires à celles obtenues par la moyenne. Dans cette thèse nous nous intéressons à l'estimation de différentes sources de variabilité que l'on peut considérer entre les essais d'un même sujet. Nous montrons dans la suite comment ces variabilités peuvent être exploitées dans une procédure de classification.

La classification comme outil de validation des modèles proposés

Même si la classification n'est pas l'unique objectif de ce travail, elle est présente à plusieurs niveaux. D'une part, elle est présente dans la construction des modèles, adaptés à des signaux regroupés en classes distinctes. D'autre part, elle fournit un contexte simple permettant une validation quantitative des modèles proposés. Il est en effet raisonnable de penser que si un classifieur performant associé à un modèle échoue dans une tâche de détection, ceci peut s'expliquer par le fait que le modèle n'est pas adapté aux signaux étudiés. Par ailleurs, la classification en EEG est un domaine très étudié, où l'évaluation des performances est relativement simple.

Nous nous plaçons donc dans un contexte particulier de classification en introduisant des conditions propres aux interfaces cerveau-machine : nous évaluons les approches proposées dans le cas où l'on dispose de deux classes de signaux qui sont de petites tailles et déséquilibrées, les signaux étant eux mêmes généralement de grande dimension.

Problèmes des grandes dimensions, petits effectifs et classes déséquilibrées

Estimer une ou plusieurs matrices de covariance peut être une opération difficile dans le cas de petits échantillons et/ou de grandes dimensions des données. Dans ce qui suit, nous traitons ce problème dans le cas bien spécifique des signaux EEG.

Nous nous intéressons ici à l'extraction de composantes pertinentes afin de résumer l'information spatiale et temporelle contenue dans ces signaux en un petit nombre d'éléments. Pour ce faire nous introduirons deux outils d'analyse : transformation en ondelettes discrète et filtrage spatial. La première étape consiste à décomposer les signaux de types potentiels évoqués dans des bases d'ondelettes bien adaptées à ce type de données. Nous montrons que les données étudiées s'expriment de façon pertinente dans un petit nombre de coefficients issus de la décomposition. L'objectif du filtrage spatial est de concentrer les informations spatialement discriminantes et ainsi de réduire la dimension spatiale en remplaçant les électrodes par un petit nombre de canaux. Après transformation en ondelettes et réduction aux coefficients significatifs, et filtrage spatial, les dimensions spatio-temporelles sont réduites de façon significative.

Le contexte applicatif considéré nous conduit à traiter des problèmes dans lesquels les classes sont déséquilibrées, et en particulier le cas où l'une des classes est très faiblement représentée. Nous sommes donc amenés à introduire des hypothèses permettant d'obtenir des estimations fiables des matrices de covariance. Notre approche se base sur la modélisation des matrices de covariance.

Quels modèles de covariance ?

Nous nous intéressons à une approche de modélisation statistique de la matrice de covariance des signaux EEG de type potentiels évoqués qui est exploitée dans le cadre de l'analyse discriminante. Les travaux originaux de cette thèse, présentés dans les parties II et III se basent sur la définition de modèles linéaires gaussiens au travers de deux modèles particuliers.

Dans une première approche nous exploitons la structure spatio-temporelle des EEG en décrivant mathématiquement la covariance d'un signal comme le produit de Kronecker entre la covariance des électrodes et la covariance temporelle, simplifiée par une transformation en ondelettes. Il s'agit là d'exprimer mathématiquement une hypothèse de découplage entre corrélations spatiales et corrélations temporelles, justifiée par les origines différentes de ces deux types de

corrélations. Il est à noter que dans ce modèle, la covariance est supposée la même pour toutes les classes.
Nous introduisons cette structure de Kronecker dans un modèle de mélange gaussien dans le cas de deux classes. Nous formalisons dans ce cas l'équivalence entre l'analyse discriminante décisionnelle classique et l'approche matricielle, récemment développée pour la classification de données indexées par un double indice (par exemple les images ou les signaux EEG comme c'est le cas ici, avec les indices spatial et temporel).
Nous nous intéressons à la problématique de l'extraction de caractéristiques discriminantes et à son adaptation à notre contexte. L'analyse discriminante descriptive est étendue et formalisée dans le cas matriciel. Nous proposons une méthode pour extraire conjointement, à partir d'observations matricielles, les caractéristiques spécifiques aux lignes et aux colonnes. Nous nous basons pour cela sur une réécriture de l'analyse discriminante binaire comme décomposition en valeurs singulières de la différence des moyennes des classes, dans des métriques bien choisies.
Nous montrons finalement que ce formalisme permet de décrire et représenter graphiquement les signaux EEG dans des sous-espaces discriminants et d'interpréter les caractéristiques extraites en termes de signatures spatiales et temporelles.

La seconde approche consiste à modéliser plus explicitement la variabilité inter-essais. Nous proposons de décrire le signal observé sous forme d'une somme de deux composantes : une composante commune à tous les essais issus d'une même condition expérimentale et une composante aléatoire modélisant la variabilité propre à l'essai observé par rapport à cette composante commune. Ce faisant, la variabilité du signal observé se décompose en deux sources distinctes : la variabilité du bruit de fond et la variabilité du signal d'intérêt. Le modèle proposé est un modèle linéaire mixte gaussien qui permet ici de structurer la matrice de covariance et d'extraire la variabilité inter-essais. Signalons que ce modèle permet de prendre en compte très simplement des covariances différentes pour les deux classes, si nécessaire ; ce qui est fait ici. Il faut souligner l'importance des transformations préalables (ondelettes, filtrage spatial et réductions de dimensions associées), qui permettent de limiter à l'extrême le nombre de paramètres du modèle.
Décrit dans un contexte de discrimination, le modèle proposé permet d'effectuer une analyse discriminante dans laquelle la covariance dépend de la classe. Dans ce travail nous affirmons que la variabilité inter-essais, bien que faible par rapport à celle de l'activité de fond, peut être pertinente pour l'analyse et la classification en EEG.

Applications à des données réelles

Les deux approches proposées ont été appliquées sur deux jeux de données réels de type potentiels évoqués (P300 et potentiels d'erreur) qui sont naturellement constitués de classes déséquilibrées. Dans chacun des cas une procédure d'extraction de caractéristiques et de classification binaire a été mise en place et des comparaisons systématiques avec des méthodes concurrentes ont été effectuées.

Contributions

En résumé, les principales contributions de ce travail sont les suivantes :

1) Nous montrons l'importance de la réduction de dimension introduite par le filtrage spatial et temporel. Les données de type potentiels évoqués sont bien représentés dans un nombre restreints de composantes à la fois déterminées par une transformation en ondelettes et une étape de filtrage spatial. Cette dernière est formalisée comme un cas particulier de l'analyse discriminante descriptive à partir du modèle spatio-temporel introduit dans la première approche.

2) Dans une première approche nous proposons une formalisation simple de l'analyse discriminante binaire pour les données matricielles avec un modèle de covariance de type produit de Kronecker, supposant la séparabilité des variabilités des lignes et des colonnes. Dans la seconde approche, nous construisons un modèle permettant de décrire avec très peu de paramètres la variabilité inter-essais, exploitant le formalisme bien établi du modèle linéaire mixte. Ce modèle permet d'extraire la variabilité inter-essais de la variabilité du bruit de fond et de l'intégrer en analyse discriminante comme une source d'information supplémentaire et pertinente pour la classification.

3) Les méthodes proposées sont appliquées sur deux jeux de données réelles.
Les résultats de classification et l'interprétation physiologique des caractéristiques spatio-temporelles discriminantes permettent de valider la première méthode sur les données de type P300 et sur les potentiels d'erreur. L'introduction de la structure spatio-temporelle permet d'extraire des composantes qui ont du sens tant d'un point de vue de l'analyse du signal que de la discriminantion.
La modélisation de la variabilité inter-essais est quant à elle validée sur les données de potentiels d'erreur en termes d'extraction du signal d'intérêt et de classification à partir de très petits échantillons d'apprentissage. Son application aux P300 est plus décevante. Ceci montre la dépendance aux données du modèle proposé. Une étude plus approfondie de ce jeu de données devra être faite afin de mieux déterminer la façon de modéliser la variabilité inter-essais dans ce cas.

Guide de lecture

Le manuscrit se structure en trois parties.

La partie I, qui contient la présente introduction, prend la forme d'un état de l'art dans lequel les principales problématiques de cette thèse sont présentées. Les signaux EEG, leurs principales caractéristiques ainsi que les problématiques d'analyse et de traitement du signal dans un contexte de classification sont introduits dans le chapitre 2. Ceci permet de contextualiser le cadre de recherche présenté dans les parties II et III. Dans le chapitre 3, nous présentons l'analyse discriminante classique et introduisons les approches décisionnelle et descriptive. Nous mettons l'accent sur les problèmes liés à l'estimation des matrices de covariance et nous présentons (de façon non-exhaustive) différentes extensions. Les bases sont ici données pour l'analyse discriminante étendue aux données matricielles dans la partie II.

La partie II présente le modèle de covariance spatio-temporel. Dans le chapitre 4 l'approche décisionnelle de l'analyse discriminante basée sur la structure de Kronecker de la covariance est formalisée et le lien avec l'approche classique présentée dans le chapitre 3 est établi. Dans le chapitre 5 l'approche descriptive classique est étendue au cas matriciel pour l'extraction et

la représentation graphique de caractéristiques discriminantes à partir d'observations structurées sous forme matricielle, dans le cas de la classification binaire. Le chapitre 6 est consacré à l'application sur les deux jeux de données présentés dans le chapitre 2. Les caractéristiques spatio-temporelles extraites sont analysées et utilisées pour une problématique de classification binaire.

Dans la partie III nous présentons l'approche basée sur le modèle mixte. La modélisation et la procédure de classification basée sur le modèle sont décrites dans le chapitre 7 et son application aux deux jeux de données est présentée dans le chapitre 8.

Enfin, dans la partie IV sont donnés la conclusion générale et les perspectives de ce travail, ainsi que les annexes et la bibliographie.

Chapitre 2

Introduction à la problématique des signaux EEG

Sommaire

L'objet de ce chapitre est d'orienter le lecteur vers les principales problématiques explorées dans ce travail de recherche. L'étude des signaux cérébraux d'un point de vue théorique et fondamental n'est pas l'objet direct de cette thèse. Nous renvoyons par exemple le lecteur aux ouvrages très complets sur le sujet de la modélisation du système neuronal (Dayan and Abbott, 2001) et des propriétés neurophysiques de l'EEG (Nunez and Srinivasan, 2006).

Nous allons nous intéresser aux caractéristiques générales de l'EEG et aux problèmes liés au traitement de ces signaux en vue d'explorer leurs propriétés statistiques qui constitueront le fil rouge des travaux présentés par la suite.

2.1 Les signaux EEG

2.1.1 Les sources de l'EEG

L'EEG correspond à l'enregistrement de la somme de l'activité électrique de larges populations de neurones actifs de façon synchrone. L'excitation d'un neurone à travers une synapse entraîne l'ouverture de canaux ioniques au niveau de sa membrane. La composition ionique étant différente à l'intérieur et à l'extérieur des cellules neuronales, l'ouverture des canaux engendre un mouvement de particules chargées dans le milieu intra et extra cellulaire (voir figure 2.1). Ce

sont les champs magnétiques induits par les potentiels électriques des neurones qui sont mesurés par l'EEG. Ainsi, La forme du tracé EEG dépend de l'état d'activation des neurones enregistrés. Ce niveau d'activation va se traduire par des changements de fréquence et d'amplitude des ondes enregistrées.

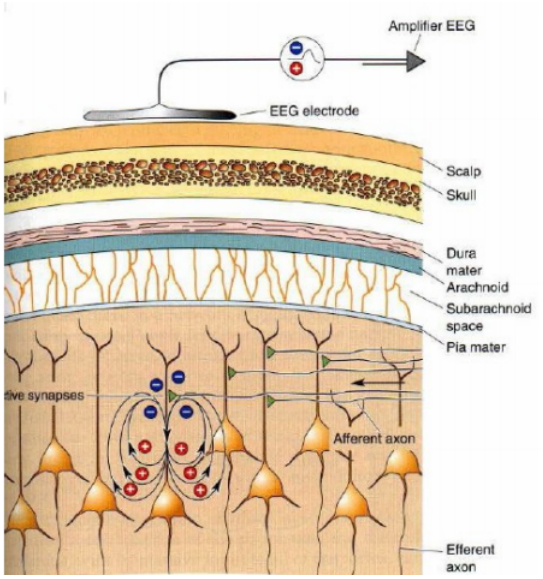


FIGURE 2.1 – Illustration de l'activité neuronale enregistrée en EEG.

2.1.2 Acquisition

Protocole d'enregistrement et utilisation

L'électroencéphalographie est l'une des techniques d'acquisition de l'activité cérébrale la plus ancienne. Le premier tracé EEG a en effet été enregistré par le neurophysiologiste allemand Hans Berger en 1929 (Berger, 1929). Cette méthode permet d'enregistrer les différences de potentiel électrique sur la surface crânienne à l'aide d'électrodes placées sur l'ensemble du scalp, le contact électrique étant assuré par un gel conducteur (on parle de méthode non-invasive). Des électrodes sont placées sur le cuir chevelu du patient ou du participant, allongé sur le dos ou assis dans un fauteuil. Les signaux sont enregistrés numériquement et stockés pour une utilisation ultérieure.

Étant donné que l'EEG mesure non pas un potentiel absolu, mais une différence de potentiels électriques, l'activité mesurée sur une électrode correspond en réalité à une mesure sur un couple d'électrodes : une électrode dite de mesure et une autre de référence. Le choix d'une électrode de référence est donc indispensable. Le plus souvent, une seule électrode est utilisée comme référence pour l'ensemble des autres. Ce choix est primordial puisqu'il peut avoir un impact sur l'ensemble de l'analyse des signaux acquis (Gençer et al., 1996).

De plus, l'activité cérébrale enregistrée à la surface du scalp étant de très faible intensité, le signal est amplifié puis filtré afin de supprimer les hautes fréquences susceptibles de ne contenir aucune information d'intérêt. Enfin, pour traiter numériquement les signaux enregistrés (par exemple, appliquer un moyennage d'enregistrements successifs), une conversion analogique-digitale (A/D) est nécessaire. Cette conversion se fait à une fréquence d'échantillonnage donnée :

le signal analogique (continu) est transformé en un signal numérique (discret) correspondant à une série de valeurs échantillonnées sur un intervalle de temps régulier.

L'EEG est utilisé pour diverses applications : dans le domaine clinique il permet par exemple d'évaluer les troubles de la conscience (coma), les troubles du sommeil ou bien encore l'épilepsie. Il est également très utilisé dans le domaine de la recherche en neurosciences où il est couramment pratiqué dans un cadre expérimental afin d'étudier certains phénomènes cognitifs. Enfin, nous citerons également son utilisation dans les protocoles d'interface Cerveau-Machine (ICM) que nous présentons dans la section 2.3.3. Selon les situations, l'EEG peut donc être exploité de façon directe dite *online*, ou bien peut être l'objet d'analyses post-acquisition (on parlera d'analyse *offline*).

Les autres méthodes d'enregistrement

Il existe de nombreuses autres méthodes d'acquisition de l'activité cérébrale. Parmi celles-ci, nous distinguons les méthodes invasives qui nécessitent l'implantation chirurgicale d'électrodes. Bien que ces méthodes permettent d'enregistrer des signaux de meilleure qualité (bonne résolution spatiale et temporelle), elles ont de nombreux inconvénients liés à leur nature. Citons par exemple l'éléctrocorticographie (EcoG) qui consiste à enregistrer l'activité cérébrale à partir d'une matrice d'électrodes directement placée à la surface du cortex cérébral.
Les méthodes non-invasives se contentent de mesures depuis l'extérieur du crâne, ayant donc un impact direct sur la résolution spatiale et/ou temporelle du signal enregistré. La magnétoencéphalographie (MEG) est la méthode d'enregistrement la plus similaire, en principe, à l'EEG, si ce n'est qu'elle mesure les champs magnétiques induits par l'activité électrique et non les potentiels électriques. Sa résolution spatiale est bien meilleure puisque le champ magnétique est moins sensible à la traversée du crâne. Par opposition, l'imagerie par résonance magnétique fonctionnelle (IRMf) dispose d'une excellente résolution spatiale mais d'une résolution temporelle plus faible. Cependant, toutes ces méthodes sont bien plus contraignantes et coûteuses

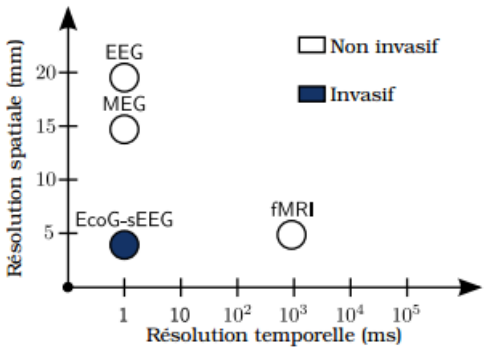


FIGURE 2.2 – – Résolution spatiale et temporelle des différentes méthodes d'enregistrement. Figure adaptée de (Fruitet, 2012).

que l'EEG. Ce dernier a en effet l'avantage d'être plus facilement transportable et plus facile à mettre en pratique au quotidien. De plus, bien que la résolution spatiale soit faible, de nombreux outils ont été proposés afin d'améliorer la localisation des phénomènes enregistrés à la surface

du scalp à partir des informations obtenues sur les électrodes (nous parlerons par exemple de filtrage spatial, voir section 2.2.1).

Qu'enregistre l'électroencéphalogramme ?

Lorsque l'on parle de signaux EEG, il faut tenir compte de l'existence de deux activités cérébrales bien différentes et qu'il est pourtant difficile de distinguer sur un tracé EEG : l'activité électroencéphalogramme spontanée et les potentiels dits évoqués. L'EEG résulte donc d'un mélange de sources qu'il est nécessaire de bien définir.

L'activité spontanée est le résultat de l'activité d'une série de "rythmes cérébraux" et qui reflète l'activation synchrone et périodique de grandes populations de neurones. Ces différentes oscillations, définies dans une large gamme de fréquences (< 4 Hz et jusqu'à plus de 100 Hz.) caractérisent différents états de conscience chez l'individu : éveil, sommeil, coma, etc. et sont également influencées par de nombreux facteurs externes comme par exemple l'état émotionnel ou bien certains stimulus.

Les potentiels évoqués désignent, par opposition à l'activité spontanée, la modification du potentiel électrique produite par le système nerveux en réponse à un stimulus. Ce stimulus peut être extérieur à l'individu (un son, une image, etc.), nous parlerons alors de potentiel évoqué exogène, ou bien il peut être lié à une activité cognitive et donc de nature endogène. Ces potentiels évoqués (PE, ou ERP : *evoked-related potentials*) sont ainsi très largement utilisés pour l'étude des différentes activités fonctionnelles neuronales et pour la description de comportements cognitifs. En particulier, les potentiels évoqués exogènes permettent de caractériser l'activité nerveuse générée par un stimulus sensoriel donné et, ainsi, d'étudier la fonction des différents systèmes sensoriels.

Les potentiels étant de très faible amplitude (de l'ordre d'une dizaine de μV) comparés à l'amplitude moyenne de l'EEG (de l'ordre d'une centaine de μV), leur étude peut s'avérer difficile. En effet, les électrodes d'enregistrement captent non seulement la réponse évoquée, mais également l'activité spontanée liée à ce qui est parfois appelé "le bruit de fond". C'est pourquoi l'analyse d'un enregistrement EEG nécessite plusieurs étapes préliminaires permettant d'améliorer sa résolution et de se concentrer sur le signal d'intérêt.

2.1.3 Moyennage et variabilité inter-essais

Une des principales limitations de l'EEG réside dans l'extraction de l'activité cérébrale d'intérêt. En effet, comme vu précédemment, ce signal d'intérêt de faible amplitude est très souvent noyé dans des activités de plus fortes amplitudes telles que l'activité spontanée (ou activité de fond). Usuellement, l'extraction de ce signal nécessite l'application d'une étape de moyennage sur un nombre important de répétitions enregistrées dans les "mêmes" conditions. Ce faisant, l'activité d'intérêt, commune à toutes les répétitions, est amplifiée par rapport aux petites fluctuations qui sont elles amoindries. Par exemple, l'extraction de potentiels évoqués se fait par moyennage de plusieurs portions d'un signal enregistrées à la suite du même stimulus (nous parlerons par la suite d'*essais*). La figure 2.3 représente le moyennage issu de plusieurs répétitions d'un même signal obtenues à partir des résultats d'un participant à une expérience de type tâche de flanker. Il s'agit des données de potentiel d'erreur présentées dans la section 2.4.1.

Cependant cette opération de moyennage n'est pas sans contrepartie. En effet, une telle moyenne déforme le signal d'intérêt et empêche une analyse précise de la dynamique des processus sous-jacents. Il est depuis longtemps reconnu que la moyenne a un impact non linéaire sur

l'estimation des temps de latence (Callaway et al., 1984; Knuth et al., 2006; Burle et al., 2008). Par exemple, le temps de latence de l'apparition d'une activité mesurée sur la moyenne sous-estime largement la moyenne réelle des onsets individuels (Smulders et al., 1996; Meyer et al., 1988; Kukleta and Lamarche, 2001). Ainsi, sa comparaison directe avec des variables chronométriques telles que le temps de réaction (*RT : Reaction Time*) est problématique (Meyer et al., 1988; Burle et al., 2008). Le procédé de moyennage empêche également l'analyse des effets d'apprentissage et/ou d'adaptation entre les répétitions (Quiroga and van Luijtelaar, 2002). Plus généralement *moyenner les essais élimine la variabilité du signal d'intérêt*, bien que cette dernière contienne des informations importantes qui peuvent être utiles dans divers contextes, de l'interprétation du signal à la classification. Nous nous intéresserons plus particulièrement à cette problématique dans la partie III de cette thèse.

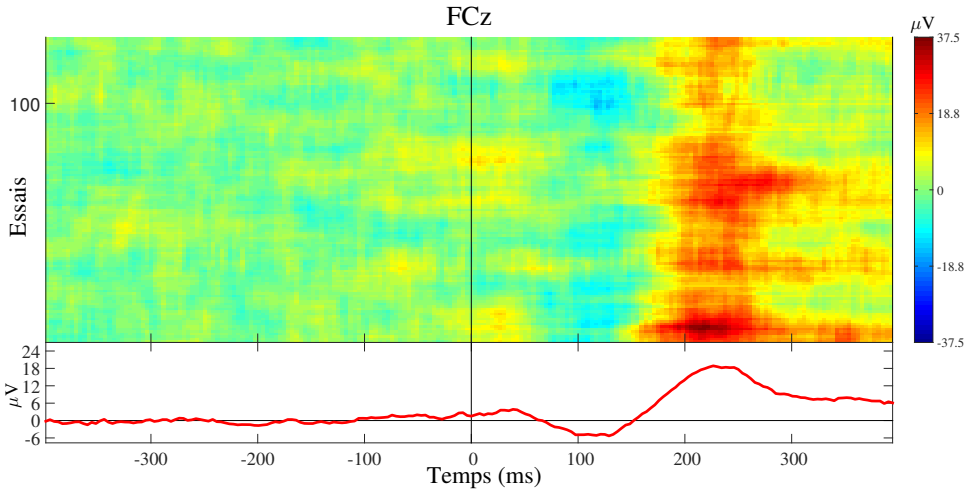


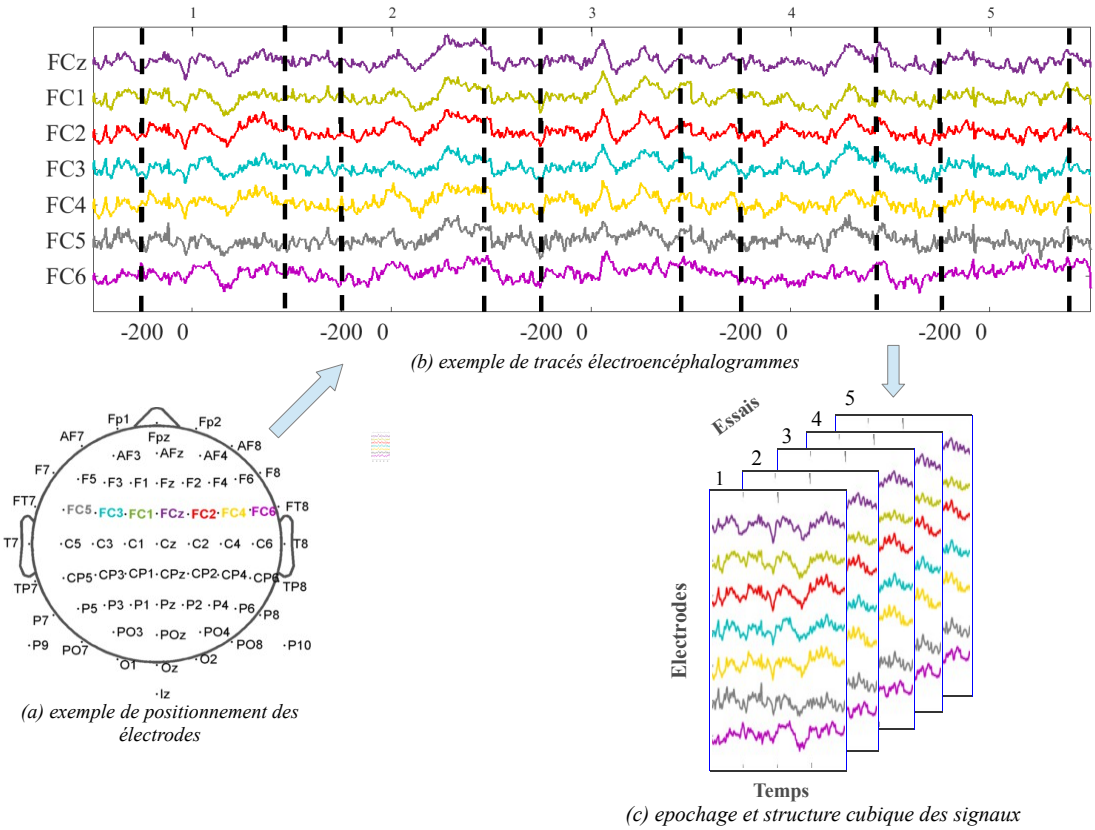
FIGURE 2.3 – Illustration du décours temporel (en abscisses) des essais uniques (potentiels d'erreur) enregistrées sur FCz (en ordonnées) et comparés à leur moyenne (en rouge). Les 130 essais représentés dans la partie haute du graphique correspondent à une portion de 800 ms comprise entre $[-400ms : +400ms]$, zéro indiquant la réponse motrice du participant à un stimulus visuel. Pour chaque essai on retrouve une activité négative entre 100 ms et 150 ms après la réponse motrice, suivie d'une activité positive. Une importante variabilité est visible. Ces deux activités se retrouvent également en moyenne.

Enfin, notons que l'analyse en essai unique est au cœur de la problématique de classification de signaux EEG, particulièrement dans le cadre des ICM, présenté dans la section 2.3.3.

Dans ce travail, nous mettons l'accent sur le développement d'outils mathématiques permettant d'améliorer l'analyse essai par essai, en particulier pour les problématiques de détection et d'estimation de potentiels évoqués. L'augmentation du rapport signal sur bruit constitue dans ce contexte un enjeu important. Un état de l'art des principales méthodes proposées est donné dans la section 2.2.

2.1.4 Structure des données

Lorsque l'on considère un ensemble de signaux EEG issus d'une même expérience, nous manipulons des données sous forme cubique : espace × temps × essais.
Chaque portion d'un signal EEG (ou essai) apparaît naturellement sous la forme d'une matrice puisque l'enregistrement de ce signal sur une séquence de temps fixée prend la forme d'un vecteur temporel. Le plus souvent, ce signal est enregistré à l'aide de multiples électrodes placées sur l'ensemble du scalp. Ainsi, lorsque l'on considère le décours temporel de dimension T d'un signal enregistré à partir de M électrodes, chaque essai est constituée de M vecteurs de dimension T. De façon générale, une expérience EEG consiste en un enchaînement rapide de stimuli de même nature et l'activité cérébrale du participant est enregistrée de manière continue. Les essais correspondent alors au découpage de segments de même taille du signal qui peuvent être par exemple localisés autour d'un évènement tel que l'apparition d'un stimulus ou bien la réponse motrice du participant (voir la figure 2.4).



(b) exemple de tracés électroencéphalogrammes

(a) exemple de positionnement des électrodes

(c) epochage et structure cubique des signaux

FIGURE 2.4 – Illustration de la structure cubique des données EEG multi-électrodes.Dans l'exemple, le signal est enregistré sur 7 électrodes (a). 5 segments (ou essais) de 600 ms sont représentés (délimités par les lignes verticales en pointillés). (b) Chaque essai est compris entre 200 ms avant et 400 ms après la réponse motrice du participant (identifiée par un zéro) . (c) les données sont segmentées de sorte que l'on considère un cube composé des 5 matrices (électrodes × temps) correspondant aux essais.

Cette structure particulière constitue un point important dans l'analyse des signaux EEG. D'une part, il est d'usage d'appliquer des traitements distincts sur les dimensions spatiales et temporelles en vue de résumer l'information d'intérêt contenu dans les signaux, améliorer la résolution ou encore pour réduire les dimensions. D'autre part, les approches de modélisation et de détection sont appliquées sur des essais vectorisés (spatio-temporels) ce qui a pour conséquence de mélanger de deux sources de variabilité (spatiale et temporelle) qui ne sont pourtant pas de même nature. Nous allons dans un premier temps nous intéresser à ces deux approches sous forme d'un état de l'art. Quelques méthodes d'extraction de caractéristiques spatiale et temporelle et de réduction de dimension parmi les plus populaires dans le cadre d'analyse et de classification en EEG sont présentées dans la suite. Dans la partie II nous nous intéresserons plus en détails à la problématique de mélange des sources de variabilités des dimensions spatiales et temporelles des signaux.

2.2 Traitement et analyse

Par nature, les signaux observés sur un EEG sont particulièrement corrélés, tant au niveau spatial qu'au niveau temporel. De ce fait, les très nombreux traitements développés en vue de réduire ces corrélations occupent une place centrale dans l'analyse de tels signaux. Dans ce qui suit, nous présentons brièvement les méthodes les plus populaires pour le traitement spatial et temporel en EEG.

2.2.1 Mélange de sources et propagation spatiale

Si l'EEG est utilisé pour analyser l'activité cérébrale, les signaux enregistrés sur le scalp ne fournissent pas d'informations directes sur les caractéristiques des sources cérébrales de profondeur, que ce soit en termes de caractéristiques temporelles et fréquentielles du signal, ou en termes de localisation et orientation spatiale de la structure neuronale génératrice. En effet, l'activité enregistrée sur une électrode ne provient pas uniquement de la zone du scalp ciblée ni des régions du cerveaux situées sous cette électrode. Le signal est en réalité le produit de la synchronisation de potentiels de champ dans de très nombreuses parties du cortex. L'activité électrique des sources se propage à travers les structures anatomiques, notamment l'os et le cuir chevelu, et les signaux de surface résultent donc d'un mélange de ces activités de profondeur et des activités dues à des perturbations électrophysiologiques autres. Ainsi, le signal EEG enregistré sur une seule électrode correspond à un mélange de signaux de sources corticales sous-jacentes. L'importante corrélation observée entre les électrodes s'explique donc par ce mélange spatial des sources d'un signal EEG. C'est pourquoi, lorsque l'on évoque l'EEG il n'est pas rare de rappeler sa mauvaise *résolution spatiale* (voir la figure 2.2).

De très nombreuses méthodes ont été développées afin de tenir compte de ces phénomènes de propagation spatiale et de mélange de sources. Les principales approches sont décrites ci-dessous.

Localisation et séparation de sources

Nous introduisons ici deux approches distinctes qui se basent sur la même formalisation du problème. Dans les deux situations, un signal X s'écrit de la façon suivante :

$$X = AS\,.$$

Les informations contenues dans les matrices A et S diffèrent d'une méthode à l'autre.

La localisation de sources permet, à partir de l'activité mesurée sur le scalp, de retrouver les dipôles liés à cette activité par résolution d'un problème inverse. Ici, X est le signal de surface, A décrit des modèles physiques et S contient les dipôles de courant qui produisent les champs de potentiels électriques. Dans ce cas, A est connue ou modélisée et représente la propagation des dipôles jusqu'au scalp (Ramirez, 2008).

Par opposition, la séparation de sources est une méthode qui n'introduit aucune hypothèse *a priori* sur la physique du phénomène. Toujours dans le cas d'un modèle $X = AS$, les algorithmes de séparation de sources cherchent à identifier à la fois A et S, aux indéterminations près. Nous citons en particulier l' Analyse en Composantes Indépendantes (ICA), ou bien l'algorithme SOBI (*second-order blind identification*) (Comon and Jutten, 2007). Ces méthodes, basées sur différentes hypothèses, permettent de "démélanger" les différentes sources d'un signal à partir des seules informations obtenues à partir du décours temporel à la surface du scalp.

Laplacien de surface

Afin d'améliorer la résolution spatiale en EEG, une méthode consiste à estimer le Laplacien de surface. Ceci permet de réduire l'effet de flou dû à la diffusion des courants à travers le crâne (Nunez and Srinivasan, 2006). De nombreuses techniques d'estimation ont été proposées et nous renvoyons à l'article (Tandonnet et al., 2005) pour une étude empirique détaillée. Il a été montré que cette méthode permet d'approximer l'activité cérébrale à la surface du cortex. Une application de cette approche est présentée dans la section 6.1.1.

2.2.2 Caractéristiques temporelles et fréquentielles

L'analyse temps-fréquence en EEG

L'activité cérébrale est constituée de nombreux rythmes cérébraux, qui se distinguent de par leurs fréquences. Les principaux rythmes sont détaillés dans la table 2.1 et pour plus de détails, voir par exemple (Sanei, 2007).

Rythmes	Fréquences
delta	0,5 Hz - 4Hz
alpha	8 Hz -13 Hz
mu	8 Hz - 13 Hz
beta	15 Hz - 30 Hz (pics jusqu'à 60 Hz)
gamma	30 Hz - 50 Hz (pics jusqu'à 120 Hz)

TABLE 2.1 – Les principaux rythmes cérébraux et leurs fréquences.

Le concept de fréquence renvoie à des propriétés de périodicité : une sinusoïde de fréquence ν Hz se répète identique à elle-même toutes les ν^{-1} secondes. Ces notions sont rendues plus précises par la transformée de Fourier (TF) qui associe à toute fonction $g : \mathbb{R} \to \mathbb{C}$ (sous certaines conditions, voir par exemple (Gasquet and Witomski, 1999)), une autre fonction $\tilde{g} : \mathbb{R} \to \mathbb{C}$:

$$\tilde{g}(\nu) = \int g(t) e^{-2i\pi\nu t} dt \,. \tag{2.1}$$

Le nombre $\tilde{g}(\nu)$ caractérise en quelque sorte le contenu fréquentiel de g à la fréquence ν, comme le montre la formule de synthèse de Fourier :

$$g(t) = \int \tilde{g}(\nu) e^{2i\pi\nu t} d\nu \,, \tag{2.2}$$

qui exprime g comme une superposition d'oscillations de toutes les fréquences $\tilde{c}_\nu : t \to e^{2i\pi\nu t}$, pondérées par $\tilde{g}(\nu)$.

Une version locale de la TF est obtenue en localisant la fonction d'intérêt $g(t)$ par une fenêtre $w(t-\tau)$, localisée au voisinage de $t = \tau$, avant le calcul de la TF. Ici, $w(t)$ est une fenêtre de référence, par exemple une fenêtre gaussienne. Ceci conduit à la transformée de Fourier à court terme (TFCT) :

$$g \to G_g : \quad G_g(\tau, \nu) = \int g(t) w(t-\tau) e^{2i\pi\nu t} dt \,, \tag{2.3}$$

où $G_g(\tau, \nu)$ fournit une mesure du contenu fréquentiel de g à la fréquence ν au voisinage de l'instant $t = \tau$. Il est possible de montrer que la fonction de deux variables G_g contient toute l'information sur g, la transformation (linéaire) $g \to G_g$ étant inversible.
Dans un contexte d'analyse de signaux EEG, la TFCT ainsi que certaines variantes (la transformée de Gabor, les bases MDCT (Modulated Discrete Cosine Transform)) a été utilisée avec succès, notamment pour l'analyse et la caractérisation de certains rythmes oscillants (voir par exemple (Villaron, 2012; Mahanta et al., 2012)) et l'outil d'analyse de référence EEGLAB (Delorme and Makeig, 2004) utilise diverses variantes de la TFCT pour l'étude des signaux EEG sous Matlab.

La TF et la TFCT sont toutefois inadaptées dans certaines situations, notamment pour l'analyse de certains phénomènes basse fréquence (potentiels évoqués) ou très localisés (e.g. pointes épileptiques). On peut parfois leur préférer la transformée en ondelettes (TO) qui permet de décomposer le signal en "atomes" de taille (et donc de résolution) variable mais de forme constante. La TO se base sur un outil d'analyse, l'ondelette mère Ψ, qui est une fonction oscillante (par exemple une dérivée de gaussienne). La TO d'une fonction g est définie par :

$$w_g(\tau, s) = \frac{1}{\sqrt{s}} \int g(t) \bar{\Psi} \left(\frac{t-\tau}{s} \right) dt \,, \tag{2.4}$$

et mesure le "contenu" de g à l'échelle s au voisinage du point $t = \tau$.

La TO permet donc de décomposer le signal à partir de fenêtres de tailles différentes selon le niveau de décomposition considéré, caractérisant ainsi différentes composantes dans le plan temps-échelle. Plusieurs problématiques émergent au sujet de l'analyse temps-fréquence des signaux de type potentiels évoqués, dont les caractéristiques fréquentielles varient dans le temps : il a été montré sur le grand moyennage que les composantes précoces sont de bien plus haute fréquence que les composantes les plus tardives. Dans ce cas, il est donc nécessaire de considérer des filtres dont les caractéristiques fréquentielles évoluent au cours du temps. De plus, les potentiels évoqués se caractérisent par des phénomènes beaucoup plus lents que les rythmes sensori-moteurs. Il s'agit donc de considérer des espaces permettant de représenter finement ces changements. C'est pourquoi l'analyse dans l'espace des ondelettes est parfaitement appropriée à cette situation. Nous verrons qu'elle permet de localiser les différentes composantes du potentiel évoqué d'intérêt et permet ainsi de les extraire de l'activité de non-intérêt.
Cette méthode a été appliquée pour la première fois à l'étude de potentiels évoqués dans les années 1990. Par exemple dans (Bartnik et al., 1992), les auteurs présentent l'intérêt de l'analyse

multi-résolution pour la localisation en temps et en fréquence des composantes du signal d'intérêt moyenné. Par la suite, l'étude en essai unique par projection dans l'espace des ondelettes a été introduite (Quiroga, 2000; Quiroga and Garcia, 2003; Wang et al., 2007). Dans ces travaux, différentes approches sont proposées pour l'extraction des potentiels évoqués du "bruit de fond".

Décompositions sur des bases d'ondelettes

Les décompositions de Fourier à court terme et en ondelettes continues, ou leurs versions discrétisées, sont généralement utilisées à des fins de représentation, d'analyse des signaux ou d'estimation de paramètres, car elles fournissent des représentations synthétiques faisant souvent apparaître clairement certaines caractéristiques. C'est dans ce contexte qu'elles ont souvent été utilisées en analyse de signaux EEG.

Il s'agit toutefois de représentations redondantes, donc peu adaptées aux problématiques de réduction de dimension auxquelles nous sommes confrontés dans ce travail. Dans ce contexte, on leur préfère généralement des versions discrètes, très peu voire pas du tout redondantes. En particulier, des transformées temps-fréquence ou temps-échelle basées sur des bases orthonormées de l'espace des signaux considéré sont souvent utiles, lorsqu'elles permettent de compacter l'information sur peu de coefficients significatifs. Par exemple, les bases MDCT ont été utilisées dans (Villaron, 2012) pour la caractérisation de rythmes *alpha* (voir table 2.1). Dans le contexte considéré ici, on utilisera plutôt des bases d'ondelettes. Nous allons voir que la décomposition des signaux sur des bases d'ondelettes permet de "concentrer" l'information pertinente sur un petit nombre de coefficients, les autres étant négligeables. De plus, il a souvent été observé que ces coefficients sont essentiellement décorrélés (Atto and Berthoumieu, 2012), ce qui simplifiera d'autant la modélisation statistique.

Étant donnée une fonction de carré intégrable ψ, la transformation en ondelettes continue se base sur des copies dilatées et translatées

$$\psi_{\tau,s}(t) = s^{-1/2}\psi((t-\tau)/s) \ ,$$

les variables de translation τ et d'échelle s variant respectivement sur l'axe réel et le demi-axe réel positif. Il est possible de montrer que si l'ondelette ψ est bien choisie, une famille discrète d'ondelettes correspondante peut former une base orthonormée de $L^2(\mathbb{R})$.

Plus précisément, en notant ψ_{jk} ces ondelettes dilatées et translatées

$$\psi_{jk}(t) = 2^{-j/2}\psi\left(2^{-j}t - k\right)$$

on montre que (Daubechies, 1992; Daubechies, 1993; Mallat, 2008) :

Théorème 1. *1. Il existe $\psi \in L^2(\mathbb{R})$ telle que la collection d'ondelettes $\{\psi_{jk} \ \ j,k \in \mathbb{Z}\}$ soit une base orthonormée de $L^2(\mathbb{R})$. Tout $x \in L^2(\mathbb{R})$ peut se décomposer de façon unique comme*

$$x = \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} d_{jk}\psi_{jk}$$

où les nombres

$$d_{jk} = \langle x, \psi_{jk} \rangle = 2^{-j/2} \int x(t)\psi(2^{-j}t - k)dt$$

sont les coefficients d'ondelette de x.

2. *Il existe des couples de fonctions (ψ, ϕ) de carré intégrable telles que tout $x \in L^2(\mathbb{R})$ puisse se décomposer de façon unique sous la forme*

$$x = \sum_{k=-\infty}^{\infty} s_{j_0 k}\phi_{j_0 k} + \sum_{j=-\infty}^{j_0}\sum_{k=-\infty}^{\infty} d_{jk}\psi_{jk}$$

où les nombres

$$s_{jk} = \langle x, \phi_{jk}\rangle = 2^{-j/2}\int x(t)\phi(2^{-j}t - k)dt$$

sont les coefficients d'échelle de x. La fonction ϕ est appelée fonction d'échelle.

Définition 1. *Dans ce qui suit, on considèrera globalement les coefficients d'ondelette et d'échelle dans un unique vecteur de coefficients, appelés coefficients multi-échelle.*

Il existe en fait une infinité de choix possibles pour l'ondelette ψ ou pour le couple (ψ, ϕ). La notion d'*analyse multirésolution* fournit un algorithme générique permettant de construire de tels couples, et autant de formules de décomposition (Daubechies, 1992; Vetterli and Kovačevic, 1995; Mallat, 2008).

L'analyse multirésolution fournit également des algorithmes rapides de calcul des coefficients multi-échelle, basés sur une paire de *filtres numériques* (respectivement passe-bas et passe-bande), qui caractérisent à la fois ψ et ϕ.

Remarque 1. *Le choix de j_0 relève de l'utilisateur. Il faut comprendre que le terme $\sum_{k=-\infty}^{\infty} s_{j_0 k}\phi_{j_0 k}$ représente les composantes "basse-fréquence" de x, correspondant aux fréquences inférieures à une fréquence de référence ν_0 proportionnelle à 2^{-j_0}. Dans le même ordre d'idées, les termes $\sum_{k=-\infty}^{\infty} d_{jk}\psi_{jk}$ représentent des composantes de x dans des bandes de fréquence centrées sur une fréquence proportionnelle à 2^j. Ainsi, un choix de j_0 correspond à une volonté de regrouper toutes les composantes fréquentielles de x de fréquence inférieure à la fréquence ν_0.*

Remarque 2. *Les signaux que nous avons à traiter sont des signaux numériques, c'est à dire discrets et de longueur finie. L'analyse multirésolution peut s'adapter à ce schéma, de différentes façons, en particulier en identifiant les échantillons du signal aux coefficients d'échelle à l'échelle la plus fine considérée, et en considérant les signaux de longueur finie comme les restrictions à un intervalle de longueur finie de signaux périodiques de période égale à cette longueur (après zero-padding si nécessaire). C'est ce que fait l'implémentation proposée par Buckheit dans Wavelab (Buckheit et al., 2005), que nous utilisons dans cette thèse.*

Ainsi, nous pouvons considérer la transformation qui associe à un signal $x \in \mathbb{R}^n$ un vecteur de coefficients multi-échelle comme une transformation linéaire orthogonale, généralement appelée transformation en ondelettes discrète. Elle peut donc être représentée matriciellement, même si ça n'est pas sous cette forme qu'elle est implémentée pratiquement.

D'après le théorème 1, notons que les coefficients multi-échelle caractérisent tout signal en temps continu dans $L^2(\mathbb{R})$. Dans le cas discret, il existe un algorithme rapide de reconstruction similaire à l'algorithme de décomposition (on parle de transformation en ondelettes inverse) qui est implémenté dans WaveLab.

2.3 Classification supervisée

Nous présentons dans cette section les principales problématiques liées à la classification supervisée de signaux EEG. Dans ce contexte, la procédure de détection se fait à partir d'un ensemble de signaux issus de deux classes ou plus qu'il est d'usage de partager en deux parties : l'échantillon d'apprentissage et l'échantillon test. Dans ce travail nous nous intéressons aux méthodes de classification basées sur une modélisation des signaux. Ainsi, le modèle est calibré sur l'échantillon d'apprentissage et ses performances sont évaluées sur l'échantillon test. Nous nous réfèrerons à l'article de revue (Lotte et al., 2007a) dans lequel les différentes problématiques liées à la classification en EEG sont identifiées et analysées.

2.3.1 Résumer le signal : extraction de caractéristiques

Lorsque l'on s'intéresse à l'analyse et à la classification de données EEG, une des premières problématiques est celle de la réduction de dimension. En effet, si l'on considère un signal EEG enregistré sur un nombre M d'électrodes pouvant aller jusqu'à plus de 256 selon les configurations standards, et sur un intervalle de temps d'une longueur souvent comprise entre $500ms$ et quelques secondes, plus d'une dizaine de milliers de paramètres doit être considérée. Dans une problématique de classification supervisée, il n'est pas rare que le nombre de paramètres à estimer soit bien plus important que le nombre d'observations disponibles. Ainsi, l'estimation des paramètres et en particulier de la matrice de covariance obtenue dans ce cas risque d'être particulièrement mauvaise (et dans de nombreux cas, ceci renvoie à des problèmes de singularité, rendant son utilisation impossible pour la mise en œuvre d'une règle de classification). Ce problème, bien connu dans de nombreux domaines d'application et notamment en biologie porte le nom de *malédiction de la dimension* (Hastie et al., 2009).

Dans ce contexte et afin de résumer l'information spatiale et/ou temporelle d'un signal EEG, de nombreuses méthodes d'extraction de caractéristiques peuvent être utilisées. Comme nous l'avons vu précédemment, ces méthodes peuvent être très variées et dépendent largement de la nature du signal observé (potentiels évoqués, rythmes sensori-moteurs, etc.) ainsi que de l'objectif visé (classification, reconstruction de sources, etc.). Que l'on s'intéresse aux dimensions spatiales ou temporelles, l'extraction de caractéristiques consiste à *résumer l'information pertinente contenue dans le signal*. Afin de réduire la dimension temporelle, des méthodes d'analyse temps-fréquence comme décrites dans la section 2.2.2 peuvent être utilisées. La réduction de la dimension spatiale prend quant à elle la forme d'un filtrage spatial dont nous présentons quelques unes des méthodes les plus populaires développées pour la classification en EEG.

2.3.2 Filtrage spatial

Cette technique consiste à définir des combinaisons linéaires de signaux enregistrés sur différentes électrodes afin de résumer l'information spatiale en un nombre réduit de composantes pertinentes. Il s'agit d'un problème qui dépend à la fois du type de signaux considérés et de l'application envisagée. Cette technique regroupe un ensemble varié de méthodes adaptées pour différents types de données. Nous pouvons citer les plus populaires.

XDawn. Spécialement conçue pour le protocole P300 Speller (voir section 2.4.2), cette méthode consiste à estimer des sous-espaces liés aux potentiels évoqués de type P300, permettant, par projection des signaux bruts dans ces sous-espaces, d'améliorer leur résolution spatiale et

temporelle (Rivet et al., 2009). Elle consiste en l'estimation conjointe des signatures temporelles et de la distribution spatiale associée aux potentiels évoqués. De plus, elle permet d'obtenir des filtres spatiaux pour les données issues du protocole P300 maximisant le rapport signal sur bruit. Dans un article récent, les auteurs ont développé une analyse théorique de la méthode, démontrant son optimalité asymptotique (Rivet and Souloumiac, 2013). La particularité de cette méthode est directement liée à la particularité des signaux issus des protocoles P300. En effet, les potentiels liés à plusieurs stimuli visuels consécutifs peuvent se chevaucher sur plusieurs essais (voir section 2.4.2 ci-après pour plus de détails). L'algorithme proposé tient compte de ce phénomène de périodicité dans le calcul des filtres.

Common Spatial Pattern (CSP). Cette méthode, développée dans le cadre de la discrimination de deux tâches d'imagination motrice est très utilisée dans des paradigmes ICM basé sur cette tâche (Ramoser et al., 2000; Blankertz et al., 2008). Elle consiste à déterminer des filtres spatiaux discriminants entre deux conditions (par exemple mouvements de la main droite ou de la main gauche) par diagonalisation conjointe des matrices de covariance des électrodes obtenues à partir des essais dans chacune des classes. Les filtres ainsi obtenus permettent de représenter les signaux EEG dans un sous-espace qui maximise l'écart entre les covariances spatiales des deux classes.

LDA L'analyse discriminante basée sur le discriminant de Fisher est utilisée pour l'estimation de filtres spatiaux pour des données de type potentiels évoqués (Huang et al., 2009; Gutiérrez and Escalona-Vargas, 2010; Spinnato et al., 2012). Cette approche est plus largement présentée et formalisée dans la section 6.1.

2.3.3 Les Interfaces Cervau-Machine

Une application directe et bien connue de la classification de signaux EEG est le paradigme d'Interface Cerveau Machine (ICM, ou *BCI : Brain Computer Interface*). Il s'agit d'une interface permettant une communication "directe" entre le cerveau et un ordinateur dont le but principal (et très général) est de décoder les signaux cérébraux afin de les retranscrire en commande ou en action pour un système informatisé. La définition d'une ICM est donc très large et concerne un grand nombre de dispositifs dont nous ne discuterons dans ce travail. Pour une présentation détaillée des ICMs nous pouvons citer (Wolpaw et al., 2000; Lebedev and Nicolelis, 2003; Lotte et al., 2007b) ainsi que les thèses de Margaux Perrin (Perrin, 2012) et de Joan Fruitet (Fruitet, 2012). Le schéma général d'une ICM est détaillé dans la figure 2.5.

Nous nous intéressons ici aux ICM basées sur les EEG et notamment celles qui s'appuient sur les potentiels évoqués. Nous parlerons d'ICM exogènes puisque les commandes de l'utilisateur sont basées sur ses réponses produites de manière automatique en réponse à un évènement. De la même façon il existe les ICM endogènes pour lesquelles les commandes sont spontanément générées par l'utilisateur (e.g. rythmes sensori-moteurs pour l'imagination motrice (Wolpaw et al., 2002; Fruitet, 2012)). Nous ne discuterons pas de ces ICM puisque l'ensemble du travail présenté ici a été directement appliqué à des potentiels évoqués permettant de réaliser ou d'améliorer une ICM exogène. En particulier, tout ce qui sera présenté par la suite se situe dans les catégories "extraction de caractéristiques" et "classification" du schéma général de l'ICM présenté dans la figure 2.5. L'ICM constitue donc une application (directe) de l'ensemble des méthodes que nous présenterons mais elle n'est pas la problématique principale.

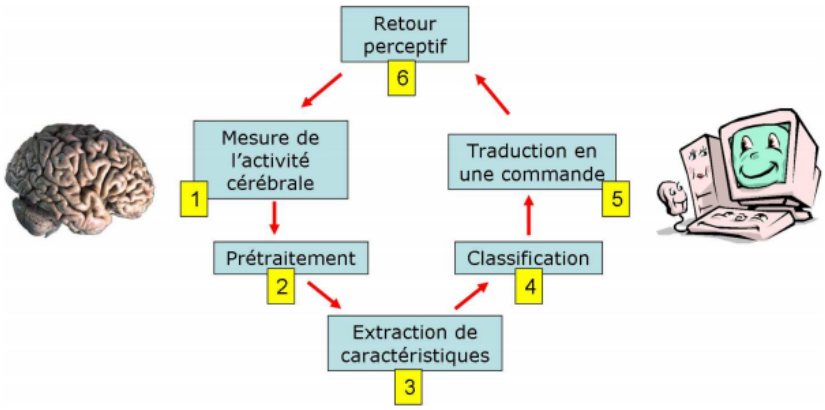


FIGURE 2.5 – Schéma général de fonctionnement d'une ICM. Figure extraite de (Lotte et al., 2007b).

2.4 Données étudiées

Dans le cadre de cette thèse, deux jeux de données ont été utilisés. Ils vont permettre d'illustrer la pertinence des différentes approches de modélisation de la variabilité proposées pour des signaux EEG du type potentiels évoqués. Le premier jeu inclut des potentiels d'erreurs et le second des potentiels liés au paradigme du P300 Speller. Ces jeux de données ont été fournis respectivement par le Laboratoire de Neuroscience Cognitive de Marseille (UMR 7291) et par l'équipe DYCOG du Centre de Recherche en Neurosciences de Lyon (INSERM). Ils ont été exploités dans le cadre du projet de recherche de l'ANR CoAdapt (ANR-09-DEFIS-002).

2.4.1 Jeu 1 : les potentiels d'erreur

Le premier jeu de données a été acquis par C. Roger au cours de sa thèse en neuroscience au LNC. Nous présentons ici de façon assez concise l'expérience ainsi les principales caractéristiques des signaux étudiés. Pour plus de détails, nous renvoyons le lecteur à l'article de référence, publié par C. Roger et ses collaborateurs (Roger et al., 2010).

Expérience et acquisition du signal

Le jeu de données inclut dix participants devant effectuer une tâche de flanker (ou dite d'Eriksen) (Eriksen and Eriksen, 1974). Sur un écran s'affichent trois lettres dont celle du milieu doit être identifiée tandis que les lettres adjacentes servent à distraire l'attention du sujet. Chaque participant doit identifier une lettre cible en pressant une manette soit de la main gauche, soit de la main droite (par exemple la main gauche si la cible est un B, la main droite s'il s'agit d'un G). Chaque essai correspond donc à l'identification d'une lettre "cible". 20 blocs expérimentaux de 128 essais ont été effectués. Le signal a été enregistré à partir de 64 électrodes placées sur l'ensemble du scalp et après une étape de pré-traitement (échantillonnage à 1024Hz et suppression des artefacts), les données ont été sous-échantillonnées à 256Hz. Les essais sélectionnés ont ensuite été segmentés en époques de 800 ms de -400 ms à $+400$ ms, où le zéro correspond au début de l'électromyogramme (EMG), déclenché par la réponse motrice du participant.

Pour chacun des participants, deux classes sont considérées : *les erreurs* (la réponse directionnelle ne correspond pas à la cible) et *les réponses correctes* (la réponse directionnelle correspond à la cible). Dans ce travail nous ne tenons pas compte de la troisième classe présentée dans Roger *et al.* correspondant aux ébauches d'erreur (réponse correcte, mais accompagnée d'une réaction de la main associée à la mauvaise réponse).
Chaque jeu de données se présente sous forme cubique puisque pour chaque classe $c = 0$ (réponse correcte) et $c = 1$ (erreur), n_c essais sont enregistrés sur 64 électrodes et 204 pas de temps. La figure 2.6 représente la différence entre les moyennes des deux classes sur des données monopolaires : pour les erreurs, on distingue une large activité négative qui apparaît environ 150 ms après la réponse motrice associée à une erreur. On parle de négativité d'erreur, qui est une activité fronto-centrale particulièrement visible sur les électrodes FCz, FC1 et FC2.

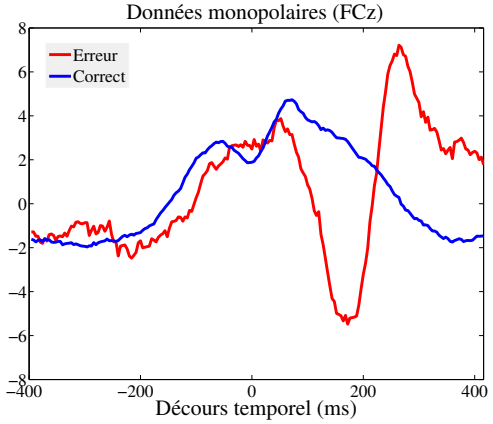


FIGURE 2.6 – Moyenne des données erreur (rouge) et correct (bleu) sur l'ensemble des 10 sujets de l'expérience présentée dans (Roger et al., 2010).

Le nombre d'essais erreur-correct diffère de façon significative entre les différents participants. Dans la table 2.2, le nombre total d'essais est donné pour chaque classe et chaque participant. Notons que nous sommes dans un cas d'*un jeu de données particulièrement déséquilibré*. En effet, le pourcentage d'erreur (qui reflète la performance des participants à la réalisation de la tâche de Flanker), varie approximativement de 2% à 12%. Par la suite, la classe erreur sera donc définie comme la classe minoritaire, ou encore comme *l'évènement rare à détecter*.

	Participant									
	A	B	C	D	E	F	G	H	I	J
Erreurs	130	105	43	39	18	28	94	145	100	63
Réponses correctes	1376	1575	1467	1735	690	759	1844	1238	1167	907

TABLE 2.2 – Nombre total d'essais par classe et par participant

Dans le cadre des ICM, les signaux d'erreurs ont une utilité particulière puisqu'ils permettent la détection et la correction automatique d'une erreur dans la décision prise par l'ordinateur (Perrin et al., 2012).

Projection dans l'espace des ondelettes

La transformation en ondelette discrète a été appliquée aux ondes d'erreur sur 5 niveaux de décomposition. La fréquence d'échantillonnage ayant été fixée à 256 Hz, les 5 niveaux correspondent respectivement aux bandes de fréquences : $64-128$ Hz, $32-64$ Hz, $16-32$ Hz, $8-16$ Hz et $4-8$ Hz. Ces fréquences sont caractérisées par les coefficients d'ondelette, tandis que la bande de fréquence la plus faible $0-4$ Hz est encodée dans les coefficients d'échelle.

La dimension temporelle est ensuite réduite à l'aide d'une approche de type filtrage passe-bas : sachant que les composantes liées à l'onde d'erreur sont de basse fréquence, la suppression des trois premiers niveaux de décomposition, qui contiennent les phénomènes de faible amplitude et de haute fréquence, est automatiquement appliquée pour toutes les données. Ceci peut être défini comme une projection dans un sous-espace approprié permettant de conserver l'information pertinente relative au signal d'intérêt (Abramovich et al., 2000).
En pratique, pour appliquer l'algorithme proposé dans Wavelab (Buckheit et al., 2005), les signaux doivent être complétés par des zéros (*zero-padding*) afin d'obtenir une taille appropriée (i.e. la taille du signal doit être une puissance de 2). Ainsi, les coefficients de bord générés par ces zéros sont également supprimés. En particulier, les 2 coefficients d'échelle de bords, les 2 coefficients d'ondelettes de bords du niveau le plus grossier et les 4 coefficients de bord du niveau supérieur sont supprimés de l'analyse.
Finalement, à l'issue de cette étape, il reste $K = 24$ coefficients multi-échelle dont 6 coefficients d'échelle et $12 + 6$ coefficients d'ondelette des niveaux 4 et 5 respectivement.

La figure 2.7 illustre l'effet de la transformation en ondelette et de la sélection des coefficients multi-échelle. La représentation du décours temporel des 24 coefficients moyens a été obtenue par une transformation en ondelettes inverse. D'après la figure 2.7(a), nous pouvons voir que les moyennes pour les classes erreur et correct obtenues à partir des coefficients sélectionnés sont très similaires aux vraies moyennes obtenues sur les données brutes. De plus, la figure illustre les différents rôles tenus par les coefficients d'échelle et d'ondelette. Les premiers capturent la forme générale du signal (b) tandis que les seconds concentrent les détails (c).

Remarque 3 (Sur le choix du filtre)**.** *Le choix du filtre D6 ici est le résultat d'un compromis entre localisation et lissage de l'ondelette. En effet, les ondelettes définies sur des supports plus larges génèrent des effets de bords bien plus important tandis que des filtres plus courts produisent des ondelettes et donc des reconstructions moins lisses (voir par exemple l'annexe C.3.4).*

Évaluation des classifieurs : détecter la classe minoritaire

Dans le cas de classes déséquilibrées, la méthodologie pour l'évaluation d'un classifieur doit être choisie précautionneusement puisque la classe majoriataire (ici les réponses correctes) peut masquer le comportement obtenu sur la classe rare (ici les erreurs). Ce problème a été étudié dans différents travaux(voir par exemple (Weiss and Provost, 2001; Weiss, 2004; Menardi and Torelli, 2014) et les références citées dans ces articles). Dans ce travail nous proposons d'utiliser différentes mesures de performance afin de comparer plusieurs classifieurs pour l'étude de la classification des potentiels d'erreurs. En effet, dans le jeu de données considéré ici, les erreurs représentent l'évènement rare (entre 2% et 12% des données selon les participants) à détecter. Nous nous intéresserons donc à des méthodes d'évaluation qui ne sont pas globales mais qui au contraire mesurent les performances classe par classe.

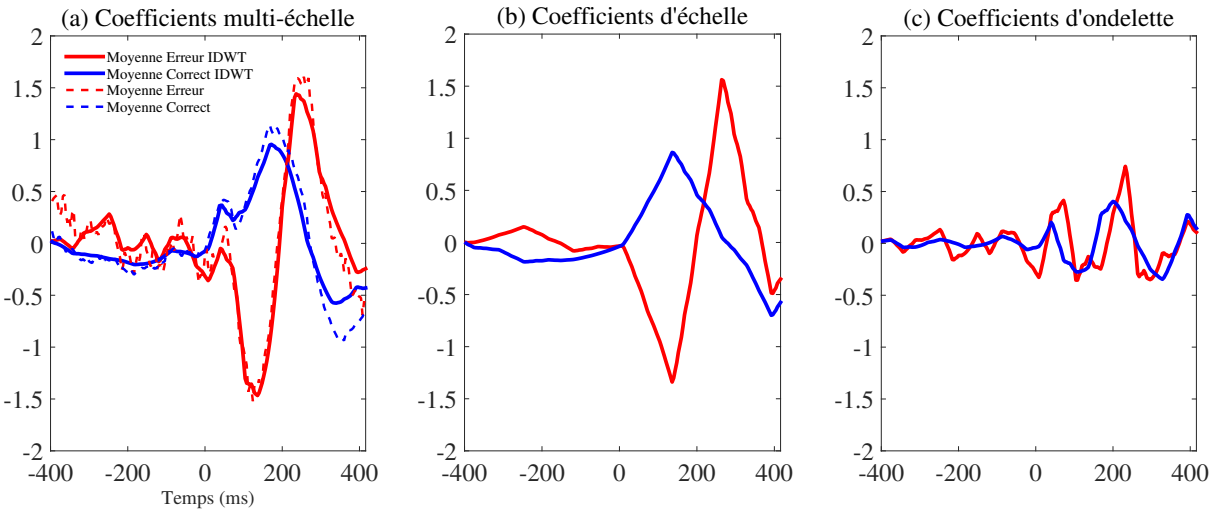


FIGURE 2.7 – Participant A. Illustration de l'effet de la transformation en ondelettes et de la sélection de coefficients multi-échelle pour les signaux d'erreur. Le filtre utilisé est un filtre de Daubechies D6. Les figures (a), (b) et (c) représentent les réponses moyennes pour la classe erreur (rouge) et correcte (bleu) sur l'électrode FCz. (a) Moyennes des deux classes obtenues à partir des 24 coefficients sélectionnés projetés dans le domaine temporel et comparé avec le décours temporel moyen de chaque classe (en pointillés sur le graphique). Les reconstructions partielles basées sur uniquement les coefficients d'échelle (b) et sur les coefficients d'ondelette (c) sont représentés afin d'illustrer le rôle de chaque coefficient.

La matrice de confusion est un des outils d'évalution de base. En classification binaire, les résultats d'un classifieur sur un ensemble de test sont catégorisés en 4 groupes : les vrais positifs (VP), les faux positifs (FP), les vrais négatifs (VN) et les faux négatifs (FN). Par exemple, si une observation est issue de la classe correct mais qu'elle est classée comme une erreur (au sens de potentiel d'erreur et non pas de l'erreur de prédiction), celle-ci sera définie comme un faux positif. La matrice de confusion s'écrit dans notre cas :

		Classe prédite	
		erreur	*correct*
Vraie classe	*erreur*	VP	FN
	correct	FP	VN

TABLE 2.3 – Matrice de confusion pour la classification de potentiels d'erreur.

A partir de cette matrice, deux autres indicateurs peuvent être définis :

— **Taux de bonne classification.** Dans le cas particulier de classes déséquilibrées, le taux de bonne classification doit être calculé dans chaque classe indépendamment. En effet, un taux de bonne classification global peut amener à une mauvaise interprétation des résultats. Par exemple, dans notre problème, les erreurs représentent environ 10% des données. Dans ce cas, le taux global de bonne classification d'un classifieur naïf qui consisterait à classer l'ensemble des données test dans la classe correct serait de 90% alors qu'il est clair que ce classifieur ne permet pas de détecter les évènements rares que sont

les erreurs. En classification binaire les taux de bonnes classifications dans chaque classe peuvent se calculer à partir de la matrice de confusion :

$$TVP = VP/(VP + FN) * 100 \quad \text{et} \quad TVN = VN/(VN + FP) * 100\,, \tag{2.5}$$

où TVP est le taux de vrais positifs (pourcentage de potentiels d'erreurs correctement détectés) et TVN est le taux de vrais négatifs (pourcentage de réponses correctes classés comme telles).

— **Le score de Pierce.** Toujours à partir de la matrice de confusion, un deuxième indicateur permet d'évaluer la performance du classifieur pour la classe rare à partir du taux de bonne prévision des erreurs H et du taux de fausse alarme F respectivement définis par :

$$H = VP/(FP + VP) \quad \text{et} \quad F = FP/(VN + VP)\,. \tag{2.6}$$

Le score de Pierce (PSS : *Pierce's Skill Score*) se définit alors comme

$$PSS = H - F\,. \tag{2.7}$$

Par construction, $PSS \in [-1 : +1]$. Ce score évalue la qualité d'un modèle de prévision pour des évènements rares. Si ce score est supérieur à 0, le taux de bonne classification est supérieur à celui des fausses alarmes (erreurs détectées comme une réponse correcte) et plus il est proche de 1, meilleur est le modèle.

2.4.2 Jeu 2 : le P300 Speller

Le second jeu de données étudiées dans ce travail a été produit dans le cadre de la thèse de Margaux Perrin et pour plus de détails sur l'expérience, nous renvoyons le lecteur à l'article (Perrin et al., 2011).

Présentation du paradigme

Ces signaux EEG ont été acquis dans le cadre du paradigme ICM *P300 Speller* dont le but est d'épeler des lettres en décodant l'activité cérébrale (Farwell and Donchin, 1988) . Ce dispositif consiste à afficher une matrice 6×6 composée de l'alphabet et des chiffres (figure 2.8). Le participant souhaitant écrire une lettre doit se concentrer sur la case correspondante dans la matrice, dont les lignes et les colonnes sont successivement et aléatoirement *flashées*. Lorsque la ligne ou la colonne contenant la lettre cible est *flashée*, plusieurs composantes électrophysiologiques sont générées et en particulier une composante visuelle (l'onde négative N1) associée au stimulus visuel, suivie d'une seconde composante liée à la détection du stimulus (l'onde positive P300). Ces deux composantes résultent des modulations attentionnelles générées dans ce type de paradigme. Les principales composantes associées à ce paradigme sont présentées dans la figure 2.9. Notons que l'onde N1 est produite en réponse à un stimulus visuel. A chaque flash correspond donc l'apparition d'une N1 qui sera plus large lorsque la ligne ou la colonne flashée contient la lettre cible. C'est pourquoi, bien que le marqueur principal de cette ICM soit la P300, la réponse visuelle précoce N1 est généralement prise en compte pour la classification.

FIGURE 2.8 – Matrice de 6×6 contenant tous les caractères pouvant être épelés avec le P300 Speller. Ici, la troisième ligne est flashée.

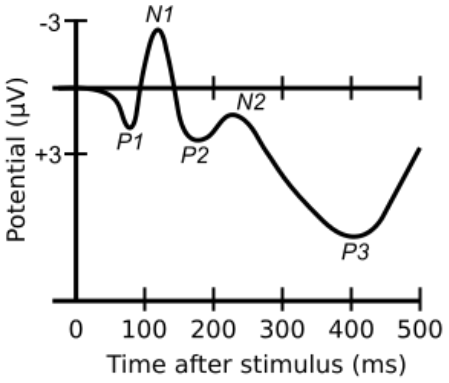


FIGURE 2.9 – Principales composantes évoquées associées au protocole du P300 Speller (source : Wikipédia).

Acquisition

Le jeu de données présenté ici est composé de 22 participants. L'expérience a été menée sur un ensemble de calibration basé sur 25 caractères et sur trois sessions tests de 120 lettres chacune. Deux classes sont donc considérées : les cibles ($c = 1$) et les non-cibles ($c = 2$). Afin d'épeler une lettre, chaque ligne et chaque colonne est *flashée* 3 fois, ce qui correspond à un total de 36 $flashs$ dont 6 sont des cibles. Nous disposons donc au total de 12960 essais pour chaque participant. Enregistrés avec une fréquence d'échantillonnage de 1000 Hz à l'aide de 32 électrodes placées sur l'ensemble du scalp, les essais correspondent au signal époché sur l'intervalle de temps [−100 ms ;+924 ms] où zéro représente l'instant du flash.

Pour un participant donné, on considère donc un ensemble de données correspondant aux 360 lettres épelées à l'issue des trois sessions. Notons que pour l'ensemble des analyses présentées dans la suite, il sera important de tenir compte des sessions au cours desquelles les performances des participants peuvent être impactées par des facteurs externes tels que la fatigue. Sur ce jeu de données il a été démontré dans (Perrin, 2012) une dégradation significative des résultats sur l'ensemble des sujets au cours des 3 sessions.

Pré-traitement des données et analyse dans l'espace des ondelettes

Les pré-traitements appliqués aux données brutes sont ceux préconisés dans (Rivet et al., 2009). Premièrement, les données non-épochées de chaque participant (enregistrement continu des 3 sessions) ont été filtrées à partir d'un filtre passe-bande de Butterworth (reference) d'ordre 4 pour ne conserver que les fréquences comprises entre 1 Hz et 20 Hz. Ensuite, les signaux ont été normalisés sur chaque électrode (moyenne nulle et variance à 1). Contrairement à Rivet *et al.*, les données ont été segmentées sans l'application de l'algorithme XDawn, la périodicité du signal étant prise en compte dans la transformation en ondelette appliquée par la suite. Les segments du signal correspondent à une période d'enregistrement de 1024 ms, de [−100 ms ; +924 ms], 0 correspondant au flash d'une ligne ou d'une colonne. Avec une fréquence d'échantillonnage de 1000 Hz, chaque essai correspond à 1024 pas de temps. Enfin, une transformation en ondelettes discrètes est appliquée afin de résumer l'information temporelle. Nous appliquons une transfor-

mation sur 6 niveaux de décomposition. La fréquence d'échantillonnage de ces données étant de 1000 Hz, les niveaux de décomposition des coefficients d'ondelette correspondent respectivement aux bandes de fréquence : $250-500$ Hz, $125-250$ Hz, $62,5-125$ Hz, $31-62$ Hz, $15-30$ Hz, $8-15$ Hz. La dernière bande de fréquence $0-8\ Hz$ est encodée par les coefficients d'échelle. La figure 2.10 représente l'ensemble des coefficients multi-échelle moyens correspondant aux bandes de fréquence allant de 0 à $62,5$ Hz pour les classes cible et non cible respectivement. Nous notons que les coefficients les plus significatifs sont concentrés sur les basses fréquences, et en particulier entre 0 Hz et 8 Hz. Notons enfin une différence importante de l'intensité des coefficients entre la classe cible et la classe non-cible.

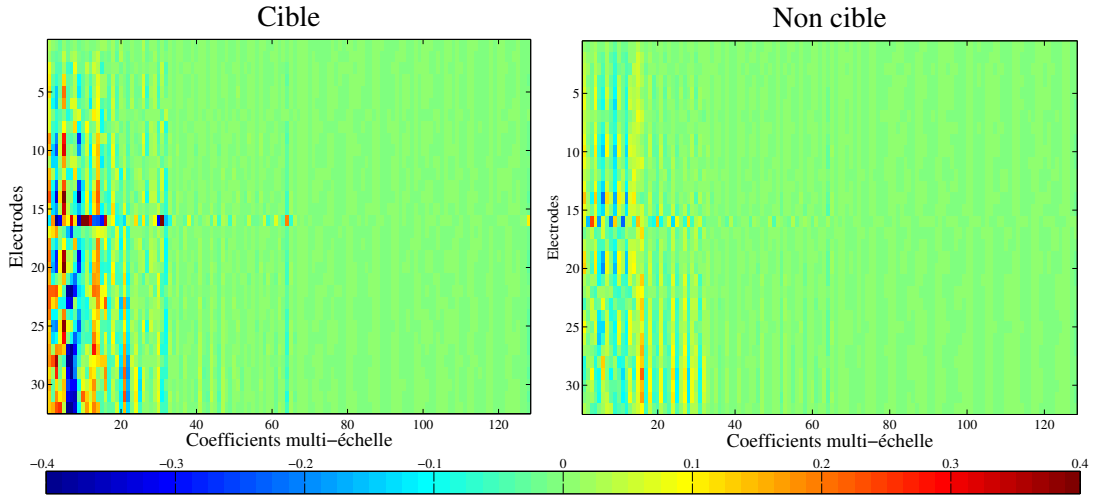


FIGURE 2.10 – Participant 4. Représentation des coefficients multi-échelle obtenus sur la moyenne des 120 lettres de la session 1 pour les classes cible (gauche) et non cible (droite). Les coefficients sont représentés en abscisses et les niveaux sont représentés du plus grossier au plus fin de gauche à droite : les 16 premiers coefficients sont les échelles, les 16 coefficients suivants sont les ondelettes de la bande $8-15$ Hz, etc.

Une réduction de dimension est appliquée afin de ne conserver que les coefficients multi-échelle les plus pertinents. Pour ces donnés, seuls les coefficients supérieurs (en valeur absolue) à la moyenne au carré calculée sur toutes les électrodes ont été sélectionnés. La figure 2.11 représente un exemple de la reconstruction temporelle obtenue pour les moyennes des classes cible et non cible obtenue à partir des coefficients multi-échelle sélectionnés.

Évaluation des classifieurs : taux de lettres correctement détectées

La procédure d'évaluation d'un classifieur pour le P300 Speller se base sur la comparaison du pourcentage de lettres correctement détectées. Une lettre correspond à un ensemble de 12 flashs (6 lignes et 6 colonnes). Selon les expériences, le choix d'une lettre se fait à partir de N_r répétitions des 12 flashs (dans l'expérience considérée ici, chaque ligne et chaque colonne est flashée aléatoirement 3 fois). En phase de test, la détection d'une lettre se fait à partir de la moyenne des N_r flashs d'une même colonne ou d'une même ligne qui sera ensuite classée. Ainsi, chaque ligne et chaque colonne dispose d'un score de classification qui sont ensuite comparés. La décision finalement prise sur la lettre est le résultat du croisement entre la ligne et la colonne

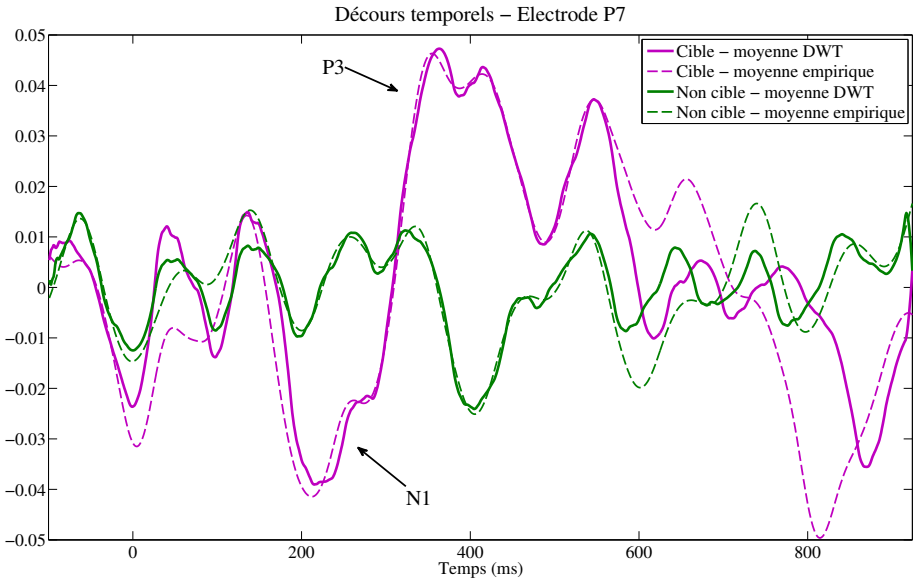


FIGURE 2.11 – Participant 4. Illustration de l'effet de la transformation en ondelette discrète et de la sélection de coefficients multi-échelle. Représentation du décours temporels obtenue par IDWT des coefficients multi-échelle sélectionnés et moyennés sur les essais dans les classes cible (violet) et non-cible (vert). Les moyennes obtenues sont assez similaires aux moyennes sur les essais bruts. En particulier on observe dans les deux cas les composantes N1 et P300 pour les essais cibles.

dont les scores discriminants sont maximums.
Généralement, les performances des classifieurs peuvent être comparées en fonction i) du nombre N_r de répétitions, ii) du nombre de lettres considérées dans le jeu d'apprentissage et iii) de la session.
Notons que dans le jeu de données présenté dans cette thèse, le nombre de répétitions est fixé à 3. L'ensemble des résultats présentés dans la suite sera donné donc pour $N_r = 3$. Cependant, il a été montré que plus le nombre de répétitions augmentait et plus les performances des classifieurs sont bonnes (Krusienski et al., 2006; Rivet et al., 2009; Manyakov et al., 2011). Nous nous plaçons ici dans un situation déforable puisque pour 3 répétitions le taux de lettres correctement classées peut être assez faible selon les jeux de données (Dans (Rivet et al., 2009) le taux de classification obtenu à partir d'une LDA à partir de filtres XDawn est de 65%). Pour le jeu de données étudié, les résultats moyens sur l'ensemble des participants sont de $83,25\%$ avec un écart-type de 11% (à partir d'une jeu d'apprentissage de 24 lettres et en utilisant l'algorithme XDawn ainsi qu'un classifieur naïf de Bayes). Dans ce travail nous évaluerons donc de la même façon les performances des modèles proposés.

Chapitre 3

L'analyse discriminante vectorielle : approche décisionnelle et descriptive en EEG

Sommaire

3.1 L'analyse discriminante en EEG

3.1.1 Contexte

Dans ce chapitre l'analyse discriminante est présentée selon deux points de vue distincts : le point de vue décisionnel et le point de vue descriptif (Hastie et al., 2009).

L'approche décisionnelle, connue sous les termes d'analyse discriminante linéaire (LDA) ou quadratique (QDA), est largement utilisée dans le domaine de l'apprentissage automatique et plus particulièrement pour l'analyse de signaux biologiques tels que les signaux EEG. Du fait de sa simplicité et de son efficacité en termes de classification, la LDA est un algorithme très populaire dans le domaine des interfaces cerveau-machine (Lotte et al., 2007a) et de l'analyse des signaux EEG (Parra et al., 2005; Blankertz et al., 2011). Pour un exemple d'application de l'analyse discriminante quadratique dans le cas des EEG, nous pouvons citer (Vidaurre et al., 2006; Kolodziej et al., 2012) qui utilisent cette méthode dans le contexte des ICM . Dans un

premier temps, nous allons présenter les principes de l'analyse discriminante décisionnelle et développer les problématiques associées à de tels algorithmes dans le cas des signaux EEG. Plus précisément, nous nous intéresserons à la différence entre l'analyse discriminante linéaire et quadratique, à la question de l'estimation de la ou des matrices de covariance et enfin nous introduirons le contexte des classes déséquilibrées.

La seconde approche, connue sous le nom d'analyse discriminante de Fisher, est moins répandue dans le cadre des signaux EEG. Cependant la version matricielle de cette méthode est parfois utilisée pour construire des filtres spatiaux comme c'est le cas dans nos applications (voir section 6.1). Il s'agit d'une méthode factorielle permettant, à partir de vecteurs d'observations, de définir des combinaisons linéaires discriminantes des variables de départ. L'objectif de la méthode est d'effectuer une réduction de dimension des données tout en préservant autant d'information discriminante que possible par la minimisation d'un critère, appelé critère de Fisher.

3.1.2 Vectorisation des signaux et notations

Dans ce chapitre, l'ensemble des méthodes d'analyse discriminante présentées ont été développées pour des données de type vectoriel. Concernant les signaux de type EEG, il est d'usage de les considérer sous forme de vecteur. Pour ce faire, chaque essai X prenant la forme d'une matrice *temps* $\times$ *espace* est réorganisé sous forme d'un vecteur x par concaténation verticale de ses colonnes. Ceci s'écrit formellement à partir de l'opérateur vec défini ci-après et dont les propriétés sont données dans l'annexe A.

Définition 2 (L'opérateur vec)**.** *Soit X une matrice de dimension $m \times n$, $x = \mathsf{vec}(X)$ est le vecteur de dimension mn obtenu par concaténation verticale des colonnes de X.*

Remarque 4. *Les termes "espace" et "temps" sont très génériques puisqu'ils se réfèrent aux dimensions spatiale et temporelle des signaux EEG. Il peut s'agir des signaux bruts, auquel cas les électrodes et les décours temporels seront les deux dimensions, comme des signaux projetés dans des sous-espaces définis par filtrage spatial et/ou par une transformation en ondelettes discrète.*

3.2 Mélange de gaussiennes multivariées et classification

Considérons n vecteurs d'observations $x_i \in \mathbb{R}^d$, pour $i = 1, \ldots, n$. Le but principal de la classification consiste à prédire la classe d'appartenance de chaque observation x_i. Nous nous intéressons dans ce cas à l'approche probabiliste qui se base sur l'estimation des probabilités d'appartenance de x_i à chaque classe G^c, $c = 1, \ldots, \mathcal{C}$. Le nombre d'observations appartenant à la classe c est noté n_c, et $\sum_{c=1}^{\mathcal{C}} n_c = n$.
Pour simplifier les notations, nous désignerons par la même lettre les variables aléatoires et leurs réalisations.

3.2.1 Formalisme de la règle de Bayes

Notons p^c la probabilité *a priori* qu'une observation x_i appartienne à la classe c, $\sum_{c=1}^{\mathcal{C}} p^c = 1$. Pour simplifier, on identifiera ici cette probabilité à la proportion de la classe c.

La distribution de probabilité d'une observation x_i dans la classe c est donnée par la densité de probabilité conditionnelle notée $f^c(x_i)$:

$$x_i|c \sim f^c(x_i)\,. \tag{3.1}$$

A partir de la formule des probabilités totales, la densité de probabilité de l'observation x_i, notée $f(x_i)$ est définie par la loi de mélange

$$f(x) = \sum_{c=1}^{\mathcal{C}} p^c f^c(x)\,. \tag{3.2}$$

D'après la formule de Bayes, la probabilité *a posteriori* d'appartenance de l'observation x_i à la classe c est alors donnée par :

$$\mathbb{P}(Classe = c|x_i) = \frac{p^c f^c(x_i)}{f(x_i)}\,. \tag{3.3}$$

La règle bayésienne de classification consiste à attribuer x_i à la classe dont la probabilité *a posteriori* est maximale. Le dénominateur étant constant au travers des classes, la règle est équivalente à maximiser le terme $p^c f^c(x_i)$ en fonction de c :

$$\hat{c} = \text{argmax}_c\, p^c f^c(x_i)\,, \tag{3.4}$$

Proposition 1 (Optimalité de la règle de Bayes)**.** *La règle de Bayes, aussi appelée règle du Maximum à Posteriori (MAP), est optimale au sens de la minimisation du risque moyen de classement, défini comme la probabilité moyenne de mauvaise classification (Saporta, 2011).*

Afin de mettre en œuvre une telle règle de décision, il est primordial de définir une loi de probabilité sur x. La densité de probabilité la plus souvent utilisée en classification est la loi gaussienne. Sous l'hypothèse de normalité, le problème décrit en (3.4) correspond au cas de l'analyse discriminante.

3.2.2 Analyse discriminante linéaire ou quadratique

Sous l'hypothèse de normalité, la densité de probabilité de x dans la classe c s'écrit :

$$f^c(x) = \frac{1}{(2\pi)^{d/2}|\Sigma^c|^{1/2}} \exp\left[-\frac{1}{2}(x-\mu^c)'(\Sigma^c)^{-1}(x-\mu^c)\right], \tag{3.5}$$

où $\mu^c \in \mathbb{R}^d$ est la moyenne de la classe c, $\Sigma^c \in \mathbb{R}^{d\times d}$ est la matrice de covariance de la classe c, supposée inversible et dont le déterminant est noté $|\Sigma^c|$.

Remarque 5. *A partir de la formule (3.2), notons qu'à présent chaque observation x est considérée comme la réalisation d'un mélange gaussien.*

Règle de classement théorique

Soit $\delta^c(x) = -2\log p^c f^c(x)$ la fonction discriminante de x dans la classe c. Dans le cas gaussien, la règle bayésienne (3.4) peut s'écrire sous la forme d'une minimisation de la fonction discriminante $\delta^c(x)$:

$$\hat{c} = \text{argmin}_c\, \delta^c(x), \tag{3.6}$$

avec

$$\delta^c(x) = (x-\mu^c)'(\Sigma^c)^{-1}(x-\mu^c) + \log|\Sigma^c| - 2\log p^c\,, \tag{3.7}$$

où $(x-\mu^c)'\Sigma^{-1}(x-\mu^c)$ est le carré de la distance entre l'observation x et μ^c le centre de la classe c, selon la métrique de Mahalanobis Σ^{-1}.

La règle de classification donnée en (3.6) définit des frontières de décisions quadratiques entre chaque classe. La méthode considérée correspond à une *analyse discriminante quadratique* (QDA : *quadratic discriminant analysis*).

Un cas particulier de l'analyse discriminante apparaît lorsque les matrices de covariance sont supposées égales :

$$\Sigma^c = \Sigma \quad \text{pour tout} \quad c = 1, \ldots, \mathcal{C}. \tag{3.8}$$

Dans cette situation, puisque le terme $\log|\Sigma^c|$ est constant à travers les classes, la fonction discriminante devient alors :

$$\delta^c(x) = (x-\mu^c)'\Sigma^{-1}(x-\mu^c) - 2\log p^c\ . \tag{3.9}$$

La méthode correspond alors à une *analyse discriminante linéaire* (LDA : *linear discriminant analysis*).

Dans le cas d'égalité des matrices de covariance, nous montrons que la règle bayésienne (3.6) peut s'écrire de la façon suivante :

$$\min_c\{(x-\mu^c)'\Sigma^{-1}(x-\mu^c) - 2\log(p^c)\} \tag{3.10a}$$

$$\iff \min_c\{-2x'\Sigma^{-1}\mu^c + (\mu^c)'\Sigma^{-1}\mu^c - 2\log(p^c)\} \tag{3.10b}$$

$$\iff \max_c\{x'\Sigma^{-1}\mu^c - \frac{1}{2}(\mu^c)'\Sigma^{-1}\mu^c + \log(p^c)\}. \tag{3.10c}$$

Lorsque les probabilités *a priori* p^c sont égales, la règle de Bayes est donc donnée par :

$$\max_c\{x'\Sigma^{-1}\mu^c - \frac{1}{2}(\mu^c)'\Sigma^{-1}\mu^c\}. \tag{3.11}$$

Classification binaire Considérons à présent le cas de deux groupes $c = \{1, 2\}$ avec égalité des covariances ($\Sigma^1 = \Sigma^2 = \Sigma$). La règle de décision est alors la suivante :
x est affectée à la classe 2 si

$$x'\Sigma^{-1}\mu^2 - \frac{1}{2}(\mu^2)'\Sigma^{-1}\mu^2 + \log(p^2) > x'\Sigma^{-1}\mu^1 - \frac{1}{2}(\mu^1)'\Sigma^{-1}\mu^1 + \log(p^1)\,. \tag{3.12}$$

Ou encore, en définissant le score

$$S(x) = x'\Sigma^{-1}(\mu^1 - \mu^2) - \frac{1}{2}(\mu^1 + \mu^2)'\Sigma^{-1}(\mu^1 - \mu^2) + s\,, \tag{3.13}$$

où $s = \log\left(\frac{p^2}{p^1}\right)$:
x est affectée à la classe 1 si $S(x) > 0$ et à la classe 2 sinon.

Erreur de classement théorique Considérons la probabilité *a posteriori* d'appartenance à la classe 2. Dans ce cas, l'erreur théorique de classement associée à l'observation x s'écrit :

$$e_r(2) = \mathbb{P}\left[S(x) > 0 \,|\, x \sim \mathcal{N}(\mu^2, \Sigma)\right]. \tag{3.14}$$

Afin de calculer cette probabilité d'erreur, nous introduisons la loi de probabilité associée à la variable aléatoire $S(x)$. Comme $S(x)$ s'écrit comme une combinaison linéaire des d composantes indépendantes de $x \sim \mathcal{N}(\mu^2, \Sigma)$, alors $S(x)$ est une variable aléatoire gaussienne. Son espérance et sa variance se définissent de la façon suivante (Saporta, 2011) :

$$\begin{aligned}
\mathbb{E}\left(S(x)\right) &= \mathbb{E}\left(x'\Sigma^{-1}(\mu^1 - \mu^2) - \frac{1}{2}(\mu^1 + \mu^2)'\Sigma^{-1}(\mu^1 - \mu^2) + s\right) && \text{(3.15a)}\\
&= \mu^2\Sigma^{-1}(\mu^1 - \mu^2) - \frac{1}{2}(\mu^1 + \mu^2)'\Sigma^{-1}(\mu^1 - \mu^2) + s && \text{(3.15b)}\\
&= -\frac{1}{2}(\mu^1 - \mu^2)'\Sigma^{-1}(\mu^1 - \mu^2) + s && \text{(3.15c)}\\
&= -\frac{1}{2}\Delta_d^2 + s, && \text{(3.15d)}
\end{aligned}$$

où $\Delta_d^2 = (\mu^1 - \mu^2)'\Sigma^{-1}(\mu^1 - \mu^2)$ est le carré de la distance de Mahalanobis entre les deux centres des classes.

$$\begin{aligned}
Var\left(S(x)\right) &= Var\left(x'\Sigma^{-1}(\mu^1 - \mu^2)\right) && \text{(3.16a)}\\
&= (\mu^1 - \mu^2)'\Sigma^{-1}\Sigma\Sigma^{-1}(\mu^1 - \mu^2) && \text{(3.16b)}\\
&= (\mu^1 - \mu^2)'\Sigma^{-1}(\mu^1 - \mu^2) && \text{(3.16c)}\\
&= \Delta_d^2, && \text{(3.16d)}
\end{aligned}$$

Nous en déduisons donc que $S(x) \sim \mathcal{N}\left(-\frac{1}{2}\Delta_d^2 + s, \Delta_d^2\right)$ et que

$$e_r(2) = \mathbb{P}\left(U > \frac{\Delta_d}{2} - \frac{s}{\Delta_d}\right) \quad \text{où} \quad U \sim \mathcal{N}(0,1). \tag{3.17}$$

Ainsi l'erreur de classification sera d'autant plus faible que la distance de Mahalanobis entre les deux centres sera grande.

3.2.3 Analyse discriminante pénalisée

Lorsque l'on dispose d'hypothèses ou d'informations *a priori* sur les données, il peut être intéressant de les exploiter afin d'améliorer l'analyse discriminante (en termes d'interprétabilité et de classification). C'est ainsi que peut se définir l'approche pénalisée de l'analyse discriminante (PDA : *Penalized discriminant analysis*), qui consiste à intégrer de l'information *a priori* dans la matrice de covariance Σ. Dans (Hastie et al., 1995), les auteurs proposent de remplacer la matrice de covariance Σ par une version pénalisée donnée sous la forme :

$$\Sigma + \alpha\Omega, \tag{3.18}$$

où Ω est la matrice de pénalisation qui "code" l'information *a priori* et α est un hyperparamètre permettant de régler l'influence de Ω sur la matrice de covariance. Dans leurs travaux, Hastie *et*

al. introduisent cette pénalisation dans le contexte du traitement d'images et proposent de renforcer le lissage spatial au travers de la matrice Ω. Dans le cadre des signaux EEG un tel *a priori* peut s'avérer intéressant, en particulier pour "lisser" la variabilité des signaux obtenus à partir des électrodes de surface. Ce point sera développé dans cette thèse et en particulier dans la section 4.2.4 pour une extension au cas de l'analyse discriminante matricielle et dans la section 6.1.1 pour une application à la problématique du filtrage spatial.

3.3 Estimation

3.3.1 Estimation de la règle de classement

Le problème d'estimation des matrices de covariance est central dans le contexte de classification considéré ici. En effet, les fonctions discriminantes (3.7) ou (3.9) utilisées dans les règles de décision se basent directement sur les paramètres du modèle de mélange gaussien μ^c et Σ^c (ou Σ dans le cas d'homoscédasticité). Ces paramètres étant inconnus, ils doivent être "remplacés" par leurs estimations. On parlera donc d'estimateur *plug-in* de la règle de classement.

La moyenne μ^c est estimée dans chacune des classes c par la moyenne empirique :

$$\bar{x}^c = \frac{1}{n_c} \sum_{i=1}^{n_c} x_i^c. \tag{3.19}$$

Pour la covariance, la règle quadratique implique l'estimation d'une matrice de covariance Σ^c par classe :

$$V^c = \frac{1}{n_c} \sum_{i=1}^{n_c} (x_i^c - \bar{x}^c)(x_i^c - \bar{x}^c)' \tag{3.20}$$

Dans le cas particulier de la LDA, les matrices de covariance étant toutes supposées égales, une seule matrice de covariance Σ est à estimer. Comme pour la QDA, l'estimateur classiquement introduit est la matrice de covariance empirique définie par :

$$V = \sum_{c=1}^{\mathcal{C}} p^c V^c. \tag{3.21}$$

Si p^c est estimé par le rapport $\frac{n_c}{n}$, alors la matrice de covariance empirique V s'écrit :

$$V = \frac{1}{n} \sum_{c=1}^{\mathcal{C}} \sum_{i=1}^{n_c} (x_i^c - \bar{x}^c)(x_i^c - \bar{x}^c)'. \tag{3.22}$$

Les estimateurs V^c et V sont biaisés et les estimateurs sans biais de Σ^c et Σ sont donnés par :

$$\hat{\Sigma^c} = \frac{n_c}{n_c - 1} V^c = \frac{1}{n_c - 1} \sum_{i=1}^{n_c} (x_i^c - \bar{x}^c)(x_i^c - \bar{x}^c)', \tag{3.23}$$

et

$$\hat{\Sigma} = \frac{n}{n - \mathcal{C}} V = \frac{1}{n - \mathcal{C}} \sum_{c=1}^{\mathcal{C}} \sum_{i=1}^{n_c} (x_i^c - \bar{x}^c)(x_i^c - \bar{x}^c)'. \tag{3.24}$$

Remarque 6. *Notons que dans le cas de la QDA, le nombre de paramètres à estimer est beaucoup plus important que pour la LDA, puisqu'il est multiplié par le nombre de classes. En effet, elle requiert* $\frac{\mathcal{C}d(d+1)}{2}$ *paramètres de covariances à estimer contre* $\frac{d(d+1)}{2}$ *paramètres de covariance dans le cas d'égalité des covariances.*

3.3.2 Tester l'égalité des matrices de covariance Σ^c

Comme les matrices de covariances Σ^c sont, en pratique, inconnues, leur égalité peut être testée *via* la mise en œuvre d'un test statistique (Anderson, 2003).

Nous introduisons ici le test de Box (e.g. (Saporta, 2011)), qui est une extension du test de Bartlett d'égalité des variances dans le cas d'un mélange gaussien multivarié.

Nous souhaitons tester les hypothèses suivantes :

$$\mathsf{H_0}: \Sigma^1 = \Sigma^2 = \ldots = \Sigma^{\mathcal{C}} = \Sigma \quad \text{contre} \quad \mathsf{H_1}: \text{Il existe au moins un couple } (l,q) \text{ tel que } \Sigma^l \neq \Sigma^k .$$

Chaque observation d'une classe c est supposée être la réalisation d'une loi normale $\mathcal{N}(\mu^c, \Sigma^c)$, où la matrice de covariance inconnue Σ^c est estimée par $\hat{\Sigma}^c$ défini en (3.23).

Soit M la statistique du test de Box définie par :

$$M = \left(1 - \frac{2d^2+3d-1}{6(d+1)(\mathcal{C}-1)}\right)\left[\left(\sum_{c=1}^{\mathcal{C}} \frac{1}{n_c-1} - \frac{1}{n-\mathcal{C}}\right)(n-\mathcal{C})\log|\hat{\Sigma}| - \sum_{c=1}^{\mathcal{C}}(n_c-1)\log|\hat{\Sigma}^c|\right] \tag{3.25}$$

Sous $\mathsf{H_0}$ (égalité des matrices de covariance), M suit approximativement une loi du χ^2 à ν degrés de liberté, avec :

$$\nu = \frac{d(d+1)(\mathcal{C}-1)}{2} . \tag{3.26}$$

Il est important de noter que le test de Box est sensible à la normalité des données considérées. Ainsi, lorsque cette hypothèse n'est pas vérifiée, le rejet de $\mathsf{H_0}$ peut correspondre au rejet de la normalité des données et non pas au rejet de l'égalité des covariances. De plus, ce test n'est pas très puissant et n'est efficace que pour des échantillons de taille assez grande (supérieure à d). Ainsi, lorsque les échantillons sont de petite taille, la statistique de test est très peu robuste et quelle que soit la conclusion du test il sera souvent malgré tout préférable de choisir la règle linéaire. De plus, la règle linéaire ne nécessite l'estimation que d'une seule matrice de covariance commune aux $\mathcal{C}$ classes. Tandis que dans le cas de la QDA, il est nécessaire d'estimer une matrice de covariance par classe, augmentant de manière significative le nombre de paramètres à estimer. Nous reviendrons plus en détails sur ce dernier point dans la section suivante.

Remarque 7. *En pratique, il arrive que la statistique de test ne puisse pas être calculée du fait d'un échantillon trop petit au regard de d. En effet, si la matrice de covariance estimée est singulière, alors* $\log|\frac{n_c}{n_c-1}V^c|$ *diverge.*

3.3.3 Le problème des grandes dimensions

Les estimations des matrices de covariance et de leur inverse sont au centre de la mise en œuvre de l'analyse discriminante. Cependant, dans le cadre de données de grandes dimensions, le procédé d'estimation peut être compromis . En effet, dans de nombreuses applications telles que l'étude de gènes ou bien l'analyse de signaux EEG, il n'est pas rare que le nombre de paramètres

à estimer soit bien largement supérieur au nombre d'observations, et dans de telles situations la matrice de covariance empirique est singulière.

Lorsque $n < d$, la matrice de covariance empirique n'est pas inversible et même lorsque $n \geq d$ mais n et d sont du même ordre de grandeur, la matrice de covariance empirique peut être inversible mais mal conditionnée (le déterminant de la matrice n'est que légèrement différent de zéro). Dans ce cas, certaines valeurs propres étant très petites, la matrice inverse a donc des valeurs propres très élevées ce qui entraîne de fortes instabilités dans la règle de décision.

C'est pour cette raison que l'hypothèse d'homogénéité des matrices de covariance est souvent faite même si elle n'est pas réaliste, ceci permettant l'estimation de la matrice de covariance sur l'ensemble des données d'apprentissage toute classe confondue. En effet les matrices de covariances empiriques calculées dans chaque classe sont souvent singulières ou mal conditionnées notamment dans les classes de petit effectif.

3.3.4 Un compromis entre la LDA et la QDA ?

Nous avons pu voir que le choix entre une approche linéaire ou une approche quadratique de l'analyse discriminante n'est pas un choix trivial. En effet, même si les matrices de covariance sont supposées différentes ou bien si le test de Box confirme cette hypothèse, les problématiques liées à leur estimation demeurent contraignantes et notamment dans le cas où le rapport $d/n_c \geq 1$.

Motivée par une telle situation, une extension de l'analyse discriminante a été proposée (Friedman, 1988). Elle effectue un compromis entre l'approche linéaire et l'approche quadratique, en introduisant l'estimateur

$$V^c(\lambda) = \frac{S^c(\lambda)}{W^c(\lambda)}, \tag{3.27}$$

où

$$S^c(\lambda) = (1-\lambda)S^c + \lambda S \ , \tag{3.28}$$

avec

$$S^c = \sum_{i=1}^{n_c} (x_i^c - \bar{x}^c)(x_i^c - \bar{x}^c)' \ , \quad S = \sum_{c=1}^{\mathcal{C}} S^c \ , \tag{3.29}$$

et

$$W^c(\lambda) = (1-\lambda)(n_c - 1) + \lambda(n - \mathcal{C}). \tag{3.30}$$

$\lambda \in \{0; 1\}$ est le paramètre de régularisation. Il contrôle le degré de régularisation de la matrice de covariance empirique de la classe c par rapport à la matrice de covariance intra-classe. Lorsque $\lambda = 0$, l'expression (3.27) correspond à la QDA (hétéroscédasticité) tandis que lorsque $\lambda = 1$, on retrouve la LDA (homoscédasticité). Pour des valeurs de λ comprises strictement entre 0 et 1, $V^c(\lambda)$ correspond à un compromis plus ou moins strict entre les deux approches.

Si cette approche régularisée présente l'avantage d'un compromis entre les deux méthodes, elle n'en reste pas moins sensible au cas où $n_c < d$. Dans ce cas la matrice de covariance empirique de la classe c est singulière et la seule régularisation permettant de pallier ce problème est de considérer $\lambda = 1$. De même, si $n = \sum_{c=1}^{\mathcal{C}} n_c$ est plus petit que ou comparable à d, la LDA elle-même devient mal posée. C'est pourquoi d'autres approches de régularisation ont été proposées afin d'obtenir de meilleures estimations des matrices de covariance, y compris dans le cas de petits échantillons.

Notons que cette méthode ne sera pas appliquée dans la suite de ces travaux. Elle a été présentée afin de fournir un état de l'art des méthodes les plus populaires développées dans le cadre de l'analyse discriminante.

3.3.5 Approches régularisées

Régularisation de la matrice de covariance empirique

Le principe de régularisation de la matrice de covariance peut être vu comme une façon de contrôler la variance de l'estimateur, généralement au prix d'une augmentation de son biais (Hastie et al., 2009). Dans cette section, nous nous limitons au cas d'égalité des matrices de covariance, mais la discussion peut être transposée sans difficulté au cas général.

A partir du formalisme utilisé dans (Ledoit and Wolf, 2003) et repris ensuite de façon plus large dans (Schafer and Strimmer, 2005), nous introduisons les trois matrices suivantes :

1) La matrice de covariance $\Sigma \in \mathbb{R}^{d\times d}$, inconnue, supposée inversible.
2) Son estimateur sans biais $\hat{\Sigma} \in \mathbb{R}^{d\times d}$ défini en (3.24).
3) Une matrice de covariance "cible" F. Cette matrice de covariance est un estimateur biaisé de Σ. Le choix de la structure de F dépend des connaissances sur les données. De façon générale la structure introduite a pour avantage d'avoir moins de paramètres à estimer et donc F est plus facilement calculable que $\hat{\Sigma}$ dans le cas où n est petit ou équivalent à d.

Différentes structures peuvent être considérées pour F, par exemple :
— $F = \text{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_d^2)$
— $F = \sigma^2 I_d$

Définir une version régularisée de la matrice de covariance consiste à proposer une combinaison convexe de son estimateur sans biais $\hat{\Sigma}$ et de la matrice de covariance cible F. En notant $\Sigma^\star$ l'estimateur régularisé, celui-ci s'écrit

$$\Sigma^\star = (1-\gamma)\hat{\Sigma} + \gamma F\,, \tag{3.31}$$

où $\gamma \in [0,1]$ est la constante de régularisation.

Cette approche est très générale puisque le choix de la structure imposée à la covariance "cible" dépend de l'expérimentateur et/ou de la problématique considérée. Nous nous référons par exemple à (Schafer and Strimmer, 2005) dans lequel six cibles de régularisation couramment utilisées sont étudiées. En particulier nous retrouvons la matrice cible de la forme suivante :

$$F = \frac{\text{Tr}(\hat{\Sigma})}{d} I_d\,. \tag{3.32}$$

Dans ce cas, l'estimateur sans biais est régularisé par ajout d'une constante (égale à sa "valeur propre moyenne") à ses éléments diagonaux $\hat{\Sigma}_{k,k}$. Cette approche a été initialement proposée par Friedman (Friedman, 1988) dans le contexte de l'analyse discriminante qui conjugue l'approche régularisée de $\hat{\Sigma}^c$ avec le compromis LDA/QDA présenté dans la section 3.3.4 :

$$\hat{\Sigma}^c(\lambda,\gamma) = (1-\gamma)\hat{\Sigma}^c(\lambda) + \gamma \frac{\text{Tr}\left[\hat{\Sigma}^c(\lambda)\right]}{d} I_d\,, \tag{3.33}$$

où

$$\hat{\Sigma}^c(\lambda) = \frac{1}{(1-\lambda)(n_c-1)+\lambda(n-\mathcal{C})}(1-\lambda)\Sigma^c + \lambda\Sigma\,. \tag{3.34}$$

La fonction discriminante estimée s'écrit alors, à partir de l'estimateur *plug-in* (3.33) :

$$\hat{\delta}^c(x) = (x-\hat{\mu}^c)'(\hat{\Sigma}^c(\lambda,\gamma))^{-1}(x-\hat{\mu}^c) + \log|\hat{\Sigma}^c(\lambda,\gamma)| - 2\log p^c. \tag{3.35}$$

Cette méthode est connue sous le nom d'analyse discriminante régularisée (RDA : regularized discriminant analysis). Dans la suite de ce travail, nous nous intéresserons à un cas particulier de cette approche, en considérant $\lambda = 1$ (et donc $\hat{\Sigma}^c(\lambda) = \hat{\Sigma}$). Dans ce cas, la régularisation sera définie uniquement sur les termes diagonaux de $\hat{\Sigma}$ en fonction de la valeur du paramètre de régularisation $\gamma \in [0;1]$.

Détermination du paramètre de régularisation

Nous nous intéressons ici au choix de la valeur optimale pour le paramètre de régularisation γ. Deux méthodes sont tout particulièrement étudiées : la validation croisée et une construction explicite de l'hyperparamètre.

Validation croisée Cette méthode consiste à définir le ou les hyperparamètre(s) à partir d'un échantillon de validation sur lequel le modèle est évalué à partir d'un critère. En classification, ce critère est communément le taux d'erreur de classement (Friedman, 1988).

Construction explicite de l'hyperparamètre Nous nous intéressons ici à l'approche introduite par (Ledoit and Wolf, 2003) et reprise dans de nombreux travaux (voir en particulier (Schafer and Strimmer, 2005) et (Blankertz et al., 2011) pour une application en EEG). Elle consiste à déterminer une bonne valeur de γ de façon analytique par la minimisation de l'erreur quadratique moyenne de l'estimateur $\Sigma^\star$ défini en (3.31).

La fonction de coût de l'estimateur $\Sigma^\star$ est définie par

$$L(\lambda) = \|\Sigma^\star - \Sigma\|_F^2\ , \tag{3.36}$$

où $\|Z\|_F^2$ est la norme de Frobenius au carré de la matrice Z.

Afin de définir la valeur optimale de régularisation, les auteurs proposent de minimiser par rapport à γ l'erreur quadratique moyenne, définie comme l'espérance de la fonction de coût : $R(\gamma) = \mathbb{E}(L(\gamma))$. Dans ce cas, la valeur optimale de γ s'obtient en résolvant $R'(\gamma) = 0$. Pour le détail des calculs et pour les propriétés asymptotiques de l'hyperparamètre nous renvoyons à (Ledoit and Wolf, 2003; Ledoit and Wolf, 2004).

Dans cette thèse, nous avons opté pour l'approche proposée par Blankertz *et al.*, correspondant à la covariance régularisée définie en (3.33), avec $\lambda = 1$. Dans ce cas, la matrice cible est celle définie en (3.32) et l'hyperparamètre $\gamma^\star$ est défini de façon analytique (pour les détails des calculs voir l'article (Blankertz et al., 2011).

3.3.6 Cas de deux classes déséquilibrées

Comme nous l'avons vu dans le chapitre 2, les problèmes de classification binaire $c = 1, 2$, avec des classes déséquilibrées ($n_1 \ll n_2$) constituent une situation très fréquente dans le cadre des signaux EEG.

Positionnement du problème

Dans le cas binaire, l'estimation de la matrice de covariance Σ est donnée par :

$$\hat{\Sigma} = \frac{n_2 - 1}{n - 2}\hat{\Sigma}^2 + \frac{n_1 - 1}{n - 2}\hat{\Sigma}^1 . \tag{3.37}$$

Il est ainsi clair que dans le cas déséquilibré, la matrice de covariance sera principalement déterminée par la classe majoritaire. Il est nécessaire de distinguer deux cas de figure.

Cas 1 : homoscédasticité. Lorsque $\Sigma^1 = \Sigma^2 = \Sigma$, la règle de décision se base sur le signe de $w'x_i$, où $w = \hat{\Sigma}^{-1}(\hat{\mu}^2 - \hat{\mu}^1)$. La qualité de la décision dépend largement de la qualité de l'estimation de la matrice de covariance et de son inverse. Dans le cas d'égalité des matrices de covariance le déséquilibre des classes n'introduit pas de difficulté supplémentaire puisque Σ est estimée à partir du jeu de données complet.

Cas 2 : hétéroscédasticité. Dans ce cas, deux matrices de covariance doivent être estimées, ce qui amplifie les problèmes présentés précédemment. En particulier, lorsque le jeu de données est petit et très déséquilibré, au moins l'une des matrices de covariance sera mal estimée. Ceci explique pourquoi la LDA est généralement préférée, même dans le cas où les *vraies* matrices de covariance sont différentes. Cependant, lorsque le jeu de données est déséquilibré, l'estimation de Σ (3.37) a tendance à se concentrer sur la classe majoritaire et à ignorer la classe rare. Ainsi, $\hat{\Sigma}$ est dominée par la classe la plus grande de sorte que $\hat{\Sigma} \approx \hat{\Sigma}^2$ (dans le cas où $n_1 \ll n_2$).

Sous- et sur-échantillonnage

Dans une telle situation, une solution communément appliquée consiste à rééquilibrer les jeux de données en utilisant des méthodes de sur- ou de sous-échantillonnage (Xie and Qiu, 2006; Xue and Titterington, 2008; Weiss, 2004).
Le sur-échantillonnage consiste à augmenter la taille de la classe minoritaire mais sans gain réel d'information. Le sous-échantillonnage permet quant à lui de diminuer la taille de la classe majoritaire. Dans le cas où la classe minoritaire est vraiment de petite taille, cette opération de sous-échantillonnage peut avoir pour conséquence une perte significative d'information et une dégradation de l'estimation de la covariance. Ces procédures peuvent s'avérer utiles dans des contextes particuliers et il faut donc être précautionneux quant à leur utilisation (Kotsiantis et al., 2006).

Le cas des matrices de covariance à dominante diagonale

Covariance diagonale Lorsque les variables observées sont supposées décorrélées, leur matrice de covariance est diagonale et seulement les d variances doivent être estimées. La LDA diagonale (Dudoit et al., 2002) est basée sur cette hypothèse :

$$\Sigma = \mathrm{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_d^2) . \tag{3.38}$$

Sous l'hypothèse de normalité, ceci est équivalent à supposer l'indépendance des variables.

Cependant, il a été montré que même lorsque l'on considère des variables qui ne sont pas décorrélées, de meilleurs résultats de classification peuvent être obtenus en remplaçant $\hat{\Sigma}$ par sa

diagonale (3.38) (ceci correspond au classifieur naïf de Bayes, voir *e.g.* (Bickel and Levina, 2004) et les références citées dans cet article).
Dans des situations telles que celle exposée dans cette thèse, une transformation peut être appliquée aux données afin de se rapprocher de cette hypothèse de décorrélation. Comme nous le verrons, c'est notamment le cas de la transformation en ondelettes discrète (voir section 2.2.2) que nous utiliserons dans nos modèles.
Dans la situation d'hétéroscédasticité, le même type d'hypothèse peut être faite sur chacune des classes (on parlera alors de QDA diagonale) (Dudoit et al., 2002).

Régularisation La régularisation de la matrice de covariance présentée dans la section 3.3.5 permet en prenant pour matrice cible une matrice diagonale de renforcer les termes diagonaux de l'estimateur sans biais $\hat{\Sigma}$. Cette méthode est notamment utile dans le cas des petits échantillons.

3.4 Approche descriptive pour l'extraction de variables discriminantes

Nous abordons à présent l'extraction de caractéristiques par l'analyse discriminante de Fisher. Nous rappelons que nous sommes dans le cas vectoriel. L'analyse se fait donc à partir de vecteurs d'observations de $\mathbb{R}^d$, où d est le nombre de variables considérées. En EEG, une observation correspond à un essai vectorisé (voir la section (3.1.2)).

3.4.1 Réduction de dimensions et variables discriminantes

L'approche descriptive de l'analyse discriminante consiste à trouver un sous-ensemble de variables permettant de discriminer les observations. Il s'agit d'une méthode factorielle cherchant les combinaisons linéaires des variables permettant de séparer le mieux possible les $\mathcal{C}$ classes. Elle correspond à une analyse en composantes principales (ACP) particulière sur les centres de chaque classe avec la métrique de Mahalanobis $\hat{\Sigma}^{-1}$. En ce sens, l'approche descriptive de la LDA ne fait appel à aucune loi de probabilité mais elle se base sur la notion de projection.

Définitions

Soit M une matrice symétrique définie positive. On munit l'espace $\mathbb{R}^d$ de la métrique M. Étant données n observations $x_i \in \mathbb{R}^d$, $i = 1, \ldots, n$, l'inertie totale du nuage des observations autour de son centre de gravité est donnée par

$$I_g = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})' M (x_i - \overline{x}), \tag{3.39}$$

où $\overline{x}$ est le centre de gravité des observations, défini par

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \,. \tag{3.40}$$

Plus l'inertie est grande, plus le nuage est dispersé, et au contraire plus elle est petite, plus le nuage est concentré autour de son centre de gravité.

Considérons à présent que les observations x_i sont réparties en $\mathcal{C}$ classes constituant $\mathcal{C}$ sous-nuages disjoints deux à deux. Soit $p^c = n_c/n$ la proportion de la classe c. L'inertie intra-classe est la somme pondérée des inerties des sous-nuages :

$$I_{intra} = \sum_{c=1}^{\mathcal{C}} p^c I_c \,, \tag{3.41}$$

où I_c est l'inertie de nuage des observations de la classe c définie par :

$$I_c = \frac{1}{n_c} \sum_{i=1}^{n_c} (x_i^c - \overline{x}^c)' M (x_i^c - \overline{x}^c), \tag{3.42}$$

et $\overline{x}^c$ est le centre de gravité des observations dans cette classe :

$$\overline{x}^c = \frac{1}{n_c} \sum_{i=1}^{n_c} x_i^c \,. \tag{3.43}$$

L'inertie totale d'un nuage n'est généralement pas égale à la somme des inerties des classes qui la composent, c'est-à-dire à l'inertie intra-classe car il faut tenir compte également de la dispersion des classes par rapport au centre de gravité du nuage global. L'inertie inter-classe mesure "l'éloignement" des centres de gravité des classes entre eux :

$$I_{inter} = \sum_{c=1}^{\mathcal{C}} p^c (\overline{x}^c - \overline{x})' M (\overline{x}^c - \overline{x}). \tag{3.44}$$

Théorème 2 (Décomposition de Huygens)**.** *L'inertie totale d'un nuage de points composé de différentes classes disjointes deux à deux est la somme de son inertie intra-classe et de son inertie inter-classe (Saporta, 2011) :*

$$I_g = I_{inter} + I_{intra}. \tag{3.45}$$

A partir de ces notions d'inertie, nous pouvons définir les matrices de covariance intra-classe et inter-classe. Pour simplifier, nous nous plaçons dans le cas où $M = I_d$, la matrice identité de dimension d.

La matrice de covariance intra-classe est la moyenne pondérée des matrices de covariance empiriques de chaque classe c :

$$S_W = \sum_{c=1}^{\mathcal{C}} p^c V^c = \frac{1}{n} \sum_{c=1}^{\mathcal{C}} \sum_{i=1}^{n_c} (x_i^c - \bar{x}^c)(x_i^c - \bar{x}^c)', \tag{3.46}$$

où $p^c = n_c/n$ et V^c est la matrice de covariance empirique de la classe c, définie en (3.20).

Propriété 1. *L'inertie intra-classe est égale à la trace de S_W :*

$$I_{intra} = \mathrm{Tr}\,(S_W)\,. \tag{3.47}$$

La matrice de covariance inter-classe est la matrice de covariance des $\mathcal{C}$ centres de gravité :

$$S_B = \sum_{c=1}^{\mathcal{C}} p^c (\bar{x}^c - \bar{x})(\bar{x}^c - \bar{x})' = \frac{1}{n} \sum_{c=1}^{\mathcal{C}} n_c (\bar{x}^c - \bar{x})(\bar{x}^c - \bar{x})'. \tag{3.48}$$

Propriété 2. *L'inertie inter-classe est égale à la trace de S_B :*

$$I_{inter} = \mathrm{Tr}\,(S_B)\,. \tag{3.49}$$

Le discriminant de Fisher

Comme brièvement évoqué en introduction de cette section, l'intérêt de la LDA d'un point de vue descriptif est de définir un sous-ensemble de variables discriminantes permettant de séparer au mieux les classes. En particulier, nous cherchons la projection discriminante telle que l'inertie inter-classe des observations projetées soit maximale et leur inertie intra-classe minimale. Soit $u \in \mathbb{R}^d$, le vecteur de projection. Notons Δ_u l'axe engendré par u.

Définition 3 (Critère de Fisher)**.** *L'analyse discriminante descriptive consiste à trouver le vecteur u qui maximise le critère de Fisher, défini comme le quotient :*

$$J(u) = \frac{u' S_B u}{u' S_W u}\,, \tag{3.50}$$

où $u'S_Bu$ et $u'S_Wu$ sont respectivement l'inertie inter-classe et l'inertie intra-classe du nuage projeté sur Δ_u.

Nous nous intéressons à présent à trouver le vecteur $u^\star \in \mathbb{R}^d$ permettant de définir l'axe de projection qui maximise le discriminant de Fisher :

$$u^\star = \arg\max_{u \in \mathbb{R}^d} J(u)\,. \tag{3.51}$$

Théorème 3. *Le critère de Fisher est maximisé lorsque*

$$S_B u = \lambda S_W u, \tag{3.52}$$

où u est un vecteur propre généralisé de S_B et S_W associé à la plus grande valeur propre non nulle λ.

Démonstration.

$$\frac{\delta J(u)}{\delta u} = 0 \tag{3.53a}$$

$$\iff \quad \frac{2(u'S_Wu)S_Bu - 2(u'S_Bu)S_Wu}{(u'S_Wu)^2} = 0 \tag{3.53b}$$

$$\iff \quad (u'S_Wu)S_Bu - (u'S_Bu)S_Wu = 0 \tag{3.53c}$$

$$\iff \quad S_Bu - \frac{(u'S_Bu)}{(u'S_Wu)} S_Wu = 0 \tag{3.53d}$$

$$\iff \quad S_Bu = \lambda S_W u\ , \tag{3.53e}$$

où $\lambda = \frac{(u'S_Bu)}{(u'S_Wu)} = J(u)$. Ainsi en prenant λ égal à la plus grande valeur propre on maximise le critère. Maximiser le critère de Fisher est donc bien équivalent à résoudre le problème aux valeurs propres généralisé et choisir un vecteur propre $u_1^\star$ associé à la plus grande valeur propre (on choisit généralement $u_1^\star$ S_W-normé ($u_1^{\star\prime} S_W u_1^\star = 1$)). □

Les composantes de $u_1^\star$ sont les coefficients de la combinaison linéaire la plus discriminante des variables de départ. On cherche ensuite un vecteur u_2 orthogonal à $u_1^\star$ selon la métrique S_W ($u_1^{\star\prime} S_W u_2 = 0$) qui maximise $J(u)$ et ainsi de suite.

Plus formellement, le problème de Fisher conduit au problème d'optimisation suivant. Il s'agit de résoudre séquentiellement

$$u_l^\star = \arg\max_{u_l \in \mathbb{R}^d} \{u_l' S_B u_l \quad \text{sous contraintes} \;\; u_l' S_w u_l \leq 1 \;\; \text{et} \;\; u_l' S_w u_i^\star = 0 \; \forall i < l\} \; . \tag{3.54}$$

La solution au problème est donnée par les vecteurs propres généralisés de S_B et S_W associés aux valeurs propres non nulles classées par ordre décroissant. Si S_W est inversible, la solution correspond aux vecteurs propres de la matrice $S_W^{-1} S_B$. Soit $k = \mathsf{rang}(S_W^{-1} S_B) \leq d$. Notons que la matrice $S_W^{-1} S_B$ est une matrice S_W-symétrique et qu'il existe une base S_W-orthonormée de vecteurs prores $(u_1^\star, \ldots, u_k^\star)$ associé aux k valeurs propres non nulles $\lambda_1 \geq \lambda_2 \geq \ldots \lambda_k > 0$.

Dans la suite la solution de ce problème est présentée pour le cas binaire ($\mathcal{C} = 2$) puis dans le cas de l'analyse discriminante multiple ($\mathcal{C} > 2$).

Cas de deux classes

Dans le cas de deux classes $c \in \{1; 2\}$, les matrices de covariances inter-classe et intra-classe s'écrivent :

$$S_B = \frac{n_1 n_2}{n^2} (\bar{x}^1 - \bar{x}^2)(\bar{x}^1 - \bar{x}^2)' \, , \tag{3.55}$$

et

$$S_W = p^1 V^1 + p^2 V^2 \, . \tag{3.56}$$

Alors, si S_W est inversible ($\mathsf{rang}(S_W) = d$), la solution optimale du problème de Fisher (3.50) est donnée par :

$$u^* = S_W^{-1} (\bar{x}^1 - \bar{x}^2) \, , \tag{3.57}$$

avec $J(u^\star) = (\bar{x}^1 - \bar{x}^2)' S_W^{-1} (\bar{x}^1 - \bar{x}^2)$, le carré de la distance de Mahalanobis entre les centres des deux classes.

Remarque 8 (Sur le rang de S_W)**.** *Dans le cas binaire,* $\mathsf{rang}(S_B) = 1$*, donc le rang de la matrice* $S_W^{-1} S_B$ *est aussi égal à 1. Le vecteur propre* u^**, associé à la valeur propre* $\lambda = J(u^\star)$ *définit l'unique direction discriminante.*

L'analyse discriminante multiple $\mathcal{C} > 2$

Lorsque l'on considère plus de 2 classes avec $d > 1$, le problème de Fisher (3.54) implique en général plus d'une direction discriminante. Notons que $\mathsf{rang}(S_B) \leq \mathcal{C} - 1$ donc lorsque S_W est inversible ($\mathsf{rang}(S_W) = d$) on a $\mathsf{rang}(S_W^{-1} S_B) \leq \min(\mathcal{C} - 1, d)$. Par conséquent lorsque $\mathcal{C} - 1 < d$, $k = \mathsf{rang}(S_W^{-1} S_B) \leq \mathcal{C} - 1$. Il y a donc au plus $\mathcal{C} - 1$ valeurs propres non-nulles associées au critère de Fisher.

Remarque 9. *Alternativement, ces résultats peuvent être obtenus à partir d'un critère de Fisher généralisé,*

$$J(U) = \frac{|U' S_B U|}{|U' S_W U|} \, , \tag{3.58}$$

où $U \in \mathbb{R}^{d \times k}$ *est cette fois la matrice dont les colonnes sont les vecteurs discriminants recherchés. L'optimisation de* $J(U)$ *produit une matrice* $U^\star = [u_1^\star, u_2^\star, \ldots, u_k^\star]$ *dont les colonnes sont les* k *vecteurs propres associés aux* k *valeurs propres non nulles de la matrice* $S_W^{-1} S_B$*.*

3.4.2 Approches régularisées

L'analyse discriminante peut être modifiée de différentes façons, notamment par des régularisations introduisant des informations a priori soit sur la matrice de covariance intra-classe soit directement sur les vecteurs discriminants recherchés.

Par exemple, dans (Witten and Tibshirani, 2011), une pénalisation ℓ_1 (LASSO) est introduite pour générer des vecteurs discriminants parcimonieux, et une pénalité de type *Fused LASSO* (Tibshirani et al., 2005) est aussi proposée pour combiner parcimonie et régularité (variation lente des composantes des vecteurs). Cette approche demande que la matrice de covariance intra-classe S_W soit inversible, et cette dernière y est remplacée par l'estimateur diagonal $diag(S_W)$ (Dudoit et al., 2002; Bickel and Levina, 2004).

Dans (Clemmensen et al., 2011), la pénalisation ℓ_1 sur les vecteurs discriminants est combinée à une régularisation de la matrice S_W effectuée par ajout d'une matrice définie positive. Une formulation alternative du problème utilisant des *scoring vectors* (voir (Hastie et al., 2009)) conduit à un problème d'optimisation de type *Elastic net* (Zou and Hastie, 2005), pour lequel il existe des algorithmes efficaces.

Conclusion Partielle

Dans cette première partie nous avons décrit les principales problématiques liées à l'analyse de signaux EEG :

— signaux multi-capteurs, de grandes dimensions,
— petits échantillons d'apprentissage,
— classes généralement déséquilibrées.

Nous avons également présenté certains outils de traitement qui seront importants par la suite (ondelettes, filtrage spatial), ainsi que les méthodes de l'analyse discriminante décisionnelle et descriptive et leurs adaptations à ce contexte.

Les performances de l'analyse discriminante dépendent fortement de la qualité de l'estimation des matrices de covariance, ce qui est problématique ici. L'approche que nous avons choisi de développer conjugue réduction de dimension et modélisation afin de réduire le nombre de paramètres de covariance à estimer.

Nous allons plus précisément considérer deux types de modélisation de la covariance :

— Décomposition en produit de Kronecker de covariances spatiale et temporelle
— Décomposition en composantes via le modèle mixte.

La décomposition en ondelettes et le filtrage spatial jouent un rôle fondamental dans ces deux approches. La transformation en ondelettes sera utilisée en amont dans les deux cas, permettra une importante réduction de dimension dans le domaine temporel et produira des coefficients multi-échelle quasiment décorrélés. Le filtrage spatial, commun lui aussi aux deux approches, apparaîtra comme un pré-traitement dans la seconde approche, alors qu'il sera un résultat de la première.

Ces deux approches font l'objet des deux parties qui suivent.

Deuxième partie

Analyse discriminante matricielle pour l'étude et la classification des signaux EEG

Préambule

Les signaux EEG peuvent être naturellement représentés sous la forme d'une matrice composée du signal enregistré sur M électrodes et sur un temps T fixé. Comme présenté précédemment, une approche classique pour l'analyse et la classification en EEG consiste à vectoriser chaque signal de sorte que l'on manipule des vecteurs spatio-temporels. De cette façon, les algorithmes de classification ou d'extraction de caractéristiques classiques présentés dans le chapitre 3 peuvent être directement appliqués sur ces données. Cependant, la question est ici de savoir *comment ne pas perdre la structure spatio-temporelle ?* En effet, si nous n'introduisons pas cette structure dans l'estimation de la matrice de covariance et si nous utilisons les méthodes d'estimation présentées dans le chapitre 3, les sources de variabilité spatiales et temporelles seront mélangées. C'est pourquoi nous nous intéressons à présent aux approches permettant de conserver cette structure. Un des apports de cette partie est de montrer que l'approche dite matricielle de l'analyse discriminante équivaut à considérer des vecteurs spatio-temporels en conservant leur structure *via* une hypothèse sur la matrice de covariance. Pour ce faire, nous exploitons les idées largement développées dans le domaine de l'estimation de sources et l'extraction de caractéristiques discriminantes des signaux cérébraux.

La première approche consiste à estimer les sources à l'origine des signaux observés enregistrés sur une multitude d'électrodes de surface. Dans ce contexte, de nombreux auteurs se sont intéressés à l'estimation des sources spatio-temporelles. Pour ce faire, un modèle sur la matrice de covariance est introduit. Dans (Huizenga et al., 2002; de Munck et al., 2002; Bijma et al., 2003), les auteurs proposent d'exprimer la matrice de covariance sous la forme d'un produit de Kronecker d'une matrice de covariance temporelle et d'une matrice de covariance spatiale. L'estimation de la matrice de covariance permet alors une analyse des sources spatiales et temporelles de façon indépendante. Cette approche permet également de réduire le nombre de paramètres à estimer : prenons l'exemple d'un signal enregistré sur 64 électrodes et 200 pas de temps, la structure de Kronecker implique l'estimation de 2.10^4 paramètres de covariance contre près de 8.10^7 dans le cas classique.
De nombreuses contributions exploitant cette modélisation ont été apportées dans le cadre de l'analyse et du débruitage des signaux EEG/MEG. Nous pouvons citer en particulier (Bijma et al., 2005) dans lequel le modèle est étendu à la somme de plusieurs produits de Kronecker. Cette approche a été ensuite adaptée à l'analyse spatio-spectrale (Bijma and de Munck, 2008; Villaron and Torrésani, 2010; Villaron, 2012).
Dans le chapitre 4, nous nous plaçons dans ce même cadre où la matrice de covariance se factorise sous la forme d'un produit de Kronecker, et nous proposons de modéliser explicitement la matrice des données EEG à partir d'un modèle linéaire gaussien. Nous en déduisons une règle de classification des données vectorisées permettant de tenir compte de leur structure matricielle.

La seconde approche est quant à elle principalement développée dans le contexte des ICM. Basée sur la notion d'extraction d'informations pertinentes et de réduction de dimension, elle est utilisée pour la classification entre différentes classes d'états cérébraux. Dans le cadre matriciel développé ici, elle consiste à tenir compte de la structure spatio-temporelle des données pour l'extraction de ces informations (voir par exemple (Parra et al., 2008), les auteurs y font la revue des méthodes linéaires d'extraction de caractéristiques pour les dimensions spatiales et temporelles en EEG). Dans d'autres travaux (Hasson-Meir et al., 2011; Farquhar, 2009) des

méthodes d'extraction de caractéristiques spatio-temporelles et spatio-spectrales sont mises en œuvre et utilisées par la suite dans un contexte d'une classification *via* un modèle de régression logistique. De même, des méthodes de classification tenant compte de la structure matricielle des signaux EEG, ont été proposées dans (Dyrholm and Parra, 2005; Dyrholm et al., 2007). Les travaux présentés dans ces articles se basent sur une projection (bi)linéaire des signaux EEG sous forme matricielle. Les espaces de projections considérés permettent d'exprimer les signaux en termes de caractéristiques spatiales et temporelles distinctement. Les auteurs proposent dans ce cas d'estimer les matrices de projection à partir d'une méthode du maximum de vraisemblance basée sur une régression logistique.
Les travaux que nous proposons dans le chapitre 5 se basent sur le même principe : nous cherchons à étudier les signaux EEG en termes de composantes spatiales et temporelles. Pour ce faire, contrairement aux méthodes présentées dans ce préambule, nous formalisons le problème de Fisher en découplant les analyses des espaces ligne et colonne des données matricielles. Ainsi, les matrices de projections définies par cette méthodes permettent de définir des caractéristiques discriminantes pour un problème de classification.

Chapitre 4

Approche décisionnelle

Sommaire

Nous nous intéressons ici à l'adaptation de la règle de décision bayésienne dans la situation où les données se représentent sous forme de matrices. Nous cherchons à conserver l'information temps × espace présente dans les signaux EEG. Pour ce faire, nous allons nous intéresser à la façon dont l'ensemble des données, initialement sous forme cubique, peut être réorganisé afin de permettre une analyse découplée de la dimension ligne (temps) et de la dimension colonne (espace). Dans ce qui suit, nous parlerons d'analyse ligne et d'analyse colonne.
Dans la suite, le $i^{\text{ème}}$ essai EEG d'une classe c de taille n_c correspond à une matrice $X_i^c \in \mathbb{R}^{K\times J}$ où K dénote la dimension temporelle (dans notre cas, le nombre de coefficients multi-échelle considérés) et J la dimension spatiale (le nombre d'électrodes). Le vecteur $x_i^c \in \mathbb{R}^{KJ}$ s'obtient par vectorisation de $(X_i^c)'$ obtenue par concaténation de ses colonnes. On notera

$$x_i^c = \mathsf{vec}\left((X_i^c)'\right) \ .$$

La définition et les propriétés de l'opérateur vec sont données dans l'annexe A et X' est la transposée de la matrice X.

Notations. Soit $\mathbb{R}^{K\times J}$ l'espace vectoriel des matrices réelles à K lignes et J colonnes. Soit $X \in \mathbb{R}^{K\times J}$. Les coefficients de X sont notés $X(k,j)$ où $k = 1,\dots,K$ est l'indice des lignes et $j = 1,\dots,J$ est l'indice des colonnes. Ainsi, $X(k,.)$ est un vecteur de $\mathbb{R}^J$ correspondant à la $k^{\text{ème}}$ ligne de la matrice X et $X(.,j)$ est un vecteur de $\mathbb{R}^K$ correspondant à la $j^{\text{ème}}$ colonne de X.
On considère donc K lignes correspondant à des vecteurs de $\mathbb{R}^J$. Nous appellerons par la suite $\mathbb{R}^J$ *l'espace des lignes*. De même, nous appellerons $\mathbb{R}^K$ *l'espace des colonnes*.

Pour simplifier les notations, nous notons de façon identique les variables aléatoires et leurs réalisations.

Remarque 10 ($\mathbb{R}^d = \mathbb{R}^{d\times 1}$)**.** *Par convention, tout vecteur $x \in \mathbb{R}^d$ (qu'il soit associé à une ligne ou une colonne) est représenté en colonne. Dans ce qui suit, par abus de langage, nous appelerons "ligne" le vecteur associé à la ligne.*

4.1 Modèle matriciel gaussien

4.1.1 Analyse ligne-colonne et hypothèse de séparabilité

Soit $X_1, \ldots, X_n$ des réalisations d'une matrice aléatoire $X \in \mathbb{R}^{K\times J}$. L'analyse sur l'espace des lignes $\mathbb{R}^J$ ou sur l'espace des colonnes $\mathbb{R}^K$ nous conduit naturellement à organiser ces données de différentes façons (voir la figure 4.1) :

- **Dans $\mathbb{R}^{\mathbf{K}}$, l'espace des colonnes (A)**, les observations matricielles X_i, $i = 1, \ldots, n$, sont concaténées selon les K lignes de sorte que l'on obtient un tableau rectangulaire avec K lignes et nJ colonnes. La matrice de covariance des lignes est estimée à partir des nJ colonnes.

- **Dans $\mathbb{R}^{\mathbf{J}}$, l'espace des lignes (B)**, les observations X_i, $i = 1, \ldots, n$ sont concaténées selon les J colonnes de sorte que l'on obtient un tableau rectangulaire avec nK lignes et J colonnes et la matrice de covariance des J colonnes est estimée sur les nK lignes .

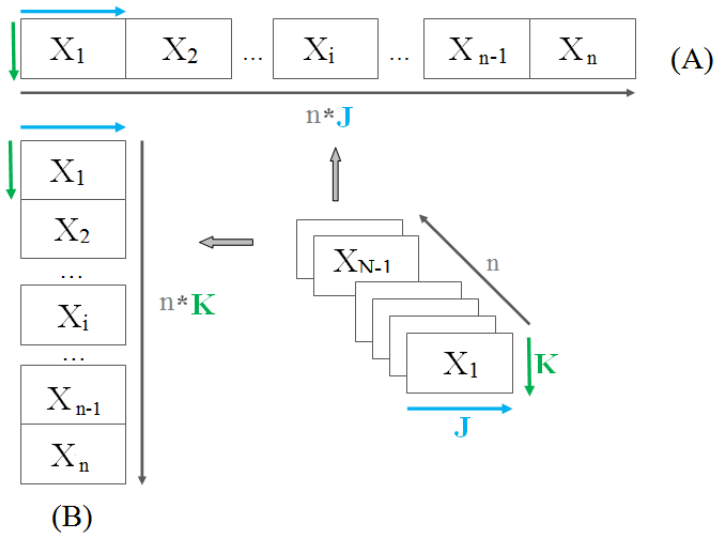


FIGURE 4.1 – Illustration des espaces des colonnes (A) et des lignes (B).

Dans ce travail, nous supposons que la matrice de covariance de $x = \mathsf{vec}(X')$, notée $\Sigma \in \mathbb{R}^{KJ\times KJ}$, peut se décomposer sous forme d'un produit de Kronecker de la covariance des lignes et de la covariance des colonnes. Pour ce faire, nous introduisons l'hypothèse suivante.

Hypothèse de séparabilité : La matrice de covariance Σ s'écrit :

$$\Sigma = \Sigma_L \otimes \Sigma_R \,, \tag{4.1}$$

où $\Sigma_L \in \mathbb{R}^{K\times K}$ est la matrice de covariance des lignes et $\Sigma_R \in \mathbb{R}^{J\times J}$ est la matrice de covariance des colonnes.

La définition du produit de Kronecker et ses principales propriétés sont données dans l'annexe A.

En particulier, la matrice Σ est inversible si et seulement si les matrices Σ_L et Σ_R sont inversibles. Dans ce cas son inverse se décompose lui aussi sous forme d'un produit de Kronecker :

$$\Sigma^{-1} = \Sigma_L^{-1} \otimes \Sigma_R^{-1} . \tag{4.2}$$

Dans la suite, nous nous plaçons dans ce cas.
Notons également que le déterminant de Σ peut s'écrire en fonction du déterminant de Σ_L et de Σ_R :

$$\det(\Sigma_L \otimes \Sigma_R) = \det(\Sigma_L)^J \det(\Sigma_R)^K . \tag{4.3}$$

Remarque 11 (Sur le problème d'identifiabilité)**.** *La décomposition* (4.1) *n'est pas l'unique décomposition de* Σ *comme produit de Kronecker. En effet, pour tout scalaire* $\kappa \neq 0$*, notons que* $\Sigma_L \otimes \Sigma_R = \kappa\Sigma_L . \otimes \kappa^{-1}\Sigma_R$*. Nous abordons cette question dans la section 4.2.2.*

La structure de covariance introduite en (4.1) se base sur le principe que les variabilités des facteurs (ici espace et temps) peuvent être caractérisées indépendamment l'une de l'autre. Cette hypothèse a été utilisée dans divers contextes, tels que l'étude de mesures répétées (Roy and Khattreeb, 2005), de données longitudinales (Chaganty and Naik, 2002) ou bien de données spatio-temporelles (Fuentes, 2004). Dans ce dernier cas, on suppose que les dimensions spatiales et temporelles sont décorrélées. Dans le cas des signaux EEG, cette structure se justifie puisque les corrélations spatiales et temporelles sont de natures très différentes et il fait sens de les distinguer dans l'analyse.

Notations et définitions.

Rappelons que par convention, lignes et colonnes sont représentées par des vecteurs colonnes (voir Remarque 10).

Définition 4 (Métriques des espaces des lignes et des colonnes)**.** *Soient* $G \in \mathbb{R}^{J\times J}$ *et* $D \in \mathbb{R}^{K\times K}$ *deux matrices symétriques et définies positives.*

a) L'espace des lignes $\mathbb{R}^J$ *est dit muni de la métrique* G *si le produit scalaire sur cet espace est défini par*

$$\langle x, y\rangle_G = x'Gy ,$$

pour tout $x, y \in \mathbb{R}^J$*.*

b) De manière similaire, l'espace des colonnes $\mathbb{R}^K$ *est muni de la métrique* D *si le produit scalaire sur cet espace est défini par*

$$\langle x, y\rangle_D = x'Dy ,$$

pour tout $x, y \in \mathbb{R}^K$ *.*

c) L'espace des matrices $\mathbb{R}^{K\times J}$ *est muni du produit scalaire suivant*

$$\langle X, Y\rangle_{G,D} = \mathrm{Tr}(XGY'D) . \tag{4.4}$$

La norme associée à ce produit scalaire et appelée norme-trace est :

$$\|X\|^2_{G,D} = \mathrm{Tr}(XGX'D) , \tag{4.5}$$

où Tr *désigne l'opérateur trace.*

Sachant que $\mathrm{Tr}(A'BCD) = \mathsf{vec}(A)'(D \otimes B)\mathsf{vec}(C)$, où $\mathsf{vec}(A)$ désigne la vectorisation de la matrice A par concaténation de ses colonnes et $\otimes$ le produit de Kronecker, on obtient l'équivalence entre les produits scalaires définis sur $\mathbb{R}^{K\times J}$ l'espace vectoriel des matrices de dimension $K \times J$ est sur $\mathbb{R}^{KJ}$ l'espace des matrices vectorisées :

Propriété 3. *Le produit scalaire défini en (4.4) sur $\mathbb{R}^{K\times J}$ est équivalent au produit scalaire défini sur l'espace $\mathbb{R}^{KJ}$ muni de la métrique $D \otimes G$:*

$$\langle X, Y\rangle_{G,D} = \mathsf{vec}(X')'(D \otimes G)\mathsf{vec}(Y') = \langle \mathsf{vec}(X'), \mathsf{vec}(Y')\rangle_{D\otimes G}\,. \tag{4.6}$$

4.1.2 Modèle gaussien

L'hypothèse de séparabilité énoncée en (4.1) a largement été exploitée dans le cadre gaussien (Dutilleul, 1999; Lu and Zimmerman, 2005; Glanz and Carvalho, 2013). Dans ce cas, le modèle considéré correspond à une distribution matricielle gaussienne (Dawid, 1981; Gupta and Nagar, 2000) définie de la façon suivante :

Définition 5 (Distribution matricielle gaussienne)**.** *Soit $M \in \mathbb{R}^{K\times J}$ et soient $\Sigma_L \in \mathbb{R}^{K\times K}$ et $\Sigma_R \in \mathbb{R}^{J\times J}$ deux matrices symétriques définies positives. La matrice aléatoire $X \in \mathbb{R}^{K\times J}$ suit une distribution gaussienne matricielle de moyenne M et de matrices de covariance Σ_L et Σ_R notée*

$$X \sim \mathcal{MN}_{K,J}\left(M, \Sigma_L, \Sigma_R\right)\,, \tag{4.7}$$

si

$$x = \mathsf{vec}(X') \sim \mathcal{N}_{KJ}\left(\mathsf{vec}(M'), \Sigma_L \otimes \Sigma_R\right)\,. \tag{4.8}$$

A partir des propriétés du produit de Kronecker (4.2) et (4.3), l'équivalence des densités de probabilité des distributions vectorielles et matricielles est aisément démontrée :

Proposition 2. *Si $X \sim \mathcal{MN}_{K,J}\left(M, \Sigma_L, \Sigma_R\right)$, où Σ_L et Σ_R sont supposées inversibles, alors la densité de probabilité de X est définie par :*

$$f(X) = (2\pi)^{-\frac{1}{2}KJ}\,|\Sigma_L|^{-\frac{1}{2}J}|\Sigma_R|^{-\frac{1}{2}K}\exp\left[\mathrm{Tr}\left\{-\frac{1}{2}(X-M)\Sigma_R^{-1}(X-M)'\Sigma_L^{-1}\right\}\right]\,, \tag{4.9}$$

et elle est égale à la densité de probabilité du vecteur gaussien $x \sim \mathcal{N}(\mu, \Sigma_L \otimes \Sigma_R)$, avec $x = \mathsf{vec}(X') \in \mathbb{R}^{KJ}$ et $\mu = \mathsf{vec}(M') \in \mathbb{R}^{KJ}$.

Démonstration. La densité de probabilité du vecteur gaussien x est donnée par :

$$\begin{aligned}
f\left(x\right) &= (2\pi)^{-\frac{1}{2}KJ}\,|\Sigma|^{-\frac{1}{2}}\exp\left[-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\right] && \text{(4.10a)}\\
&= (2\pi)^{-\frac{1}{2}KJ}\,|\Sigma_L \otimes \Sigma_R|^{-\frac{1}{2}}\exp\left[-\frac{1}{2}(x-\mu)'(\Sigma_L \otimes \Sigma_R)^{-1}(x-\mu)\right] && \text{(4.10b)}\\
&= (2\pi)^{-\frac{1}{2}KJ}\,|\Sigma_L|^{-\frac{1}{2}J}|\Sigma_R|^{-\frac{1}{2}K}\times && \text{(4.10c)}\\
&\qquad \exp\left[-\frac{1}{2}\mathsf{vec}(X'-M')'\mathsf{vec}\left[(\Sigma_R)^{-1}(X'-M')(\Sigma_L)^{-1}\right]\right]\\
&= (2\pi)^{-\frac{1}{2}KJ}\,|\Sigma_L|^{-\frac{1}{2}J}|\Sigma_R|^{-\frac{1}{2}K}\times && \text{(4.10d)}\\
&\qquad \exp\left[\mathrm{Tr}\left\{-\frac{1}{2}(X-M)(\Sigma_R)^{-1}(X-M)'(\Sigma_L)^{-1}\right\}\right]\,.\\
&= f(X)\,. && \text{(4.10e)}
\end{aligned}$$

□

4.1.3 Décorrélation ligne-colonne

Définition 6 (Décomposition de Cholesky)**.** *Soit $A \in \mathbb{R}^{N\times N}$ une matrice symétrique définie positive, la décomposition de Cholesky consiste à trouver (par exemple) une matrice triangulaire supérieure $S \in \mathbb{R}^{N\times N}$ telle que*

$$A = S'S \ . \tag{4.11}$$

Nous introduisons ici une décomposition de Cholesky de Σ_L^{-1} et de Σ_R^{-1} :

$$\Sigma_L^{-1} = L'L \, . \tag{4.12a}$$

$$\Sigma_R^{-1} = RR' \, , \tag{4.12b}$$

où L (resp. R) est triangulaire supérieure (resp. inférieure).

En multipliant les colonnes de la matrice gaussienne X par L', nous obtenons la densité de probabilité suivante :

$$f(L'X) = (2\pi)^{-\frac{1}{2}KJ} \, |I_K|^{-\frac{1}{2}J} |\Sigma_R|^{-\frac{1}{2}K} \exp\left[\mathrm{Tr}\left\{-\frac{1}{2}(X-M)\Sigma_R^{-1}(X-M)'\right\} .\right] \tag{4.13}$$

On en déduit que

Proposition 3. *La matrice $L'X$ est une matrice gaussienne de loi :*

$$L'X \sim \mathcal{MN}_{K,J}(L'M, I_K, \Sigma_R) \, . \tag{4.14}$$

Démonstration. En effet, $\mathsf{vec}((L'X)') = \mathsf{vec}(X'L) = (L' \otimes I_J)\mathsf{vec}(X')$ est un vecteur gaussien de moyenne $(L' \otimes I_J)\mathsf{vec}(M') = \mathsf{vec}(L'M)$, et de matrice de covariance $(L' \otimes I_K)\Sigma_L \otimes \Sigma_R(L \otimes I_K) = L'\Sigma_L L \otimes I_J\Sigma_R I_J = I_K \otimes \Sigma_R$.
Ainsi, on a montré que $\mathsf{vec}((L'X)') \sim \mathcal{N}(\mathsf{vec}((L'M)'), I_K \otimes \Sigma_R)$, et à partir de la définition 5, on en déduit bien l'expression (4.14). □

De même, en multipliant les lignes de la matrice gaussienne X par R, nous obtenons la densité de probabilité suivante :

$$f(XR) = (2\pi)^{-\frac{1}{2}KJ} \, |\Sigma_L|^{-\frac{1}{2}J} |I_J|^{-\frac{1}{2}K} \exp\left[\mathrm{Tr}\left\{-\frac{1}{2}(X-M)(X-M)'\Sigma_L^{-1}\right\}\right] \tag{4.15}$$

de sorte que

Proposition 4. *La matrice XR est une matrice gaussienne de loi :*

$$XR \sim \mathcal{MN}_{K,J}(MR, \Sigma_L, I_J) \, . \tag{4.16}$$

L'application de ces transformations permet ainsi de découpler les analyses sur l'espace des lignes $\mathbb{R}^J$ et sur l'espace des colonnes $\mathbb{R}^K$ afin d'estimer les matrices de covariance $\Sigma_L \in \mathbb{R}^{K\times K}$ et $\Sigma_R \in \mathbb{R}^{J\times J}$.

Soit $X_1, \ldots, X_n$, des matrices aléatoires i.i.d de loi $\mathcal{MN}(M, \Sigma_L, \Sigma_R)$ et telles que $X_i \in \mathbb{R}^{K\times J}$ pour tout $i = 1, \ldots, n$.

Analyse sur $\mathbb{R}^{\mathbf{J}}$ l'espace des lignes. Les K lignes de la matrice $L'(X_i - M) \in \mathbb{R}^{K\times J}$ sont indépendantes et identiquement distribuées (i.e. $\Sigma_L = I_K$) et $\Sigma_R \in \mathbb{R}^{J\times J}$ la matrice de covariance des colonnes peut alors être estimée de manière classique à partir de nK observations issues de (4.14) (voir la figure 4.1.1 (B)) ;

Analyse sur $\mathbb{R}^{\mathbf{K}}$ l'espace des colonnes. Les J colonnes de la matrice $(X_i - M)R \in \mathbb{R}^{K\times J}$ sont indépendantes et identiquement distribuées (i.e. $\Sigma_R = I_J$) et $\Sigma_L \in \mathbb{R}^{K\times K}$ la matrice de covariance des lignes peut être estimée de manière classique à partir de nJ observations issues de (4.16) (voir la figure 4.1.1 (A)) ;

4.2 Estimation des matrices de covariance

4.2.1 Estimation par maximum de vraisemblance

Soit l'échantillon de n matrices aléatoires $(X_1, \ldots, X_n)$ de la loi $\mathcal{MN}_{K,J}(M, \Sigma_L, \Sigma_R)$. La vraisemblance de l'échantillon est donnée par :

$$(2\pi)^{-\frac{1}{2}nKJ} |\Sigma_L|^{-\frac{1}{2}nJ} |\Sigma_R|^{-\frac{1}{2}nK} \exp\left[\mathrm{Tr}\left\{-\frac{1}{2}\sum_{i=1}^{n}(X_i - M)\Sigma_R^{-1}(X_i - M)'\Sigma_L^{-1}\right\}\right]. \tag{4.17}$$

L'estimateur du maximum de vraisemblance de la moyenne matricielle M est donné par :

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i\,. \tag{4.18}$$

Sachant Σ_R, l'estimateur du maximum de vraisemblance Σ_L est donné par

$$\hat{\Sigma}_L = \frac{1}{nJ}\sum_{i=1}^{n}(X_i - \bar{X})(\Sigma_R)^{-1}(X_i - \bar{X})'\,, \tag{4.19}$$

et $\hat{\Sigma} = \hat{\Sigma}_L \otimes \Sigma_R$ est l'estimateur du maximum de vraisemblance de Σ.

De même, sachant Σ_L, l'estimateur du maximum de vraisemblance de Σ_R est donné par

$$\hat{\Sigma}_R = \frac{1}{nK}\sum_{i=1}^{n}(X_i - \bar{X})'(\Sigma_L)^{-1}(X_i - \bar{X})\,, \tag{4.20}$$

et $\hat{\Sigma} = \Sigma_L \otimes \hat{\Sigma}_R$ est l'estimateur du maximum de vraisemblance de Σ.

Remarque 12. *L'estimateur de $\Sigma_L \in \mathbb{R}^{K\times K}$ correspond à l'estimateur empirique de la matrice de covariance de K vecteurs (correspondant ici aux lignes) sur $n \times J$ observations indépendantes après avoir appliqué la transformation (4.16) sur les colonnes :*

$$\frac{1}{nJ}\sum_{i=1}^{n}(X_iR - \bar{X}R)(X_iR - \bar{X}R)' = \frac{1}{nJ}\sum_{i=1}^{n}(X_i - \bar{X})RR'(X_iR - \bar{X}R)' \tag{4.21a}$$

$$= \frac{1}{nJ}\sum_{i=1}^{n}(X_i - \bar{X})(\Sigma_R)^{-1}(X_i - \bar{X})' \tag{4.21b}$$

$$= \hat{\Sigma}_L \tag{4.21c}$$

De même, l'estimateur de $\Sigma_R \in \mathbb{R}^{J\times J}$ correspond à l'estimateur empirique de la matrice de covariance de J vecteurs (correspondant aux colonnes) sur $n \times K$ observations indépendantes après avoir appliqué la transformation (4.14) sur les lignes.

Remarque 13. *La moyenne étant inconnue et estimée par $\bar{X}$ dans (4.19) (respectivement (4.20)), sachant Σ_R (respectivement Σ_L), l'estimateur du maximum de vraisemblance $\hat{\Sigma}_L$ (respectivement $\hat{\Sigma}_R$) est biaisé. Les estimateurs sans biais sont définis respectivement par :*

$$\frac{n}{n-1}\hat{\Sigma}_L = \frac{1}{nJ-J}\sum_{i=1}^{n}(X_i-\bar{X})(\Sigma_R)^{-1}(X_i-\bar{X})'\,, \tag{4.22}$$

et

$$\frac{n}{n-1}\hat{\Sigma}_R = \frac{1}{nK-K}\sum_{i=1}^{n}(X_i-\bar{X})'(\Sigma_L)^{-1}(X_i-\bar{X})\,. \tag{4.23}$$

Lorsque aucune des deux matrices de covariance Σ_L et Σ_R n'est connue, nous utilisons les estimateurs définis par :

$$S_L = \frac{1}{nJ}\sum_{i=1}^{n}(X_i-\bar{X})(S_R)^{-1}(X_i-\bar{X})'\,, \tag{4.24}$$

et

$$S_R = \frac{1}{nK}\sum_{i=1}^{n}(X_i-\bar{X})'(S_L)^{-1}(X_i-\bar{X})\,. \tag{4.25}$$

Dans le cas où les matrices Σ_R et Σ_L sont toutes deux inconnues, nous choisirons comme estimateur de $\Sigma \in \mathbb{R}^{KJ\times KJ}$ le produit de Kronecker des deux estimateurs (4.24) et (4.25), noté

$$S = S_L \otimes S_R\,. \tag{4.26}$$

En pratique le calcul de S_L et de S_R se fait de manière itérative, en mettant en œuvre un algorithme appelé *flip-flop* (Dutilleul, 1999; Lu and Zimmerman, 2005; Srivastava et al., 2008).

Remarque 14 (Sur le nombre de paramètres à estimer)**.** *L'hypothèse de séparabilité permet de réduire de manière significative le nombre de paramètres à estimer pour la matrice de covariance. En effet, lorsque l'on considère la covariance classique $\Sigma \in \mathbb{R}^{KJ\times KJ}$, $\frac{KJ(KJ+1)}{2}$ paramètres doivent être estimés. En revanche, lorsque l'on considère le produit de Kronecker de $\Sigma_L \in \mathbb{R}^{K\times K}$ et de $\Sigma_R \in \mathbb{R}^{J\times J}$, le nombre de paramètres à estimer est alors seulement de $\frac{K(K+1)}{2} + \frac{J(J+1)}{2}$.*

4.2.2 L'algorithme flip-flop

Nous rappelons que l'estimation des matrices Σ_L et Σ_R telle que définie en (4.24) et (4.25) ne fournit pas de solution unique (voir remarque 11). Afin de traiter ce problème d'identifiabilité nous imposons la contrainte $\|\Sigma_R\|_F = 1$. L'estimation de Σ_L et de Σ_R est obtenue à partir des estimateurs définis en (4.24) et (4.25).

Algorithme 1 Algorithme flip-flop pour l'estimation de Σ_L et Σ_R

Entrées: $\mathbf{X} \in \mathbb{R}^{K \times J \times n}$

$\bar{X} \leftarrow \frac{1}{n} \sum_{i=1}^{n} X_i$

$S_L^{(1)} \leftarrow I_K$

$l \leftarrow 1$

tant que $\|S_L^{(l)} - S_L^{(l-1)}\|_F / \|S_L^{(l-1)}\|_F > s$ **et** $\|S_R^{(l)} - S_R^{(l-1)}\|_F / \|S_R^{(l-1)}\|_F > s$ **:**

$l \leftarrow l + 1$

$S_R^{(l)} \leftarrow \frac{1}{nK} \sum_{i=1}^{nK} (X_i - \bar{X})' (S_L^{(l-1)})^{-1} (X_i - \bar{X})$

$S_R^{(l)} \leftarrow S_R^{(l)} / \|S_R^{(l)}\|_F$

$S_L^{(l)} \leftarrow \frac{1}{nJ} \sum_{i=1}^{nJ} (X_i - \bar{X}) (S_R^{(l)})^{-1} (X_i - \bar{X})'$

fin tant que

Sorties: $\bar{X}$, S_L et S_R

L'algorithme flip-flop : L'initialisation se fait en fixant $S_L = I_K$. A chaque itération les deux matrices sont réestimées à partir des formulations (4.24) et (4.25). La contrainte $\|S_R\|_F = 1$ est également imposée à chaque itération. Le processus d'estimation est ainsi réitéré jusqu'à ce que la différence relative (mesurée via la norme de Frobenius) entre les estimations de deux itérations consécutives soit inférieure à un seuil s. L'algorithme d'estimation est détaillé ci-dessus.

Remarque 15.

1. *Il est possible de démontrer (Srivastava et al., 2008) que si le nombre d'observations n est assez grand, plus précisément tel que $n \geq \max(K, J)$, l'imposition de la contrainte $\|S_R\|_F = 1$ à chaque itération permet d'assurer la convergence de l'algorithme vers la solution unique du maximum de vraisemblance sous cette contrainte.*

2. *Signalons que la contrainte utilisée ici diffère de celle de (Srivastava et al., 2008) (c'est à dire, $S_R(K, K) = 1$), mais ceci ne modifie pas le résultat. Notons aussi qu'il est possible de faire porter la contrainte sur S_L plutôt que S_R.*

3. *La condition $n \geq \max(K, J)$ est à comparer à la condition $n \geq KJ$ nécessaire pour l'estimation de Σ par maximum de vraisemblance en l'absence d'hypothèse de séparabilité. Ceci montre que cette hypothèse peut permettre de réduire drastiquement la taille de l'échantillon d'apprentissage dans un contexte de classification.*

4.2.3 Tests

Plusieurs tests du type rapport de vraisemblance ont été introduits permettant de tester la décomposition de la matrice de covariance en produit de Kronecker sous la forme générale ou avec des matrices lignes ou colonnes particulières, à partir d'un échantillon gaussien multivariée (Lu and Zimmerman, 2005; Mitchell et al., 2006; Srivastava et al., 2008).

Nous choisissons de présenter les tests donnés dans (Srivastava et al., 2008) permettant de prendre en compte la contrainte d'identifiabilité assurant l'unicité des estimateurs du maximum de vraisemblance (voir remarque ci-dessus). Soient donc $x_1, x_2, \ldots, x_n$, un échantillon de vecteurs gaussiens de loi $\mathcal{N}_{KJ}(\mu, \Sigma)$ avec la matrice $\Sigma \in \mathbb{R}^{KJ \times KJ}$ symétrique et définie positive. Rappelons que $x_i = \mathsf{vec}(X_i')$.

Tester la séparabilité de la matrice de covariance

Soient les hypothèses suivantes à tester

$$\mathsf{H}_0 : \Sigma = \Sigma_L \otimes \Sigma_R^\star \quad \text{contre} \quad \mathsf{H}_1 : \Sigma \neq \Sigma_L \otimes \Sigma_R^\star \,,$$

où $\Sigma_R^\star$ vérifie la condition d'identifiabilité.

L'expression simplifiée du rapport de vraisemblance de H_0 contre H_1 s'écrit

$$LR = \frac{|V|^{\frac{n}{2}}}{|S_L|^{\frac{nJ}{2}} |S_R^\star|^{\frac{nK}{2}}} \,, \tag{4.27}$$

où V est l'estimateur du maximum de vraisemblance de Σ sous H_1 (Σ n'a pas de structure particulière), donné par

$$V = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})' \,. \tag{4.28}$$

et S_L et $S_R^\star$ sont les estimateurs du maximum de vraisemblance de Σ_L et $\Sigma_R^\star$ sous l'hypothèse H_0.

Proposition 5. *Soit T la statistique de test définie par :*

$$T = -2 \log LR = nJ \log |S_L| + nK \log |S_R^\star| - n \log |V| \,. \tag{4.29}$$

Alors sous H_0 , T suit asymptotiquement une loi du χ^2 à ν degrés de liberté, avec

$$\nu = \frac{KJ(KJ+1)}{2} - \frac{K(K+1)}{2} - \frac{J(J+1)}{2} + 1 \,. \tag{4.30}$$

Remarque 16. *Le nombre de degrés de liberté ν est égal au nombre de paramètres à estimer sous H_1 moins le nombre de paramètres à estimer sous H_0. Sous H_0 en imposant la contrainte d'identifiabilité, il y a un paramètre de moins à estimer.*

Tester la corrélation des lignes ou des colonnes

Il peut également être intéressant de tester s'il y a corrélation des lignes ou des colonnes. Nous présentons ici un cas particulier d'un test introduit dans (Srivastava et al., 2008).

Soient les hypothèses suivantes à tester :

$$\mathsf{H}_0 : \Sigma = \Sigma_L \otimes \Sigma_R^\star \,,\ \Sigma_L = \sigma^2 I_K \quad \text{contre} \quad \mathsf{H}_1 : \Sigma = \Sigma_L \otimes \Sigma_R^\star \,, \Sigma_L \neq \sigma^2 I_K \,,$$

où $\Sigma_R^\star$ vérifie la condition d'identifiabilité.

L'expression simplifiée du rapport de vraisemblance de H_0 contre H_1 est donnée par :

$$LRc = \frac{|S_L|^{\frac{nJ}{2}} |S_R^\star|^{\frac{nK}{2}}}{|\sigma^2 I_K|^{\frac{nJ}{2}} I |V_R^\star|^{\frac{nK}{2}}} \,, \tag{4.31}$$

où $V_R^\star$ et $\hat{\sigma}^2$ sont les estimateurs du maximum de vraisemblance de $\Sigma_R^\star$ et σ^2 sous H_0 et S_L et $S_R^\star$ les estimateurs du maximum de vraisemblance de Σ_L et $\Sigma_R^\star$ sous l'hypothèse H_1.

Proposition 6. *Soit Θ la statistique de test définie par :*

$$\Theta = -2 \log LRc = nJK\sigma^2 + nK(\log |V_R^\star| - \log |S_R^\star|) - nJ \log |S_L| \,. \tag{4.32}$$

Alors Θ suit asymptotiquement une loi du χ^2 à ν degrés de liberté, avec

$$\nu = \frac{K(K+1)}{2} - 1 \,. \tag{4.33}$$

De manière similaire on peut tester la corrélation des colonnes en imposant la contrainte d'identifiabilité sur la matrice Σ_L.

4.2.4 Extensions du modèle de séparabilité

Nous avons vu que l'hypothèse de séparabilité permet une importante diminution du nombre de paramètres à estimer . Cependant, elle offre aussi la possibilité d'estimer la matrice de covariance globale en introduisant des régularisations et/ou pénalisations sur les estimateurs dans les dimensions ligne et colonne distinctement. Dans ce cas, les méthodes décrites dans le chapitre 3 peuvent s'appliquer à l'une ou l'autre dimension. Le modèle de Kronecker pourra ainsi être étendu à des situations très variées dépendant des informations *a priori* (parcimonie, corrélation, etc.), des problématiques étudiées ou encore du contexte (faibles échantillons d'apprentissage par exemple).

Un compromis entre V et $S_L \otimes S_R$?

Rappelons que V est l'estimateur sans biais usuel de Σ (sans structure particulière) défini en (4.28). Lorsque l'hypothèse de séparabilité n'est pas pertinente, une méthode de régularisation peut être utilisée. Inspirée de l'approche régularisée de l'analyse discriminante proposée dans (Friedman, 1988), une méthode d'estimation de la matrice de covariance permettant de relâcher l'hypothèse de séparabilité a été proposée (Mahanta et al., 2013). Les auteurs introduisent un compromis entre V et $S_L \otimes S_R$, pondéré par un paramètre de régularisation $\gamma \in [0; 1]$:

$$S(\gamma) = (1-\gamma)(S_L \otimes S_R) + \gamma V \,, \tag{4.34}$$

où γ est sélectionné par validation croisée. Lorsque $\gamma = 1$, la matrice de covariance est estimée sans structure particulière. Lorsque $\gamma = 0$, la matrice de covariance est estimée sous l'hypothèse de séparabilité. Enfin, lorsque $\gamma \in]0; 1[$, l'estimation de la matrice de covariance est un compromis entre les deux approches.

Approche pénalisée

Nous nous intéressons à présent à l'exploitation d'une information *a priori* dans l'estimation de la matrice de covariance Σ_L ou bien de Σ_R.
A partir de l'analyse discriminante pénalisée (PDA) proposée par Hastie et al. (Hastie et al., 1995), nous introduisons les matrices de covariance pénalisées :

$$S_L^P = S_L + \alpha_L \Omega_L \quad \text{et} \quad S_R^P = S_R + \alpha_R \Omega_R \,. \tag{4.35}$$

où $\Omega_L \in \mathbb{R}^{K \times K}$ et $\Omega_R \in \mathbb{R}^{J \times J}$ sont deux matrices symétriques définies positives qui encodent les informations *a priori* des dimensions ligne et colonne respectivement, et $\alpha_L \in \mathbb{R}$ et $\alpha_R \in \mathbb{R}$

sont les deux hyperparamètres permettant de calibrer l'impact des matrices Ω_L et Ω_R.

Dans le cas des signaux EEG, nous pouvons par exemple introduire un terme supplémentaire à la matrice de covariance des électrodes permettant de tenir compte de la notion de proximité spatiale de l'information enregistrée sur des électrodes adjacentes. Dans la section 6.1.1 nous exploitons cette approche et proposons une version pénalisée de l'analyse discriminante matricielle permettant d'extraire des combinaisons linéaires d'électrodes discriminantes tenant compte de la notion de voisinage entre les électrodes.

4.3 Règle de Bayes dans le cas matriciel

Plaçons nous à présent dans un contexte de classification à $\mathcal{C}$ classes comme présenté dans la section 3.2.1 abordé maintenant dans le cas matriciel. Considérons les n réalisations de la matrice X, notées $X_i \in \mathbb{R}^{K\times J}$ pour tout $i = 1, \ldots, n$. La probabilité d'une classe G^c est notée p^c et l'on rappelle que l'on a $\sum_{c=1}^{\mathcal{C}} p^c = 1$.

L'ensemble du formalisme donné dans la section 3.2.1 s'étend au cas de données matricielles de la façon suivante :

La distribution de probabilité associée à l'observation X_i dans la classe c est donnée par la densité de probabilité conditionnelle notée f^c :

$$X_{|G^c} \sim f^c\,. \tag{4.36}$$

La densité de probabilité associée à l'observation X_i, notée f est définie par la loi de mélange

$$f(X) = \sum_{c=1}^{\mathcal{C}} p^c f^c(X)\,. \tag{4.37}$$

D'après la formule de Bayes, la probabilité *a posteriori* d'appartenance de l'observation X_i à la classe G^c est donnée par :

$$\mathbb{P}(Classe = c|X_i) = \frac{p^c f^c(X_i)}{f(X_i)}\,. \tag{4.38}$$

Comme dans le cas vectoriel, la règle de Bayes est alors équivalente à maximiser le terme au numérateur de l'équation (4.38). Soit $\Delta^c(X) = -2\log p^c f^c(X)$ la fonction discriminante de X dans la classe c. La règle bayésienne dans le cas matriciel consiste à minimiser la fonction discriminante :

$$\hat{c} = \text{argmin}_c\, \Delta^c(X)\,. \tag{4.39}$$

Dans le cas où les matrices de covariance sont différentes dans les $\mathcal{C}$ classes, la fonction discriminante s'écrit :

$$\Delta^c(X) = \text{Tr}\left[(X - M^c)(\Sigma_R^c)^{-1}(X - M^c)'(\Sigma_L^c)^{-1}\right] + J\log|\Sigma_L^c| + K\log|\Sigma_R^c| - 2\log(p^c)\,. \tag{4.40}$$

Lorsque l'on suppose l'égalité des matrices de covariance, elle se simplifie sous la forme suivante :

$$\Delta^c(X) = \text{Tr}\left[(X - M^c)(\Sigma_R)^{-1}(X - M^c)'(\Sigma_L)^{-1}\right] - 2\log(p^c)\,. \tag{4.41}$$

A partir des propriétés du produit de Kronecker, on montre que sous l'hypothèse de séparabilité (4.1), la fonction discriminante $\Delta^c(X)$ est égale à la fonction discriminante $\delta^c(x)$ définie dans le cas vectoriel (3.7) :

$$\delta^c(x) = (x - \mu^c)' \left(\Sigma_L^{-1} \otimes \Sigma_R^{-1}\right) (x - \mu^c) + \log|\Sigma^c| - 2\log(p^c) \tag{4.42}$$

Proposition 7. *Soit $X \in \mathbb{R}^{K\times J}$. Sous l'hypothèse de séparabilité,*

$$\Delta^c(X) = \delta^c(x) \quad \textit{où} \quad x = \mathsf{vec}(X') \in \mathbb{R}^{KJ} \tag{4.43}$$

La règle matricielle est donc optimale au sens de la minimisation de l'erreur de classification.

Démonstration. En notant que $\mathrm{Tr}(DX'EXB) = (\mathsf{vec}(X))'D'B' \otimes E\mathsf{vec}(X)$, on a :

$$\begin{aligned}
\Delta^c(X) &= \mathrm{Tr}\left[(X - M^c)(\Sigma_R)^{-1}(X - M^c)'(\Sigma_L)^{-1}\right] + J\log|\Sigma_L^c| + K\log|\Sigma_R^c| - 2\log(p^c) && (4.44)\\
&= (\mathsf{vec}(X - M^c)')'\Sigma_L^{-1} \otimes \Sigma_R^{-1}(\mathsf{vec}(X - M^c))' + \log|\Sigma_L^c \otimes \Sigma_R^c| - 2\log(p^c) && (4.45)\\
&= \delta^c(x)\,. && (4.46)
\end{aligned}$$

□

Discussion : apport du formalisme matriciel

A ce niveau, le formalisme matriciel n'apporte aucun bénéfice par rapport à l'approche vectorielle sous l'hypothèse de séparabilité. L'apport interviendra lorsque l'approche matricielle sera utilisée pour sélectionner des caractéristiques qui pourront être analysées, et utilisées pour la classification.

En termes d'analyse, nous verrons que les caractéristiques peuvent être représentées graphiquement, et interprétées en termes de composantes du signal (décours temporel et représentation spatiale).

En termes de classification, même si la règle énoncée ci-dessus est optimale dans le cadre du modèle de mélange gaussien sous hypothèse de séparabilité, nous verrons que cette régle de classification appliquée à un sous-ensemble des caractéristiques peut s'avérer supérieure dans le cas de données réelles, auxquelles le modèle n'est sans doute qu'imparfaitement adaptée.

La construction des caractéristiques est traitée dans le chapitre suivant.

Chapitre 5

Approche descriptive pour l'extraction de caractéristiques

Sommaire

5.1 État de l'art

De très nombreux travaux se sont intéressés à l'application du critère de Fisher dans le contexte de données matricielles, et plus généralement tensorielles (notamment pour des applications en traitement d'images et en reconnaissance faciale, voir par exemple (Lu et al., 2009)). Toutes les méthodes développées dans ce domaine consistent en l'extraction de caractéristiques discriminantes par optimisation d'un critère de Fisher et se distinguent soit par la construction des matrices de covariance intra-classe et inter-classe (Luo et al., 2009) soit par le critère utilisé pour la recherche des sous-espaces. Les premières approches ont été développées afin de définir des transformations linéaires des données matricielles de façon uni-directionnelle, sur l'espace des lignes ou des colonnes uniquement (Li and Yuan, 2005; Xiong et al., 2006).

La première extension au cas bi-directionnel a été proposée par (Ye et al., 2004). Les auteurs ont introduit un algorithme itératif permettant de définir des transformations linéaires sur les espaces des lignes et des colonnes qui maximisent la séparation des classes. A chaque itération, une LDA uni-directionnelle est mise en œuvre sur l'espace des lignes ou des colonnes séparément. Cependant, il a été démontré par la suite (Inoue and Urahama, 2006; Luo et al., 2009) que l'algorithme proposé par Ye *et al.* ne garantit aucune convergence. Afin de régler ce problème, (Luo et al., 2009) ont introduit une nouvelle fonction coût. Cependant la mise en œuvre de leur algorithme est plus coûteuse ce qui le rend plus difficilement utilisable. Dans le même temps, une LDA bi-directionnelle a été proposée (Noushath et al., 2006; Inoue and Urahama, 2006). Dans ces travaux, les auteurs définissent des matrices de covariance intra-classe et inter-classe ligne (resp. colonne), qui ne tiennent pas compte des corrélations entre colonnes (resp. lignes), et en déduisent les sous-espaces associés par l'optimisation de critères de Fisher correspondants. Cependant, en pratique cette méthode ne fournit pas de bons résultats, ce qui peut s'expliquer par la non prise en compte de ces corrélations (Zhao et al., 2012). Notons en effet que les algorithmes présentés

ci-dessus ne se basent pas sur une hypothèse de séparabilité de la variabilité des lignes et de celle des colonnes. L'absence de modèle explicite ne permet pas d'assoir l'algorithme d'estimation des covariances sur un fondement solide et d'en faire une analyse.

Une des premières modélisations de la structure de covariance par un produit de Kronecker pour l'optimisation du critère de Fisher a été introduite dans (Zhao et al., 2012). Les auteurs proposent une LDA matricielle séparable en introduisant une structure de Kronecker sur les matrices de covariance intra et inter-classe et en proposant des estimateurs tenant compte de cette séparabilité. Un tel formalisme permet ainsi d'inscrire l'approche matricielle de la LDA dans un cadre théorique clair à partir duquel certaines méthodes uni-directionnelles et bi-directionnelles ont pu être réinterprétées. En particulier, les auteurs montrent que la LDA uni-directionnelle correspond au cas où l'une des deux matrices de covariance intra-classe est égale à la matrice identité.
Une LDA matricielle basée sur l'hypothèse de séparabilité a été introduite dans le cas des signaux EEG (Mahanta et al., 2012; Mahanta et al., 2013), modélisés sous forme de matrices aléatoires gaussiennes (Gupta and Nagar, 2000). Dans ces travaux les auteurs introduisent l'hypothèse de séparabilité aussi bien sur les matrices de covariance intra que sur les matrices inter-classe mais n'utilisent pas les mêmes métriques pour la construction des deux types de matrices.

Dans ce travail, nous nous intéressons au développement de l'analyse discriminante descriptive matricielle sous l'hypothèse de séparabilité et exploitons son lien avec l'approche vectorielle au travers du produit de Kronecker. En particulier, dans le cas binaire, notre approche repose sur la décomposition en valeurs singulières (DVS) de la différence des moyennes en munissant les espaces des lignes et des colonnes de la métrique de Mahalanobis. Ceci nous permet d'extraire conjointement les combinaisons linéaires *les plus discriminantes* dans ces deux espaces et d'en déduire des représentations graphiques de ces composantes ainsi que des visualisations des données matricielles dans des sous-espaces de dimension réduite. Une application à l'étude de caractéristiques discriminantes de potentiels évoqués est ainsi proposée et l'évaluation des transformations obtenues est faite par l'utilisation de ces caractéristiques à une problématique de classification.
Dans (Spinnato et al., 2014), nous présentons l'extraction de caractéristiques discriminantes espace-temps-échelle des potentiels d'erreur et dans (Spinnato et al., 2015b) nous avons étudié l'apport de la méthode descriptive pour les données issues du protocole P300 Speller.

5.2 Analyse discriminante descriptive sous l'hypothèse de séparabilité

5.2.1 Discriminant de Fisher

Pour chaque classe c, $c = 1, \dots, \mathcal{C}$ on considère n_c données matricielles $X_i^c \in \mathbb{R}^{K \times J}$, $i = 1, \dots n_c$ avec $n = \sum_{c=1}^{\mathcal{C}} n_c$. Nous nous plaçons ici dans le cas d'homoscédasticité c'est à dire dans le cas où la variabilité est la même dans les $\mathcal{C}$ classes.

Critère de Fisher avec la métrique P

Soit P une matrice symétrique définie positive. Si nous munissons l'espace $\mathbb{R}^d$ de la métrique P, l'analyse discriminante présentée dans la section 3.4 consiste à rechercher des vecteurs P-orthonormés maximisant le rapport entre l'inertie intra-classe et l'inertie inter-classe. Soit $w \in \mathbb{R}^d$

un vecteur P-normé ($w'Pw = 1$), l'inertie du nuage projeté sur l'axe Δ_w engendré par w est égale à $w'PS_BPw$, où S_B est la matrice de covariance inter-classe (3.48). De même, l'inertie intra-classe projetée sur Δ_w est $w'PS_WPw$, où S_W est la matrice de covariance intra-classe (3.46), supposée inversible. On cherche donc le vecteur w P-normé qui maximise le critère suivant :

$$\max_w \frac{w'PS_BPw}{w'PS_WPw}\,. \tag{5.1}$$

Le vecteur propre P-normé de la matrice $(PS_WP)^{-1}(PS_BP)$ associé à la plus grande valeur propre est solution du problème.

Comme nous l'avons vu dans la section 3.4, le problème de Fisher conduit au problème d'optimisation suivant. Il s'agit de résoudre séquentiellement

$$w_l^\star = \arg\max_{w_l\in\mathbb{R}^d} \{w_l'PS_BPw_l \quad \text{sous contraintes} \quad w_l'PS_WPw_l \leq 1 \quad \text{et} \quad w_l'PS_WPw_i^\star = 0\ \forall i < l\}\ . \tag{5.2}$$

Critère de Fisher dans l'espace des lignes et l'espace des colonnes

Sous l'hypothèse de séparabilité (4.1), les matrices intra-classe des colonnes et des lignes notées $S_W^L \in \mathbb{R}^{K\times K}$ et $S_W^R \in \mathbb{R}^{J\times J}$ (supposées inversibles) sont données par :

$$S_W^L = \frac{1}{nJ}\sum_{c=1}^{\mathcal{C}}\sum_{i=1}^{n_c}(X_i^c - \bar{X}^c)(S_W^R)^{-1}(X_i^c - \bar{X}^c)'\,, \tag{5.3}$$

$$S_W^R = \frac{1}{nK}\sum_{c=1}^{\mathcal{C}}\sum_{i=1}^{n_c}(X_i^c - \bar{X}^c)'(S_W^L)^{-1}(X_i^c - \bar{X}^c)\,, \tag{5.4}$$

$$\text{avec} \quad \bar{X}^c = \frac{1}{n_c}\sum_{i=1}^{n_c} X_i^c\,. \tag{5.5}$$

Nous définissons les matrices inter-classe des colonnes et des lignes notées $S_B^L \in \mathbb{R}^{K\times K}$ et $S_B^R \in \mathbb{R}^{J\times J}$ par :

$$S_B^L = \frac{1}{n}\sum_{c=1}^{\mathcal{C}} n_c(\bar{X}^c - \bar{X})(S_W^R)^{-1}(\bar{X}^c - \bar{X})'\,, \tag{5.6}$$

$$S_B^R = \frac{1}{n}\sum_{c=1}^{\mathcal{C}} n_c(\bar{X}^c - \bar{X})'(S_W^L)^{-1}(\bar{X}^c - \bar{X})\,, \tag{5.7}$$

$$\text{avec} \quad \bar{X} = \frac{1}{n}\sum_{c=1}^{\mathcal{C}}\sum_{i=1}^{n_c} X_i^c\,. \tag{5.8}$$

Rechercher les combinaisons linéaires des colonnes les plus discriminantes dans l'espace des lignes $\mathbb{R}^J$ muni de la métrique $G = (S_W^R)^{-1}$ (respectivement des lignes dans l'espace des colonnes $\mathbb{R}^K$ muni de la métrique $D = (S_W^L)^{-1}$) revient à résoudre le problème de Fisher (5.2) avec la métrique G (respectivement la métrique D).

Étant donné que la matrice $S_B^L(S_W^L)^{-1}$ est D-symétrique ($(DS_B^L(S_W^L)^{-1})' = DS_B^L(S_W^L)^{-1}$), une base de vecteurs D-orthonormés existe et on a le résultat suivant :

Proposition 8. *Soit $\mathbb{R}^K$ l'espace des colonnes muni de la métrique D et $Q_L = \mathsf{rang}(S_B^L(S_W^L)^{-1})$. La solution au problème de Fisher (5.2) est donnée par les vecteurs propres D-orthonormés u_q, $q = 1, \ldots, Q_L$, de la matrice $S_B^L(S_W^L)^{-1}$ associés aux Q_L valeurs propres non nulles rangées dans l'ordre décroissant.*

Remarque 17. *Les vecteurs propres D-orthonormés u_q sont définis au signe près.*

De même dans l'espace des lignes, on a :

Proposition 9. *Soit $\mathbb{R}^J$ l'espace des lignes muni de la métrique G et $Q_R = \mathsf{rang}(S_B^R(S_W^R)^{-1})$. La solution au problème de Fisher (5.2) est donnée par les vecteurs propres D-orthonormés v_q, $q = 1, \ldots, Q_R$, de la matrice $S_B^R(S_W^R)^{-1}$ associés aux Q_R valeurs propres non nulles rangées dans l'ordre décroissant.*

Séparabilité du problème

D'après les propositions précédentes et les propriétés du produit de Kronecker (voir propriété (7) de l'annexe A), on obtient le résultat suivant :

Proposition 10. *Soient $Q_L = \mathsf{rang}(S_B^L(S_W^L)^{-1})$ et $Q_R = \mathsf{rang}(S_B^R(S_W^R)^{-1})$.*
Soient u_l, $l = 1, \ldots Q_L$, les vecteurs propres D-orthonormés de la matrice $S_B^L(S_W^L)^{-1}$ associées aux Q_L valeurs propres non-nulles λ_l et soient v_r, $r = 1, \ldots Q_R$, les vecteurs propres G-orthonormés de la matrice $S_B^R(S_W^R)^{-1}$ associé aux Q_R valeurs propres non-nulles ψ_r.
Alors les vecteurs $u_l \otimes v_r$ sont les vecteurs propres $D \otimes G$-orthonormés de la matrice

$$S_B^L(S_W^L)^{-1} \otimes S_B^R(S_W^R)^{-1}$$

associés aux $Q_L Q_R$ valeurs propres non nulles $\lambda_l \psi_r$.

En utilisant $(A \otimes B)(C \otimes D) = AC \otimes BD$ et en posant

$$S_W = S_W^L \otimes S_W^R \quad \text{et} \quad S_B = S_B^L \otimes S_B^R\,, \tag{5.9}$$

nous obtenons :

$$\begin{aligned} S_B(S_W)^{-1} &= (S_B^L \otimes S_B^R)(S_W^L \otimes S_W^R)^{-1} && (5.10)\\ &= (S_B^L \otimes S_B^R)\left((S_W^L)^{-1} \otimes (S_W^R)^{-1}\right) && (5.11)\\ &= S_B^L(S_W^L)^{-1} \otimes S_B^R(S_W^R)^{-1}. && (5.12) \end{aligned}$$

Ainsi dans R^{KJ}, en prenant $S_W = S_W^L \otimes S_W^R$ et $S_B = S_B^L \otimes S_B^R$ dans le critère de Fisher (5.2) avec la métrique $P = S_W^{-1}$, maximiser ce critère revient à maximiser séparément le critère de Fisher dans l'espace des lignes et dans l'espace des colonnes sous l'hypothèse de séparabilité.

Nous pouvons donc en conclure que dans le cas de données matricielles dont les variabilités ligne et colonne sont supposées séparables, l'extraction de "variables" discriminantes peut se faire par l'intermédiaire de la maximisation de deux critères de Fisher distincts : le critère de Fisher de la dimension ligne et le critère de Fisher de la dimension colonne.

Cas binaire

Soient n_1 et n_2 les effectifs des classes 1 et 2. Posons $n = n_1 + n_2$. Dans ce cas, les matrices de covariance intra-classe sont données par les expressions (5.3) et (5.4) avec $\mathcal{C} = 2$ et les matrices de covariance inter-classe définies en (5.6) et (5.7) ont une écriture simplifiée :

$$S_B^L = \frac{n_1 n_2}{(n)^2} \left(\bar{X}^1 - \bar{X}^2\right) (S_W^R)^{-1} \left(\bar{X}^1 - \bar{X}^2\right)' , \tag{5.13}$$

$$S_B^R = \frac{n_1 n_2}{(n)^2} \left(\bar{X}^1 - \bar{X}^2\right)' (S_W^L)^{-1} \left(\bar{X}^1 - \bar{X}^2\right) . \tag{5.14}$$

Remarque 18. *La manière dont nous définissons les matrices de covariance inter-classe dans* (5.13) *et* (5.14) *(valable également dans le cas où $\mathcal{C} > 2$) diffère de celle des approches proposées dans (Mahanta et al., 2012; Zhao et al., 2012). A notre connaissance, dans le cadre de l'analyse discriminante matricielle les matrices inter-classe n'ont jamais été définies de cette façon.*

Dans ce qui suit, nous nous allons étudier en détails le cas binaire et montrer que dans ce cas particulier, courant en pratique, des propriétés intéressantes apparaissent. Premièrement, nous allons montrer que déterminer les axes discriminants par les critères de Fisher dans l'espace des lignes et dans l'espace des colonnes est équivalent à décomposer en valeurs singulières la différence des moyennes des deux classes. Cette approche donne des résultats théoriques intéressants et permet de formaliser le problème d'extraction de caractéristiques discriminantes pour des données matricielles. Ceci permet également de montrer la pertinence des matrices de covariance intra-classe (5.13) et (5.14) utilisées. De plus, cette approche fait apparaître une propriété de dualité entre l'analyse sur l'espace des lignes et sur l'espace des colonnes qui réduit le coût de mise en œuvre de la méthode.

5.2.2 Décomposition en valeurs singulières de la différence des moyennes

Nous rappelons que l'espace des lignes $\mathbb{R}^J$ est muni de la métrique $G = (S_W^R)^{-1}$, l'espace des colonnes $\mathbb{R}^K$ de la métrique $D = (S_W^L)^{-1}$ et l'espace $\mathbb{R}^{KJ}$ de la métrique $D \otimes G = S_W^{-1}$.

Nous proposons une décomposition en valeurs singulières (DVS) de la différence des moyennes empiriques des deux classes $\bar{X}^1$ et $\bar{X}^2$ afin de rechercher les combinaisons linéaires des lignes et des colonnes les plus discriminantes. Soit Q le rang de la matrice $\bar{X}^1 - \bar{X}^2 \in \mathbb{R}^{K \times J}$ on cherche la matrice Z_r de rang $r \leq Q$ solution du problème d'optimisation suivant :

$$\min_{Z_r} \|(\bar{X}^1 - \bar{X}^2) - Z_r\|_{G,D}^2 , \tag{5.15}$$

où $\|\cdot\|_{G,D}$ est la norme-trace définie en (4.5).

Proposition 11. *Soit Q le rang de la matrice $\bar{X}^1 - \bar{X}^2 \in \mathbb{R}^{K \times J}$. La décomposition en valeurs singulières de la différence des deux moyennes s'écrit :*

$$\bar{X}^1 - \bar{X}^2 = U \Lambda^{\frac{1}{2}} V' = \sum_{q=1}^{Q} \sqrt{\lambda_q} u_q v_q' , \tag{5.16}$$

où,

- *$U \in \mathbb{R}^{K\times Q}$ a pour colonnes les vecteurs propres $u_q \in \mathbb{R}^K$ D-orthonormés de la matrice D-symétrique semi-définie positive $(\bar{X}^1 - \bar{X}^2)G(\bar{X}^1 - \bar{X}^2)'D$.*
Le vecteur propre u_q est associé à la $q^{ème}$ valeur propre non nulle λ_q rangée par ordre décroissant dans la matrice diagonale $\Lambda \in \mathbb{R}^{Q\times Q}$.
- *$V \in \mathbb{R}^{J\times Q}$ a pour colonnes les vecteurs propres v_q G-orthonormés de la matrice G-symétrique semi-définie positive $(\bar{X}^1 - \bar{X}^2)'D(\bar{X}^1 - \bar{X}^2)G$, associés aux mêmes valeurs propres non nulles λ_q. De plus, V peut s'écrire à partir de la matrice U :*

$$V = (\bar{X}^1 - \bar{X}^2)'DU\Lambda^{-\frac{1}{2}}\,. \tag{5.17}$$

Notons que (5.17) exprime une propriété de dualité entre l'espace ligne $\mathbb{R}^K$ et l'espace colonne $\mathbb{R}^J$. Cette propriété a déjà été démontrée par le formalisme de l'analyse discriminante matricielle dans le cas binaire, voir la proposition 1 de (Spinnato et al., 2014).

Considérons à présent l'espace $\mathbb{R}^{KJ}$. En rappelant que $u_q \otimes v_q = \mathsf{vec}(u_q v_q')$, nous obtenons une décomposition de la différence des moyennes vectorisées à partir de la DVS de la différence des moyennes sous forme matricielle (5.16) :

Proposition 12. *La différence des moyennes vectorisées des deux classes peut s'écrire sous la forme suivante :*

$$\mathsf{vec}((\bar{X}^1)' - (\bar{X}^2)') = \sum_{q=1}^{Q} \sqrt{\lambda_q}(u_q \otimes v_q)\,, \tag{5.18}$$

où $u_q \otimes v_q$, $q = 1, \dots Q$ sont des vecteurs S_W^{-1}-orthonormés de $\mathbb{R}^{KJ}$.

Démonstration. En utilisant $\mathsf{vec}(AXB) = (B' \otimes A)\mathsf{vec}(X)$, et en notant que $\mathsf{vec}((\bar{X}^1)' - (\bar{X}^2)') = \mathsf{vec}\left((\bar{X}^1 - \bar{X}^2)'\right)$, nous pouvons écrire, d'après (5.16) :

$$\mathsf{vec}((\bar{X}^1)' - (\bar{X}^2)') = \mathsf{vec}(V\Lambda^{\frac{1}{2}}U') = (U \otimes V)\mathsf{vec}(\Lambda^{\frac{1}{2}}) = \sum_{q=1}^{Q} \sqrt{\lambda_q}(u_q \otimes v_q)\,.$$

□

Proposition 13. *Le carré de la distance de Mahalanobis entre les moyennes s'écrit alors comme la somme des valeurs propres non-nulles λ_q :*

$$\|\mathsf{vec}((\bar{X}^1)' - (\bar{X}^2)')\|^2_{S_W^{-1}} = \sum_{q=1}^{Q} \lambda_q\,, \tag{5.19}$$

avec

$$\lambda_q = |\langle \mathsf{vec}((\bar{X}^1)' - (\bar{X}^2)'), u_q \otimes v_q \rangle_{S_W^{-1}}|^2\,. \tag{5.20}$$

Démonstration. En utilisant $\mathrm{Tr}(A'BCD) = \mathsf{vec}(A)'(D \otimes B)\mathsf{vec}(C)$ et en écrivant $\|\mathsf{vec}((\bar{X}^1)' - (\bar{X}^2)')\|^2_{S_W^{-1}} = \mathsf{vec}((\bar{X}^1)' - (\bar{X}^2)')'S_W^{-1}\mathsf{vec}((\bar{X}^1)' - (\bar{X}^2)')$ le carré de la distance de Mahalanobis entre les moyennes vectorisées, on trouve :

$$\mathsf{vec}((\bar{X}^1)' - (\bar{X}^2)')'(D \otimes G)\mathsf{vec}((\bar{X}^1)' - (\bar{X}^2)') = \mathrm{Tr}\left((\bar{X}^1 - \bar{X}^2)G((\bar{X}^1)' - (\bar{X}^2)')D\right)\,,$$

et d'après la décomposition (5.16), on sait que

$$\mathrm{Tr}\left((\bar{X}^1 - \bar{X}^2)G((\bar{X}^1)' - (\bar{X}^2)')D\right) = \sum_{q=1}^{Q} \lambda_q \,.$$

Notons enfin que d'après (5.18), la coordonnée du projeté de $\mathsf{vec}((\bar{X}^1)' - (\bar{X}^2)')$ sur la droite engendrée par $u_q \otimes v_q$ est $\sqrt{\lambda_q}$. Il vient alors directement l'expression (5.20). □

Lien avec l'analyse discriminante matricielle descriptive

Dans le cas binaire, nous montrons ici l'équivalence entre l'analyse discriminante matricielle présentée dans la section 5.2.1 et la DVS proposée.

Proposition 14. *Soient $u_1, \ldots, u_Q \in \mathbb{R}^K$ les vecteurs propres D-orthonormés définis dans la DVS de la différence des moyennes des deux classes (5.16). Pour tout $q = 1, \ldots, Q$, le vecteur propre u_q associé à la valeur propre non-nulle λ_q est aussi vecteur propre de la matrice $S_B^L(S_W^L)^{-1} \in \mathbb{R}^{K\times K}$ associé à la valeur propre non-nulle $c\lambda_q$, avec $c = n_1 n_2/(n_1+n_2)^2$.*
De même les vecteurs propres G-orthonormés v_q, $q = 1, \ldots, Q$ définis dans la DVS sont aussi vecteurs propres de la matrice $S_B^R(S_W^R)^{-1} \in \mathbb{R}^{J\times J}$ associés aux valeurs propres $c\lambda_q$.

Démonstration. En utilisant l'expression (5.13) nous avons pour tout $q = 1, \ldots, Q$:

$$(\bar{X}^1 - \bar{X}^2)(S_W^R)^{-1}(\bar{X}^1 - \bar{X}^2)'(S_W^L)^{-1}u_q = \lambda_q u_q \tag{5.21}$$

$$\iff \quad \frac{n_1 n_2}{(n_1+n_2)^2}(\bar{X}^1 - \bar{X}^2)(S_W^R)^{-1}(\bar{X}^1 - \bar{X}^2)'(S_W^L)^{-1}u_q = \frac{n_1 n_2}{(n_1+n_2)^2}\lambda_q u_q \tag{5.22}$$

$$\iff \quad S_B^L(S_W^L)^{-1}u_q = \frac{n_1 n_2}{(n_1+n_2)^2}\lambda_q u_q \,. \tag{5.23}$$

De même, on montre que

$$S_B^R(S_W^R)^{-1}v_q = \frac{n_1 n_2}{(n_1+n_2)^2}\lambda_q v_q \,. \tag{5.24}$$

□

De plus, nous montrons également qu'il est équivalent de rechercher les vecteurs propres de la matrice $S_B^L(S_W^L)^{-1}$ (resp. $S_B^R(S_W^R)^{-1}$) et ceux de la matrice $(S_W^L)^{-1}S_B^L$ (resp. $(S_W^L)^{-1}S_B^L$). Donnons le résultat dans l'espace des colonnes $\mathbb{R}^K$:

Proposition 15. *Soient les vecteurs propres D-orthonormés u_q, $q = 1, \ldots, Q$ de la matrice $S_B^L(S_W^L)^{-1} \in \mathbb{R}^{K\times K}$ associés aux valeurs propres non nulles λ_q. Si nous posons*

$$\tilde{u}_q = Du_q \,, \ \textit{pour } q = 1, \ldots Q \,, \tag{5.25}$$

alors les vecteurs $\tilde{u}_q$ sont les vecteurs propres D^{-1}-orthonormés de la matrice $(S_W^L)^{-1}S_B^L$ associés aux mêmes Q valeurs propres non nulles λ_q.

Démonstration. Rappelons que $D = (S_W^L)^{-1}$.

$$(S_W^L)^{-1}S_B^L\tilde{u}_q = (S_W^L)^{-1}S_B^L(S_W^L)^{-1}u_q \tag{5.26}$$

$$\iff \quad (S_W^L)^{-1}\lambda_q u_q = \lambda_q \tilde{u}_q \,. \tag{5.27}$$

□

Soit $\tilde{U}_r \in \mathbb{R}^{K\times r}$ la matrice dont les colonnes sont les vecteurs propres $\tilde{u}_q$, $q = 1, \dots r$ avec $r \leq Q$ associée aux r plus grandes valeurs propres. On a donc la relation

$$\tilde{U}_r = (S_W^L)^{-1} U_r \ , \tag{5.28}$$

où U_r est la matrice dont les colonnes sont les r vecteurs propres u_q, $q = 1, \dots r$. Il en est de même dans l'espace des lignes $\mathbb{R}^J$.

Proposition 16. *Soient les vecteurs propres G-orthonormés v_q, $q = 1, \dots, Q$ de la matrice $S_B^R (S_W^R)^{-1} \in \mathbb{R}^{J\times J}$ associés aux valeurs propres non nulles λ_q. Si nous posons*

$$\tilde{v}_q = G v_q \,, \ \textit{pour } q = 1, \dots Q \,, \tag{5.29}$$

alors les vecteurs $\tilde{u}_q$ sont les vecteurs propres G^{-1}-orthonormés de la matrice $(S_W^R)^{-1} S_B^R$ associés aux Q mêmes valeurs propres non nulles λ_q.

Soit $\tilde{V}_r \in \mathbb{R}^{K\times r}$ la matrice dont les colonnes sont les vecteurs propres $\tilde{v}_q$, $q = 1, \dots r$ avec $r \leq Q$ associée aux r plus grandes valeurs propres. On a donc la relation

$$\tilde{V}_r = (S_W^R)^{-1} V_r \ , \tag{5.30}$$

où V_r est la matrice dont les colonnes sont les r vecteurs propres v_q, $q = 1, \dots r$.

5.2.3 Représentations des données : caractéristiques et représentations graphiques

Nous exploitons ici les propriétés de la décomposition en valeurs singulières présentée ci-dessus pour la description des données au travers de leurs composantes discriminantes. La séparabilité des espaces ligne et colonne permet de représenter les données dans les différents espaces $\mathbb{R}^K$, $\mathbb{R}^J$ et $\mathbb{R}^{KJ}$.

Caractéristiques ligne-colonne

Les caractéristiques ligne-colonne sont définies par projection dans l'espace ligne-colonne $\mathbb{R}^{KJ}$. Soit E_r le sous-espace engendré par les r premiers vecteurs propres $u_q \otimes v_q$. Un résultat important de la DVS est donné par l'approximation de rang faible de la différence de moyennes.

Proposition 17 (Théorème d'Eckart-Young (Eckart and Young, 1936))**.** *Soit $Q = \mathsf{rang}(\bar{X}^1 - \bar{X}^2)$, et soit $r \leq Q$. La projection S_W^{-1}-orthogonale de la différence des moyennes vectorisées sur E_r est donnée par*

$$\mathsf{proj}_{E_r}\left(\mathsf{vec}((\bar{X}^1)' - (\bar{X}^2)')\right) = \sum_{q=1}^{r} \sqrt{\lambda_q}(u_q \otimes v_q) \,. \tag{5.31}$$

L'erreur d'approximation de la différence des moyennes projetée sur E_r est alors :

$$\left\| \mathsf{vec}((\bar{X}^1)' - (\bar{X}^2)') - \mathsf{proj}_{E_r}\left(\mathsf{vec}((\bar{X}^1)' - (\bar{X}^2)')\right) \right\|^2_{S_W^{-1}} = \lambda_{r+1} + \lambda_{r+2} + \dots + \lambda_Q \,. \tag{5.32}$$

Cette dernière proposition permet de mesurer la perte d'information induite par la sélection des r premiers vecteurs dans l'analyse à partir de la décomposition en valeurs singulières (5.18).

Considérons les observations $x_i = \mathsf{vec}(X_i') \in \mathbb{R}^{KJ}$, où $X_i \in \mathbb{R}^{K\times J}$. Les coordonnées des projetés de x_i sur les axes engendrés par les vecteurs $u_q \otimes v_p \in \mathbb{R}^{KJ}$ pour $q, p = 1, \ldots, r$ sont données par

$$\langle x_i, u_q \otimes v_p \rangle_{S_W^{-1}} = x_i' S_W^{-1} (u_q \otimes v_p)\,, \tag{5.33}$$

où $S_W^{-1} = (S_W^L)^{-1} \otimes (S_W^R)^{-1}$. A partir de la relation (5.29) on en déduit

$$\langle x_i, u_q \otimes v_p \rangle_{S_W^{-1}} = x_i' (S_W^L)^{-1} u_q \otimes (S_W^R)^{-1} v_p = x_i' (\tilde{u}_q \otimes \tilde{v}_p)\,. \tag{5.34}$$

Définition 7. *Le vecteur des* caractéristiques ligne-colonne $y_i \in \mathbb{R}^{r^2}$ *est le vecteur des coordonnées des projetés de* x_i *sur les* r^2 *axes engendrés par* $\tilde{u}_q \otimes \tilde{v}_p$, $q, p = 1, \ldots, r$ *:*

$$y_i = (\tilde{U}_r \otimes \tilde{V}_r)' x_i\,. \tag{5.35}$$

De manière équivalente on définit la matrice des caractéristiques ligne-colonne $Y_i \in \mathbb{R}^{r\times r}$ *par*

$$Y_i = \tilde{U}_r' X_i \tilde{V}_r\ , \tag{5.36}$$

avec $y_i = \mathsf{vec}(Y_i')$.

Caractéristiques ligne et caractéristiques colonne

La DVS (5.16) effectuée sur la différence des moyennes matricielles permet aussi d'étudier les projections des observations dans l'espace des lignes ($\mathbb{R}^J$) et dans l'espace des colonnes ($\mathbb{R}^K$) séparément.

Pour tout $k = 1, \ldots, r$, la coordonnée du projeté du $k^{\text{ème}}$ vecteur-ligne de X_i, notée $X_i(k, .) \in \mathbb{R}^J$, sur l'axe engendré par le vecteur v_q selon la métrique $G = (S_W^R)^{-1}$ est donnée par :

$$\langle X_i(k, .), v_q \rangle_G = X_i(k, .)' G v_q = X_i(k, .)' \tilde{v}_q\,, \tag{5.37}$$

où $\tilde{v}_q$ est le vecteur propre de $(S_W^R)^{-1} S_B^R$. associé à la valeur propre λ_q.

Définition 8. *Les* r *colonnes de la matrice* $X_i \tilde{V}_r \in \mathbb{R}^{K\times r}$ *sont appelées caractéristiques discriminantes des* K *lignes de* X_i*.*

Pour tout $j = 1, \ldots, J$, la coordonnée du projeté du $j^{\text{ème}}$ vecteur-colonne de X_i, notée $X_i(., j) \in \mathbb{R}^K$, sur l'axe engendré par le vecteur v_q selon la métrique $D = (S_W^L)^{-1}$ est donnée par :

$$\langle u_q, X_i(., j) \rangle_D = u_q' D X_i(., j) = \tilde{u}_q' X_i(., j)\,, \tag{5.38}$$

où $\tilde{u}_q$ est le vecteur propre de $(S_W^L)^{-1} S_B^L$ associé à la valeur propre λ_q. Nous en déduisons la définition suivante :

Définition 9. *Les* r *lignes de la matrice* $\tilde{U}_r' X_i \in \mathbb{R}^{r\times J}$ *sont appelées caractéristiques discriminantes des* J *colonnes de* X_i*.*

Représentations graphiques

Par analogie avec l'analyse en composantes principales, le formalisme développé ci-dessus permet de donner plusieurs types de représentations graphiques des données, fournissant des informations complémentaires :

— la représentation des observations $x_i \in \mathbb{R}^{KJ}$ relativement aux r premiers axes engendrés par les vecteurs $\tilde{u}_q \otimes \tilde{v}_q$,

— la représentation des colonnes (vecteurs de $\mathbb{R}^K$) sur les axes engendrés par les vecteurs $\tilde{u}_q$,

— la représentation des lignes (vecteurs de $\mathbb{R}^J$) sur les axes engendrés par les vecteurs $\tilde{v}_q$.

Dans l'application aux signaux EEG présentée dans le chapitre qui suit, ceci fournira des représentations des essais dans le domaine spatial (colonnes), le domaine temporel (lignes) ainsi que le domaine spatio-temporel.

5.3 Classification dans l'espace des caractéristiques

L'approche descriptive décrite ci-dessus permet de dégager les caractéristiques discriminantes entre deux classes. Les composantes extraites par projection des observations dans les sous-espaces ligne et colonne sélectionnés peuvent être également exploitées pour la classification.

Nous nous replaçons dans le cadre du modèle de mélange gaussien avec hypothèse de séparabilité décrit dans le chapitre précédent, en supposant l'égalité des matrices de covariance des classes. Dans chaque classe c, les vecteurs x_i^c sont des réalisations d'un vecteur gaussien de loi $\mathcal{N}(\mu^c, \Sigma_L \otimes \Sigma_R)$. Les vecteurs des caractéristiques spatio-temporelles $y_i^c = (\tilde{U}_r \otimes \tilde{V}_r)' x_i^c$ obtenus par projection des x_i^c sont donc des réalisations d'un vecteur gaussien de loi

$$y_i^c \sim \mathcal{N}\left(\tilde{\mu}^c, \tilde{\Sigma}\right) , \tag{5.39}$$

où

$$\tilde{\mu}^c = (\tilde{U}_r \otimes \tilde{V}_r)' \mu^c \in \mathbb{R}^{r^2} , \quad \text{et} \quad \tilde{\Sigma} = \left(\tilde{U}_r' \Sigma_L \tilde{U}_r\right) \otimes \left(\tilde{V}_r' \Sigma_R \tilde{V}_r\right) \in \mathbb{R}^{r^2 \times r^2} . \tag{5.40}$$

L'expression de $\tilde{\Sigma}$ ci-dessus a été obtenue en utilisant les propriétés des vecteurs gaussiens et du produit tensoriel :

$$\tilde{\Sigma} = (\tilde{U}_r \otimes \tilde{V}_r)' \left(\Sigma_L \otimes \Sigma_R\right) (\tilde{U}_r \otimes \tilde{V}_r) = \left(\tilde{U}_r' \Sigma_L \tilde{U}_r\right) \otimes \left(\tilde{V}_r' \Sigma_R \tilde{V}_r\right) .$$

La règle de décision de Bayes (LDA) peut donc être appliquée au vecteur des caractéristiques y_i de chaque nouvelle observation. Dans ce cas, l'analyse discriminante se base sur la minimisation de la fonction discriminante $\delta^c(y_i)$ définie par :

$$\delta^c(y_i) = (y_i - \tilde{\mu}^c)' \, \tilde{\Sigma}^{-1} \, (y_i - \tilde{\mu}^c) - 2 \log(p_c) . \tag{5.41}$$

Notons que lorsque l'on ne fait pas de sélection, i.e. $\tilde{U}_r = \tilde{U}$ et $\tilde{V}_r = \tilde{V}$, la règle de décision basée sur la fonction discriminante (5.41) est équivalente à la règle de décision (4.42) appliquée aux observations x_i.

Nous avons vu que cette dernière est optimale au sens de la minimisation du risque moyen d'erreur. Nous verrons cependant dans le chapitre suivant que les résultats de classification sont

généralement meilleurs lorsqu'une sélection des caractéristiques associées aux premiers axes a été effectuée. Notre interprétation est que ceci résulte d'une adéquation imparfaite du modèle aux données, plus particulièrement du fait que la "vraie" matrice de covariance n'est qu'imparfaitement modélisée par un produit de Kronecker. Dans ce contexte, la sélection des caractéristiques les plus importantes est comparable à un débruitage, par projection sur un sous-espace dans lequel l'hypothèse de séparabilité est mieux satisfaite.

Chapitre 6

Applications

Sommaire

6.1 Construction de filtres spatiaux discriminants

Nous pouvons maintenant exploiter les techniques d'analyse discriminante matricielle développées dans le chapitre 5 pour la construction de *filtres spatiaux*. Dans le contexte des EEG, le filtrage spatial vise à substituer aux électrodes des combinaisons linéaires, appelées canaux, qui optimisent la discrimination entre les classes (nous sommes ici dans le cas binaire). Il s'agit d'une appellation générique qui recouvre différentes approches. Nous construisons des filtres spatiaux en nous basant sur la factorisation de la matrice de covariance en produit de Kronecker, comme développé par exemple par (Gutiérrez and Escalona-Vargas, 2010) dans des cas particuliers.

Nous illustrons cette approche, et la comparons à d'autres sur les deux jeux de données décrits dans la section 2.4. Dans les deux cas, notre point de départ est les données après transformation en ondelette et réduction de la dimension temporelle.

6.1.1 Formalisation

Soit $X_i \in \mathbb{R}^{K\times J}$, où K est le nombre de coefficients multi-échelle et J est le nombre d'électrodes. Les matrices de covariance $S_W^R \in \mathbb{R}^{J\times J}$ et $S_B^R \in \mathbb{R}^{J\times J}$ (resp. $S_W^L \in \mathbb{R}^{K\times K}$ et $S_B^L \in \mathbb{R}^{K\times K}$) sont les matrices de covariance intra- et inter- classe des électrodes (resp. des coefficients multi-échelle). L'analyse discriminante descriptive mise en œuvre ici visera donc à déterminer les combinaisons linéaires les plus discriminantes de signaux enregistrés sur les électrodes, associées aux vecteurs propres S_W^R-orthonormés $\tilde{v}_q$ définis en (5.29). Chaque vecteur $\tilde{v}_q \in \mathbb{R}^J$ est appelé filtre spatial, d'après la terminologie couramment utilisée (McFarland et al., 1997; Blankertz et al., 2008; Blankertz et al., 2011; Parra et al., 2005).

Une fois déterminés, ces filtres spatiaux sont appliqués à la matrice X_i par projection de ses K lignes sur le sous-espace engendré par les vecteurs $\tilde{v}_1, \ldots, \tilde{v}_r$ comme dans la définition 8.

Le nombre r de filtres $\tilde{v}_q$ conservés pour l'analyse est choisi selon le critère du pourcentage cumulé des valeurs propres λ_q.

Approche classique

Lorsqu'ils sont déterminés par analyse discriminante, les filtres spatiaux sont classiquement définis comme les vecteurs propres de la matrice $(\tilde{S}_W^R)^{-1}\tilde{S}_B^R \in \mathbb{R}^{J\times J}$ (voir par exemple (Huang et al., 2008; Huang et al., 2009; Gutiérrez and Escalona-Vargas, 2010; Spinnato et al., 2012)), où :

$$\tilde{S}_B^R = \frac{1}{nK}\sum_{c=1}^{2} n_c \left(\bar{X}^1 - \bar{X}^2\right)' \left(\bar{X}^1 - \bar{X}^2\right), \tag{6.1}$$

$$\tilde{S}_W^R = \frac{1}{nK}\sum_{c=1}^{2}\sum_{i=1}^{n_c} \left(X_i^c - \bar{X}^c\right)' \left(X_i^c - \bar{X}^c\right). \tag{6.2}$$

Notons que les matrices de covariance inter-classe et intra-classe des électrodes $\tilde{S}_B^R$ et $\tilde{S}_W^R$ coïncident avec les matrices S_B^R et S_W^R telle que définie en (5.14) et (5.4) dans le cas où $S_W^L = I_K$. Définir le filtrage spatial de cette manière revient implicitement à faire l'hypothèse que les K coefficients de $X(.,j) \in \mathbb{R}^K$ sont i.i.d. Cette hypothèse sur les lignes peut être testée en utilisant le test présenté dans la section 4.2.3.

Dans la section suivante, cette méthode sera nommée "LDA matricielle I_K".

Normalisation de la variabilité temporelle

L'hypothèse de décorrélation des coefficients multi-échelle n'est en général pas complètement réaliste, il importe donc de prendre en compte ces corrélations. Sous hypothèse de séparabilité, le formalisme développé dans le chapitre précédent, et notamment l'estimation conjointe de S_W^L et S_W^R via l'algorithme *flip-flop* permet de construire des filtres spatiaux tenant compte des corrélations temporelles, filtres toujours définis comme vecteurs propres S_W^R-orthonormés de la matrice $(S_W^R)^{-1}S_B^R$.

Dans la section suivante, cette méthode sera nommée "LDA matricielle".

Pénalisation spatiale

En EEG, il est bien connu que les signaux enregistrés sur des électrodes proches sont très corrélés entre eux. On s'attend donc à ce que les filtres spatiaux optimaux exhibent une propriété similaire, et présentent un aspect "lisse". En d'autres termes, l'on s'attend à ce qu'une information présente sur une électrode se retrouve aussi au niveau des électrodes proches. Cette propriété peut être confortée par l'introduction d'un terme de régularisation dans le calcul des filtres, suivant l'analyse discriminante pénalisée proposée par Hastie et al. (Hastie et al., 1995) et présentée dans la section 4.2.4. Nous introduisons ainsi $\widetilde{C}_W^R$, version pénalisée de la matrice de covariance S_W^R, définie par

$$\widetilde{C}_W^R = S_W^R + \alpha\Omega \ , \tag{6.3}$$

où $\Omega \in \mathbb{R}^{J\times J}$ est une matrice semi-définie positive qui encode l'information spatiale donnée *a priori* et α est un hyperparamètre permettant de régler le poids de cette information sur la matrice de covariance S_W^R. Afin de tenir compte de la distribution des électrodes à la surface du scalp, nous utilisons pour Ω un Laplacien de surface (Huiskamp, 1991) (déjà évoqué dans la section 2.2.1) basé sur une matrice de voisinage du premier ordre des électrodes :

Définition 10. *Soient deux électrodes e_i et e_j, $i, j = 1, \ldots, J$, la matrice Laplacienne Ω est définie par*

$$\Omega_{ij} = \begin{cases} \mathsf{deg}(e_i) & si \quad i = j \\ -1 & si \quad e_i \text{ et } e_j \text{ sont adjacentes} \\ 0 & sinon \end{cases} \tag{6.4}$$

où $\mathsf{deg}(e_i)$ est le nombre d'électrode adjacentes à e_i (au premier ordre).

Cette approche repose donc sur le choix de deux ingrédients :

1. La construction de la matrice de voisinage se base donc sur la notion d'adjacence entre les électrodes, et il n'est pas facile de donner un sens physiologique à cette notion. Nous avons fait ici un choix arbitraire de la notion de voisinage en décidant de considérer la proximité au premier ordre comme on pourrait le faire dans un réseau carré (les voisines d'une électrode donnée sont les quatre électrodes les plus proches) sans tenir compte de la forme volumique de la tête. Une seconde difficulté rencontrée dans la construction d'une telle matrice consiste est la définition des voisins des électrodes de bord. Ce choix peut évidemment être discuté, et amélioré (par exemple en exploitant des notions de géométrie riemannienne comme dans (Barachant and Bonnet, 2011)). Nous avons choisi de nous en tenir au choix le plus simple.
2. La détermination de l'hyperparamètre α dans (6.3) nécessite un ensemble de validation. Dans ce travail, la classification peut être considérée comme un critère de validation de la méthode. C'est pourquoi α_{opt} est choisi par validation croisée comme la valeur de α qui maximise le taux de classification sur l'ensemble de validation.

Dans la section suivante, cette méthode sera nommée "LDA matricielle pénalisée".

6.1.2 Étude des filtres spatiaux : application aux potentiels d'erreur

Mise en œuvre. Les résultats présentés ci-dessous ont été obtenus à partir de 660 essais (dont 60 erreurs). La phase de détermination du paramètre de pénalisation spatiale α est composée de 60 essais (dont 30 erreurs). Une procédure de *leave-one-out* a été mise en place : pour une valeur donnée $\alpha \geq 0$, à chaque itération, un essai est enlevé du jeu de validation. Les filtres sont appris sur les 59 essais restants et appliqués à l'essai enlevé. La valeur α optimale est celle qui minimise l'erreur moyenne de classification obtenue sur les 60 essais testés les uns après les autres.

Les résultats sont présentés ici pour le participant A du jeu de données. Des résultats équivalents pour deux autres sujets sont présentés dans l'annexe C.1. .

Éboulis des valeurs propres. La figure 6.1 représente l'éboulis des valeurs propres associé à chacune des trois méthodes étudiées.
Les valeurs propres de l'analyse discriminante matricielle ("LDA matricielle") sont plus grandes mais décroissent bien moins vite que celles des deux autres méthodes ("LDA matricielle I_k" et "LDA matricielle pénalisée"). Pour conserver 75% de l'information discriminante, 9 filtres sont nécessaires à la "LDA matricielle" contre 2 pour la "LDA matricielle pénalisée" (avec $\alpha_{opt} = 0,2$). Les valeurs propres associées à l'approche classique ("LDA matricielle I_k") sont très faibles. Afin de conserver 75% de l'information discriminante il faudrait ici conserver les 8 premiers filtres.

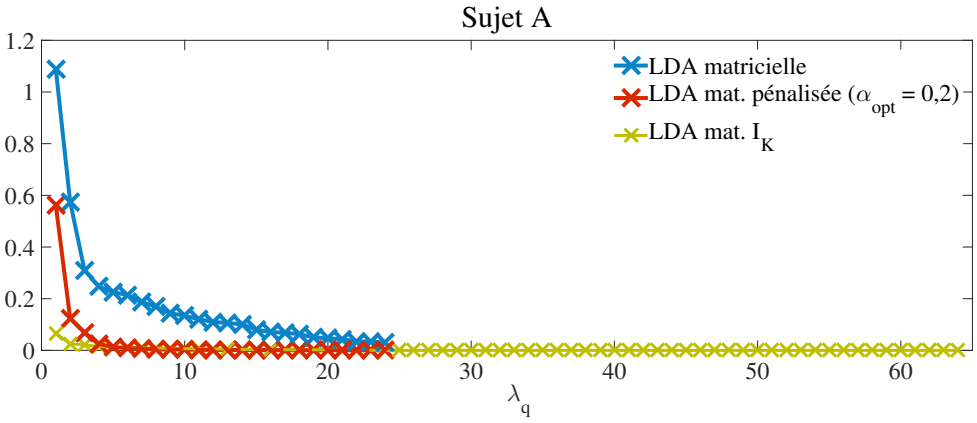


FIGURE 6.1 – Participant A : éboulis des valeurs propres pour les trois méthodes : LDA matricielle (bleu), LDA matricielle pénalisée (rouge) et LDA matricielle I_K (jaune). Par construction, les deux premières méthodes ont $K = 24$ valeurs propres non-nulles (voir propriété de dualité) tandis que la troisième est de rang inférieur ou égal à 64.

Topographies des filtres. Les 3 premiers filtres spatiaux les plus discriminants sont représentés pour chaque méthode dans la figure 6.2. Rappelons aussi que les filtres spatiaux considérés sont orthogonaux selon la métrique S_W^R (i.e. $\tilde{v}_q S_W^R \tilde{v}_j = 0$ pour tout $j \neq q$, avec $j, q = 1, \ldots, 64$), il est donc difficile d'interpréter physiologiquement ces combinaisons linéaires, excepté pour $\tilde{v}_1$. La comparaison de ce dernier filtre entre les 3 méthodes s'avère cependant intéressante. Les filtres semblent indiquer que la zone fronto-centrale joue un rôle important dans la discrimination des potentiels d'erreur. En effet, cette zone est largement présente dans les filtres construits selon les trois méthodes. Ceci est cohérent avec ce que l'on connaît des potentiels d'erreur (Falkenstein et al., 2000).
Pour l'analyse discriminante matricielle, ce filtre contient 25% de l'information spatiale discriminante totale, contre 68% pour la méthode pénalisée et 30% pour l'approche classique.
Le rôle joué par la pénalisation spatiale est évident. Les filtres représentés sont clairement lissés et les activités discriminantes sont beaucoup moins localisées que pour les deux autres méthodes.

Discussion Nos résultats semblent donc montrer l'intérêt de l'hypothèse de séparabilité conjuguée à une régularisation spatiale. Les filtres spatiaux obtenus reproduisent des propriétés de localisation connues (fronto-centrale) dans ce type de données, et la régularisation spatiale améliore le pouvoir discriminant du premier filtre en comparaison des autres.
Les mêmes observations peuvent être faites pour les participants B et H dont les résultats sont présentés dans l'annexe C.1. Pour le participant B, la valeur de pénalisation optimale étant égale à zéro dans l'exemple, la LDA matricielle pénalisée est équivalente à la LDA matricielle. Le 1er filtre concentre 21% de l'information spatiale discriminante contre 30% pour la LDA matricielle I_K. Pour le participant H, la topographie du 1er filtre représente clairement des activités fronto-centrales similaires à celles observées chez le participant A. De plus, le pouvoir discriminant de ce filtre obtenu par l'analyse discriminante pénalisée est significativement plus important (53%) que celui du filtre obtenue par la LDA matricielle (24%) et la LDA matricielle I_K (31%).

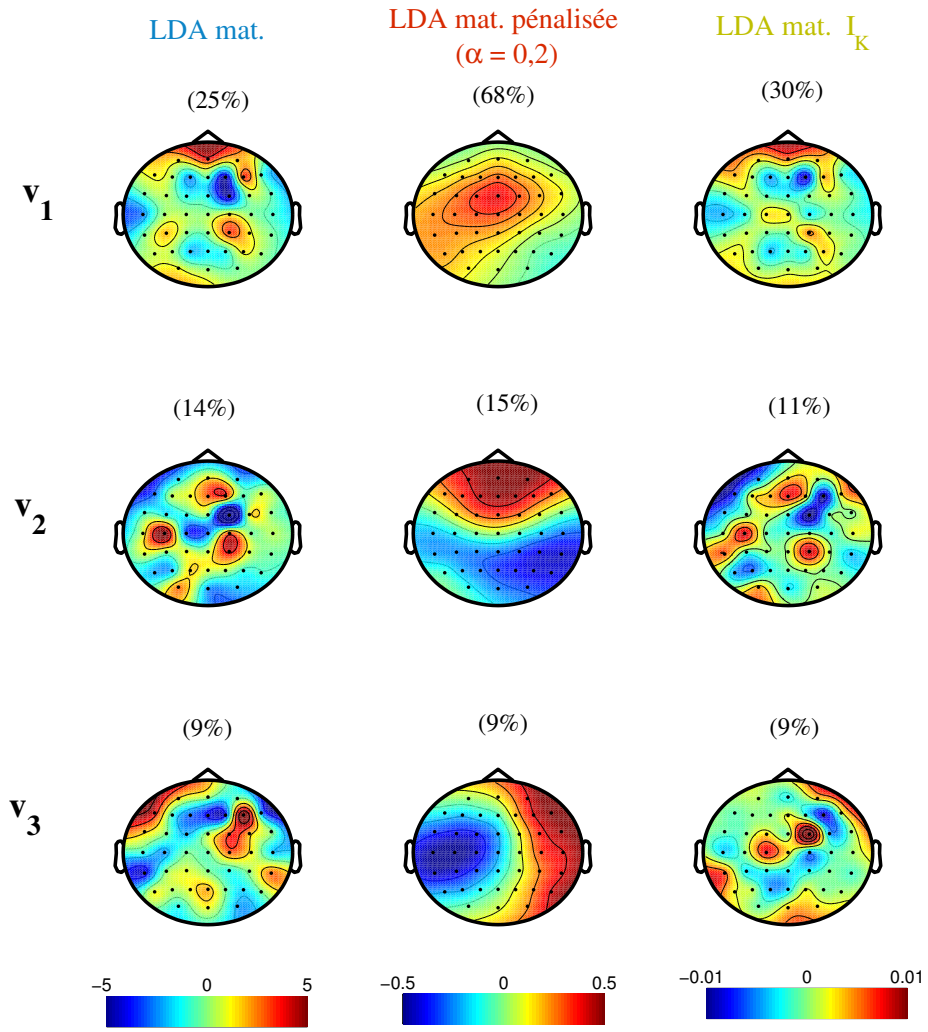


FIGURE 6.2 – Participant A. Représentation des 3 filtres spatiaux les plus discriminants ($\tilde{v}_1$, $\tilde{v}_2$ et $\tilde{v}_3$) obtenus à partir des deux méthodes proposées : LDA matricielle (1ère colonne), LDA matricielle pénalisée (2ème colonne), et de l'approche classique (3èmecolonne). Le pourcentage d'information discriminante contenue dans chacun des filtres est donné entre parenthèses. Les barres horizontales indiquent l'échelle de valeurs pour la colonne correspondante. Chaque topographie correspond au schéma d'une tête, les oreilles et le nez indiquant son sens. Chaque point correspond à l'emplacement d'une des 64 électrodes.

6.2 Extraction de caractéristiques espace-temps-échelle

Intéressons nous à présent à l'extraction conjointe de caractéristiques discriminantes spatiales et temporelles (ou plutôt multi-échelle puisque les données de départ sont issues d'une transformation en ondelettes discrète suivie d'une réduction de dimension correspondante).

Ici, les caractéristiques discriminantes des lignes (définition 8) correspondent aux **caractéristiques discriminantes temps-échelle**. La $q^{\text{ème}}$ caractéristique temps-échelle de l'essai i est constituée des coordonnées des K lignes correspondant aux coefficients multi-échelle sur l'axe engendré par $\tilde{v}_q \in \mathbb{R}^J$.

Les caractéristiques discriminantes des colonnes (définition 9) correspondent aux **caractéristiques discriminantes spatiales**. La $q^{\text{ème}}$ caractéristique spatiale de l'essai i est constituée des coordonnées des J colonnes correspondant aux électrodes sur l'axe engendré par $\tilde{u}_q \in \mathbb{R}^K$.

Nous appelons aussi **caractéristiques espace-temps-échelle** les caractéristiques ligne-colonne de l'essai i.

6.2.1 Étude des potentiels d'erreurs

Nous rappelons que les signaux ont été enregistrés à partir de 64 électrodes et qu'après sélection 24 coefficients multi-échelle ont été conservés (voir détails section 2.4.1). On a donc $X_i^c \in \mathbb{R}^{24\times 64}$, la $i^{\text{ème}}$ observation de la classe c, avec $i = 1, \ldots, n_c$. Nous considérons les deux classes *Erreur* ($c = 1$) et *Correct* ($c = 2$). Les matrices de covariance temporelle et spatiale sont respectivement $\Sigma_L \in \mathbb{R}^{24\times 24}$ et $\Sigma_R \in \mathbb{R}^{64\times 64}$.

Les données vectorisées $x_i^c = \mathsf{vec}((X_i^c)')$ sont ici de dimension $64 * 24 = 1536$. L'analyse discriminante classique demanderait donc d'estimer une matrice de covariance $\Sigma \in \mathbb{R}^{1536\times 1536}$, soit près de 11.10^5 paramètres, ce qui n'est pas réalisable ici. Nous utiliserons plutôt ici deux méthodes dérivées de la LDA et qui permettent dans cette situation d'estimer une matrice de covariance : i) une LDA qui utilise la pseudo-inverse de Moore-Penrose plutôt que l'inverse de la matrice de covariance et ii) la LDA diagonale (3.38) (voir par exemple (Bickel and Levina, 2004)).

Dans la suite, le jeu d'apprentissage est constitué de 60 erreurs et de 600 réponses correctes, le jeu de données pour l'apprentissage du paramètre de pénalisation spatiale est constitué de 60 essais dont 30 erreurs, tirés au hasard, et le reste des essais constituent le jeu test. Ce tirage est effectué 30 fois de façon aléatoire et à chaque itération les performances des différentes méthodes sont mesurées. Notons que les participants dont le nombre d'erreurs produites est inférieures à 90 n'ont pas été retenus dans cette étude (voir table 2.2).

Analysons dans une premier temps les vecteurs propres obtenus en termes de caractéristiques discriminantes. Pour ce faire nous utilisons les outils de visualisation des données et des variables discriminantes décrits dans la section 5.2.3. Nous présentons les résultats obtenus sur le participant A.

Analyse des caractéristiques spatio-temporelles

Les figures ci-dessous représentent la projection des essais *Erreur* et *Correct* ainsi que les moyennes des classes dans le premier plan discriminant engendré par $\tilde{u}_1 \otimes \tilde{v}_1$, $\tilde{u}_2 \otimes \tilde{v}_2$. Ces vecteurs ont été calculés soit par LDA matricielle (figure 6.3, graphiques de gauche), soit par la version pénalisée (figure 6.3, graphiques de droite, ici $\alpha_{opt} = 0,6$). Chaque nuage de points est accompagné de l'éboulis des valeurs propres associées aux vecteurs propres. Dans les deux

cas, ces représentations suggèrent que les deux composantes résument la plus grande part de l'information discriminante contenue dans les données. Notons que pour la version pénalisée, un étude du nuage de points montre clairement que la première composante $\tilde{u}_1 \otimes \tilde{v}_1$ suffit à discriminer les deux classes. Pour la LDA matricielle non-pénalisée, la deuxième composante $\tilde{u}_2 \otimes \tilde{v}_2$ participe également à discriminer les classes. Une représentation des observations dans les plans engendrés par les vecteurs suivants ($\tilde{u}_3 \otimes \tilde{v}_3$ et $\tilde{u}_4 \otimes \tilde{v}_4$) permettrait de compléter cette analyse.

Cette représentation est particulièrement intéressante pour étudier le comportement essai par essai, visualiser la distance de chaque point par rapport à la moyenne de sa classe et identifier d'éventuels comportements atypiques. Par exemple quelques essais *Erreur* (en rouge) sont beaucoup plus proches de la moyenne de la classe *Correct* (carré jaune) que de la moyenne de leur propre classe. Nous pouvons penser que ces mêmes essais seront très probablement mal classés dans une procédure de détection des erreurs. La même remarque peut être faite pour certains essais *Correct* dont le comportement semble être plus proche de celui d'une erreur.

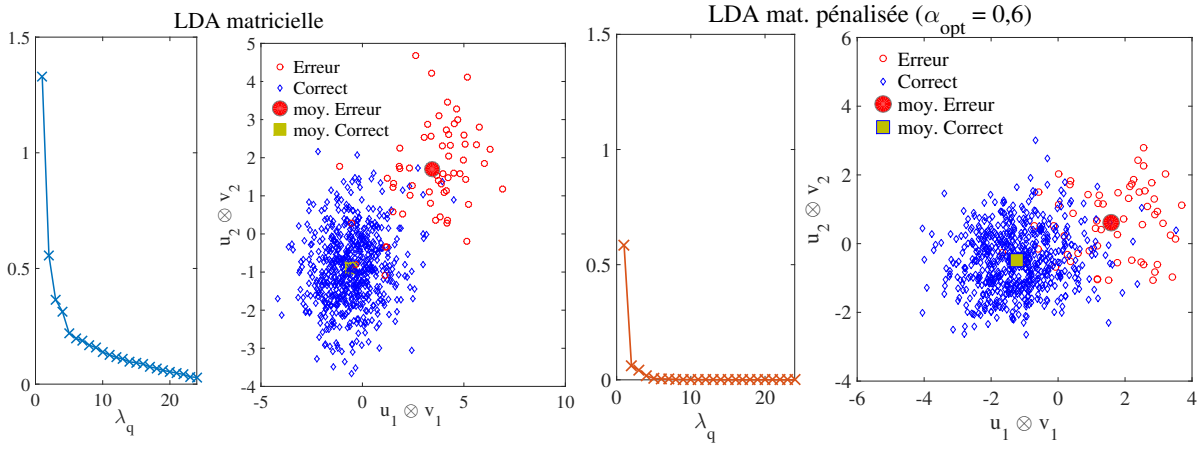


FIGURE 6.3 – Participant A. Éboulis des valeurs propres et nuages des observations projetées sur le premier plan discriminant engendré par les vecteurs $\tilde{u}_1 \otimes \tilde{v}_1$ et $\tilde{u}_2 \otimes \tilde{v}_2$. Comparaison de la LDA matricielle (gauche) et de sa version pénalisée (droite) avec $\alpha = 0,6$.

Analyse des caractéristiques discriminantes spatiales et temporelles

Les caractéristiques discriminantes sont ici analysées à partir de la projection de la différence des moyennes sur les sous-espaces engendrés par les premiers vecteurs propres. Les résultats sont toujours donnés pour le participant A. D'autres résultats seront donnés en annexes.

Les graphiques du haut de la figure 6.4 représentent (a) la projection des différences des moyennes des classes mesurées sur les 64 électrodes et (b) les trois premières caractéristiques temporelles discriminantes obtenues par synthèse d'ondelettes à partir des projections des différences des coefficients moyens. Dans ce cas, les vecteurs propres ont été déterminés par la LDA matricielle. Les deux graphiques du bas de la même figure représentent quant à eux les mêmes résultats obtenus à partir de la LDA matricielle pénalisée sur le même jeu de données (avec $\alpha_{opt} = 0,6$).

Pour la LDA matricielle comme pour sa version pénalisée, la première composante spatiale discriminante semble correspondre à la négativité d'erreur (Ne) suivie d'un positivité qui pour-

rait s'apparenter à la positivité d'erreur (Pe). Toujours dans le cas de la LDA matricielle, la seconde caractéristique temporelle reflète les mêmes activités mais en plus faible amplitude et avec une différence de latence. La troisième composante se caractérise par des fluctuations plus faibles à qui il est difficile de donner du sens. Pour la version pénalisée, la première caractéristique semble suffire à décrire la négativité d'erreur et les deux autres caractéristiques ne correspondent à aucune activité particulière, ce qui est cohérent avec l'éboulis des valeurs propres présentés dans la figure 6.3 (la première valeur propre concentre à elle seule plus de 70% de l'information discriminante contenue dans le signal).
Les deux premières caractéristiques spatiales représentées sur la même figure indiquent des groupes distincts d'électrodes associés aux caractéristiques temporelles. On retrouve en particulier l'électrode FCz et ses électrodes voisines (FC1,FC2,Fz, F1 et F2) pour la première composante discriminante. Elles correspondent à des activités fronto-centrales.

Les résultats sont comparés aux caractéristiques spatiales et temporelles désignées dans la littérature comme pertinentes pour les potentiels d'erreurs (voir par exemple (Dal Seno et al., 2009; Roger et al., 2010) et les références incluses). Les potentiels d'erreur sont des activités cérébrales principalement actives dans les régions fronto-centrales et il est admis que le pic négatif apparaît environ 100 ms après la réponse du participant (on parle de *négativité d'erreur*, Ne) et qu'il est souvent suivi d'une activité positive (la *positivé d'erreur*, Pe). Ceci est en accord avec les caractéristiques extraites par nos méthodes.

Ces caractéristiques discriminantes sont à présent évaluées en termes de performance de classification.

Classification dans l'espace des caractéristiques

La procédure appliquée est basée sur la méthode proposée dans la section 5.3.
Sur un ensemble d'apprentissage, les caractéristiques sont déterminées à partir de l'analyse discriminante matricielle descriptive. Les paramètres des modèles matriciels gaussiens $M^1 \in \mathbb{R}^{24\times64}$, $M^2 \in \mathbb{R}^{24\times64}$, $\Sigma_L \in \mathbb{R}^{24\times24}$ et $\Sigma_R \in \mathbb{R}^{64\times64}$ sont estimés et les vecteurs propres $\tilde{u}_q \in \mathbb{R}^{24}$ et $\tilde{v}_q \in \mathbb{R}^{64}$ déterminant les coefficients des combinaisons linéaires discriminantes des coefficients multi-échelle et des électrodes sont calculés.
Sur l'ensemble de test, les matrices de caractéristiques discriminantes sont obtenues par projection des données tests sur les vecteurs $\tilde{u}_q$ et $\tilde{v}_q$ conjointement (5.35) et elles sont labellisées *Erreur* ou *Correct* en fonction du résultat de la fonction discriminante (5.41) dans chaque classe.
Un troisième jeu de validation est également utilisé pour la détermination de la valeur optimale α_{opt} du paramètre de régularisation dans le cas de l'analyse discriminante matricielle pénalisée sur l'espace des électrodes $\mathbb{R}^{64}$. La pénalisation appliquée est présentée dans la section (6.1.1).

La figure 6.5 représente l'évolution du taux de classification de la LDA matricielle et de la LDA matricielle pénalisée pour le participant A en fonction du nombre r de vecteurs propres $\tilde{u}_r \in \mathbb{R}^{24}$ et $\tilde{v}_r \in \mathbb{R}^{64}$ sélectionnées. Les résultats de classification sont donnés pour les essais *Erreur* et *Correct* distinctement. Nous rappelons en effet que nous sommes dans une situation de classes déséquilibrées et que dans ce cas l'analyse des résultats doit être faite classe par classe (voir section 2.4.1).
Pour la détection de la classe rare *Erreur*, les performances des deux méthodes se dégradent à mesure que le nombre de vecteurs propres sélectionnés augmente. Pour $r = 1$, les performances de la LDA matricielle et de la LDA matricielle pénalisée sont équivalentes et près de 80% des

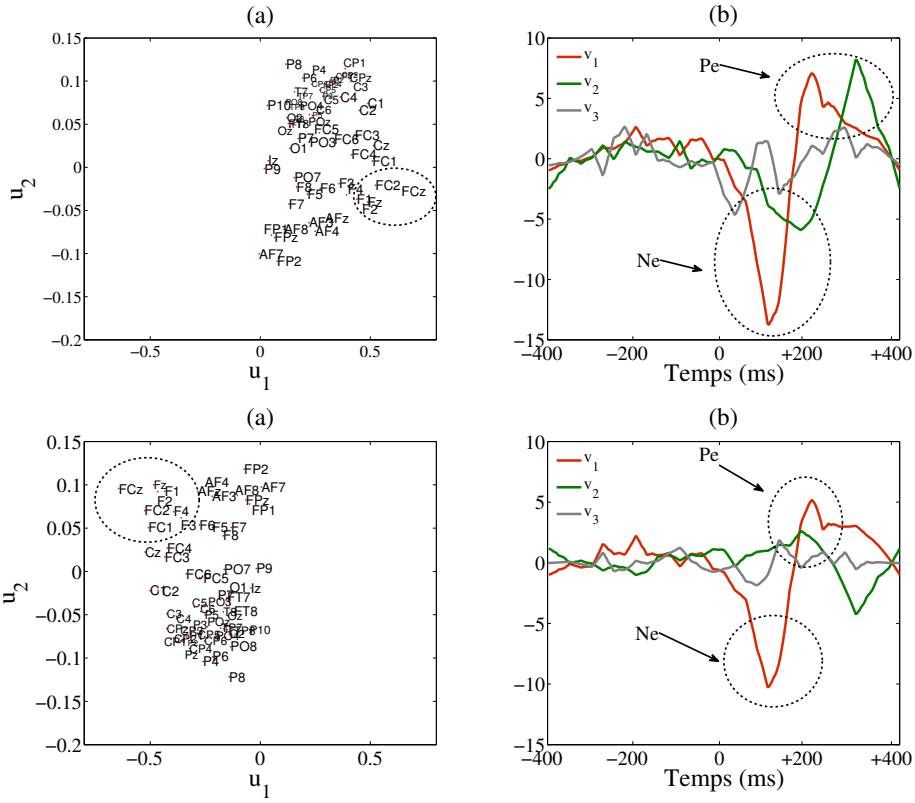


FIGURE 6.4 – Participant A. caractéristiques discriminantes spatiales (a) et temporelles (b) de la différence des moyennes des deux classes, obtenues par la LDA matricielle (graphiques du haut) et par la LDA matricielle pénalisée (graphiques du bas). Les composantes identifiées comme similaires à la négativité et à la positivité d'erreur sont respectivement désignées par "Ne" et "Pe". Sur les graphiques de gauche les électrodes fronto-centrales sont encerclées.

essais *Erreur* sont correctement classés. Entre $r = 1$ et $r = 3$, les performances de l'approche pénalisée se dégradent plus rapidement que celles de la LDA matricielle. Puis entre $r = 3$ et $r = 10$ la méthode pénalisée regagne en performance pour finalement présenter des résultats équivalents à la version non-pénalisée lorsque qu'aucune sélection n'est faite. Les erreurs sont alors détectées correctement dans un peu moins de 50% des cas. Pour la classe majoritaire des réponses correctes, les performances de classification sont bien meilleures (autour de 90%) et moins variables (des écarts-types beaucoup plus faibles) que pour les erreurs. De plus, les résultats sont influencés par le nombre de caractéristiques sélectionnées, mais en bien moindre mesure.

Nous allons à présent comparer les deux méthodes proposées aux méthodes de classification dérivées de la LDA vectorielle. Au vu des résultats présentés dans la figure 6.5, nous allons étudier deux cas : le cas où l'on ne fait aucune sélection pour les méthodes matricielles (et qui semble défavorable) et le cas où l'on ne conserve que les premiers vecteurs propres $\tilde{u}_1$ et $\tilde{v}_1$.

Sans sélection Dans le cas où aucune sélection d'un sous-espace vectoriel n'a été faite pour la LDA matricielle et sa version pénalisée on a donc $\tilde{U}_r = \tilde{U} \in \mathbb{R}^{24\times 24}$ et $\tilde{V}r = \tilde{V} \in \mathbb{R}^{64\times 24}$.

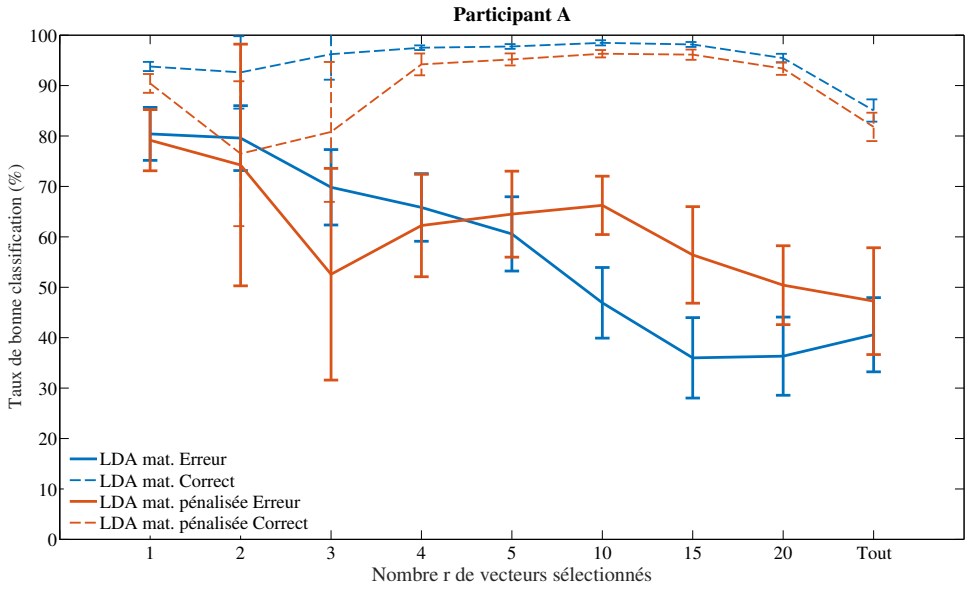


FIGURE 6.5 – Evolution du taux de bonne classification en fonction du nombre r de caractéristiques sélectionnées. Les courbes correspondent au taux moyen de bonne classification calculé sur 30 itérations. Les barres verticales correspondent à l'écart-type. Les résultats sont donnés pour les classes *Erreur* et *Correct* distinctement et l'approche matricielle est comparée à sa version pénalisée.

Les résultats pour les six sujets disposant de suffisamment d'essais *Erreur* (voir table 2.2) sont donnés dans la table 6.1.

Dans ce cas et pour l'ensemble des participants, la LDA diagonale est la méthode qui présente en moyenne la meilleure performance de classification pour la classe *Erreur*. Pour les participants A, G et H, plus de 70% des essais de la classe rare sont correctement classés avec cette méthode, contre moins de la moitié avec les méthodes de LDA matricielle. La LDA utilisant la pseudo-inverse présente quant à elle des résultats en moyenne légèrement meilleurs (autour de 50% pour tous les participants). Enfin notons que les écarts-types sont très larges pour les résultats des *Erreur* pour toutes les méthodes et tous les participants et en particulier pour le participant G. Ceci indique une importante variabilité dans le comportement des classifieurs et ne permet donc pas de conclure de façon catégorique quant à la comparaison des différentes méthodes.
En revanche, ceci n'est pas le cas pour la classification des réponses correctes. Pour tous les participants et toutes le méthodes, les taux moyens de classification sont beaucoup plus importants que pour ceux de la classe *Erreur*, et notamment pour la LDA matricielle et la LDA utilisant la pseudo-inverse dont les résultats de classification varient autour de 90%. Le taux de classification de la LDA matricielle pénalisée et de la LDA diagonale varient autour de 80%. Ceci s'explique en particulier par le nombre très important d'essais *Correct* considéré en phase d'apprentissage (dans cette expérience on en considère 600) et qui permet d'estimer de façon très précise les matrices de covariance qui ici représentent bien mieux les réponses correctes que les erreurs (voir section 3.3.6).

Participants		Méthodes			
		LDA matricielle	LDA mat. pénalisée	LDA (pinv)	LDA (diag)
A	*Err*	40,6 (7,4)	47,2 (10,6)	60,7 (7,5)	80,7 (6,1)
	Corr	85,1 (2,2)	81,8 (2,8)	95,9 (0,9)	89,2 (1,2)
B	*Err*	38,9 (11,9)	44,4 (12,8)	49,8 (10,3)	56,9 (14,1)
	Corr	80,7 (2)	79,6 (1,2)	93,6 (1,1)	74,7 (2)
G	*Err*	46,7 (22,5)	43,3 (25,4)	58,3 (27,3)	60,8 (26,8)
	Corr	82,5 (2,4)	81,8 (2,6)	96,3 (27,3)	82,1 (26,8)
H	*Err*	37,7 (7,6)	44,5 (10,5)	49,1 (7,3)	64,5 (6)
	Corr	81,6 (3)	78,8 (3,2)	94,9 (1)	80,5 (2)
I	*Err*	46,3 (14)	47 (14,2)	55,6 (17,1)	75,7 (11)
	Corr	82,1 (3)	81,1 (3)	95,9 (0,9)	86,1 (1,5)
J	*Err*	37,8 (28,7)	−−	54,4 (25,5)	70 (33,2)
	Corr	81,5 (2,9)	−−	95,1 (1,2)	82,5 (2,2)

TABLE 6.1 – Résultats de classification sans sélection. Pour chaque participant les taux de bonnes classification moyens ainsi que les écarts-types (entre parenthèses) calculés sur 30 itérations sont donnés dans chaque classe. Pour le participant J la méthode pénalisée n'a pas été calculée par manque d'essais *Erreur* pour un jeu de validation.

Sélection de $\tilde{u}_1$ et $\tilde{v}_1$ Dans la table 6.2, les performances sont à présent comparées dans le cas où les méthodes matricielles sont testées à partir des données projetées sur les premiers vecteurs propres $\tilde{u}_1$ et $\tilde{v}_1$ uniquement.

La même tendance que celle observée pour le participant A et représentée dans la figure 6.5 est observée pour tous les participants. Lorsque les méthodes matricielles de classification sont appliquées sur les premiers vecteurs propres $\tilde{u}_1 \in \mathbb{R}^{24}$ et $\tilde{v}_1 \in \mathbb{R}^{64}$, la détection des essais *Erreur* est très significativement meilleure que dans le cas sans sélection. Pour tous les participants, les résultats sont compris entre 70% et 90% des erreurs correctement classées pour les deux méthodes matricielles.

Les résultats pour la classe *Correct* restent quant à eux de très bonne qualité pour toutes les méthodes et tous les participants. Pour la LDA matricielle et sa version pénalisée, ils sont du même ordre que ceux présentés dans la table 6.1.

Discussion

Les résultats obtenus sur les données de potentiels d'erreur permettent clairement de valider l'approche matricielle proposée, tant au niveau des performances de classification que pour ce qui concerne la pertinence des caractéristiques discriminantes extraites. En particulier, nous avons montré que la projection des données sur le premier vecteur propre $\tilde{u}_1 \otimes \tilde{v}_1$ permet d'augmenter de façon très substantielle la détection des essais de la classe *Erreur*. De plus, une analyse de ses caractéristiques spatiales et temporelles permet de montrer qu'elles coïncident avec des comportements reconnus dans la littérature comme caractéristiques des potentiels d'erreur. Afin de compléter cette étude il serait intéressant d'étudier l'évolution des performances de classification en fonction du nombre de vecteurs propres sélectionnés, comme cela a été fait pour le participant A (voir figure 6.5).

Participants		Méthodes			
		LDA matricielle	LDA mat. pénalisée	LDA (pinv)	LDA (diag)
A	*Err*	80,4 (5,3)	79,2 (6,1)	57,7 (7,1)	80,3 (6,7)
	Corr	93,8 (0,9)	90,4 (1,9)	86,4 (0,8)	89,5 (1,3)
B	*Err*	71,6 (1,8)	60,2 (17,6)	52,4 (12,8)	61,8 (16,3)
	Corr	86 (1,7)	79,4 (6,9)	94 (1,2)	75,3 (1,9)
G	*Err*	90 (18,1)	83,3 (17,8)	64,2 (23,3)	74,2 (24,1)
	Corr	94,6 (0,8)	88,9 (5,9)	96,1 (0,8)	81,6 (1,7)
H	*Err*	74,9 (1,9)	71,6 (8)	47,8 (7,3)	64,2 (6)
	Corr	92,7 (1,5)	84,8 (5,9)	94,8 (1,2)	81 (2,2)
I	*Err*	78,7 (16,6)	84,7 (11,4)	59 (15)	75,3 (15,5)
	Corr	92,6 (1,2)	88,9 (2,9)	95,6 (1,1)	87,1 (1,2)
J	*Err*	75,5 (21,3)	−−	54,4 (25,5)	66,7 (35)
	Corr	95,2 (0,8)	−−	95 (1,1)	82,5 (2,2)

TABLE 6.2 – Résultats de classification avec sélection des premières caractéristiques les plus discriminantes $\tilde{u}_1$ et $\tilde{v}_1$.

6.2.2 Caractéristiques discriminantes pour le P300 Speller

Illustrons à présent les apports de la méthode proposée sur le deuxième jeu de données étudié dans cette thèse. Nous nous intéressons donc ici à l'extraction de caractéristiques discriminantes pour l'analyse et la classification de signaux acquis au cours d'un protocole de P300 Speller. Les données étudiées ont été acquises à partir de 32 électrodes et une transformation en ondelettes discrète a été appliquée, en utilisant un filtre de Daubechies D8. Après sélection des coefficients multi-échelle significatifs (voir section 2.4.2), nous sélectionnons K coefficients (dans cette application, K varie de 24 à 29 selon les participants). Nous considérons donc la matrice de covariance des coefficients $\Sigma_L \in \mathbb{R}^{K\times K}$ et la matrice de covariance des électrodes $\Sigma_R \in \mathbb{R}^{32\times 32}$. Les résultats présentés dans cette section correspondent à ceux du sujet 4.

Projection de données dans l'espace spatio-temporel.

La figure 6.6(a) représente l'éboulis des 28 valeurs propres non nulles. Dans un objectif de réduction de dimension, l'identification d'un "coude" suggère de choisir le sous-espace de projection E_r de dimension $r = 4$. La figure 6.6 (b) représente la projection des observations sur le premier plan factoriel engendré par $\tilde{u}_1 \otimes \tilde{v}_1$ et $\tilde{u}_2 \otimes \tilde{v}_2$ ainsi que les deux moyennes des cibles et des non-cibles. Cette représentation permet de visualiser la séparation des deux classes dans ce sous-espace et de distinguer les observations atypiques.

Analyse des caractéristiques discriminantes.

La figure 6.7(b) représente les trois premières caractéristiques temporelles discriminantes. Elles ont été obtenues par synthèse d'ondelettes à partir des projections des différences des coefficients moyens entre les classes cible et non-cible sur les axes engendrés par $\tilde{v}_1$, $\tilde{v}_2$ et $\tilde{v}_3$. Les deux principales caractéristiques associées respectivement à v_1 et v_2, deux vecteurs $(S_W^R)^{-1}$-orthogonaux, apparaissent similaires à l'onde N1 et à l'onde P300. Ceci semble donc indiquer une forme de découplage de ces deux caractéristiques. La figure 6.7(a) représente la projection

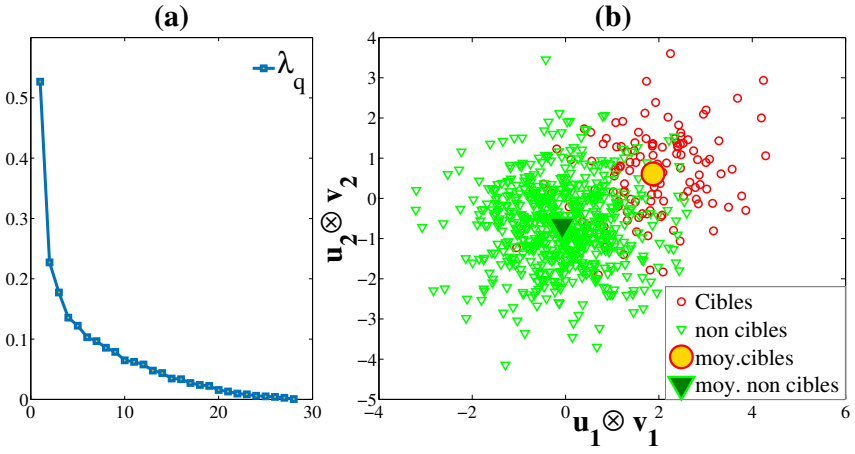


FIGURE 6.6 – Sujet 3. Valeurs propres (a) et représentation des observations et des moyennes dans le premier plan spatio-temporel (b).

dans le plan $(\tilde{u}_1, \tilde{u}_2)$ des différences des moyennes des classes mesurées sur chaque électrode. A partir de cette représentation des groupes distincts d'électrodes peuvent être associés à chacune des caractéristiques précédentes : les électrodes de la zone occipitale (Oz, O1 et O2) sont liées à la N1 et les électrodes centrales-pariétales (CP1, CP2 principalement) sont quant à elles souvent associées à l'onde P300. De plus, on peut détecter des comportements atypiques, comme celui de l'électrode T8. Le comportement singulier de cette électrode située sur la tempe droite est connu (artefacts) et il n'est pas rare de devoir l'exclure des analyses.

Classification pour la détection de lettres dans le protocole P300 Speller

Dans cette application, la procédure d'évaluation des classifieurs est la suivante : pour chaque participant, la phase d'apprentissage est constituée de 20 lettres et les différentes méthodes sont testées sur 40 lettres. Ainsi, en rappelant qu'une lettre correspond à 36 flashs dont 6 flashs *cibles*, le jeu d'apprentissage est constitué de 600 essais *non-cibles* et de 120 essais *cibles*. Une méthode de validation croisée a été mise en œuvre afin de tirer de façon aléatoire 30 jeux apprentissage/test.

Contrairement à l'application sur les potentiels d'erreur, un filtrage spatial par l'approche classique de la LDA matricielle (voir section 6.1.1) a été effectué pour l'utilisation des méthodes comparatives : LDA, LDA diagonale et LDA régularisée, qui, sans cette étape, fournissaient des résultats de classification très faibles. Dans ce filtrage spatial, le nombre r de filtres conservés est choisi de façon à conserver 75% de l'inertie. Nous comparons ces méthodes à notre méthode et étudions en particulier l'évolution de taux de lettres correctement classées en fonction du nombre de vecteurs propres $\tilde{u}_r \in \mathbb{R}^K$ et $\tilde{v}_r \in \mathbb{R}^{32}$ conservés pour l'extraction de caractéristiques. Notons que dans cet exemple nous nous somme concentrés sur l'approche descriptive de l'analyse discriminante sans la version pénalisée. Cette dernière est en effet assez coûteuse à mettre en œuvre du fait du choix du paramètre de pénalisation et elle est donc difficilement applicable dans le contexte des ICM.

Les résultats pour quatre sujets de l'expérience sont représentés dans la figure 6.8. Notons

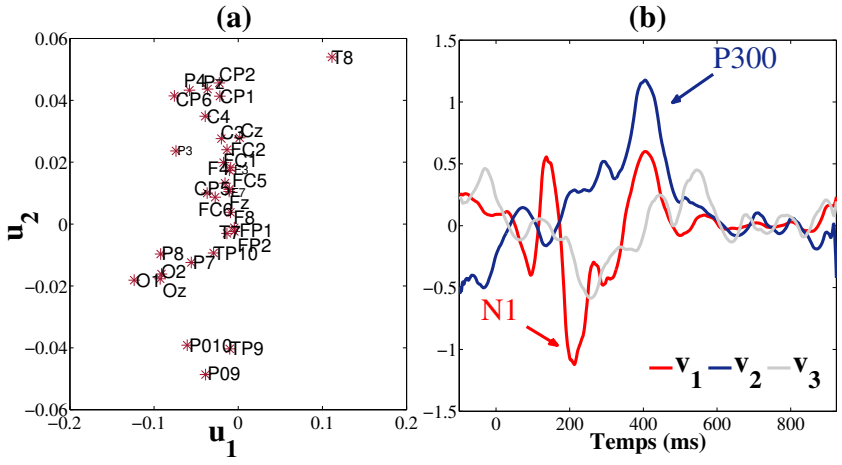


FIGURE 6.7 – Sujet 3. caractéristiques discriminantes spatiales (a) et temporelles (b) de la différence des moyennes des 2 classes.

tout d'abord la grande diversité de performances entre les sujets. Pour le sujet 2 par exemple, aucune méthode étudiée ne fournit un taux de lettres correctement classées supérieur à 60%, alors que pour le sujet 3 le taux avoisine les 100% pour tous les classifieurs. Dans ces deux cas, les résultats de la LDA matricielle ne diffèrent pas de façon significative selon le nombre de vecteur propre conservés.
Pour les sujets 4 et 7, deux comportements se distinguent : pour le sujet 4 notre approche fournit des résultats supérieurs à ceux obtenus par LDA, LDA diagonale et LDA régularisée et ce d'autant plus que le nombre de vecteurs diminue. En particulier, pour $r = 2$, le taux de lettres bien classées se situe autour de 80%, contre 70% lorsque 15 vecteurs sont conservés. Pour le sujet 7 au contraire, les performances de notre méthodes augmentent à mesure que le nombre de vecteurs conservés augmente jusqu'à atteindre des performances équivalentes aux autres méthodes lorsque tous les vecteurs propres sont conservés pour l'analyse (autour de 90%).

Ces résultats illustrent bien la performance de notre méthode appliquée à des données issues d'un protocole P300 Speller. Elle fournit des résultats meilleurs ou équivalent à des méthodes qui ont montré leur efficacité dans ce contexte. Les résultats obtenus dans les mêmes conditions pour 4 autres sujets sont présentés dans l'annexe C.2.

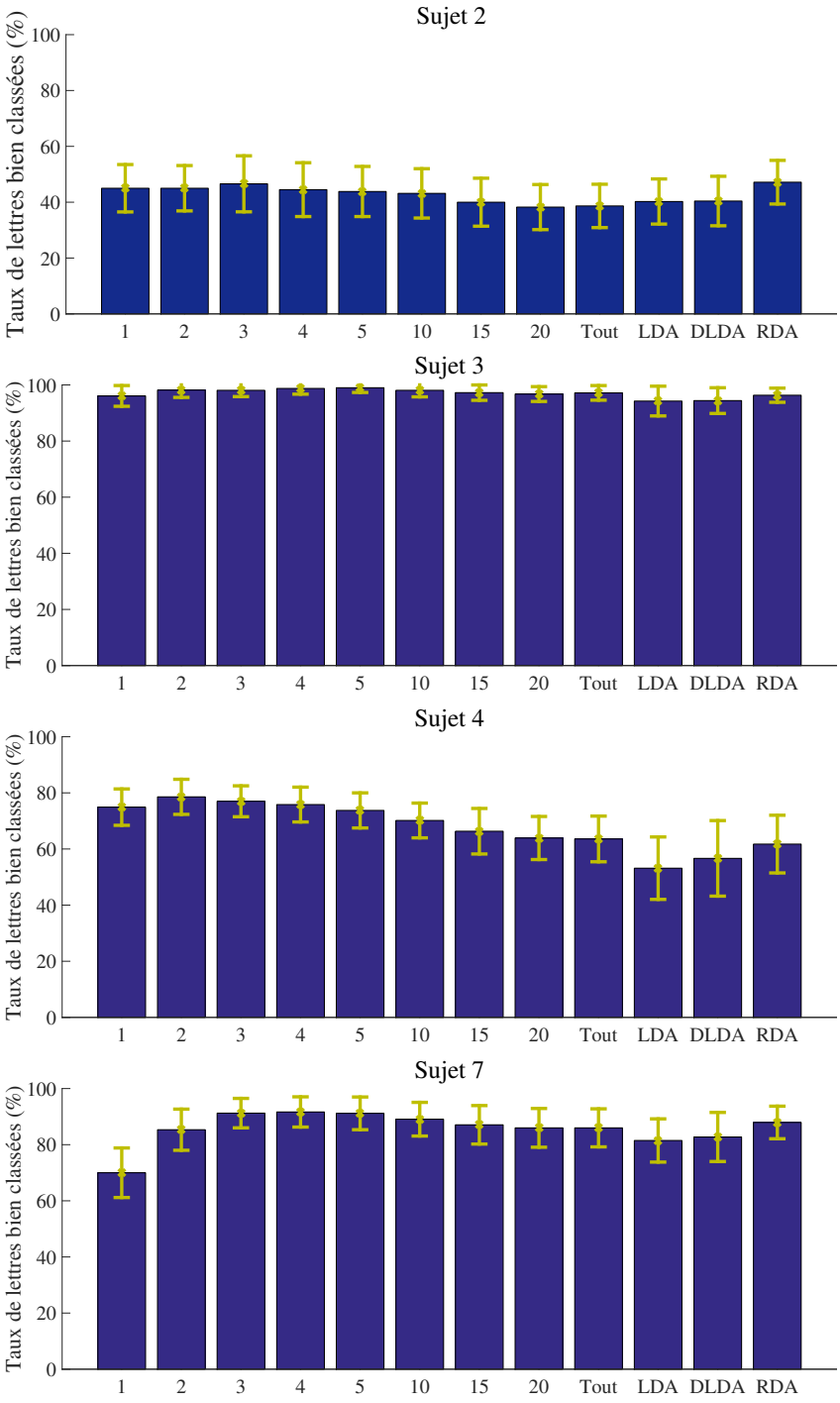


FIGURE 6.8 – Taux de lettres correctement classées. Chaque barre correspond au taux moyen calculé sur 30 itérations et les axes verticaux représentent l'écart-type. Les résultats de la LDA matricielle sont présentés en fonction du nombre r de caractéristiques discriminantes sélectionnées $\tilde{u}_r$ et $\tilde{v}_r$ (9 premières barres) et comparés à la LDA, la LDA diagonale (DLDA) et à la LDA régularisée (RDA).

Conclusion Partielle

Dans cette partie nous nous sommes intéressés à la structure matricielle des signaux EEG (espace × temps) et à la façon dont celle-ci peut être exploitée dans le contexte de l'analyse discriminante. Chaque signal EEG vectorisé est modélisé à partir d'une distribution gaussienne dont la matrice de covariance est factorisée sous la forme d'un produit de Kronecker entre la matrice de covariance spatiale et la matrice de covariance temporelle.
Les principaux avantages de cette modélisation de la matrice de covariance sont les suivants. Elle permet notamment

- de réduire de façon significative le nombre de paramètres à estimer,
- d'établir un lien avec l'analyse discriminante matricielle, la loi gaussienne matricielle étant définie comme la loi gaussienne de la matrice vectorisée,
- dans le cas de signaux EEG, d'estimer les sources de variabilité spatiales et temporelles et d'affiner l'analyse de ce type de données.
- d'introduire de façon naturelle des a priori séparément sur les dimensions spatiale et temporelle.

Les résultats obtenus sur les jeux de données réels ont démontré la pertinence de l'hypothèse de séparabilité tant d'un point de vue descriptif qu'en termes de classification binaire.

En effet, pour les données de potentiels d'erreur, l'étude des caractéristiques spatiales et temporelles discriminantes a permis de retrouver des composantes définies dans la littérature comme jouant un rôle prépondérant dans la différence entre les réponses correctes et les erreurs. La projection des données sur des sous-espaces de dimension réduite avant classification permet d'améliorer de façon spectaculaire la détection de la classe minoritaire (ici les erreurs). Il est à noter qu'en l'absence de cette projection, le taux d'erreur de classification est élevé dans la classe minoritaire, laissant penser que la variabilité de cette classe diffère de celle de la classe majoritaire, contrairement à l'hypothèse faite.

Pour les données du protocole P300 Speller, il faut noter que les deux premières composantes discriminantes obtenues semblent correspondre à l'onde P300 et l'onde N1, toutes les deux reconnues comme importantes dans l'étude de ce type de signaux. La sélection d'un sous-ensemble de caractéristiques pour la classification permet dans ce cas aussi d'améliorer les performances en détection, même si le résultat est moins marqué que dans le cas des potentiels d'erreur.

Le fait que de meilleurs résultats soient obtenus en limitant la classification à un sous-ensemble de caractéristiques spatio-temporelles semble indiquer que le modèle (mélange gaussien avec hypothèse de séparabilité et matrices de covariance égales dans les deux classes), n'est pas complètement en adéquation avec les données. En effet, si tel était le cas, les meilleurs résultats devraient être obtenus sans sélection, ce cas correspondant à la règle de Bayes optimale. Par exemple, la structure de Kronecker pourrait ne donner qu'une approximation, acceptable mais imparfaite de la "vraie" matrice de covariance.

Notons enfin que dans le modèle de covariance considéré ici, nous modélisons la variabilité globale du signal, sans distinguer l'activité de fond et le signal d'intérêt. Nous développons dans ce qui suit une approche dans laquelle ces deux composantes sont modélisées séparément, et permettant aux deux classes d'avoir des variabilités différentes.

Troisième partie

Un modèle linéaire mixte pour l'analyse et la classification en EEG

Chapitre 7

Modélisation de la variabilité des signaux EEG

Sommaire

7.1 Positionnement du problème

7.1.1 Présentation du modèle linéaire mixte gaussien

Les modèles linéaires mixtes gaussiens sont des extensions des modèles linéaires classiques dans lesquels sont considérés des facteurs à effets fixes (qui interviennent au niveau de la moyenne) et des facteurs à effets aléatoires (qui vont intervenir au niveau de la variance) (Rao and Kleffe, 1988; McCulloch et al., 2008). Un facteur est dit à effets mixtes lorsqu'il influe sur la moyenne et la variance. Ces modèles sont couramment utilisés pour modéliser des données que l'on sait influencées par des phénomènes aléatoires (supposés) connus. Les effets aléatoires introduits dans le modèle seront donc caractérisés par une matrice de covariance qu'il faudra estimer en plus de celle de l'erreur (qui correspond quant à elle à la variabilité non expliquée par le modèle). Nous parlerons de *composantes de la variance*.

Par souci de simplicité nous présentons tout d'abord ce modèle dans le cas standard où la variable à expliquer est une variable aléatoire réelle. Soit $y = (y_1, y_2, \ldots, y_N)' \in \mathbb{R}^N$, le vecteur constitué de N observations (qui ne sont pas forcément indépendantes) d'une variable aléatoire

réelle à expliquer. Un modèle linéaire mixte gaussien intégrant p effets fixes et q effets aléatoires s'écrit :

$$y = X\beta + Zu + \varepsilon\,, \tag{7.1}$$

où

1) la partie des effets fixes $X\beta$ où $X \in \mathbb{R}^{N\times p}$ est la matrice *design* (supposée connue) associée au paramètre des effets fixes $\beta \in \mathbb{R}^p$ inconnu ; les colonnes de X correspondent aux variables explicatives de la moyenne.
2) la partie des effets aléatoires Zu où $Z \in \mathbb{R}^{N\times q}$ est la matrice *design* (supposée connue) associée au vecteur des effets aléatoires $u \in \mathbb{R}^q$. On suppose que u est un vecteur gaussien centré de matrice de covariance $\Lambda \in \mathbb{R}^{q\times q}$ inconnue.
3) le vecteur aléatoire des erreurs du modèle $\varepsilon \sim \mathcal{N}(0, Q)$, où $Q \in \mathbb{R}^{N\times N}$ est inconnue. Il est supposé indépendant du vecteur des effets aléatoires u.

Selon ce modèle, y est donc un vecteur gaussien de moyenne $E(y) = X\beta$ et de matrice de covariance $\Sigma = Z\Lambda Z' + Q$.

Les paramètres à estimer sont donc β, Λ et Q. Le vecteur aléatoire u des effets aléatoires non directement observé sera prédit à partir des observations y_i à l'aide de la loi jointe de (y, u).

A partir d'un tel modèle, les démarches usuelles d'estimation et de prédiction en régression sont modifiées par l'introduction de composantes de la variance supplémentaires. Les méthodes usuelles d'estimation des paramètres fixes et des composantes de la variance ainsi que la prédiction des effets aléatoires sont présentées dans l'annexe B.

Remarque 19. *Dans ce travail nous nous concentrons sur l'approche linéaire classique du modèle mixte. Cependant, son extension au cas du modèle linéaire généralisé (McCulloch et al., 2008) et au modèle non linéaire (Davidian and Giltinan, 1995) illustre bien l'étendue de son champ d'application.*

Le modèle (7.1) peut être facilement étendu au cas multidimensionnel où la variable à expliquer est un vecteur aléatoire comme nous allons le voir dans ce qui suit, étant donné que nous nous souhaitons modéliser des données matricielles vectorisées. En effet, nous nous intéressons à la construction d'un tel modèle pour des signaux EEG avec pour objectif la classification supervisée dans le cadre de classes déséquilibrées et de petits échantillons.

7.1.2 Le modèle linéaire mixte en électrophysiologie

Les modèles linéaires mixtes offrent un cadre de travail riche et flexible pour décrire et quantifier différentes sources de variabilité. Les données expérimentales, comme celles étudiées dans cette thèse, ont souvent une structure hiérarchique et il est intéressant de pouvoir intégrer dans le modèle la variabilité des différents niveaux. Pour ces raisons, l'utilisation de tels modèles pour des applications en neurosciences a été considérée dans des contextes très divers. Citons par exemple (Baayen et al., 2008) qui fournit une introduction au modèle mixte et développe les applications possibles en psychologie et en neuroimagerie.

Le modèle mixte a également été utilisé dans un cadre bayésien pour des applications en électrophysiologie. Dans (Davidson, 2009), un modèle mixte fonctionnel est introduit pour l'analyse de potentiels évoqués. Ce modèle inclut une transformation en ondelettes discrète ainsi qu'une contrainte de parcimonie introduite par une distribution *a priori*. Dans (Fazli et al., 2011), la variabilité inter-sujets est modélisée pour une application aux ICM. Un classifieur multi-sujet est

proposé en utilisant un modèle linéaire mixte avec une hypothèse de parcimonie introduite par une contrainte de type ℓ_1. La variabilité intra et inter sujet est modélisée au travers des effets fixes et aléatoires et la méthode permet des transferts d'informations entre les sujets. Enfin, nous nous référons aux travaux de Huang *et al.* (Huang et al., 2008; Huang et al., 2009) qui proposent un modèle pour chaque sujet d'une expérience et qui tient compte de la variabilité entre les répétitions d'une même condition expérimentale (les essais). L'application principale présentée par les auteurs est la classification de potentiels évoqués dans le cadre des ICM. Ces travaux ont servi de point de départ pour la méthode proposée dans ce chapitre. Cependant, celle-ci se distingue de celle de Huang *et al.* en trois points : i) nous proposons une modélisation des signaux EEG dans l'espace des ondelettes, ce qui permet d'une part de réduire la dimension temporelle de façon significative en concentrant l'information pertinente dans un faible nombre de coefficients multi-échelle, et d'autre part les propriétés de décorrélation des ondelettes introduisent une structure diagonale dominante de la matrice de covariance ; ii) le modèle proposé est formalisé comme un modèle d'analyse de variance, ce qui fournit un cadre théorique simple et solide ; iii) nous nous plaçons dans le cas binaire déséquilibré où l'une des deux classes est de taille très faible ; cette situation étant assez courante en EEG.

7.1.3 Modéliser la variabilité inter-essais

Comme déjà discuté dans la section 2.1.3, l'étude des signaux EEG essai par essai est au cœur des problématiques d'analyse et de classification. Afin de réduire l'impact du bruit et de l'activité de fond, des stratégies communes d'extraction de composantes plus élémentaires contenant le signal d'intérêt sont largement utilisées (voir section 2.3.1). Ces méthodes cherchent à supprimer la variabilité induite par l'activité de fond des signaux en espérant que la variabilité restante soit attribuable au signal d'intérêt.

Des approches alternatives se basent sur une modélisation explicite du signal. Par exemple, l'analyse discriminante linéaire peut s'interpréter en termes d'un modèle très simple de mélange gaussien (voir section 3.2.2). Dans des travaux récents, Blankertz *et al.* (Blankertz et al., 2011) proposent une version régularisée de la LDA pour l'étude et la classification de potentiels évoqués. Dans le modèle proposé, chaque essai unique s'écrit comme la somme du signal d'intérêt, qui est approché comme une constante à travers tous les essais (la moyenne), et d'une activité de fond aléatoire modélisée comme un bruit gaussien. Bien que l'algorithme de détection déduit soit très efficace pour des tâches de classification binaire, il ne tient pas compte de la variabilité inter-essais qui est noyée dans la matrice de covariance du bruit de fond. Comme le bruit de fond est supposé indépendant du type d'activité cérébrale enregistrée un tel modèle suppose l'égalité des matrices de covariance dans chaque classe. Ainsi, les paramètres du modèle à estimer sont les moyennes dans chaque classe et une matrice de covariance de très grande dimension. Cette grande dimension par rapport au nombre d'observations engendre d'importants problèmes d'estimation. Pour remédier à cela, les auteurs introduisent une version régularisée de l'estimation de la matrice de covariance comme décrite dans la section 3.3.5.

Qu'en est-il si l'on souhaite modéliser la variabilité inter-essais ? Dans le cas d'un modèle linéaire comme celui présenté par Blankertz *et al.*, il est nécessaire d'ajouter un terme propre à l'essai. L'approche proposée dans la suite se base sur une modélisation de la variabilité des essais dans chaque classe et l'algorithme de classification obtenu correspond cette fois-ci à une analyse discriminante quadratique puisque l'on suppose que les matrices de covariance sont différentes selon la classe du signal EEG étudiée. Dans ce cas, il est important de noter que modéliser

directement les séries temporelles multi-électrodes peut s'avérer assez difficile pour deux raisons principales. Premièrement, en raison de l'importante corrélation spatiale et temporelle. Ainsi, la matrice de covariance de l'activité de fond sera de très grandes dimensions, ce qui la rend difficile à estimer et à traiter numériquement. Deuxièmement, les grandes dimensions nécessitent un nombre très important d'observations (supérieurs au nombre de paramètres à estimer). Pour ces raisons, des transformations préliminaires sont nécessaires afin d'obtenir des données de dimension plus faible qui peuvent être par la suite facilement modélisées par un modèle linéaire mixte dont la matrice de covariance de l'activité de fond sera diagonale dominante.

7.2 Modélisation

7.2.1 Le modèle

Dans cette section nous introduisons une analyse essai par essai des signaux EEG dans laquelle la variabilité est modélisée au travers d'un modèle linéaire mixte gaussien.

Les signaux sont enregistrés à partir de M électrodes sur une période de temps fixée de dimension T. Pour un participant donné, nous considérons n_c essais dans chaque classe c. Chaque essai prend la forme de M séries temporelles chacune constituée de T échantillons temporels.

Un point sur les notations. Soit $s^c_{i,j}(t)$ le signal de l'essai i enregistré à l'électrode j et au temps t dans la classe c. Dans le cas de M électrodes et de T pas de temps, chaque essai prend la forme d'une matrice $S^c_i \in \mathbb{R}^{T\times M}$ de terme général $s^c_{i,j}(t)$, où $j = 1,\dots,M$, $t = 1,\dots,T$ et $c \in \{1;2\}$. Nous considérons dans la suite des essais spatio-temporels obtenus par concaténation des colonnes de S^c_i. On note $s^c_i = \mathsf{vec}((S^c_i)') \in \mathbb{R}^{TM}$, où l'opérateur de vectorisation est défini en annexes (voir définition 12 de l'annexe A).

Hypothèse d'additivité. Nous supposons que de tels signaux sont modélisés comme la somme de deux composantes indépendantes :

- *Une activité de fond*, supposée stationnaire (au sens des processus aléatoires faiblement stationnaires, à savoir les deux premiers moments sont invariants par translation) et corrélée spatialement et temporellement (voir par exemple (Koopmans, 1995)),
- *Une composante associée à l'évènement* (aussi appelée signal d'intérêt), intrinsèquement non-stationnaire et caractérisée par un indice de classe qui définit la réponse cérébrale aux évènements étudiés.
  Ce signal d'intérêt est en outre modélisé comme la somme d'une composante fixe, commune à tous les essais d'une même classe, et une composante aléatoire qui dépend de l'essai (nommée par la suite *l'effet aléatoire de l'essai*).

Le modèle statistique

Pour un participant donné, nous considérons n_c essais par classe, avec $c \in \{1;2\}$ qui prennent la forme de vecteurs $s^c_i \in \mathbb{R}^{TM}$. Dans la classe c le modèle vectoriel s'écrit :

$$s^c_i = \mu^c + \Gamma^c b^c_i + \varepsilon_i \,. \tag{7.2}$$

- $\mu^c \in \mathbb{R}^{TM}$ est le vecteur moyen de la classe c,

- $b_i^c \sim \mathcal{N}(0, \tau^2)$ est une variable aléatoire gaussienne, centrée et de variance τ^2. Elle modélise l'effet aléatoire de l'essai i,
- $\Gamma^c \in \mathbb{R}^{TM}$ est un vecteur de coefficients dépendant de la classe qui module l'impact de l'essai sur chaque électrode et chaque pas de temps,
- $\varepsilon_i \sim \mathcal{N}(0, Q)$ est la partie résiduelle qui modélise l'activité de fond et où $Q \in \mathbb{R}^{TM \times TM}$ est une matrice semi-définie positive. De plus, nous supposons que le vecteur des résidus ε_i et l'effet aléatoire b_i^c sont indépendants.

Après avoir défini le vecteur de coefficients Γ^c, le modèle défini en (7.2) correspond à un modèle linéaire mixte gaussien tel que :

$$s_i^c | \Gamma^c \sim \mathcal{N}\left(\mu^c, V^c\right) \ , \qquad \text{où} \qquad V^c = \tau^2 \Gamma^c \left(\Gamma^c\right)' + Q \ . \tag{7.3}$$

Chaque essai se décompose en une partie fixe, correspondant à la moyenne de sa classe d'appartenance, un effet aléatoire propre à l'essai modulé par le vecteur Γ^c et un effet aléatoire modélisant la variabilité de l'activité de fond.

Remarque 20. 1) Puisque la partie fixe du modèle (7.2) correspond à la moyenne dans chaque classe, sachant Γ^c, le modèle correspond au modèle d'une analyse de variance (Searle et al., 1992) où la classe est le facteur de l'effet fixe et l'essai est le facteur de l'effet aléatoire. Ceci permet de considérer un modèle simple basé sur une théorie bien connue.

2) Dans le cas où la variabilité inter-essai est négligeable, la partie de l'effet aléatoire $\Gamma^c b_i^c$ est inutile et le modèle correspond alors à un simple modèle linéaire gaussien comme celui décrit dans (Blankertz et al., 2011).

Cependant, le modèle défini de la sorte est difficilement applicable directement sur des séries temporelles multi-électrodes et ceci pour deux raisons.
Premièrement, comme nous l'avons vu dans la section 2.2, il existe d'importantes corrélations entre électrodes et entre pas de temps. Dans ce cas, la covariance du résidu du modèle Q est une matrice pleine et de grande dimension qu'il est difficile d'estimer et de manipuler numériquement. Deuxièmement, les dimensions des données vectorisées (TM) rendent elles-aussi la modélisation difficile à mettre en œuvre et en particulier dans le cas de petits échantillons (lorsque $n < TM$).

C'est pourquoi des transformations préliminaires sont appliquées afin de réduire la dimension des données et de concentrer l'information que l'on souhaite modéliser par (7.2) dans un petit nombre de composantes pertinentes. Nous rappelons que dans cette thèse nous nous plaçons dans le cas de l'analyse de potentiels évoqués dont nous avons déjà montré qu'ils étaient bien représentés dans des bases d'ondelettes (voir section 2.2.2). Nous allons donc considérer par la suite des séries temporelles exprimées sous la forme de coefficients multi-échelles et sur lesquelles une méthode de filtrage spatial va être appliquée. Cette approche a été décrite dans le chapitre 5 et les filtres discriminants ont été formalisés dans la section 6.1.

Réduction de dimension et décorrélation

Transformation en ondelette. Nous rappelons ici que les bases d'ondelette sont définies par dilatation et translation d'une forme d'onde particulière notée ici ψ. Lorsque cette fonction est bien choisie (en fonction du signal considéré), la famille associée $\{\psi_\lambda\}_{\lambda \in \Lambda}$ forme alors une base orthonormale de l'espace des signaux considérés (Daubechies, 1992; Vetterli and Kovačevic,

1995; Mallat, 2008). Notons qu'ici $\lambda \in \Lambda$ est un double indice (temps-échelle) numérotant les coefficients multiéchelle (coefficients d'ondelette et coefficients d'échelle). La taille de Λ est égale à T, la dimension des séries temporelles (pour une application de la DWT sous Wavelab par exemple (Buckheit et al., 2005), T doit être une puissance de 2). Chaque signal $S^c_{i,j} \in \mathbb{R}^T$ peut alors s'écrire comme

$$S^c_{i,j} = \sum_{\lambda \in \Lambda} x^c_{i,j}(\lambda)\psi_\lambda \,, \tag{7.4}$$

et il est donc caractérisé par les coefficients multi-échelle $x^c_{i,j}(\lambda)$ déterminés par

$$x^c_{i,j}(\lambda) = \langle S^c_{i,j}, \psi_\lambda \rangle \,. \tag{7.5}$$

Une réduction de dimension est faite par la suppression de coefficients qui ne sont pas pertinents pour l'analyse, c'est-à-dire les coefficients $x^c_{i,j}$ qui sont numériquement négligeables sur toutes les électrodes $j = 1, \ldots, M$. De plus, la variabilité des coefficients dépend du niveau de décomposition, c'est pourquoi une normalisation niveau par niveau est appliquée sur toutes les électrodes afin de corriger ces différences. Notons K le nombre de coefficients conservés pour la modélisation. La matrice des coefficients multi-échelles de l'essai i conservés pour les M électrodes s'écrit $X^c_i \in \mathbb{R}^{K \times M}$, avec $K \ll T$.

Filtrage spatial Afin de réduire la dimension spatiale, nous utilisons ici une méthode présentée dans la section 6.1. A partir des matrices $X^c_i \in \mathbb{R}^{K \times M}$, $i = 1, \ldots, n_c$, les vecteurs propres $\tilde{v}_q \in \mathbb{R}^M$, $q = 1, \ldots, M$ sont déterminés en optimisant un critère de Fisher comme défini dans la section 6.1.1. Soit $\tilde{V}$ la matrice contenant les J premiers vecteurs propres $\tilde{v}_q$, les canaux de l'essai i sont définis par projection de $X^c_i(k) \in \mathbb{R}^J$ sur $\tilde{V}$ (voir définition 8) :

$$Y^c_i(k) = X^c_i(k)\tilde{V} \,. \tag{7.6}$$

Après avoir appliqué ces deux transformations et réduit les dimensions spatiales et temporelles, nous considérons une matrice $Y^c_i \in \mathbb{R}^{K \times J}$ où K est le nombre de coefficients multi-échelle et J est le nombre de canaux retenus. Notons que par ces transformations les dimensions spatiales et temporelles sont significativement réduites (i.e $K \ll T$ et $J \ll M$) mais aussi et surtout une décorrélation est introduite permettant de renforcer la structure diagonale dominante de la matrice de covariance Q.

Soit $y^c_i \in \mathbb{R}^{KJ}$ le vecteur des essais s^c_i transformés, défini par $y^c_i = \mathsf{vec}((Y^c_i)')$. Les transformations appliquées étant linéaires, alors, d'après (7.3), y^c_i est un vecteur gaussien. Par souci de simplicité, nous allons utiliser les mêmes notations que celles définies dans (7.2) et (7.3) pour désigner les paramètres de la loi et du modèle de y^c_i :

$$y^c_i = \mu^c + \Gamma^c b^c_i + \varepsilon_i \,. \tag{7.7}$$

- $\mu^c \in \mathbb{R}^{KJ}$ est le vecteur moyen de la classe c,
- $b^c_i \sim \mathcal{N}(0, \tau^2)$ est une variable aléatoire gaussienne, centrée et de variance τ^2. Elle modélise l'effet aléatoire de l'essai i,
- $\Gamma^c \in \mathbb{R}^{KJ}$ est un vecteur de coefficient dépendant de la classe qui module l'impact de l'essai sur chaque électrode et chaque pas de temps,

- $\varepsilon_i \sim \mathcal{N}(0, Q)$ est la partie résiduelle qui modélise l'activité de fond, où $Q = \sigma^2 \mathbf{I}_{KJ}$, avec $\mathbf{I}_{KJ}$ la matrice identité de dimension KJ. De plus, nous supposons que le vecteur des résidus ε_i et l'effet aléatoire b_i^c sont indépendants.

et

$$y_i^c | \Gamma^c \sim \mathcal{N}\left(\mu^c, V^c\right) \ , \qquad \text{où} \qquad V^c = \tau^2 \Gamma^c \left(\Gamma^c\right)' + \sigma^2 \mathbf{I}_{KJ} \ . \tag{7.8}$$

Remarque 21. *Les transformations introduites ont permis de réduire les dimensions spatiales et temporelles et de modéliser l'activité de fond comme un bruit blanc gaussien. Le modèle proposé en (7.7) permet d'obtenir des estimations robustes de ses paramètres à partir d'un faible nombre d'observations. En effet, après avoir spécifié Γ^c, il ne reste plus que 4 paramètres à estimer : 2 paramètres de dimension KJ, les deux vecteurs de moyenne μ^c, $c = \{1; 2\}$ et 2 paramètres réels, les variances τ^2 et σ^2.*

Rappelons que la simplicité du modèle proposé est motivée par son utilisation dans les cas où peu d'essais sont disponibles pour l'estimation des paramètres. C'est le cas dans certaines conditions expérimentales dans lesquelles le nombre d'essais d'intérêt est faible, soit du fait du protocole expérimental ou bien du fait de comportements rares des sujets (par exemple les erreurs dans des tâches de temps de réaction sont parfois très peu nombreuses).

Un autre exemple se trouve dans les données exploitées dans les protocoles ICM et pour lesquels l'ensemble d'apprentissage doit être de taille raisonnable. Dans ce contexte, des modèles trop complexes sont difficilement applicables.

7.2.2 Écriture matricielle

Le modèle linéaire mixte présenté peut s'écrire sous la forme matricielle usuelle (voir par exemple (Searle et al., 1992)) :

$$Y = X\beta + Zu + \varepsilon \, , \tag{7.9}$$

où

- Le vecteur d'observations $Y \in \mathbb{R}^{nKJ}$ est obtenu par concaténation des $n = n_1 + n_2$ essais $y_i^c \in \mathbb{R}^{KJ}$, $c \in \{1, 2\}$,
- $X = \begin{pmatrix} X^1 & 0 \\ 0 & X^2 \end{pmatrix}$ est la matrice design (connue) des effets fixes, avec $X^c = \mathbf{1}_{n_c} \otimes \mathbf{I}_{KJ}$ pour $c \in \{1; 2\}$, $\mathbf{1}_{n_c}$ est le vecteur (colonne) unité de taille n_c, $\mathbf{I}_{KJ}$ est la matrice identité de dimension KJ et $\otimes$ dénote le produit de Kronecker,
- $\beta = \begin{pmatrix} \mu^1 \\ \mu^2 \end{pmatrix}$ est le vecteur-paramètre (inconnu) des effets fixes, où $\mu^c \in \mathbb{R}^{KJ}$ est le vecteur moyenne de la classe c,
- $Z = \begin{pmatrix} Z^1 & 0 \\ 0 & Z^2 \end{pmatrix}$ est la matrice design (connue) associée aux effets aléatoires et où $Z^c = \mathbf{I}_{n_c} \otimes \Gamma^c \in \mathbb{R}^{n_c KJ \times n_c}$, avec $\Gamma^c \in \mathbb{R}^{KJ}$ le vecteur des coefficients défini *a priori*,

- $u = \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix} \sim \mathcal{N}\left(0, \tau^2 \mathbf{I}_n\right)$ est le vecteur gaussien des effets aléatoires associés aux n essais, avec τ^2 inconnu,
- $\varepsilon \sim \mathcal{N}\left(0, \sigma^2 \mathbf{I}_{nKJ}\right)$ est le vecteur des résidus gaussiens i.i.d. et indépendant du vecteur des effets aléatoires u avec σ^2 inconnu.

L'implémentation de ce modèle pour l'estimation des paramètres et la prédiction des effets aléatoires a été faite à l'aide de la fonction MATLAB `mixed.m` écrite par Witkovský (Witkovský, 2002), fonction disponible en ligne. Cet algorithme utilise les équations du modèle mixte d'Henderson et offre un large choix de méthodes d'estimation des paramètres.

7.3 Classification

A partir de la loi du vecteur $y_i^c \in \mathbb{R}^{KJ}$, nous introduisons ici une procédure de classification binaire basée sur l'approche probabiliste de l'analyse discriminante comme définie dans la section 3.2.

7.3.1 Une QDA particulière

Soit $y_i \in \mathbb{R}^{KJ}$ un essai dont on souhaite déterminer la classe d'appartenance $c \in \{1; 2\}$. On considère le mélange gaussien suivant :

$$y_i \sim p_1 \, \mathcal{N}(\mu^1, V^1) + p_2 \, \mathcal{N}(\mu^2, V^2)\,, \tag{7.10}$$

où, p_c est la probabilité *a priori* de la classe c, $\mu^c \in \mathbb{R}^{KJ}$ est la moyenne et $V^c \in \mathbb{R}^{KJ \times KJ}$ est la matrice de covariance de la classe c.
Á partir du modèle mixte proposé (7.7), la matrice de covariance V^c est définie comme la somme de la variabilité de l'effet aléatoire et de celle du résidu : $V^c = \tau^2 \Gamma^c \left(\Gamma^c\right)' + \sigma^2 \mathbf{I}_{KJ}$. Notons donc que le modèle proposé est un cas particulier de l'analyse discriminante quadratique (QDA) présentée dans la section 3.2.2. Dans chaque classe la matrice de covariance V^c est modélisée au travers des composantes de la variance du modèle mixte. Ainsi, après avoir spécifié $\Gamma^c \in \mathbb{R}^{KJ}$ pour chacune des classes, il ne reste seulement que les deux composantes de la variance τ^2 et σ^2 à estimer. Ceci permet d'envisager l'estimation des paramètres sur de petits échantillons d'apprentissage.

Finalement, chaque essai y_i à classer est assigné à la classe c d'après le maximum *a posteriori* calculé à partir de la formule de Bayes (voir proposition 1) :

$$\max_{c \in \{1,2\}} \mathbb{P}(\text{Classe} = c \,|\, y_i)\,, \tag{7.11}$$

et le classifieur optimal correspond à celui de l'analyse discriminante quadratique, basée sur la fonction discriminante suivante :

$$g_c(y_i) = (y_i - \mu^c)'(V^c)^{-1}(y_i - \mu^c) - 2\ln(p_c) + \ln|V^c|\,, \tag{7.12}$$

dans chaque classe c. Ceci conduit donc au classifieur binaire suivant :

$$\delta(y_i) = g_2(y_i) - g_1(y_i)\,, \tag{7.13}$$

l'essai y_i étant assigné à la classe 2 si $\delta(y_i) < 0$, et la classe 1 sinon.

7.3.2 Une méthodologie spécifique pour la classification des essais uniques

Dans cette section l'ensemble des étapes de la procédure de classification basée sur le modèle mixte est résumé. Comme toute procédure de classification supervisée, deux phases sont considérées : une phase d'apprentissage au cours de laquelle les paramètres du modèle sont estimés et une phase de test pour classer chaque nouvel essai dans une des deux classes $c \in \{1; 2\}$.

Nous rappelons ici que nous souhaitons minimiser la taille de l'ensemble d'apprentissage tout en garantissant que les paramètres du modèle soient estimés de façon précise.

A) Phase d'apprentissage

1) **Traitement des données**
   Dans un premier temps, la décorrélation et la réduction de la dimension temporelle est effectuée à travers de la décomposition en ondelettes discrète. Les K coefficients multi-échelle sont normalisés de sorte que la variabilité des coefficients soit indépendante du niveau de décomposition.
   Ensuite, la dimension spatiale est réduite à son tour. Les filtres spatiaux sont estimés à partir du discriminant de Fisher matriciel sur les M électrodes et seulement les J combinaisons les plus discriminantes (les canaux) sont sélectionnées ($J \ll M$). Les propriétés d'orthogonalité des filtres construits pas analyse discriminante permettent de décorréler la dimension spatiale.
2) **Modélisation dans l'espace des ondelettes**
   Les coefficients multi-échelle et multi-canaux sont modélisés dans l'espace des ondelettes par le modèle proposé (7.7), où la moyenne dans la classe est la composante fixe et la composante aléatoire dépend de la classe, du canal, du coefficient et de l'essai. Dans la partie aléatoire, la dépendance à la classe, au canal et au coefficient est donnée par le vecteur $\Gamma^c \in \mathbb{R}^{KJ}$ qui doit être défini *a priori*. Ce dernier point est étudié dans la section 7.4.
3) **Estimation des paramètres du modèle**
   Après avoir fixé Γ^c, l'ensemble des paramètres, noté $\Theta = (\mu^1, \mu^2, \tau^2, \sigma^2) \in \mathbb{R}^{2KJ+2}$ du modèle mixte peut être estimé par des méthodes basées sur le maximum de vraisemblance en utilisant l'hypothèse selon laquelle b et ε sont indépendants et distribués selon des lois gaussiennes multivariées. En pratique la vraisemblance du modèle est calculée à partir des n essais de la phase d'apprentissage et le modèle utilisé est le modèle matriciel défini en (7.9).
   Les paramètres estimés sont notés $\hat{\Theta} = (\hat{\mu}^1, \hat{\mu}^2, \hat{\tau}^2, \hat{\sigma}^2)$.

B) Phase de test

1) **Décorrélation et réduction de dimension**
   Les mêmes transformations que celles définies en phase d'apprentissage sont appliquées à chaque nouvel essai à classer.

2) **Règle de classification**
Le classifieur $\hat{\delta}(y_i)$ est obtenu par *plug-in* des estimations de μ^1, μ^2,τ^2,σ^2 (calculées au cours de la phase d'apprentissage) dans l'expression (7.13).
A partir des estimations des composantes de la variance $\hat{\tau}^2$ et $\hat{\sigma}^2$, l'estimation de la matrice de covariance de la classe c est donnée par :

$$\widehat{V}^c = \hat{\tau}^2 \Gamma^c (\Gamma^c)' + \hat{\sigma}^2 \mathbf{I}_{KJ} \ . \tag{7.14}$$

7.3.3 Reconstruction des signaux en essais uniques

Rappelons que l'une des problématiques importantes de la régression linéaire est la prédiction. Dans le cas du modèle mixte, une prédiction de l'effet aléatoire b_i^c de l'essai i dans la classe c est obtenue à partir des équations d'Henderson (définies par la proposition B.3.2 dans l'annexe B) :

$$\hat{b}_i^c = \hat{\tau}^2 (\Gamma^c)' (\widehat{V}^c)^{-1} (y_i - \hat{\mu}^c) \ . \tag{7.15}$$

La prédiction de l'effet aléatoire de l'essai i dans la classe c permet de prédire le signal d'intérêt de ce même essai de la façon suivante :

$$\hat{y}_i^c = \hat{\mu}^c + \Gamma^c \hat{b}_i^c \, . \tag{7.16}$$

Notons que, contrairement à Huang *et al*, cette prédiction est un sous-produit du modèle qui permet de visualiser les signaux d'intérêt construits par le modèle et ne joue aucun rôle dans la procédure de classification. Ces prédictions du signal d'intérêt sont illustrées dans l'application au chapitre suivant sur les deux jeux de données étudiés.

7.4 Choix du vecteur design de la partie aléatoire

Afin de spécifier intégralement le modèle, le vecteur de coefficients $\Gamma^c \in \mathbb{R}^{KJ}$ (vecteur design) doit être défini. Nous rappelons que ce vecteur module l'impact des effets aléatoires b_i^c des essais selon les coefficients et les canaux de la même façon dans chaque classe. L'introduction de ce vecteur design permet de simplifier le modèle en considérant un effet aléatoire ne dépendant que des essais ; la difficulté étant de spécifier ce vecteur (comme nous le verrons dans la partie application, le choix de Γ^c dépend du type de données considéré). Pour ce faire il est parfois possible de se baser sur une analyse exploratoire préalable de la variabilité des données, comme nous le montrons dans le cas de l'étude des potentiels d'erreur. Alternativement, nous proposons une méthode heuristique de détermination de ce vecteur, utilisable dès lors qu'une première estimation des paramètres du modèle est disponible.

7.4.1 Analyse exploratoire de la variabilité

L'analyse exploratoire consiste à rechercher des comportements particuliers de la variabilité. Il est par exemple naturel d'étudier le lien entre la moyenne des coefficients sur les canaux et leur écart-type. Nous nous intéressons dans ces travaux à la construction de la matrice design des effets aléatoires pour des données de type potentiels évoqués.

Par exemple, dans le cas du jeu de données de potentiels d'erreur (voir section 2.4.1), une étude préliminaire a permis de faire ressortir une relation monotone entre la variabilité et l'intensité des coefficients multi-échelle moyens tout au moins pour les coefficients de très grande

amplitude. En revanche, une telle relation n'a pu être visualisée par une analyse exploratoire des données issues du protocole P300 Speller.

Il est à noter que l'exploration des données pour la modélisation du signal d'intérêt est un problème en soi puisque les données observées correspondent au signal d'intérêt noyé dans une activité de fond de plus grande amplitude. Ainsi, si les hypothèses émises sur le lien entre amplitude et variabilité sont raisonnables, elles ne permettent pas de tirer des conclusions claires à ce sujet.

7.4.2 Une heuristique pour la détermination de Γ^c

Le choix du vecteur de coefficients Γ^c pouvant s'avérer difficile, nous proposons une méthode heuristique pour sa détermination.
Soit $y_i^c \in \mathbb{R}^{KJ}$ le $i^{\text{ème}}$ le vecteur de coefficients multi-échelles et multi-canaux de la classe c, $i = 1, \ldots, n_c$. D'après le modèle mixte défini dans la section 7.2.1, en supposant la moyenne dans chaque classe et les effets aléatoires b_i^c connus, rechercher le vecteur Γ^c revient à résoudre un problème de régression. En effet si on pose $\tilde{y}_i^c = y_i^c - \mu^c$, on obtient :

$$\tilde{y}_i^c = \Gamma^c b_i^c + \varepsilon_i \,, \tag{7.17}$$

où les $\tilde{y}_i^c$ sont les observations centrées, les b_i^c sont supposés connus, Γ^c le paramètre inconnu de $\mathbb{R}^{KJ}$ à estimer et les résidus ε_i sont des vecteurs gaussiens i.i.d. centrés de matrice de covariance $\sigma^2 I_{KJ}$.

Ainsi la loi conditionnelle de $\tilde{y}_i^c$ sachant b_i^c est égale à $\mathcal{N}(\Gamma^c b_i^c, \sigma^2 I_{KJ})$. Nous cherchons donc ici à estimer dans chaque classe les coefficients du vecteur Γ^c.

Remarque 22. *Afin de simplifier les notations dans ce paragraphe, nous posons $d = KJ$.*

Soit $\tilde{y}^c \in \mathbb{R}^{n_c d}$ le vecteur défini par la concaténation des n_c vecteurs $\tilde{y}_i^c \in \mathbb{R}^d$ correspondant aux essais de la classe c. Le modèle (7.17) s'écrit sous la forme matricielle d'un modèle de régression linéaire gaussien :

$$\tilde{y}^c = X^c \Gamma^c + E \,, \tag{7.18}$$

où

- $X^c \in \mathbb{R}^{dn_c \times d}$ est la matrice connue définie par $b^c \otimes I_d$, avec $b^c = (b_1^c, b_2^c, \ldots, b_{n_c}^c)'$;
- $\Gamma^c \in \mathbb{R}^d$ est le paramètre à estimer ;
- $E \in \mathbb{R}^{dn_c}$ est le vecteur aléatoire des dn_c erreurs i.i.d. du modèle, $E \sim \mathcal{N}(0, \sigma^2 I_{dn_c})$.

Le paramètre Γ^c peut donc être estimé par la méthode des moindres carrés ordinaires (MCO) :

$$\hat{\Gamma}^c_{MCO} = ((X^c)' X^c)^{-1} (X^c)' \tilde{y}^c \,. \tag{7.19}$$

Si le vecteur b^c supposé connu est normalisé ($(b^c)' b^c = 1$) alors la matrice X^c est orthogonale :

$$((X^c)' X^c) = (b^c \otimes I_d)'(b^c \otimes I_d) = (b^c)' b^c \otimes I_d = I_d \,, \tag{7.20}$$

et

$$\hat{\Gamma}^c_{MCO} = (X^c)' \tilde{y} = \sum_{i=1}^{n_c} b_i^c \tilde{y}_i^c \,. \tag{7.21}$$

Mais d'autres techniques peuvent être utilisées par exemple les méthodes de type LASSO prenant en compte une hypothèse de parcimonie dans l'estimation de Γ (Tibshirani, 1996).

Dans les applications nous utiliserons l'estimateur LASSO qui est une version régularisée de l'estimateur MCO permettant de forcer à zéro des coefficients de Γ éliminant ainsi des sources de variabilité peu significatives dans le modèle mixte considéré par la suite.

Estimateur LASSO du vecteur design. La méthode LASSO consiste à minimiser le critère des moindres carrés avec pénalisation ℓ_1 :

$$\min_{\Gamma^c \in \mathbb{R}^d} \left\{ \|\tilde{y}^c - X^c \Gamma^c\|^2 + 2\lambda \|\Gamma^c\|_1 \right\} , \tag{7.22}$$

où $\lambda \geq 0$ est le paramètre de pénalisation et $\|\Gamma^c\|_1 = \sum_{j=1}^{d} |\Gamma_j^c|$. Lorsque $\lambda = 0$, le problème du LASSO est équivalent à celui des moindres carrés. Plus λ augmente, plus le nombre de coefficients Γ_j^c mis à zéro augmente. La solution du LASSO est alors dite parcimonieuse.

Etant donné que la matrice X^c est orthogonale ($(X^c)'X^c = I_d$), la solution de l'expression (7.22) est explicite : les estimateurs LASSO des composantes Γ_j^c du vecteur Γ^c sont, pour tout $j = 1, \ldots, d$:

$$\hat{\Gamma}^c_{j,lasso} = \mathsf{signe}(\hat{\Gamma}^c_{j,MCO})(|\Gamma^c_{j,MCO}| - \lambda)_+ , \tag{7.23}$$

où $\mathsf{signe}(A)$ donne le signe du terme A, $(A)_+$ est égale à A si $A > 0$ et à zéro sinon et $\hat{\Gamma}^c_{j,MCO} = \sum_{i=1}^{n_c} b_i^c \tilde{y}_i^c(j)$ est l'estimateur MCO.

Remarque 23. *L'estimateur LASSO défini en (7.23) effectue un seuillage doux de l'estimateur des moindres carrés ordinaires de Γ. Par ailleurs, lorsque $\lambda \geq \max_j \hat{\Gamma}^c_{j,MCO}$, alors $\hat{\Gamma}^c_{j,lasso} = 0$ pour tout $j = 1, \ldots, d$.*

Cet estimateur repose sur des effets aléatoires b_i^c fixés. Toutefois dans le modèle (7.7), ces effets aléatoires ne sont pas connus et dans une première étape nous devons les prédire à l'aide des équations du modèle mixte (7.15) construit sur l'ensemble des deux classes après avoir fixé Γ^1 et Γ^2 à une première valeur initiale.

En remplaçant b_i^c par sa prédiction $\hat{b}_i^c$ normalisée dans l'expression (**??**), on obtient donc :

$$\hat{\Gamma}^c_{j,lasso} = \mathsf{signe} \left(\frac{1}{\|\hat{b}^c\|} \sum_{i=1}^{n_c} \hat{b}_i^c \tilde{y}_i^c(j) \right) \left(\frac{1}{\|\hat{b}^c\|} \sum_{i=1}^{n_c} |\hat{b}_i^c \tilde{y}_i^c(j)| - \lambda \right)_+ . \tag{7.24}$$

Remarque 24 (Détermination du paramètre de régularisation)**.** *Le choix de λ est un problème classique (et difficile), nous renvoyons à (Tibshirani, 1996) et références incluses pour une discussion de ce problème.*

Etant donné que nous intéressons dans ce travail à définir une procédure de classification via une analyse discriminante particulière, dans la méthode LASSO, le critère de sélection pour la valeur de λ sera basé sur la classification. Le paramètre λ sera commun aux deux classes. Nous choisirons pour λ_{opt} la valeur qui minimise le taux d'erreur moyen mesuré sur un échantillon à partir d'une approche leave-one-out et obtenu à partir de la règle de décision définie en (7.13).

Soit

$$\lambda_{max} = \max_{c,j} \frac{1}{\|\hat{b}^c\|} \sum_{i=1}^{n_c} \hat{b}_i^c \tilde{y}_i^c(j) \ .$$

Dans ce cas, tous les coefficients des vecteurs Γ^1 *et* Γ^2 *sont estimés à zéro. Afin de choisir la valeur* λ_{opt}*, nous évaluons donc l'erreur moyenne de classification pour différentes valeurs de* λ*, comprises entre* 0 *(estimateurs MCO) et* λ_{max} *(pas d'effet aléatoire).*

Synthèse : une procédure heuristique pour déterminer Γ^c.

- Initialisation de Γ^c par $\Gamma^c_{(0)} \in \mathbb{R}^d$ pour $c = 1, 2$;
- Estimation des paramètres du modèle (7.7) en prenant $\Gamma^c = \Gamma^c_{(0)}$: $\hat{\tau}^2_{(0)}$, $\hat{\sigma}^2_{(0)}$, $\hat{\mu}^1_{(0)}$, $\hat{\mu}^2_{(0)}$
- Centrage des données : $\tilde{y}^c_i = y^c_i - \hat{\mu}^c_{(0)}$;
- Prédiction des effets aléatoires : $\hat{b}^c_i(0)$, pour tout $i = 1, \ldots, n_c$:

$$\hat{b}^c_i(0) = \tau^2_{(0)}\Gamma^c_{(0)}(\hat{V}^c_{(0)})^{-1}\tilde{y}^c_i \quad \text{avec} \quad \hat{V}^c_{(0)} = \hat{\tau}^2_{(0)}\Gamma^c_{(0)}(\Gamma^c_{(0)})' + \hat{\sigma}^2_{(0)}I_d\,;$$

- Dans chaque classe c, estimation de Γ^c par une méthode de régression en fixant le vecteur $b^c = (b^c_1, \cdots, b^c_{n_c})'$ des effets aléatoires des essais de la classe égal à sa prédiction normalisée $\frac{1}{\|\hat{b}^c(0)\|}\hat{b}^c(0)$;

On estime ensuite à nouveau les paramètres du modèle (7.7) en prenant $\Gamma^c = \hat{\Gamma}^c$.

Nous allons à présent étudier la procédure proposée dans ce chapitre, à la fois en termes de modélisation des essais uniques qu'en termes de classification pour les deux jeux de données étudiés dans cette thèse.

Chapitre 8

Applications

Sommaire

8.1 Détection de potentiels d'erreur

L'ensemble des résultats présentés dans cette section et dans l'annexe C.3 ont fait l'objet d'une publication et peuvent être retrouvés dans l'article (Spinnato et al., 2015a). Des résultats supplémentaires obtenus sur ce même jeu de données illustrant l'apport de l'estimation du vecteur design Γ^c par l'approche LASSO proposée dans la section 7.4.2 sont donnés dans l'annexe C.3.5.

8.1.1 Réduction de dimension et décorrélation

La première étape, brièvement résumée ici, reprend les éléments décrits dans le paragraphe du même titre de la section 7.2 du chapitre précédent. Partant des signaux temporels (constitués on le rappelle de 204 échantillons temporels, avec une fréquence d'échantillonnage de 256 Hz), complétés par des zéros à chaque extrémité pour atteindre 256 échantillons temporels, les coefficients multi-échelle correspondants sont calculés par DWT sur 5 niveaux de décomposition d'ondelettes et un niveau de fonction d'échelle, correspondant aux bandes fréquentielles 64-128 Hz, 32-64 Hz, 16-32 Hz, 8-16 Hz, 4-8 Hz et 0-4 Hz respectivement. Les résultats présentés dans ce chapitre ont été obtenus en choisissant le filtre D6 qui est le résultat d'un compromis entre localisation et lissage de l'ondelette (voir 2.4.1) . Cette transformation a pour effet d'une part de "concentrer" l'information pertinente sur un petit nombre de coefficients multi-échelle et d'autre part de "décorréler" ces coefficients (voir section 2.2.2). Les coefficients négligeables sont exclus de la modélisation, ainsi que les coefficients affectés par des effets de bord. Après cette étape, $K = 24$ coefficients seulement ont ainsi été sélectionnés ; plus précisément, les trois premiers niveaux de décomposition (hautes fréquences) ont été éliminés, et seuls subsistent 12+6 coefficients d'ondelettes des niveaux 4 et 5, plus 6 coefficients d'échelle.

Un filtrage spatial est ensuite effectué à partir d'une LDA matricielle (voir section 5.2.1 et section 6.1). Comme nous l'avons montré dans la section 6.1, ce filtrage permet de concentrer l'information spatiale discriminante par la construction de combinaisons linéaires de signaux enregistrés sur les électrodes appelées canaux. Le nombre J de canaux sélectionnés par le filtrage spatial est déterminé de sorte que 75% de l'information discriminante soit conservée (voir 6.1.2).

8.1.2 Modéliser les potentiels d'erreur

Après filtrages temporel et spatial, on considère les données matricielles $Y_i^c \in \mathbb{R}^{24\times J}$ contenant en lignes les coefficients multi-échelle et en colonnes les canaux avec $J \ll M = 64$ le nombre d'électrodes. Dans l'ensemble d'apprentissage utilisé en exemple $J = 8$. La modélisation des essais (*Erreur* et *Correct*) est effectuée sur les données vectorisées $y_i^c \in \mathbb{R}^{24J}$ ainsi que la procédure de classification.

Tester l'inégalité des matrices de covariance

La procédure de classification utilisée étant basée sur l'hypothèse que les matrices de covariance sont différentes dans les deux classes *Erreur* et *Correct*, il est naturel de tester en premier lieu si la différence observée est significative. Pour cela nous utilisons le test de Box présenté dans la section 3.3.2. Les matrices de covariance empiriques des deux classes sont représentées sur la figure 8.1. Par souci de lisibilité, les matrices sont représentées en utilisant uniquement les 2 premiers canaux. De plus, pour minimiser les biais d'une comparaison visuelle, les deux matrices ont été calculées dans les deux classes à partir d'un jeu d'apprentissage de même taille. Une première étude visuelle laisse apparaître des différences au niveau des éléments diagonaux de la matrice ainsi que la présence de termes non-diagonaux non-nuls pour la classe *Erreur*.

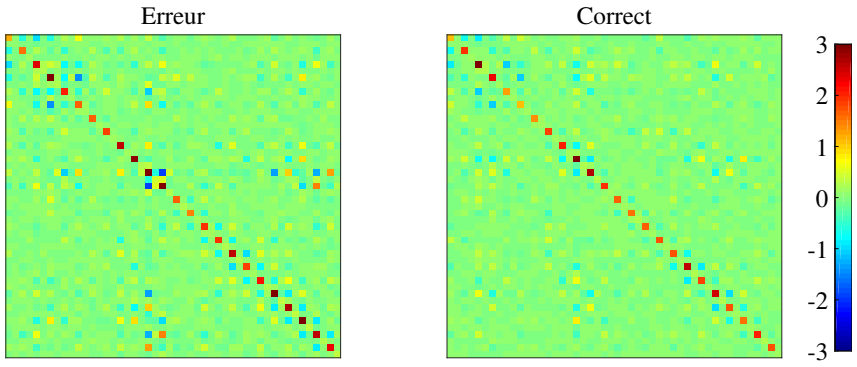


FIGURE 8.1 – Participant A. Représentation des covariances empiriques $\hat{\Sigma}^{Err}$ et $\hat{\Sigma}^{Corr}$. Dans cet exemple $J = 2$ canaux et $K = 24$, les matrices sont donc de dimension 48×48.

La différence significative de ces deux matrices est évaluée quantitativement par l'application du test de Box (Anderson, 2003) dont les résultats pour le participant A sont fournis dans la table 8.1. L'hypothèse H_0 d'égalité des matrices de covariance est ici clairement rejetée (p-valeur $\ll 10^{-6}$). Des résultats similaires sont obtenus pour l'ensemble des participants du jeu de données (voir annexe C.3.2).

Nous pouvons donc ici conclure qu'il semble raisonnable de modéliser via le modèle mixte (7.7)

MBox	Chi-sqr	g	p	dl	p-valeur
3273.05	2859.74	2	48	1176	$\ll 10^{-6}$

TABLE 8.1 – Résultats du test de Box pour tester l'inégalité des matrices de covariance des classes *Erreur* et *Correct* à partir des covariances empiriques $\hat{\Sigma}^{Err} \in \mathbb{R}^{JK\times JK}$ et $\hat{\Sigma}^{Corr} \in \mathbb{R}^{JK\times JK}$. La statistique de test prend pour valeur (3273.05) et sous H_0 l'hypothèse d'égalité des matrices, la loi de la statistique est un Khi-2 à 1176 degrés de liberté (ici $J = 2$, donc $p = JK = 24 * 2 = 48$ et $dl = p(p+1)/2 = 1176$). La p-valeur est égale à la probabilité qu'une variable du Khi-2 à 1176 degrés de liberté soit supérieure à 3273.05.

la variabilité de manière différente dans les deux classes et d'utiliser une QDA pour la classification (section 7.3.1).

Choix du vecteur design Γ^c de la partie aléatoire

Une analyse exploratoire a permis de mettre en évidence une relation monotone entre la variabilité et l'intensité des coefficients multi-échelle moyens. Ceci est illustré par la figure 8.2 pour deux participants de l'étude. Ces graphiques semblent indiquer qu'il existe un lien entre l'amplitude des coefficients et leur variabilité (avec un décalage dû à la variabilité induite par le bruit de fond).

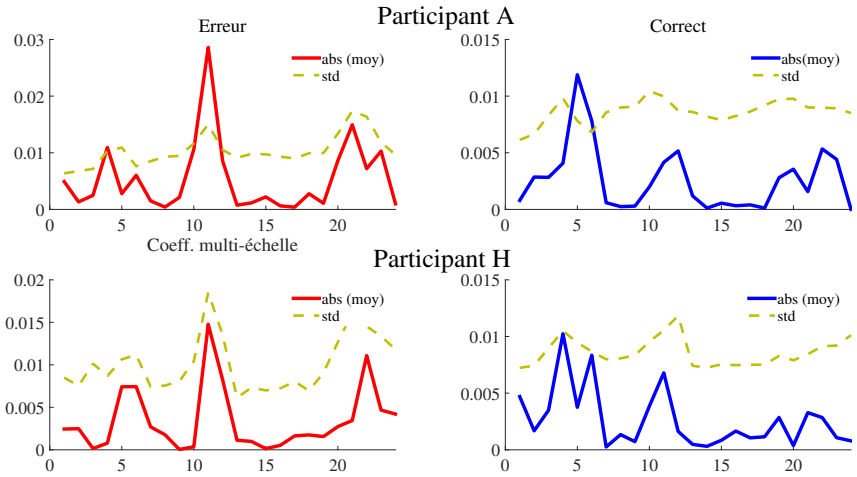


FIGURE 8.2 – Illustration de la relation monotone entre la moyenne (en valeur absolue) des coefficients multi-échelle et leur écart-type pour les participants A et H. La moyenne en valeur absolue des 24 coefficients est représentée pour les erreurs (rouge) et les réponses correctes (bleu) sur le premier canal. Les écarts-types associés sont représentés en tirets verts sur chaque graphique.

A partir de ce constat, le choix le plus simple a été fait en prenant Γ^c égal à la moyenne empirique des essais y_i^c de la classe c :

$$\Gamma^c = \frac{1}{n_c}\sum_{i=1}^{n_c} y_i^c \,. \tag{8.1}$$

Dans ce cas, dans le modèle (7.7), la partie fixe est la moyenne théorique dans la classe c et Γ^c, le vecteur design de la partie aléatoire, est définie par la moyenne empirique calculée dans la classe c. Notons cependant qu'en phase d'estimation du modèle, la moyenne théorique est estimée par la moyenne théorique.

Estimation des composantes de la variance

Nous étudions ici une de nos principales hypothèses selon laquelle le signal d'intérêt peut se décomposer sous la forme d'une somme de deux composantes : une composante fixe dépendant de la classe (la moyenne) et un composante aléatoire dépendant quant à elle de l'essai (l'effet aléatoire b_i^c). La pertinence d'introduire un effet aléatoire dû à l'essai est étudiée en testant si la variance estimée $\hat{\tau}^2$ de cet effet est significativement différente de zéro. Si cette différence n'est pas significative, l'effet aléatoire n'a pas lieu d'être dans le modèle. Dans ce cas, un modèle linéaire classique serait plus approprié comme celui proposé dans (Blankertz et al., 2011).

Les estimations des composantes de la variance $\hat{\tau}^2$ et $\hat{\sigma}^2$ obtenues par REML pour le participant A sont présentées dans la figure 8.3.

Chaque boîte à moustache correspond à des tailles d'échantillons différents (en respectant la proportion de 10% d'essais de la classe *Erreur*) et résume la distribution des variances estimées sur 100 jeux d'apprentissage différents. Notons que les valeurs de $\hat{\sigma}^2$, la variance résiduelle estimée sont beaucoup plus grandes que celles de $\hat{\tau}^2$. Globalement, la dispersion des estimations décroît à mesure que les tailles des échantillons d'apprentissage augmentent. Cependant la figure laisse apparaître que des estimations assez précises peuvent être obtenues avec des échantillons de taille assez petite (20 erreurs et 200 corrects). Dans la figure 8.3 (a) les quartiles de la distribution des $\hat{\tau}^2$ a tendance à augmenter légèrement avec la taille de l'échantillon et les valeurs minimales s'écartent de zéro .

Cette analyse exploratoire est confortée par un test statistique. Le test statistique de Wald (Wald, 1943) est utilisé et on conclut qu'en moyenne la variance de l'effet aléatoire est strictement positive (p-valeur $\ll 10^{-3}$), justifiant ainsi l'introduction du modèle mixte (7.7).

8.1.3 Détecter les potentiels d'erreur

Pour le jeu de données considéré, le nombre n_1 d'essais *Erreur* est beaucoup plus faible que le nombre n_2 d'essais *Correct*. Afin d'évaluer les performances de notre méthode par rapport à d'autres classifieurs, nous nous sommes placés dans un contexte de classes déséquilibrés en générant différents échantillons de tailles distinctes respectant une proportion de 10% d'erreurs. Ainsi, pour chaque participant, des échantillons d'apprentissage déséquilibrés ont été tirés aléatoirement (i.e $\{n_1 = 20,\ n_2 = 200\}, \{n_1 = 30,\ n_2 = 300\}, \ldots$), et pour chaque ensemble d'apprentissage tiré, l'ensemble de test est quant à lui composé de l'ensemble des essais restants. Pour chaque participant et pour chaque taille d'échantillons d'apprentissage, les différentes méthodes de classification ont été évaluées par validation croisée à partir de 100 itérations.

La méthode proposée (nommée par la suite LMM) est comparée avec plusieurs classifieurs, tous dérivés de l'analyse discriminante linéaire ou quadratique :

— L'analyse discriminante linéaire classique (LDA) ;
— La LDA régularisée (RDA) comme proposée par Blankertz *et al.* ;
— La LDA diagonale (DLDA) ;
— La QDA diagonale (DQDA).

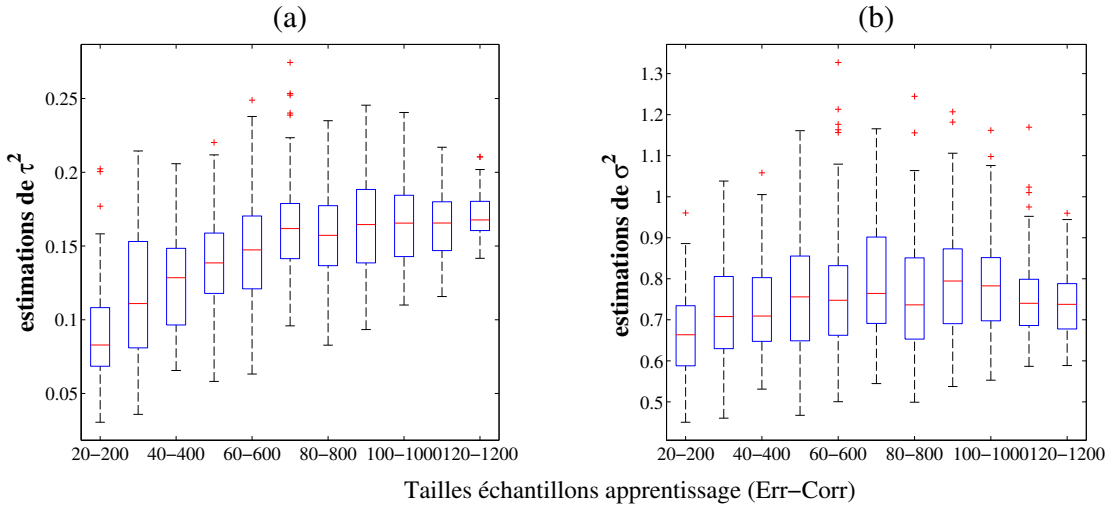


FIGURE 8.3 – Estimations des composantes de la variance. (a) Estimation de la variance τ^2 de l'effet aléatoire. (b) Estimations de la variance résiduelle σ^2. Chaque boîte à moustaches représente une série de 100 estimations pour différentes tailles des échantillons d'apprentissage Err-Corr (en abscisses). *La valeur centrale de la boîte correspond à la médiane (le* $2^{ème}$ *quartile, les limites de la boîtes correspondent au* 1^{er} *et au* $3^{ème}$ *quartile, les moustaches correspondent aux bornes au delà desquelles les valeurs sont considérées atypiques. Ces dernières sont représentées individuellement par une croix rouge.*

De plus, les trois premières méthodes sont également étudiées dans le cas du sous-échantillonnage où la taille de la classe majoritaire (*Correct*) est égale à la taille de la classe minoritaire (*Erreur*). Il est à noter que la QDA classique n'a pas pu être appliquée puisque pour toutes les tailles d'échantillons d'apprentissage, la matrice de covariance de la classe *Erreur* ne peut être correctement estimée étant donné que le nombre de coefficients de la matrice à estimer est supérieur au nombre d'essais d'apprentissage.
Afin d'évaluer la performance du modèle uniquement, toutes les méthodes ont été appliquées sur les mêmes données transformées (transformation en ondelettes et filtrage spatial) et sur les mêmes jeux d'apprentissage et jeux de test.

Résultats de classification : participant *A*. La figure 8.4 représente les résultats obtenus pour le participant *A*. Les barres d'erreur représentent les écarts-types de la distribution des taux de bonne classification pour les 100 itérations effectuées. Etant donné que les échantillons d'apprentissage et de test sont complémentaires dans notre étude, quand la taille de l'ensemble d'apprentissage augmente le nombre d'essais tests diminue et la barre d'erreur augmente. On souligne que ces barres d'erreur mesurent uniquement la dispersion des taux de bonne classification mais ne doivent pas être utilisées dans la comparaison des méthodes qui s'effectue (voir plus bas) à l'aide d'un test statistique sur échantillons appariés.

La méthode proposée fournit clairement de bien meilleurs résultats pour la détection des évènements rares (les erreurs) dans toutes les situations et tout particulièrement dans le cas de petits échantillons ($20-200$, $30-300$, $40-400$). Pour la classe dominante des réponses correctes, notre méthode permet d'obtenir des résultats équivalents à ceux des autres méthodes

(aux alentours de 95%).

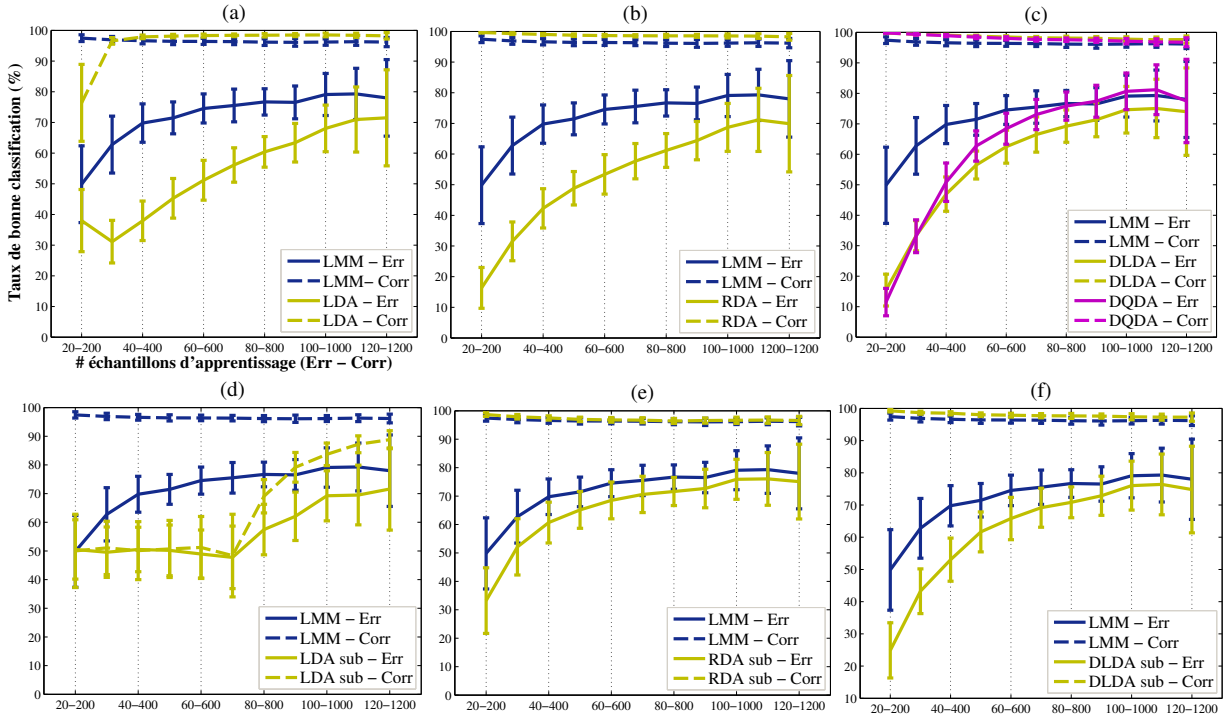


FIGURE 8.4 – Taux de bonne classification pour différentes tailles d'échantillons d'apprentissage (*Erreur* – *Correct*). Les résultats correspondent au taux moyen de bonne classification obtenu à partir de 100 itérations pour chaque taille d'échantillon. Les barres verticales représentent les écarts-types. Les résultats sont donnés pour les essais test de la classe *Erreur* (traits pleins) et de la classe *Correct* (pointillés) distinctement. Notre méthode (LMM) est comparée à trois classifieurs différents pour les situations de classes déséquilibrées et de sous-échantillonnage. Les figures (a), (b) et (c) représentent les résultats dans le cas déséquilibré (10% d'erreurs dans le jeu d'apprentissage) pour la LDA classique, la LDA régularisée (RDA), et la LDA et la QDA diagonales (resp. DLDA et DQDA). Les figures (d), (e) et (f) représentent les résultats de ces classifieurs obtenus après sous-échantillonnage de la classe *Correct*.

En complément, nous introduisons dans la table 8.2 les résultats de classification résumés par le score de Pierce (2.7). Le score de Pierce est calculé pour chaque classifieur sur les mêmes ensembles de test. Ainsi la comparaison des scores de Pierce entre notre méthode et chacune des autres méthodes peut s'effectuer par un test statistique sur échantillons appariés. Nous avons choisi d'utiliser un test non paramétrique, le test des rangs signés de Wilcoxon (Gibbons and Chakraborti, 2010; Hollander et al., 2014).

Pour le participant A, notre approche fournit des résultats significativement meilleurs que ceux obtenus par les autres classifieurs, mis à part pour la DQDA à partir de grandes tailles d'échantillons d'apprentissage (à partir de $80 - 800$) dont les performances sont équivalentes ou meilleures que celle de LMM.

	Comparaison des classifieurs			
Err-Corr	LMM-LDA	LMM-RDA	LMM-DLDA	LMM-DQDA
20-200	0.14 (10^{-12})	0.36 (10^{-18})	0.37 (10^{-18})	0.41 (10^{-18})
30-300	0.34 (10^{-18})	0.34 (10^{-18})	0.32 (10^{-18})	0.32 (10^{-18})
40-400	0.35 (10^{-18})	0.30 (10^{-18})	0.25 (10^{-18})	0.21 (10^{-18})
50-500	0.28 (10^{-18})	0.25 (10^{-18})	0.16 (10^{-18})	0.09 (10^{-16})
60-600	0.25 (10^{-18})	0.23 (10^{-18})	0.13 (10^{-18})	0.07 (10^{-13})
70-700	0.21 (10^{-18})	0.19 (10^{-18})	0.10 (10^{-18})	0.03 (10^{-4})
80-800	0.18 (10^{-18})	0.17 (10^{-18})	0.08 (10^{-17})	0.009 (0.43)
90-900	0.14 (10^{-18})	0.13 (10^{-18})	0.06 (10^{-13})	-0.01 (10^{-3})
100-1000	0.12 (10^{-17})	0.11 (10^{-17})	0.05 (10^{-10})	-0.02 (10^{-4})
110-1100	0.09 (10^{-12})	0.09 (10^{-13})	0.04 (10^{-7})	-0.02 (10^{-4})
120-1200	0.07 (10^{-4})	0.08 (10^{-6})	0.04 (10^{-2})	0.005 (0.23)

TABLE 8.2 – Participant *A*. Comparaison des classifieurs à partir du score de Pierce. Les performances de notre méthode (LMM) sont comparées à celles de chacun des autres classifieurs à partir du test des rangs signés de Wilcoxon. Pour chaque paire de classifieurs *LMM-autre méthode*, la différence ente les scores de Pierce moyens (calculés sur les 100 itérations) est présentée accompagnée de la p-valeur associée au test de Wilcoxon (entre parenthèses).

Résultats pour l'ensemble des participants. Les résultats obtenus pour tous les autres participants sont présentés dans l'annexe C.3. Les mêmes conclusions que celles tirées pour le participant *A* peuvent être faites pour les autres participants. Dans la plupart des situations, notre méthode s'avère meilleure que tous les autres classifieurs utilisés ici. Et ceci est particulièrement significatif dans le cas de petits échantillons d'apprentissage, lorsque la classe *Erreur* n'est constituée que d'un petit nombre d'essais (20, 30, 40, 50). LMM est alors significativement meilleure (p-valeur $\ll 10^{-2}$). Contrairement au cas équilibré où la LDA a généralement de meilleures performances que la QDA même lorsque les matrices de covariance sont différentes dans les deux classes, la DQDA donne de meilleurs résultats que la DLDA. Ainsi dans le cas déséquilibré, nous insistons sur le fait qu'il est préférable de prendre en compte la différence de variabilité entre les deux classes. Sachant que d'une part la performance d'un classifieur est liée à la qualité de l'estimation des matrices de covariance, et d'autre part la qualité de l'estimation dépend du nombre de paramètres à estimer par rapport au nombre d'observations, dans le cas de petits échantillons, il est important de modéliser la variabilité de chaque classe de manière à réduire le nombre de paramètres à estimer. C'est ce que nous faisons au travers de notre procédure. En effet uniquement les estimations des moyennes des classes et de deux paramètres de variance sont nécessaires pour appliquer notre procédure de classification. De plus nous pouvons considérer que les bons résultats de la classification constituent une sorte de validation indirecte du modèle proposée pour ce jeu de données.

Remarque 25. *Des résultats similaires sont obtenus en utilisant différents filtres de Daubechies. Ceci démontre la consistance de notre méthode au choix de l'ondelette pour le filtrage temporel. Les résultats sont donnés en annexe pour le participant A (voir C.3.4). Quelque soit le filtre choisi, notre méthode donne de meilleurs résultats de classification. Le filtre de Haar fournit les meilleurs résultats en terme de classification, mais celui-ci ne fournit pas un représentation satisfaisante des signaux reconstruits (voir figure C.8 de l'annexe), c'est pourquoi l'ensemble des*

résultats a été présenté ici avec le filtre D6.

8.1.4 Ajustement essai par essai

Comme décrit dans la section 7.3.3, une estimation du signal d'intérêt peut être calculée pour chaque essai à partir de l'équation (7.16). Après projection de l'essai ajusté dans l'espace des électrodes (en utilisant le pseudo-inverse de Moore-Penrose) et une transformation en ondelette inverse, nous obtenons des signaux multi-électrodes dit *ajustés* à partir du modèle estimé. Un exemple de ces signaux ajustés est donné dans la figure 8.5 pour le participant A sur l'électrode FCz. La figure 8.5 (a) représente les moyennes dans chaque classe des signaux bruts et des signaux ajustés. On constate que les deux moyennes sont très similaires. Les mêmes résultats ont été obtenus pour les autres participants (voir annexe C.3.3), parfois en moindre mesure. Ceci indique donc que ni la réduction de dimension, ni la modélisation n'ont affecté le comportement moyen des signaux. Bien que peu surprenant et souhaité, ce résultat assure que la forme globale des signaux a été conservée et que la modélisation capture bien le signal d'intérêt et amoindrit l'activité de fond. Étudions maintenant les signaux uniques ajustés par le modèle représentés dans la figure 8.5 (b). On note que dans la classe *Erreur*, aussi bien les moyennes estimées, que les signaux uniques ajustés laissent apparaître la négativité d'erreur environ 150 ms après l'EMG, comme cela a été largement décrit dans la littérature (Falkenstein et al., 2000).

La valeur ajoutée du modèle est de permettre de décrire et de quantifier la variabilité inter-essais. La méthode permet à chaque essai de s'écarter du comportement moyen et la variabilité entre les essais apparaît clairement dans la figure 8.5 (b). On observe des variations d'amplitude pour la négativité d'erreur entre chaque essai : tandis que certaines erreurs induisent une importante activité négative, d'autres présentent de potentiels de faible amplitude.

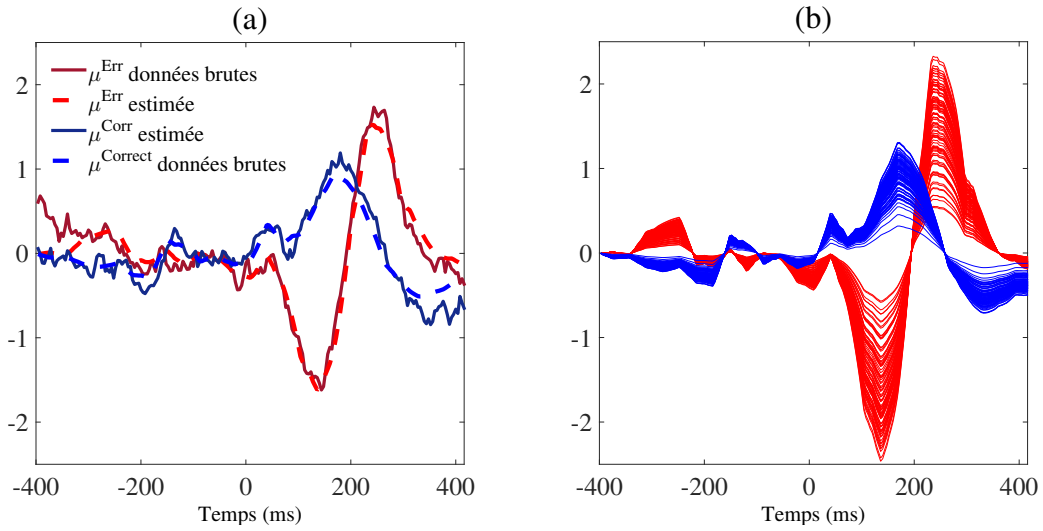


FIGURE 8.5 – Illustration de l'ajustement des signaux par le modèle proposé. (a) Moyennes des signaux des classes *Erreur* (rouge) et *Correct* (bleu) sur l'électrode FCz pour le participant A. Traits pleins : moyennes des essais bruts dans chaque classe, tirets : moyenne des essais ajustés. (b) Représentation des essais uniques ajustés par le modèle dans les deux classes.

Sur la pertinence de la méthode Les résultats obtenus sur ce jeu de données consistants au travers des participants, permettent de valider la procédure, tant du point de vue de la modélisation que de la classification. En effet, la comparaison du classifieur LMM par rapport à diverses méthodes dérivées de l'analyse discriminante linéaire et quadratique montre que la procédure proposée permet d'obtenir des résultats significativement meilleurs dans un cas généralement difficile à traiter : le cas où les classes sont déséquilibrées et jeu d'apprentissage est de petite taille.

Enfin, comme nous l'avons déjà signalé plus haut, les bons résultats obtenus pour la classification fournissent une sorte de validation du modèle proposé.
Dans le cas des potentiels d'erreur, le choix de la moyenne empirique comme vecteur design de l'effet aléatoire (i.e $\Gamma^c = \bar{y}^c$) semble très pertinent. Ceci conforte l'hypothèse que la variabilité du signal d'intérêt est proportionnelle à son amplitude moyenne. Un autre choix du vecteur Γ^c a été fait en utilisant l'heuristique développée en 7.4.1. Les résultats obtenus sont moins bons que les précédents. Ils sont présentés pour un participant en annexe (C.3.5).

8.2 Modélisation de la variabilité inter-essais pour les données du P300 Speller

Nous présentons ici les résultats obtenus sur le deuxième jeu de données. Nous rappelons que dans le protocole P300 Speller les données sont naturellement déséquilibrées : pour une lettre épelée on dispose de 6 flashs *Cible* et de 30 flashs *Non cible*. Ce déséquilibre est cependant moins important que celui rencontré dans le premier jeu de données.

Comme précédemment, les traitements pour la décorrélation et la réduction de dimension spatiales et temporelles sont appliquées. Nous rappelons que pour ce jeu de données la transformation en ondelettes est faite à partir d'un filtre de Daubechies D8 et une sélection des coefficients multi-échelle les plus pertinents est faite (voir section 2.4.2). De plus, le nombre J de canaux conservés par le filtrage spatial est déterminé de sorte que 75% de l'information discriminante soit conservée.

8.2.1 Tester l'égalité des matrices de covariance

Comme pour le premier jeu de données, une première étape de la modélisation des signaux du protocole P300 Speller est de tester l'égalité des matrices de covariance des classes *Cible* et *Non Cible*. Les résultats du test de Box sont présentés pour les 4 sujets qui sont étudiés dans cette application [1] sont donnés dans la table 8.2.1.

Pour les sujets 2, 3 et 6 l'hypothèse d'égalité des matrices de covariance peut être rejetée (p-valeur $\ll 10^{-6}$). Ainsi, nous pouvons conclure que $\Sigma^{Cible} \neq \Sigma^{nonCible}$. En revanche nous ne pouvons pas tirer cette conclusion pour le sujet 1 (p-valeur $= 0.19$).

Nous avons représenté dans les figures 8.6 et 8.7 les matrices de covariance empiriques $\hat{\Sigma}^{Cible}$ et $\hat{\Sigma}^{nonCible}$ pour les sujets 1 et 2 respectivement. Les dimensions KJ des matrices sont données dans la deuxième colonne de la table 8.2.1. Pour le sujet 1, nous observons une variabilité très importante sur les 16 premières composantes sur la matrice de covariance *Cible* comme pour la matrice de covariance *non Cible*, correspondant aux 8 coefficients d'échelle exprimés sur les 2 canaux. Le reste de la matrice de covariance est très clairement diagonale dominante. Les différences s'expriment donc sur la variabilité des coefficients d'échelle. Pour le sujet 2 les principales composantes de la variance sont aussi celles liées aux coefficients d'échelle. Contrairement au sujet 1, les matrices empiriques sont diagonales dans leur ensemble. Enfin, il est à noter que les échelles de valeurs sont très différentes d'un sujet à l'autre. La variabilité pour les covariances du sujet 1 est près de 10 fois plus importante que celle du sujet 2.

Cette première étude semble indiquer que le modèle proposé ici ne sera pas adapté à la situation du sujet 2. Nous verrons dans la section 8.2.3 que cette intuition est confirmée au vu des résultats de classification.

1. L'application présentée ici prend la forme d'une première étude exploratoire. Il faudrait dans un second temps faire une étude plus systématique de comparaison des résultats sur l'ensemble des sujets de ce jeu de données.

		Résultats du test		
Participant	KJ	Statistique	p-valeur	Inégalité des covariances ?
1	$25 \times 2 = 50$	1674.08	0.19	Non
2	$36 \times 3 = 108$	15105.75	$<< 10^{-6}$	Oui
3	$29 \times 4 = 116$	16939.23	$<< 10^{-6}$	Oui
6	$26 \times 3 = 78$	5233.93	$<< 10^{-6}$	Oui

TABLE 8.3 – Résultats du test de Box pour 4 sujets du jeu de données.

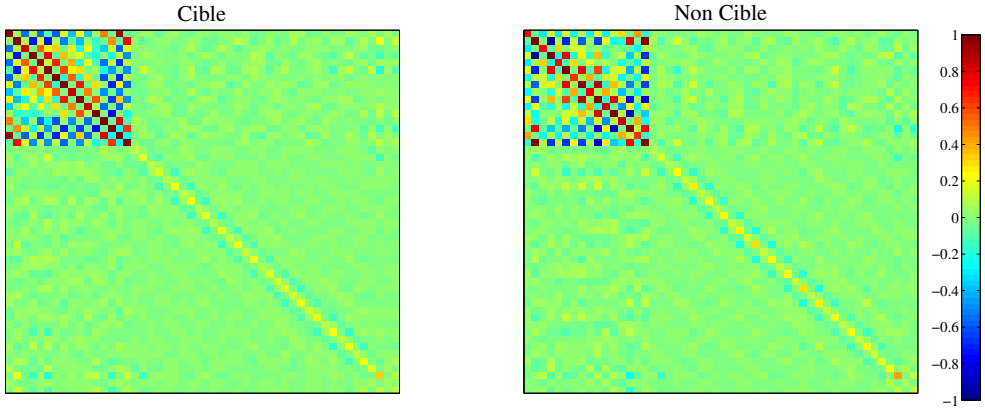


FIGURE 8.6 – Sujet 1. Représentation des matrices de covariance empiriques pour les classes *Cible* et *non Cible*.

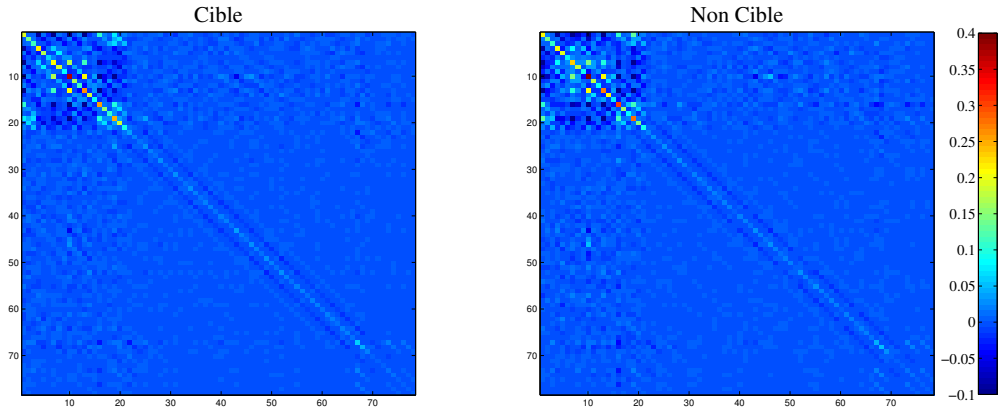


FIGURE 8.7 – Sujet 2. Représentation des matrices de covariance empiriques pour les classes *Cible* et *non Cible*.

8.2.2 Choix des vecteurs design Γ^{Cible} et $\Gamma^{nonCible}$

Contrairement aux potentiels d'erreur pour lesquelles le choix de la moyenne empirique pour le vecteur design des effets aléatoires semble pertinent, par l'analyse exploratoire de la variabilité des données du P300 Speller nous ne sommes pas arrivés à dégager de comportement caractéristique permettant d'orienter le choix du vecteur Γ^c. Pour déterminer Γ^c nous avons donc utilisé l'heuristique basée sur l'estimation LASSO présentée dans la section 7.4.1. Nous avons pris pour valeur initiale de Γ^c, $\hat{\Gamma}^c_{(0)}$ égale à la moyenne empirique des essais dans chaque classe. Ainsi, nous comparons cette approche au modèle LMM utilisant directement la moyenne empirique sans la procédure d'apprentissage permettant de seuiller les coefficients du vecteur design.

Les résultats de classification présentés dans la suite ont été obtenus à partir de 20 lettres utilisées pour la phase d'apprentissage et de 40 lettres en phase de test.

8.2.3 Taux de lettres correctement classées

Nous présentons sous forme de graphiques (voir figure 8.8) les résultats de classification obtenus à partir des modèles LMM et LMM-LASSO en les comparant à d'autres classifieurs dérivés de LDA. Les résultats sont donnés pour 4 participants.

Comme déjà observé dans l'application précédente pour ce jeu de données (voir section 6.2.2), les résultats de classification sont très différents selon le sujet observé. Par exemple, pour le sujet 1 la meilleure méthode (ici l'analyse discriminante régularisée - RDA) ne permet pas de classer correctement plus de 60% des lettres, alors que le pour le sujet 2 l'ensemble des méthodes permet de classer correctement 95% des lettres.

D'après les graphiques de la figure 8.8, pour les sujets 2 et 6, les modèles LMM et LMM-LASSO fournissent des résultats équivalents aux autres méthodes tandis que pour les sujets 1 et sujets 3 les modèles proposés apparaissent moins performants que les autres.

On note également que l'apport de la détermination du vecteur Γ^c via l'heuristique et l'estimation LASSO ne semble pas significative.
Ceci s'explique très probablement par le choix de l'initialisation qui correspond ici à la moyenne empirique et qui ne semble pas pertinente pour ces données. En effet, il faudrait effectuer une étude plus poussée permettant de comparer différentes initialisation ainsi que différentes méthodes de régression dans l'heuristique afin de déterminer un vecteur Γ^c plus adapté. D'autre part, il est à noter qu'ici le contexte choisi pour comparer les modèles proposés aux autres classifieurs ne correspond pas à la situation où la taille de l'échantillon d'apprentissage est petite. En effet ici la matrice de covariance commune aux deux classes utilisée dans la LDA peut-être estimée correctement. Il serait intéressant de comparer à nouveau les méthodes dans une situation où l'estimation de cette matrice est singulière.

Une première étude sur les données P300 Speller. Les résultats présentés dans cette section ne permettent pas de valider le modèle proposé pour la variabilité des données issues du protocole P300 Speller. Ceci étant, ce ne sont que des résultats préliminaires, et des investigations supplémentaires seraient nécessaires avant de tirer des conclusions plus fermes.

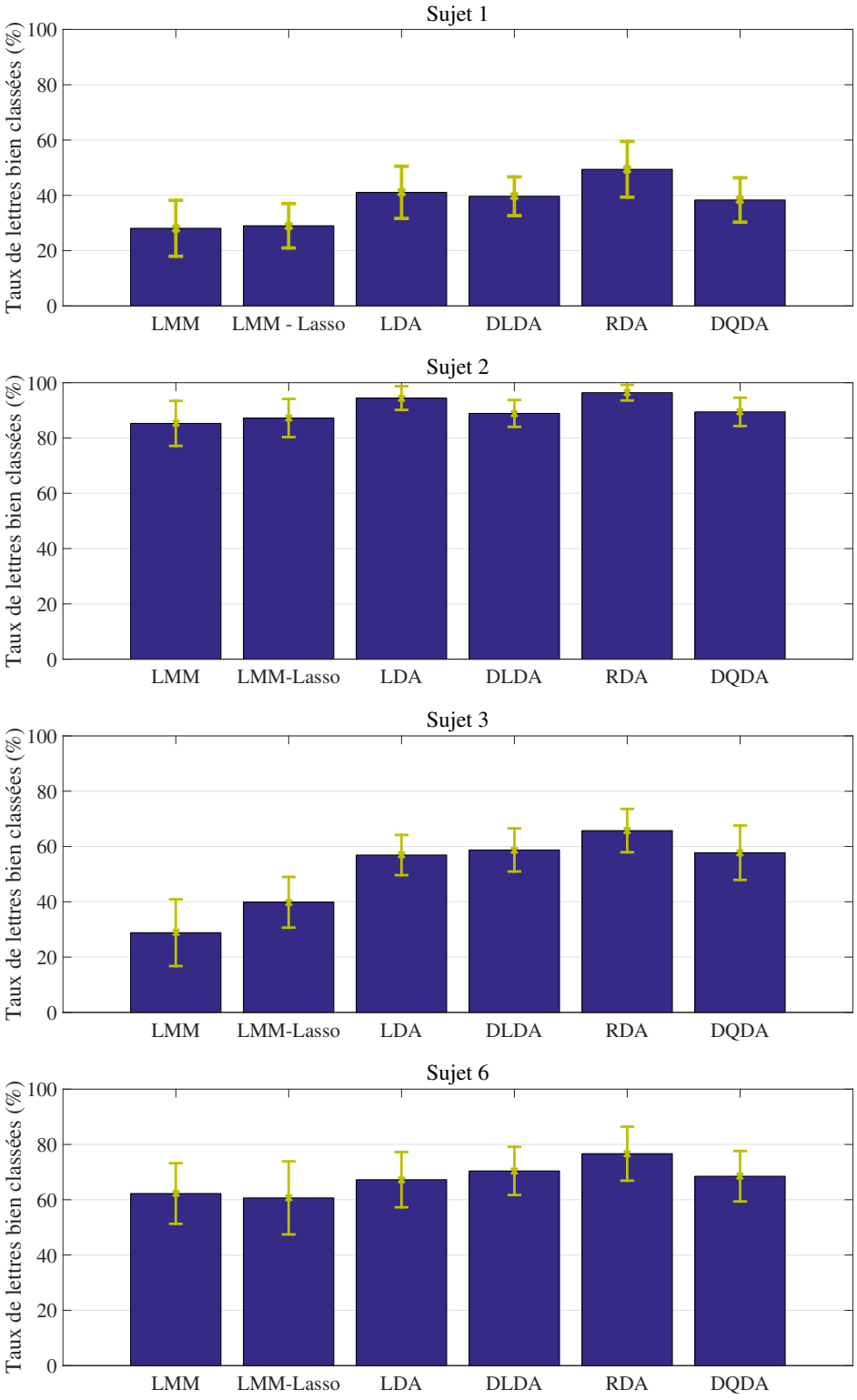


FIGURE 8.8 – Taux de lettres correctement détectées pour différents classifieurs. Les résultats sont présentés pour les sujets 1,2,3 et 6 du jeu de données.

Conclusion Partielle

Nous avons proposé dans cette partie une procédure de classification en EEG basée sur une modélisation explicite de la variabilité des signaux dans le domaine des ondelettes. Dans ce modèle chaque essai s'écrit comme la somme de deux composantes indépendantes, la première modélisant l'activité du signal d'intérêt que l'on souhaite extraire tandis que la seconde intègre l'activité de non-intérêt, plus communément appelée activité de fond. Cette hypothèse est décrite mathématiquement sous la forme d'un modèle linéaire mixte gaussien.

Ce modèle exprime la matrice de covariance comme la somme de deux matrices : la première correspondant au signal d'intérêt dépend de la classe, et la seconde correspondant à l'activité de fond est commune aux deux classes. Ces deux matrices sont caractérisées par un très petit nombre de paramètres, grâce aux réductions de dimension préalables (ondelettes et filtrage spatial), ce qui permet d'en obtenir des estimations robustes dans le contexte étudié. Il n'est donc plus nécessaire de supposer l'égalité des matrices de covariance des classes comme cela est fait usuellement pour effectuer une analyse discriminante. Les résultats obtenus en classification sur le jeu de potentiels d'erreur montre que ce choix est pertinent. Nos résultats démontrent qu'une modélisation pertinente de la variabilité permet d'utiliser la QDA même dans dans des situations de petits effectifs. L'avantage est particulièrement important dans le cas déséquilibré, c'est dans cette situation que l'approche proposée s'est révélée la plus performante en comparaison de méthodes concurrentes.

Cependant, le modèle repose sur le choix explicite d'un vecteur *design* de l'effet aléatoire des essais, qui doit être en adéquation avec la variabilité de ceux-ci. La contrepartie de la simplicité du modèle est l'intégration d'information a priori à travers ce vecteur *design*. Dans le cas de données *ErrP* ce vecteur a été déterminé à partir d'une analyse exploratoire des données, et a conduit à de très bons résultats de classification. Le vecteur *design* peut aussi être vu comme une façon d'introduire dans le modèle des hypothèses sur la variabilité du signal d'intérêt, la classification apparaissant alors comme un moyen simple de tester ces hypothèses.

Il serait cependant aussi souhaitable de définir une méthode plus systématique de détermination du vecteur *design*. Nous avons tenté ici une détermination par régression avec un a priori de parcimonie, mais les résultats se sont avérés décevants, en particulier pour le jeu de données P300. Il s'agit d'une question qui reste à approfondir.

Notons enfin que l'approche proposée permet, par le biais de l'estimation des différents paramètres du modèle, d'extraire des informations sur le signal d'intérêt. Il fournit par exemple des mesures quantitatives de la variabilité inter-essais dans chaque classe. Il permet de plus de donner une estimation du signal d'intérêt. Dans le cadre des potentiels d'erreur, ce dernier point a permis de révéler une importante variabilité entre les signaux d'intérêt estimés dans la classe *Erreur*. Ce résultat pourrait être étudié au regard d'autres variables telles que le temps de réaction ou bien l'attention visuelle.

Quatrième partie

Conclusions générales, annexes et bibliographie

Conclusion générale et perspectives

Objectifs et principales contributions

Le travail présenté dans ce manuscrit a abordé l'analyse et la classification des signaux EEG sous l'angle de la modélisation statistique de la variabilité. Une motivation importante est d'aller au delà de l'opération de moyennage très souvent effectuée, et qui n'a plus lieu d'être dans de nombreux protocoles comme par exemple ceux liés aux interfaces cerveau-machine (ICM). L'une des principales problématiques est dans ce cas de développer des outils d'analyse essai par essai.

Les signaux EEG sont des données de grandes dimensions dont on sait qu'elles sont très fortement corrélées aussi bien spatialement que temporellement. De plus, dans les problématiques actuelles, on cherche de plus en plus à réduire le nombre d'observations afin de diminuer le temps (et le coût) d'analyse des EEG permettant par exemple de s'orienter vers des approches de traitements *online* (c'est d'ailleurs très clairement l'une des ambitions des ICM). Cependant, la combinaison grandes dimensions / petits effectifs pose des problèmes lorsque l'on souhaite caractériser et estimer de façon fiable la variabilité, comme il peut être nécessaire de faire lorsque l'on s'intéresse à de l'essai par essai.

Afin de réduire la difficulté de ce problème, le travail présenté consiste à appréhender la variabilité à partir d'un ensemble de paramètres bien plus petit que le nombre de paramètres mis en jeu dans des matrices de covariance. L'approche suivie est basée sur la modélisation de la covariance, qui cherche à capturer dans un modèle les principaux comportements des signaux et les encoder dans un nombre réduit de paramètres. La modélisation est accompagnée de transformations, préalables ou intégrées dans le modèle, qui conduisent à une réduction de dimension très substantielle : décomposition en ondelettes et projection sur un sous-espace d'ondelettes d'une part, filtrage spatial d'autre part. Il faut insister sur l'importance de ces étapes. Sans la réduction de dimension qu'elles opèrent, il est évident que les modèles de covariance développés ne pourraient pas afficher les mêmes niveaux de performance.

Les problématiques de classification et/ou de décision sont souvent centrales dans l'analyse des signaux EEG et particulièrement dans le contexte des ICM. Les protocoles sont donc généralement construits pour détecter des réponses (ou non-réponses) simples à des stimuli, ou la présence/absence de certaines activités spontanées. La problématique de classification est donc omni-présente dans ce travail, aussi bien dans la construction des modèles que dans leur analyse et leur validation.

Plus précisément, nous avons développé et étudié deux modélisations pour la covariance en EEG dans le cadre de modèles additifs gaussiens. La première approche se base sur l'hypothèse de séparabilité des sources de variabilité spatiale et temporelle, qui prend forme mathématiquement par une factorisation de la matrice de covariance (commune aux classes) en un produit de Kronecker des matrices de covariance associées à ces deux espaces.
La seconde approche consiste quant à elle à modéliser la variabilité inter-essais par un modèle à effets mixtes permettant d'extraire la variabilité propre à un essai donné par rapport à la variabilité du bruit de fond. Ce modèle, moins complexe en termes de paramétrisation, permet de supposer une covariance dépendant de la classe.

Ces deux approches ont été testées sur des jeux de données réels, et évaluées quantitativement sur des problèmes de classification. Elles ont également été évaluées plus qualitativement

grâce à des représentations graphiques issues des deux modèles, qui peuvent être interprétées et comparées à des résultats issus d'autres approches.

Quel modèle de covariance choisir : une comparaison des deux approches

Les deux modèles de covariance proposés dans ce travail sont basés sur des hypothèses fortes et bien différentes quant à la structure de la variabilité des signaux EEG.

Basée sur une hypothèse de séparabilité entre la variabilité spatiale et la variabilité temporelle, la structure de Kronecker de la matrice de covariance ainsi introduite permet de clairement formaliser théoriquement l'analyse discriminante dans le cas matriciel. Dans le cadre applicatif considéré dans cette thèse, la méthode proposée a permis de définir des caractéristiques spatiales et temporelles pertinentes définies comme discriminantes pour les données de potentiels d'erreur et pour les données de type P300 Speller. Les résultats de classification suggèrent pour les deux jeux considérés que la sélection d'un nombre limité de caractéristiques les plus pertinentes améliore la classification. Etant donné que dans le cadre du modèle la classification optimale est théoriquement obtenue en utilisant toutes les caractéristiques, ceci amène à questionner l'adéquation du modèle à ces données. En restant dans le cadre de modèles gaussiens, on pourrait par exemple imaginer que la "vraie" matrice de covariance ne se réduit pas uniquement à un produit de Kronecker, mais pourrait être la somme de plusieurs termes (correspondant d'autres composantes de moindre importance, ou à un bruit de fond). La confrontation aux données de ce modèle peut donc conduire à des hypothèses plus sophistiquées sur la nature des données.

L'introduction d'une structure de Kronecker pour la matrice de covariance permet de réduire substantiellement la dimensionalité du modèle, et donc d'améliorer son applicabilité. Pour donner un ordre d'idées, un signal long d'une seconde échantillonné à 250 Hz sur 40 électrodes représente 10^4 valeurs, et la matrice de covariance associée comporte donc environ $5\,10^7$ coefficients indépendants, ce qui n'est pas gérable. La structure de Kronecker permet de réduire ce nombre à environ $3.2\,10^4$, ce qui reste grand mais représente déjà un gain considérable. Les réductions de dimension, notamment utilisant les ondelettes, permettent de ramener ce nombre à l'ordre du millier de coefficients de matrice, ce qui devient gérable avec des échantillons de l'ordre de quelques centaines de réalisations.

Le second modèle de covariance permet de modéliser la variabilité des signaux d'une tout autre manière, en définissant une variabilité propre à chaque essai qui est différenciée de la variabilité du bruit de fond, supposé indépendant. Dès lors que l'on dispose d'un *a priori* bien défini sur la forme de cette variabilité inter-essai (essentiellement caractérisée par un vecteur *design*), le modèle présente l'avantage d'être simple, et de s'appuyer sur un ensemble de méthodes et algorithmes d'estimation bien maîtrisés. Dans l'application à la modélisation des données de potentiels d'erreur, nous avons pu proposer une forme simple pour le vecteur *design* qui s'est avérée pertinente et performante si l'on en juge par les résultats de classification. Dans ce cas le modèle est remarquablement simple, et permet de traiter des jeux de données à effectifs faibles et classes déséquilibrées, voire très déséquilibrées, avec une analyse discriminante quadratique plutôt que linéaire. Si la qualité des résultats de classification est utilisée pour qualifier la qualité de la modélisation, de bons résultats (comme dans ce cas) montrent que les hypothèses faites sur la variabilité au travers de ce vecteur *design* sont raisonnables.

La construction basée sur le vecteur *design* offre donc beaucoup de flexibilité à la modélisation, mais la détermination de ce vecteur, qui influe fortement sur les performances de la

modélisation, peut ne pas être facile. C'est ce que nous avons pu constater sur le jeu de P300, apparemment plus complexe à modéliser que les ErrP. L'heuristique que nous avons proposée pour déterminer automatiquement le vecteur *design* n'a pas permis dans ce cas d'obtenir de résultats satisfaisants. Il ne s'agissait toutefois que d'une première tentative, qui mérite d'être poussée plus avant.

Quoi qu'il en soit, cette approche, lorsque le vecteur *design* peut être déterminé de façon pertinente, permet de traiter des jeux de donnés avec des effectifs bien plus faibles que la première approche, avec des classes déséquilibrées.

Quelques perspectives

Extensions de l'analyse discriminante matricielle. Comme nous l'avons vu, le modèle de covariance structurée comme un produit de Kronecker fait probablement sens dans certaines applications EEG, mais n'est probablement qu'une approximation de modèles plus réalistes. Une extension naturelle, déjà évoquée dans la littérature, est de modéliser la covariance comme somme de produits de Kronecker (Bijma et al., 2005). Ce problème a récemment connu un net regain d'intérêt dans différents domaines du traitement du signal, et les méthodes d'estimation utilisées ici (incluant les algorithmes) ont été étendues à ce contexte (Tsiligkaridis and Hero, 2013). Un prolongement naturel de notre travail pourra être d'investiguer ces approches et d'étudier leur pertinence dans notre cadre applicatif, à la fois en termes de classification et en termes de description des signaux. On peut en attendre une amélioration des performances en classification ; il est aussi naturel de se demander si les composantes ainsi estimées pourraient être mises en correspondance avec des processus physiologiques indépendants. Ceci étant, il faut garder à l'esprit que complexifier le modèle et augmenter le nombre de paramètres à estimer se traduit aussi par une augmentation de la taille de l'échantillon nécessaire à l'estimation de ces paramètres, ce qui n'est souvent pas souhaitable. Ceci pourrait être contrôlé par l'introduction d'un *a priori* de parcimonie sur les covariances, dans l'esprit des méthodes de LASSO graphiques développées pour l'estimation de covariances sous forme Kronecker (Tsiligkaridis et al., 2013).

Une question naturelle qui vient après la modélisation de la variabilité des essais pour un sujet unique en EEG est celle de la modélisation de la variabilité inter sujets. S'il est encore difficile à ce stade d'imaginer comment le modèle matriciel pourrait s'étendre au cas multi-sujets, une des contributions de notre travail pourrait être utile dans le contexte de la construction de filtres spatiaux. L'introduction du lissage spatial par pénalisation de la matrice de covariance des électrodes pourrait permettre d'apprendre des filtres spatiaux discriminants à partir de plusieurs sujets d'une même étude. Dans le cas des interfaces cerveau-machine, cette approche permettrait de raccourcir la phase de calibration du système en exploitant, pour un nouveau participant, les informations obtenues à partir des sujets précédents.

Extensions possibles de l'approche modèle mixte. La modélisation présentée dans la partie III peut elle aussi faire l'objet d'extensions, plus ou moins directes.

Comme nous l'avons déjà signalé plus haut, le vecteur *design* joue un grand rôle dans la qualité de la modélisation. Il pilote la variabilité de la partie aléatoire du modèle, qui doit reflêter le comportement de la variabilité des signaux. Dans certains cas cette dernière peut être révélée par une analyse exploratoire et intégrée dans le modèle, mais une procédure plus automatisée est clairement souhaitable. L'heuristique que nous avons proposée, basée sur une

simple régression pénalisée, n'a pas montré sur les données considérées qu'elle pouvait apporter un plus, mais ceci demande à être étudié plus en détail, sans doute avec d'autres types de régression et de pénalisation, et peut être aussi sur d'autres jeux de données.

Compte tenu des résultats obtenus dans la partie II, il est tentant de chercher à associer l'hypothèse de séparabilité (expression de la matrice de covariance comme produit de Kronecker de deux matrices plus petites) avec le modèle mixte. En découplant l'analyse des dimensions spatiales et temporelles, comme présenté dans la partie II, une modélisation plus complexe de la composante aléatoire dans le modèle pourrait être introduite de façon à tenir compte de la variabilité spatio-temporelle inter-essais. Ceci demanderait toutefois d'importantes modifications du modèle.

Comme signalé plus haut, il est aussi tentant de chercher à tenir compte à la fois de la variabilité inter-essais et inter-sujets, comme ceci a été proposé dans (Fazli et al., 2011) dans un contexte d'interface cerveau-machine. En effet, la comparaison des potentiels d'erreur pour tous les participants semble révéler une importante similarité entre les participants, mais aussi quelques différences, à la fois en termes de topographie et de décours temporel. Il pourrait donc être envisagé d'intégrer la variabilité propre au participant au sein d'un modèle mixte plus sophistiqué, qui modéliserait l'écart du participant (en amplitude et en topographie) au comportement moyen (le grand moyennage). Le modèle pourrait inclure dans le même temps des paramètres communs et spécifiques à chaque participant. D'un point de vue statistique , l'introduction de ces paramètres communs pourrait augmenter la robustesse des estimations à partir de petits ensembles de données. Une des limitations est cependant qu'une telle approche nécessite un nombre assez important de participants afin d'estimer cette variabilité inter-participants.

Enfin, la modélisation proposée jusqu'à présent se base sur l'hypothèse sous-jacente d'additivité de l'activité de fond et du signal d'intérêt. Cependant cette hypothèse peut être remise en question. Une approche alternative serait par exemple de proposer un modèle mixte non-linéaire, modélisant le signal d'intérêt comme une modulation multiplicative de l'activité de fond.

Autres perspectives. Les deux approches proposées ont été spécialement conçues pour l'application à la classe très particulière des potentiels évoqués. Une extension à des signaux oscillants tels que les différents rythmes sensori-moteurs pourrait aussi être envisagée. Dans ce cas, il sera nécessaire de choisir des bases de décomposition des signaux plus adaptées à cette problématique, comme par exemple les bases MDCT (bases de Fourier localisées).
La modélisation spatio-fréquentielle sous forme d'un produit de Kronecker a déjà été exploitée pour l'analyse de signaux oscillants dans (Villaron, 2012; Mahanta et al., 2012). Une extension à un problème de classification pourrait être faite par l'exploitation de l'approche matricielle de l'analyse discriminante que nous avons proposée dans la partie II de cette thèse.
La seconde approche, basée sur la modélisation de la variabilité des signaux pourrait elle aussi être étendue au cas de signaux oscillants si i) la base de décomposition est bien choisie et ii) si l'analyse exploratoire des données permet de dégager des comportements particuliers pour la modélisation de la variabilité inter-essais.
En pratique, il serait par exemple intéressant de considérer des données d'imagination motrice (étude des ondes *mu* et *beta*) pour une application aux interfaces cerveau-machine.

Annexe A

Le produit de Kronecker et l'opérateur vec

Définition 11 (Le produit de Kronecker)**.** *Soient A une matrice de taille $m \times n$ et B une matrice de taille $p \times q$. Leur produit tensoriel est la matrice $A \otimes B$ de taille $mp \times nq$, définie par blocs successifs de taille $p \times q$, le bloc d'indice i, j valant $(a)_{ij} \cdot B$:*

$$A \otimes B = \begin{pmatrix} a_{11}B & \cdots & a_{1n}B \\ \vdots & \ddots & \vdots \\ a_{m1}B & \cdots & a_{mn}B \end{pmatrix}. \tag{A.1}$$

Propriété 4 (Manipuler le produit de Kronecker)**.**

$$\begin{aligned} A \otimes (B + C) &= A \otimes B + A \otimes C\,, && \text{(A.2)} \\ A \otimes B &\neq B \otimes A\,, && \text{(A.3)} \\ A \otimes (B \otimes C) &= (A \otimes B) \otimes C\,, && \text{(A.4)} \\ (A \otimes B)' &= A' \otimes B'\,, && \text{(A.5)} \\ (A \otimes B)(C \otimes D) &= AC \otimes BD\,. && \text{(A.6)} \\ & && \text{(A.7)} \end{aligned}$$

Propriété 5 (Inversibilité)**.** *Si A et B sont deux matrices inversibles, alors*

$$(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}\,. \tag{A.8}$$

Propriété 6 (Rang, déterminant et trace)**.** *Si A et B sont deux matrices carrées de dimension $m \times m$ et $p \times p$ respectivement, alors*

$$\begin{aligned} \operatorname{rang}(A \otimes B) &= \operatorname{rang}(A)\operatorname{rang}(B) = mp\,, && \text{(A.9)} \\ \det(A \otimes B) &= \det(A)^p \det(B)^m\,, && \text{(A.10)} \\ \operatorname{Tr}(A \otimes B) &= \operatorname{Tr}(A)\operatorname{Tr}(B)\,. && \text{(A.11)} \end{aligned}$$

Propriété 7. *Soient $\lambda_1, \ldots, \lambda_m \geq 0$ et $\gamma_1, \ldots, \gamma_p \geq 0$ les valeurs propres de A et B respectivement, alors pour tout $i = 1, \ldots, m$ et pour tout $j = 1, \ldots, p$, le nombre $\lambda_i \gamma_j$ est valeur propre de $A \otimes B$.*

En effet, si u_i est vecteur propre de A et v_j est vecteur de B, associés respectivement aux valeurs propres non-nulles λ_i et γ_j, alors, par associativité du produit de Kronecker, $(A \otimes B)(u_i \otimes v_j) = (\lambda_i \gamma_j)(u_i \otimes v_j)$.

Définition 12 (L'opérateur vec)**.** *Soit X une matrice de dimension $m \times n$, $x = \mathsf{vec}(X)$ est le vecteur de dimension mn obtenu par concaténation verticale des colonnes de X. Par exemple, si $X \in \mathbb{R}^{2\times 2}$, $x \in \mathbb{R}^4$, avec :*

$$X = \begin{pmatrix} x1 & x2 \\ x3 & x4 \end{pmatrix} \quad et \quad x = \begin{pmatrix} x1 \\ x3 \\ x2 \\ x4 \end{pmatrix} . \tag{A.12}$$

Propriété 8. *Pour toutes matrices A, B, C, D, E et X de dimensions correspondantes et pour toute constante α :*

$$\begin{aligned}
\mathsf{vec}(AXB) &= (B' \otimes A)\mathsf{vec}(X) && \text{(A.13)}\\
\mathsf{vec}(A+B) &= \mathsf{vec}(A) + \mathsf{vec}(B) && \text{(A.14)}\\
\mathsf{vec}(\alpha A) &= \alpha.\mathsf{vec}(A) && \text{(A.15)}\\
\mathrm{Tr}(A'B) &= \mathsf{vec}(A)'\mathsf{vec}(B) && \text{(A.16)}\\
\mathrm{Tr}(CXB) &= (\mathsf{vec}(C'))'(I \otimes X)\mathsf{vec}(B) && \text{(A.17)}\\
\mathrm{Tr}(DX'EXB) &= (\mathsf{vec}(X))'(D'B' \otimes E)\mathsf{vec}(X) && \text{(A.18)}\\
&= ((\mathsf{vec}(X))'(BD \otimes E')\mathsf{vec}(X) && \text{(A.19)}
\end{aligned}$$

Annexe B

Modèle linéaire mixte gaussien

Un modèle mixte est un modèle qui contient des effets fixes et aléatoires, c'est à dire qui prend en compte les facteurs influençant la moyenne (*effets fixes*) mais aussi la covariance (*effets aléatoires*) de la variable à expliquer. On parle d'un facteur à *effet mixte* s'il influe sur la moyenne et la covariance. A l'heure actuelle pratiquement toutes les formes de modèle statistique classique ont été étendues pour accueillir des effets aléatoires. Le modèle linéaire mixte gaussien est une extension du modèle linéaire gaussien. Par souci de simplicité nous présentons ce modèle dans le cas standard où la variable Y à expliquer est une variable aléatoire réelle.

B.1 Le modèle

Soient $y = (y_1, y_2, \ldots, y_N)' \in \mathbb{R}^N$, le vecteur constitué de N observations (pas forcement indépendantes) d'une variable aléatoire réelle à expliquer Y. Un modèle linéaire mixte gaussien intégrant p effets fixes et q effets aléatoires est un modèle de régression linéaire gaussien s'écrit :

$$y = X\beta + Zu + \varepsilon\,, \tag{B.1}$$

où

- la partie des effets fixes $X\beta$ où $X \in \mathbb{R}^{N\times p}$ est la matrice *design* (supposée connue) associée au paramètre des effets fixes $\beta \in \mathbb{R}^p$ inconnu ; les colonnes de X correspondent aux variables explicatives de la moyenne ;
- la partie des effets aléatoires Zu où $Z \in \mathbb{R}^{N\times q}$ est la matrice *design* (supposée connue) associée au vecteur des effets aléatoires $u \in \mathbb{R}^q$. On suppose que le vecteur u non directement observé est un vecteur gaussien centré de matrice de covariance $\Lambda \in \mathbb{R}^{q\times q}$ inconnue ;
- le vecteur aléatoire des erreurs du modèle $\varepsilon \sim \mathcal{N}(0, Q)$, où $Q \in \mathbb{R}^{N\times N}$ est inconnue. Il est supposé indépendant du vecteur des effets aléatoires u.

Pour simplifier la présentation, on suppose qu'il y a K facteurs à effets aléatoires indépendants, c'est à dire on décompose le terme des effets aléatoires Zu en la somme de K termes aléatoires indépendants. De plus on suppose que les résidus ε_i sont i.i.d. Le modèle se réécrit alors sous la forme suivante :

$$y = X\beta + \sum_{k=1}^{K} Z^k u^k + \varepsilon\,, \tag{B.2}$$

où

- $Z^k \in \mathbb{R}^{N\times q_k}$, $k = 1, \ldots, K$ sont des matrices connues avec $\sum_{k=1}^{K} q_k = q$;
- $u^k \in \mathbb{R}^{q_k}$, $k = 1, \ldots, K$ sont des vecteurs aléatoires indépendants de loi $\mathcal{N}(0, \sigma_k^2 I_N)$ avec les variances σ_k^2 inconnues ;
- la matrice de covariance des erreurs $Q = \sigma_\varepsilon^2 I_N$, avec la variance σ_ε^2 inconnue.

On supposera de plus la matrice X de rang plein (ceci n'est pas très restrictif car l'on peut généralement se ramener à une telle situation) .

Remarque 26. *La correspondance entre les modèles* (B.1) *et* (B.2) *est établie en posant :*

$$Z = (Z^1|Z^2|\cdots|Z^K)\,, \qquad u = ((u^1)', \cdots, (u^K)')'\,,$$

$$\Lambda = diag(\sigma_1^2 I_{q_1}, \sigma_2^2 I_{q_2}, \ldots, \sigma_q^2 I_{q_K})\,.$$

Loi du modèle

Comme $y - X\beta$ s'écrit comme la somme de vecteurs aléatoires gaussiens indépendants, le vecteur y est un vecteur gaussien. Comme $\mathbb{E}(u) = \mathbb{E}(\varepsilon) = 0$, l'espérance de y est donnée par :

$$\mathbb{E}(y) = \mathbb{E}(X\beta + Zu + \varepsilon) = X\beta\,. \tag{B.3}$$

De plus, par l'indépendance des u^k entre-eux et avec ε, la matrice de covariance de y est donnée par :

$$V = \sum_{k=1}^{K} \sigma_k^2 Z_k Z_k' + \sigma_\varepsilon^2 I_N = Z\Lambda Z' + \sigma_\varepsilon^2 I_N\,. \tag{B.4}$$

La loi de y est donc :

$$y \sim \mathcal{N}(X\beta, Z\Lambda Z' + \sigma_\varepsilon^2 I_N)\,. \tag{B.5}$$

Remarque 27. *La loi de y permet d'illustrer de quelle manière les facteurs fixes et les effets aléatoires interviennent dans la modélisation : la partie fixe définit un comportement "général" (les effets fixes intervenant au niveau de la moyenne) tandis que la partie aléatoire permet de tenir compte de différentes sources de variabilité de y. Ici selon notre modèle* (B.2)*, on décompose la variabilité de y en $K+1$ sources de variabilité indépendantes : les K facteurs plus le bruit (résidu). Chaque vecteur aléatoire u^k correspond à q_k effets aléatoires du même facteur k, considéré comme une source de la variabilité de y. On parle de décomposition de la variance en différentes composantes. Les variances σ_k^2, $k = 1, \ldots, K$, et σ_ε^2 sont appelées composantes de la variance.*

Les paramètres du modèle

D'après la définition du modèle linéaire mixte gaussien (B.2), les paramètres à estimer sont les suivants :

- le vecteur des effets fixes β associés à la matrice *design* X ;
- Les $K+1$ variances de la partie aléatoire : les K variances σ_k^2 des effets aléatoires et la variance σ_ε^2 des résidus du modèle.

Remarque 28. *Soit $q = \sum_{k=1}^{K} q_k$. Les q effets aléatoires sont régit par des lois de probabilité. On ne parlera donc pas d'estimation, mais de prédiction des effets aléatoires à partir des observations y.*

B.2 Estimation des paramètres

Dans le modèle mixte, nous devons conjointement estimer les paramètres des effets fixes et les composantes de la variance associées aux différents effets aléatoires, la difficulté résidant principalement dans l'estimation des composantes de la variance. Nous nous intéressons tout d'abord à l'estimation des effets fixes sous l'hypothèse que la matrice V est connue ; l'expression de l'estimateur β ne changeant pas lorsque V est inconnue (on remplace V par son estimateur $\hat{V}$). Puis nous considérons le problème plus difficile de l'estimation des composantes de la variance.

Le maximum de vraisemblance (ML : *Maximum likelihood*) est la méthode la plus répandue permettant d'obtenir des estimateurs dont les propriétés asymptotiques sont optimales (convergence, efficacité, normalité). Elle consiste à estimer les paramètres par les valeurs maximisant la log-vraisemblance du modèle.

La log-vraisemblance du modèle mixte gaussien (B.2) s'écrit

$$\mathcal{L}(y;\beta;V) = -\frac{1}{2}\left[\ln(2\pi) + \ln(\det(V)) + (y - X\beta)'V^{-1}(y - X\beta)\right] \tag{B.6}$$

où $V = \sum_{k=1}^{K} \sigma_k^2 Z^k (Z^k)' + \sigma_\varepsilon^2 I_N$ est la matrice de covariance de y.

B.2.1 Estimation par ML des effets fixes β dans le cas où la matrice V est supposée connue

Intéressons nous tout d'abord au cas où V est supposée connue et inversible. Il s'agit de maximiser cette fonction par rapport à β. La propriété de convexité de la log-vraisemblance $\mathcal{L}$ permet d'assurer l'unicité de la solution. La dérivée partielle de $\mathcal{L}$ par rapport à β est

$$\frac{\partial \mathcal{L}}{\partial \beta} = X'V^{-1}(y - X\beta)\,, \tag{B.7}$$

et la solution est obtenue en cherchant la valeur de β pour laquelle (B.7) s'annule. L'estimateur du maximum de vraisemblance de β est finalement :

$$\widehat{\beta} = \left(X'V^{-1}X\right)^{-1} X'V^{-1}y\,. \tag{B.8}$$

Cet estimateur est un estimateur BLUE (*Best Linear Unbiased Estimator*). Il s'agit du meilleur estimateur dans la classe des estimateurs sans biais (de variance minimale).

Remarque 29.
- *L'estimateur* (B.8) *peut-être obtenue par la méthode des moindres carrés généralisés.*
- *Lorsque* $V = \sigma_\varepsilon^2 I_N$*, (aucun effet aléatoire), on retrouve l'estimateur des moindres carrés ordinaires :* $\widehat{\beta} = (X'X)^{-1} X'y$ *.*

Variabilité et intervalles de confiance de $\widehat{\beta}$

L'estimateur BLUE des effets fixes peut s'écrire comme $\widehat{\beta} = (X'V^{-1}X)^{-1} X'V^{-1}y = Ay$ où A est une matrice fixée. Comme $y \sim \mathcal{N}(X\beta, V)$, nous pouvons donc en déduire que $\widehat{\beta}$ suit aussi une distribution gaussienne.

De plus, comme $\widehat{\beta}$ est un estimateur non biaisé de β, on a par définition :

$$\mathbb{E}(\widehat{\beta}) = \beta. \tag{B.9}$$

Sa variance est donnée par :

$$\mathbb{V}ar(\widehat{\beta}) = \left(X'V^{-1}X\right)^{-1} X'V^{-1}\mathbb{V}ar\,(y)\left(X'V^{-1}X\right)^{-1} = \left(X'V^{-1}X\right)^{-1}. \tag{B.10}$$

Sachant V, la loi de l'estimateur BLUE des effets fixes est donc :

$$\widehat{\beta} \sim \mathcal{N}\left(\beta, \left(X'V^{-1}X\right)^{-1}\right) \tag{B.11}$$

A partir de ce dernier résultat, nous pouvons construire un intervalle de confiance au niveau $1-\alpha$ pour un effet fixe β_j, pour $j = 1, \ldots, p$:

$$\beta_j \in \left[\widehat{\beta_j} - q_{1-\frac{\alpha}{2}}\sqrt{\mathbb{V}ar(\widehat{\beta_j})};\ \widehat{\beta_j} + q_{1-\frac{\alpha}{2}}\sqrt{\mathbb{V}ar(\widehat{\beta_j})}\right]. \tag{B.12}$$

où $q_{1-\frac{\alpha}{2}}$ est le quantile associé à la probabilité $1-\frac{\alpha}{2}$ d'une loi de Student à $n-p$ degrés de liberté.

Lorsque V est inconnue

Quand la matrice V n'est plus supposée connue et qu'elle est estimée, on estimera encore β à l'aide de (B.8) mais on remplacera V par son estimation $\hat{V}$:

$$\widehat{\beta}_{\hat{V}} = \left(X'\widehat{V}^{-1}X\right)^{-1}\left(X'\widehat{V}^{-1}y\right). \tag{B.13}$$

Il est clair que l'estimation de β dépendra donc de la méthode d'estimation de V. Mais l'estimateur restera un estimateur sans biais de β. En conclusion dans le modèle mixte, la difficulté au point de vue statistique n'est pas l'estimation des effets fixes mais celle des composantes de la variance.

B.2.2 Estimation des composantes de la variance

Revenons à présent sur le problème d'estimation des composantes de la variance. Il s'agit donc ici d'estimer les paramètres de variances σ_k^2, $k = 1, \ldots, K$ et σ_ε^2. Ceci revient sous nos hypothèses à estimer V la matrice de covariance de y :

$$V = \sum_{k=1}^{K} \sigma_k^2 Z_k Z_k' + \sigma_\varepsilon^2 I_N. \tag{B.14}$$

Il existe différentes méthodes d'estimation des composantes de la variance. Nous allons nous intéresser à la méthode du maximum de vraisemblance (ML) et du maximum de vraisemblance restreint (REML), largement utilisées et présentes dans tous les logiciels statistiques. Pour plus de détails nous renvoyons à (Rao and Kleffe, 1988).

Remarque 30. *Bien que moins utilisées que ML et REML, deux autres méthodes d'estimation des composantes de la variance sont disponibles dans la plupart des logiciels. Il s'agit des méthodes connues sous les noms de MINQUE (MInimium Norm Quadratic Unbiaised Estimator) (Rao, 1971a) et de MIVQUE (MInimum Variance Quadratic Unbiased Estimator) (Rao, 1971b) qui fournissent des solutions "raisonnables", à défaut d'être optimales, pour les composantes de la variance du modèle mixte. Certains logiciels proposent de les utiliser comme initialisations pour les méthodes ML et REML.*

Estimation par maximum de vraisemblance (ML)

Considérons la log-vraisemblance du modèle donnée en (B.6) et intéressons nous à l'estimation ML des paramètres de la partie fixe et de la partie aléatoire simultanément.

Du calcul des dérivées partielles selon les composantes de β, on en déduit le système de p équations (voir paragraphe précédent) :

$$\frac{\partial \mathcal{L}}{\partial \beta} = X'V^{-1}(y - X\beta) = 0\,. \tag{B.15}$$

Du calcul des dérivées partielles selon les $K+1$ composantes de la variance σ_k^2, $k = 1, \ldots, K$ et σ_ε^2 on en déduit le système à $K+1$ équations :

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \sigma_k^2} &= -\frac{1}{2}\text{Trace}(V^{-1}Z^k(Z^k)') + \frac{1}{2}(y - X\beta)'V^{-1}Z^k(Z^k)'V^{-1}(y - X\beta) = 0 \quad \text{pour } k = 1, \cdots K\,,\\
\frac{\partial \mathcal{L}}{\partial \sigma_\varepsilon^2} &= -\frac{1}{2}\text{Trace}(V^{-1}) + \frac{1}{2}(y - X\beta)'V^{-1}V^{-1}(y - X\beta) = 0\,.
\end{aligned}$$

Le système à $p + K + 1$ inconnues et à $p + K + 1$ équations non linéaires obtenu est résolu par des méthodes numériques itératives. Généralement ces procédures donnent également une estimation de la matrice de covariance asymptotique des estimateurs via le calcul de la matrice des dérivées secondes.

Remarque 31. *Les estimateurs de la variance obtenus par ML sont généralement biaisés (même s'ils sont asymptotiquement sans biais). Le fait qu'il y ait mélange de paramètres de la partie fixe et de la partie aléatoire dans cette méthode d'estimation produit un biais systématique dans l'estimation des composantes de la variance. Ce biais est similaire à celui de l'estimateur ML de la variance de l'erreur dans un modèle de régression linéaire classique. La méthode ML estime σ_ε^2 en considérant que le nombre de degrés de liberté est N le nombre d'observations alors que lorsque $\beta \in \mathbb{R}^p$ est inconnu, le nombre de degrés de liberté est $N - p$. Plus simplement ce biais est similaire à celui de l'estimateur empirique de la variance dans un série statistique (N au dénomiteur) dans le cas où la moyenne de la série est inconnue, l'estimateur non biaisé est obtenu en prenant $N - 1$ au dénominateur. La méthode du maximum de vraisemblance restreint (REML) présentée dans la suite permet de séparer l'estimation des paramètres de la partie fixe et ceux de la partie aléatoire.*

Estimation par maximum de vraisemblance restreint (REML)

L'idée est donc ici de séparer l'estimation des deux types de paramètres et de se concentrer sur l'estimation des paramètres de la variance. Pour cela on introduit une vraisemblance restreinte, correspondant à la vraisemblance du projeté de y sur un sous-espace de $\mathbb{R}^N$ sur lequel

la partie fixe s'annule.

Rappelons que nous supposons que $X \in \mathbb{R}^{N\times p}$ est de rang plein avec $N > p$, c'est à dire que $\mathsf{rang}(X) = p$. Les vecteurs colonnes de X, notés $X^j \in \mathbb{R}^N$, $j = 1, \ldots, p$, engendrent donc un sous-espace de $\mathbb{R}^N$ de dimension p, noté E_p. Le projecteur orthogonal de $\mathbb{R}^N$ dans E_p a pour matrice associée

$$H = X(X'X)^{-1}X' .$$

L'idée est donc de projeter le vecteur des observations $y \in \mathbb{R}^N$ sur le sous-espace vectoriel supplémentaire orthogonal à E_p de dimension $N - p$, noté F_{N-p} ; le projecteur orthogonal de $\mathbb{R}^N$ dans F_{N-p} ayant pour matrice $H^\perp = I_N - H$. En effet, étant donné que $H^\perp X = O_{N\times N}$, le vecteur $H^\perp y$ ne dépend plus du paramètre β. Le vecteur $H^\perp y \in \mathbb{R}^N$ est un vecteur gaussien centré ($H^\perp X\beta = 0_N$) de matrice de covariance $H^\perp V(H^\perp)' \in \mathbb{R}^{N\times N}$ de rang $N - p$. Cette matrice est donc singulière, et le vecteur projeté est un vecteur gaussien de $\mathbb{R}^N$ "dégénéré". Il doit être transformé par une matrice $A \in \mathbb{R}^{(N-p)\times N}$ de rang $N - p$ pour s'écrire comme un vecteur de $\mathbb{R}^{N-p}$. Ainsi, si on pose $T = AH^\perp$, le vecteur $Ty \in \mathbb{R}^{N-p}$ est un vecteur gaussien centré et de matrice de covariance $V^1 = TVT'$ inversible. On montre que

$$V^1 = \sum_{k=1}^{K} \sigma_k^2 TZ^k(Z^k)'T' + \sigma_\varepsilon^2 TT' .$$

La vraisemblance restreinte est la vraisemblance du vecteur gaussien Ty et ne dépend que des composantes de la variance. Notons que cette vraisemblance ne dépend du choix de la matrice A qu'à travers l'ajout d'une constante et donc ce choix n'intervient pas dans le problème de maximisation.

Ainsi une fois les composantes de la variance estimées par REML, on en déduit l'estimation de la matrice de covariance V

$$\widehat{V}_{REML} = \sum_{k=1}^{K} \hat\sigma_k^2 Z^k(Z^k)' + \hat\sigma_\varepsilon^2 TT' ,$$

et l'estimation de β par (B.13).

Remarque 32. *En pratique, quelle que soit la méthode d'estimation mise en œuvre, il peut arriver d'obtenir des valeurs nulles et même parfois négatives pour certaines composantes de la variance. Les valeurs négatives sont forcées à zéro.*

Tests de l'effet aléatoire du facteur

Pour tester si le $k$ème facteur influe significativement sur la variabilité de y, on teste la nullité de la variance σ_k^2 associée :

$$\mathsf{H_0} : \sigma_k^2 = 0 \quad \text{contre} \quad \mathsf{H_1} : \sigma_k^2 > 0 . \tag{B.16}$$

Un test classiquement utilisé est celui de Wald dont la statistique donnée par :

$$w = \frac{\hat\sigma_k^4}{\widehat{var}(\hat\sigma_k^2)} ,$$

suit asymptotiquement une loi du χ^2 à 1 degré de liberté. Il est à noter que ce test est très peu puissant et qu'il doit utilisé comme un indicateur.

Tester l'effet aléatoire d'un facteur dans le modèle mixte est un problème délicat. Une revue des principaux tests et de leurs propriétés est par exemple donnée dans (**?**).

B.3 Prédiction des effets aléatoires

B.3.1 Espérance conditionnelle et BLUP

Nous nous intéressons à présent à prédire les vecteurs des effets aléatoires $u_i \in \mathbb{R}^q$ pour $i = 1, \dots N$. Ces effets ne sont pas directement observés et nous devons les prédire *a posteriori* au vu des observations y_i. Il s'agit de trouver l'espérance conditionnelle de u_i sachant y_i qui dans le cas gaussien fournit le meilleur prédicteur au sens de l'erreur quadratique moyenne (BLUP : *Best Linear Unbiased Predictor*).

Rappels sur les vecteurs gaussiens. Soit $(U_1, U_2)'$ un vecteur gaussien de moyenne $(\mu_1, \mu_2)'$ et de matrice de covariance

$$\begin{pmatrix} V_1 & Cov_{12} \\ Cov_{21} & V_2 \end{pmatrix} .$$

La loi conditionnelle de U_1 sachant U_2 est définie par :

$$\mathcal{N}\left(\mu_1 + Cov_{12}V_2^{-1}(u_2 - \mu_2); V_1 - Cov_{12}V_2^{-1}Cov_{21}\right) . \tag{B.17}$$

Sous les hypothèses du modèle, $(u_i, y_i)'$ est un vecteur gaussien de $\mathbb{R}^{q+1}$:

$$(u_i, y_i)' \sim \mathcal{N}\left(\begin{pmatrix} 0_q \\ X(i,.)\beta \end{pmatrix}; \begin{pmatrix} \Lambda & \Lambda Z(i,.)' \\ Z(i,.)\Lambda & V(i,i) \end{pmatrix}\right) , \tag{B.18}$$

où

- $X(i,.)$ est la $i$ème ligne de la matrice X,
- $Z(i,.)$ est la $i$ème ligne de la matrice Z,
- $V(i,i)$ est la variance de y_i donnée par le $i$ème terme diagonal de la matrice de covariance de V .

A partir de (B.17) nous obtenons la loi conditionnelle de u_i sachant y_i et l'estimateur BLUP est donné par la moyenne de cette loi qui est l'espérance conditionnelle de u_i sachant y_i :

$$\mathbb{E}(u_i|y_i) = \Lambda Z(i,.)'V(i,i)^{-1}(y_i - X(i,.)\beta) , \tag{B.19}$$

où $\Lambda = diag(\sigma_1^2, \cdots, \sigma_K^2)$ et $V(i,i) = \sum_{k=1}^K \sigma_k^2 Z^k(Z^k)'(i,i) + \sigma_\varepsilon^2$.

La prédiction du vecteur des effets aléatoires u_i est obtenue en remplaçant les paramètres β, $\sigma_1^2, \cdots, \sigma_K^2$ et σ_ε^2 par leurs estimations dans (B.19) :

$$\widehat{u_i} = \widehat{\Lambda} Z(i,.)'\widehat{V}(i,i)^{-1}(y_i - X_i\widehat{\beta}) . \tag{B.20}$$

On montre que la variabilité u_i est plus grande que celle de sa prédiction.

B.3.2 Equations du modèle mixte d'Henderson

Henderdon a développé une méthode permettant d'obtenir conjointement l'estimation de β et les prédictions des effets aléatoires u_i (voir par exemple (Henderson, 1982)) par résolution d'un système linéaire. Cette approche s'effectue par l'optimisation de la log-vraisemblance conjointe de y et de u par rapport aux effets fixes β et aléatoires u.

La loi conjointe de (y, u) est donnée par

$$f(y,u) = f(y|u)f(u)\,, \tag{B.21}$$

où

$$f(y|u) = 2\pi^{-\frac{n}{2}} \det(Q)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(y - X\beta - Zu)'Q^{-1}(y - X\beta - Zu)\right)\,, \tag{B.22}$$

$$f(u) = 2\pi^{-\frac{q}{2}} \det(\Lambda)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}u'\Lambda^{-1}u\right)\,. \tag{B.23}$$

Plaçons nous dans le cas où les matrices de covariance Λ et Q sont connues.

La log-vraisemblance $\mathcal{L}(y, u)$ est obtenue par la somme des deux termes suivants :

$$\mathcal{L}(y|u) = -\frac{1}{2}(n\ln(2\pi) + \ln(\det(Q)) + (y - X\beta - Zu)'Q^{-1}(y - X\beta - Zu)) \tag{B.24}$$

$$\mathcal{L}(u) = -\frac{1}{2}(q\ln(2\pi) + \ln(\Lambda) + u'\Lambda^{-1}u) \tag{B.25}$$

Ainsi, la maximisation de la log-vraisemblance est comme précédemment obtenu *via* l'annulation des dérivées partielles :

$$\frac{\partial(\mathcal{L}(y,u))}{\partial\beta} = -2X'Q^{-1}(y - X\beta - Zu) = 0 \tag{B.26}$$

$$\frac{\partial(\mathcal{L}(y,u))}{\partial u} = Z'Q^{-1}(y - X\beta - Zu) + 2\Lambda^{-1}u = 0 \tag{B.27}$$

On obtient un système à $p + q$ équations linéaires avec $p + q$ inconnues, ces équations sont appelées les équations d'Henderson.

Les équations du modèles mixtes d'Henderson. Si les matrices de covariance Q et Λ sont connues, l'estimation de β et la prédiction de u sont les solutions du système suivant :

$$\begin{pmatrix} X'Q^{-1}X & X'R^{-1}Z \\ Z'Q^{-1}X & Z'Q^{-1}Z + \Lambda^{-1} \end{pmatrix} \begin{pmatrix} \widehat{\beta} \\ \widehat{u} \end{pmatrix} = \begin{pmatrix} X'Q^{-1}y \\ Z'Q^{-1}y \end{pmatrix} \tag{B.28}$$

Lorsque les matrices Q et Λ sont inconnues, comme c'est le cas en général, on les remplace par leur estimateurs dans la solution. L'estimateur de β et la prédiction de u obtenus via ce procédé correspondent respectivement à l'estimateur BLUE et au prédicteur BLUP présentés précédemment.

Annexe C

Résultats supplémentaires

Cette annexe contient un ensemble de résultats obtenus sur les deux jeux de données et pour les différentes méthodes proposées dans cette thèse et qui ne sont pas présentées dans le document principal.

C.1 Filtres spatiaux discriminants - Potentiels d'erreur

Nous présentons ici les résultats supplémentaires de filtres spatiaux discriminants pour les données de potentiels d'erreur présentés dans la section 6.1.2. Les éboulis des valeurs propres ainsi que les topographies des trois premiers filtres spatiaux discriminants sont représentés pour la LDA matricielle proposée, pour sa version pénalisée et pour l'approche classique ("LDA matricielle I_K"). Les résultats fournis dans la suite sont ceux des sujets B et H du jeu de données. Les filtres présentés ont été déterminés à partir d'un échantillon d'apprentissage de 600 essais *Correct* et de 60 essais *Erreur*. Le paramètre de pénalisation a été quant à lui appris sur une échantillon supplémentaire de 30 essais de chaque classe.

Éboulis des valeurs propres

Les éboulis des valeurs propres obtenues à partir des trois méthodes de filtrage spatial basé sur l'analyse discriminante : l'approche classique (LDA mat. I_K), l'approche proposée (LDA matricielle) et sa version pénalisée (LDA mat. pénalisée) sont représentés pour le sujet B et H dans la figure C.1. Notons que dans l'exemple présenté pour le sujet B, la valeur optimale du paramètre de pénalisation est $\alpha_{opt} = 0$. Ainsi l'approche pénalisation coïncide avec l'approche matricielle. Ceci explique la superposition des valeurs des deux méthodes qui sont équivalentes dans ce cas.

Topographies des filtres spatiaux

Les topographies des trois premiers filtres spatiaux discriminants associés au éboulis des valeurs propres présentées dans la figure C.1. On retrouve pour le sujet B l'équivalence entre les deux méthodes proposées puisque la valeur optimale du paramètre de pénalisation est nulle.

Discussion

Comme pour le participant A, dont les résultats sont présentés dans la section 6.1.2 du document principal, Notons que l'ensemble des filtres spatiaux semble indiquer qu'une importante part de la discrimination entre les erreurs et les réponses correctes se fait au niveau de la région fronto-centrale. De plus, l'approche pénalisée permet , lorsque le paramètre de pénalisation n'est pas nul, de concentrer l'information discriminante dans le premier filtre.

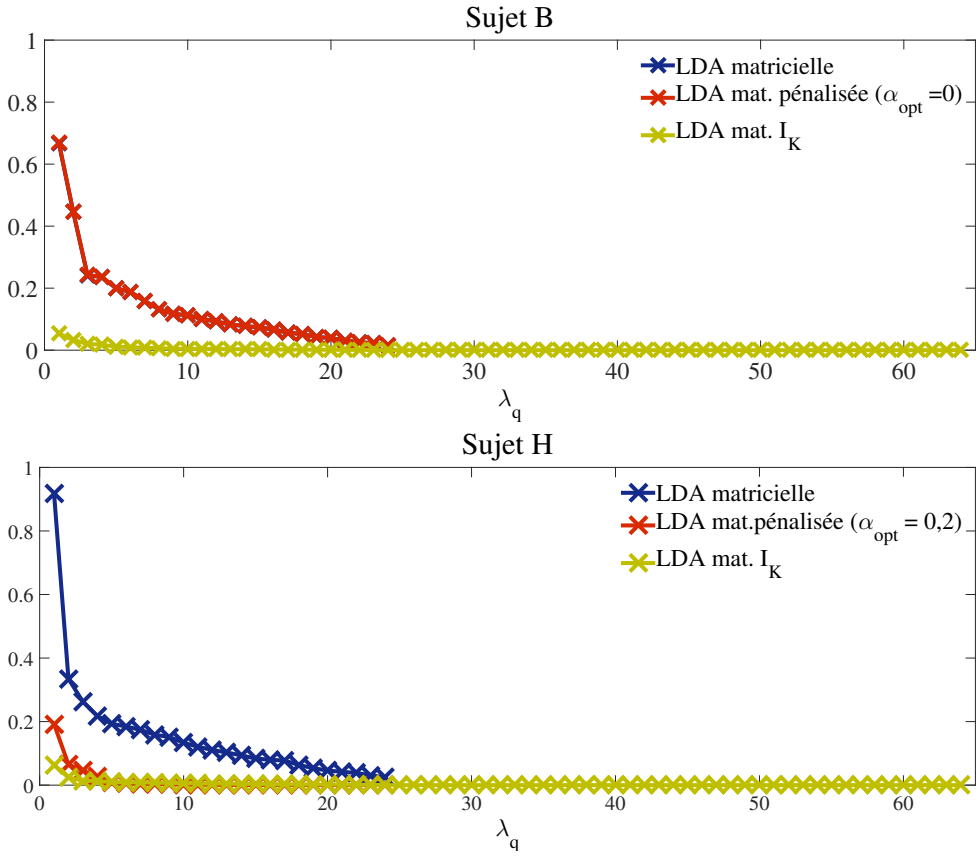


FIGURE C.1 – Éboulis des valeurs propres.

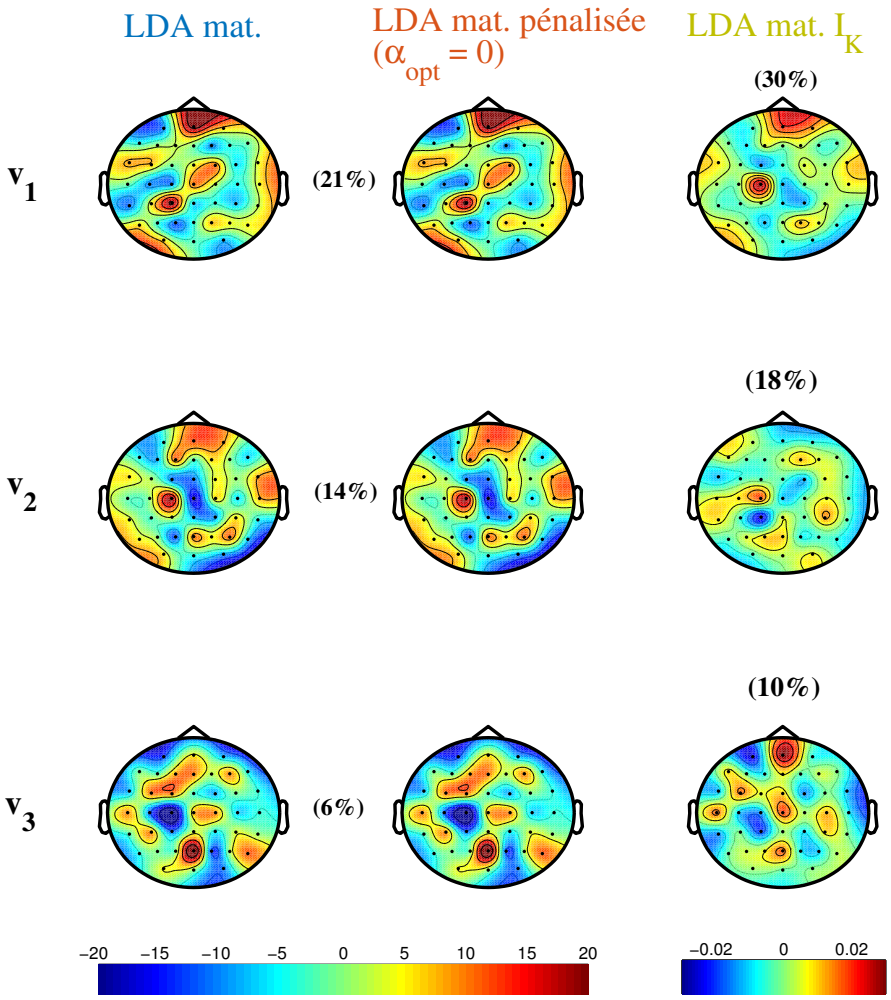


FIGURE C.2 – Sujet B. Filtres spatiaux discriminants

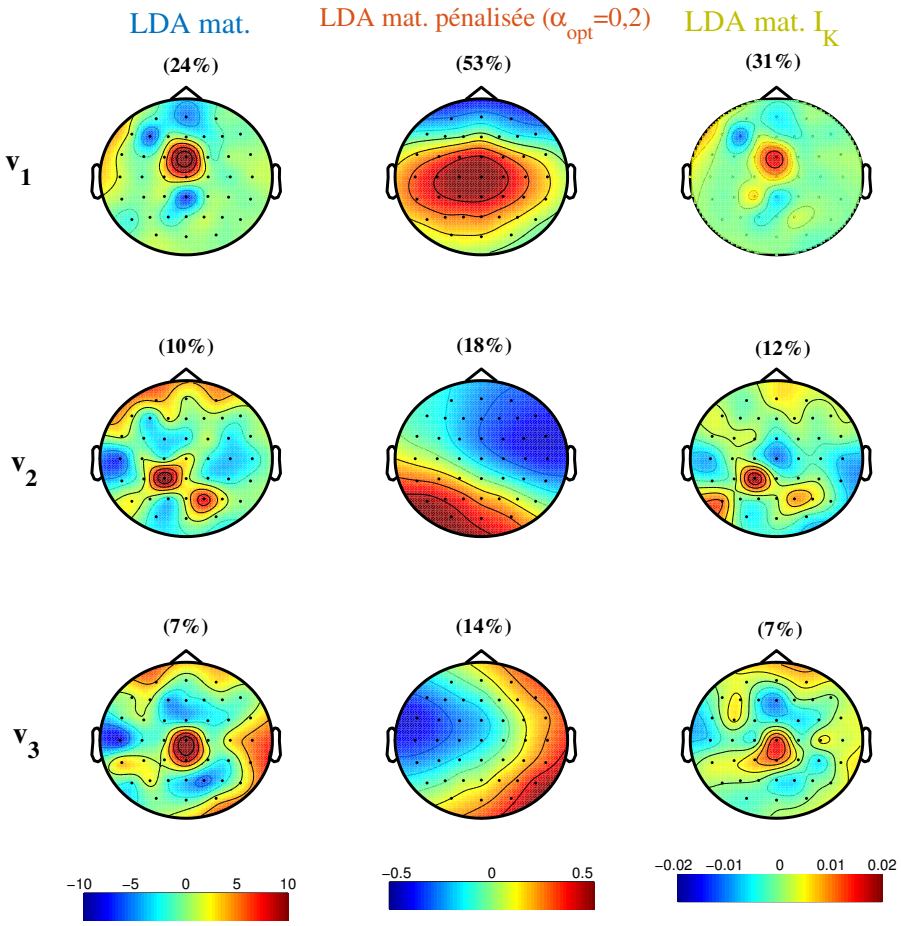


FIGURE C.3 – Sujet H. Filtres spatiaux discriminants

C.2 Résultats LDA matricielle - P300 Speller

Les résultats présentés dans cette section sont les résultats de classification obtenus sur les données du protocole P300 Speller pour les sujets 8, 9 et 10 de l'expérience présentée dans la section 2.4.2. Ils viennent en complément des résultats de classification présentés pour 4 autres sujets dans la section 6.2.2 du document principal.

Nous rappelons que chaque méthode a été évaluée à partir d'un ensemble d'apprentissage composée de 20 lettres, d'un ensemble de test de 30 lettres. L'approche matricielle que nous avons proposé pour l'extraction et la classification de caractéristiques discriminantes est systématiquement comparée à la LDA, la LDA diagonale (DLDA) et la LDA régularisée (RDA). Nous étudions également les performances de notre méthode en fonction du nombre de vecteurs propres $\tilde{u}_q$ et $\tilde{v}_q$ conservés pour la classification.

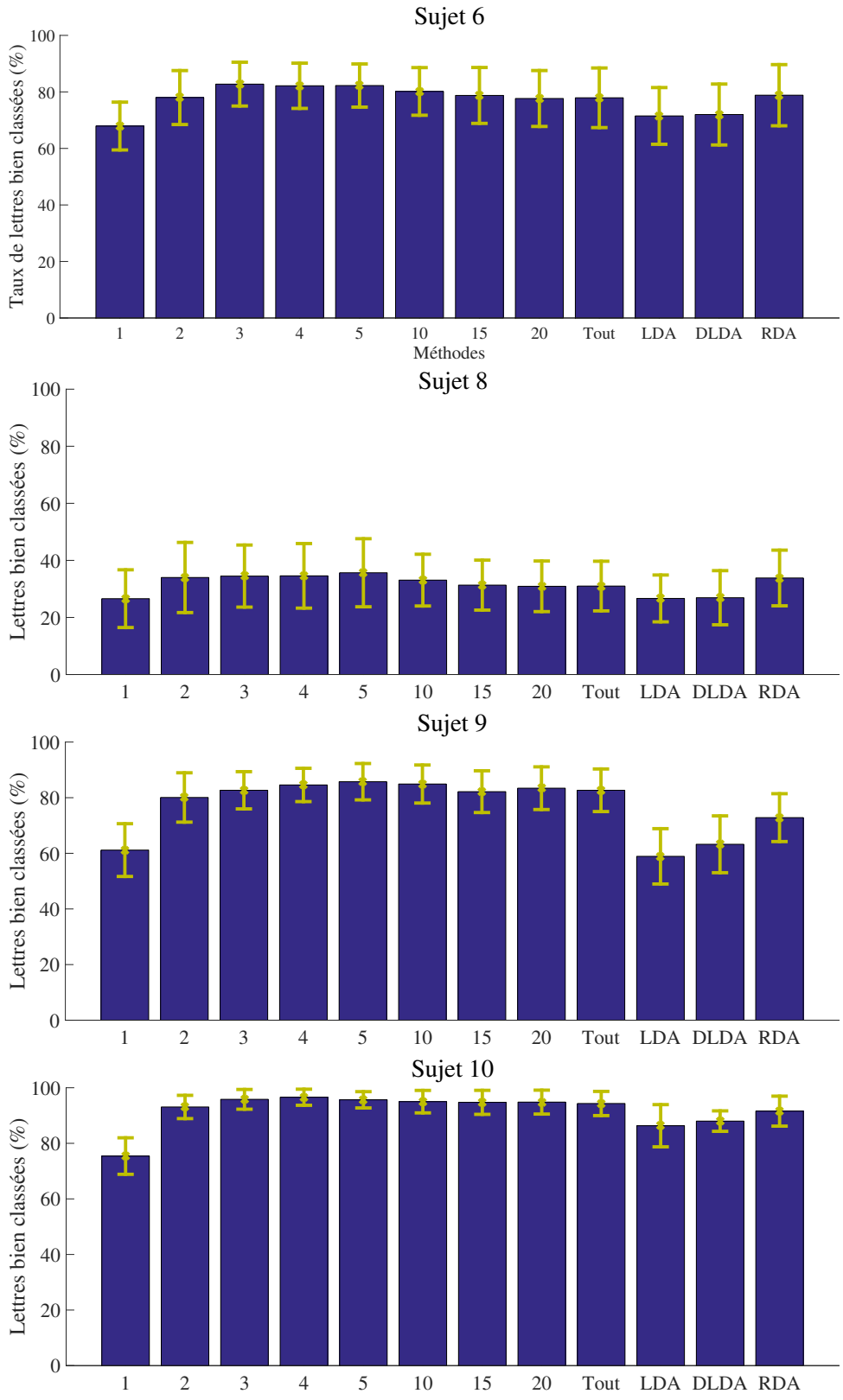


FIGURE C.4 – Taux de lettres correctement classées. Chaque barre correspond au taux moyen calculé sur 30 itérations et les axes verticaux représentent l'écart-type. Les résultats de la LDA matricielle sont présentés en fonction du nombre r de composantes discriminantes sélectionnées $\tilde{u}_q$ et $\tilde{v}_q$, $q = 1, \ldots, r$ (9 premières barres) et comparés à la LDA, la LDA diagonale (DLDA) et à la LDA régularisée (RDA).

C.3 Classification de potentiels d'erreur avec l'approche modèle mixte

C.3.1 Résultats obtenus pour l'ensemble des sujets de l'étude de la section 8.1.3

	Comparaison des classifieurs - Participant B			
$Err - Corr$	LMM-LDA	LMM-RDA	LMM-DLDA	LMM-DQDA
20-200	0.04 (10^{-5})	0.22 (10^{-18})	0.24 (10^{-18})	0.29 (10^{-18})
30-300	0.16 (10^{-17})	0.20 (10^{-18})	0.17 (10^{-18})	0.26 (10^{-18})
40-400	0.14 (10^{-17})	0.14 (10^{-18})	0.11 (10^{-17})	0.18 (10^{-18})
50-500	0.19 (10^{-18})	0.15 (10^{-18})	0.09 (10^{-16})	0.14 (10^{-18})
60-600	0.16 (10^{-18})	0.13 (10^{-7})	0.05 (10^{-7})	0.09 (10^{-14})
70-700	0.13 (10^{-16})	0.10 (10^{-14})	0.003 (0.97)	0.03 (10^{-3})
80-800	0.14 (10^{-15})	0.10 (10^{-13})	0.004 (0.28)	0.01 (0.37)
90-900	0.12 (10^{-10})	0.08 (10^{-7})	−0.01 (0.31)	0.005 (0.92)

	Comparaison des classifieurs - Participant C			
$Err - Corr$	LMM-LDA	LMM-RLDA	LMM-DLDA	LMM-DQDA
20-200	0.03 (10^{-2})	0.25 (10^{-18})	0.29 (10^{-18})	0.32 (10^{-18})
30-300	0.18 (10^{-16})	0.19 (10^{-17})	0.20 (10^{-17})	0.23 (10^{-18})

	Comparaison des classifieurs - Participant D			
$Err - Corr$	LMM-LDA	LMM-RDA	LMM-DLDA	LMM-DQDA
20-200	0.28 (10^{-17})	0.35 (10^{-18})	0.39 (10^{-18})	0.42 (10^{-18})
30-300	0.31 (10^{-18})	0.23 (10^{-18})	0.21 (10^{-17})	0.21 (10^{-16})

	Comparaison des classifieurs - Participant G			
$Err - Corr$	LMM-LDA	LMM-RDA	LMM-DLDA	LMM-DQDA
20-200	-0.02 (0.25)	0.25 (10^{-18})	0.26 (10^{-18})	0.35 (10^{-18})
30-300	0.15 (10^{-13})	0.20 (10^{-18})	0.16 (10^{-15})	0.23 (10^{-17})
40-400	0.14 (10^{-14})	0.14 (10^{-17})	0.07 (10^{-8})	0.12 (10^{-15})
50-500	0.10 (10^{-13})	0.09 (10^{-16})	0.003 (0.98)	0.03 (10^{-3})
60-600	0.04 (10^{-4})	0.05 (10^{-7})	-0.04 (10^{-6})	-0.03 (10^{-5})
70-700	0.01 (0.61)	0.02 (0.42)	-0.07 (10^{-12})	-0.08 (10^{-13})
80-800	0.02 (0.41)	0.01 (0.61)	-0.08 (10^{-13})	-0.09 (10^{-13})

	Comparaison des classifieurs - Participant H			
$Err - Corr$	LMM-LDA	LMM-RDA	LMM-DLDA	LMM-DQDA
20-200	-0.01 (0.63)	0.25 (10^{-18})	0.25 (10^{-18})	0.27 (10^{-18})
30-300	0.21 (10^{-18})	0.25 (10^{-18})	0.23 (10^{-18})	0.26 (10^{-18})
40-400	0.20 (10^{-18})	0.20 (10^{-18})	0.15 (10^{-17})	0.16 (10^{-18})
50-500	0.20 (10^{-18})	0.18 (10^{-18})	0.12 (10^{-15})	0.11 (10^{-14})
60-600	0.18 (10^{-18})	0.15 (10^{-18})	0.09 (10^{-15})	0.07 (10^{-10})
70-700	0.15 (10^{-17})	0.13 (10^{-17})	0.06 (10^{-10})	0.04 (10^{-6})
80-800	0.16 (10^{-17})	0.14 (10^{-17})	0.06 (10^{-10})	0.03 (10^{-3})
90-900	0.14 (10^{-18})	0.12 (10^{-17})	0.04 (10^{-7})	0.01 (0.68)
100-1000	0.11 (10^{-17})	0.09 (10^{-15})	0.01 (10^{-2})	-0.03 (10^{-3})
110-1100	0.10 (10^{-12})	0.08 (10^{-12})	-0.001 (0.78)	-0.04 (10^{-4})
120-1200	0.12 (10^{-14})	0.09 (10^{-13})	0.09 (0.44)	-0.04 (10^{-3})

	Comparaison des classifieurs - Participant I			
$Err - Corr$	LMM-LDA	LMM-RDA	LMM-DLDA	LMM-DQDA
20-200	0.40 (10^{-18}))	0.42 (10^{-18})	0.37 (10^{-18})	0.46 (10^{-18})
30-300	0.37 (10^{-18})	0.31 (10^{-18})	0.20 (10^{-18})	0.24 (10^{-18})
40-400	0.29 (10^{-18})	0.23 (10^{-18})	0.10 (10^{-18})	0.09 (10^{-16})
50-500	0.23 (10^{-18})	0.19 (10^{-18})	0.08 (10^{-16})	0.04 (10^{-10})
60-600	0.18 (10^{-18})	0.16 (10^{-18})	0.04 (10^{-11})	0.002 (0.24)
70-700	0.16 (10^{-18})	0.15 (10^{-18})	0.02 (10^{-2})	-0.02 (10^{-6})
80-800	0.15 (10^{-17})	0.15 (10^{-17})	0.01 (0.55)	-0.02 (10^{-7})
90-900	0.14 (10^{-12})	0.13 (10^{-12})	-0.01 (10^{-8})	-0.02 (10^{-10})

	Comparaison des classifieurs - Participant J			
$Err - Corr$	LMM-LDA	LMM-RDA	LMM-DLDA	LMM-DQDA
20-200	0.24 (10^{-17})	0.31 (10^{-18})	0.35 (10^{-18})	0.39 (10^{-18})
30-300	0.30 (10^{-18})	0.26 (10^{-18})	0.28 (10^{-18})	0.26 (10^{-18})
40-400	0.27 (10^{-18})	0.21 (10^{-18})	0.20 (10^{-17})	0.16 (10^{-16})
50-500	0.22 (10^{-16})	0.17 (10^{-16})	0.15 (10^{-15})	0.10 (10^{-7})

C.3.2 Test d'égalité des matrices de covariance

Les résultats du test de Box sont donnés pour tous les participants du jeu de données dans la table C.1. Lorsque le test n'a pas pu être calculé (trop petits échantillons pour estimer les matrices de covariance), ce est indiquer par le terme "NA" dans la table. Lorsque la statitsique de tes peut être calculée, le résultat du test permet de rejeter H_0, l'hypothèse d'égalité des matrices de covariance Σ^{Err} et Σ^{Corr}.

		Résultats du test		
Participant	n_1/n_2	Statistique	p-valeur	Inégalité des covariances ?
A	130/130	3273.05	$\ll 10^{-6}$	Oui
B	90/90	1945.26	$<< 10^{-6}$	Oui
C	40/40	NA	NA	NA
D	30/30	NA	NA	NA
H	120/120	1699.47	$<< 10^{-6}$	Oui
I	90/90	1978.9	$<< 10^{-6}$	Oui
J	50/50	2184.4	0.001	Oui

TABLE C.1 – Résultats du test de Box. $J = 2$, $K = 24$. lorsque n_1 et n_2 sont inférieurs à $KJ = 48$, la statistique de test ne peut pas être calculée (NA dans la table).

C.3.3 Ajustements essai par essai obtenus pour tous les participants (complément de la section 8.1.4)

Comme décrit dans la section 8.1.4, pour chaque essai unique une estimation du signal d'intérêt peut être obtenue à partir du modèles estimés. Après avoir projeté les données ajustées dans l'espace des électrodes et effectué une transformation en ondelettes inverse, nous obtenons le décours temporels des signaux ajustés. Dans les figures C.5 et C.6 nous représentons des exemple de signaux ajustés sur l'électrode FCz pour tous les participants de l'étude.

C.3.4 Choix du filtre pour la transformation en ondelettes

Dans la figure C.7 nous présentons des comparaisons systématiques pour 4 filtres d'ondelettes différents en termes de taux de classification pour les 5 méthodes considérées dans l'application (voir C.3) et 3 tailles différentes des échantillons d'apprentissage (cas déséquilibré). Quelle que soit la situation considérée, notre méthode (LMM) est celle qui fournit les meilleurs résultats, ce qui permet de conclure sur la consistance de notre modèle au choix du filtre pour le traitement en ondelettes.

Les ajustements des essais uniques par le modèle sont représentés dans la figure pour les 4 filtres d'ondelettes étudiés dans la figure C.7. Notons que plus la taille du filtre est grande, plus la reconstruction associée est lisse. En particulier, le filtre de Haar donne des reconstructions discontinues.

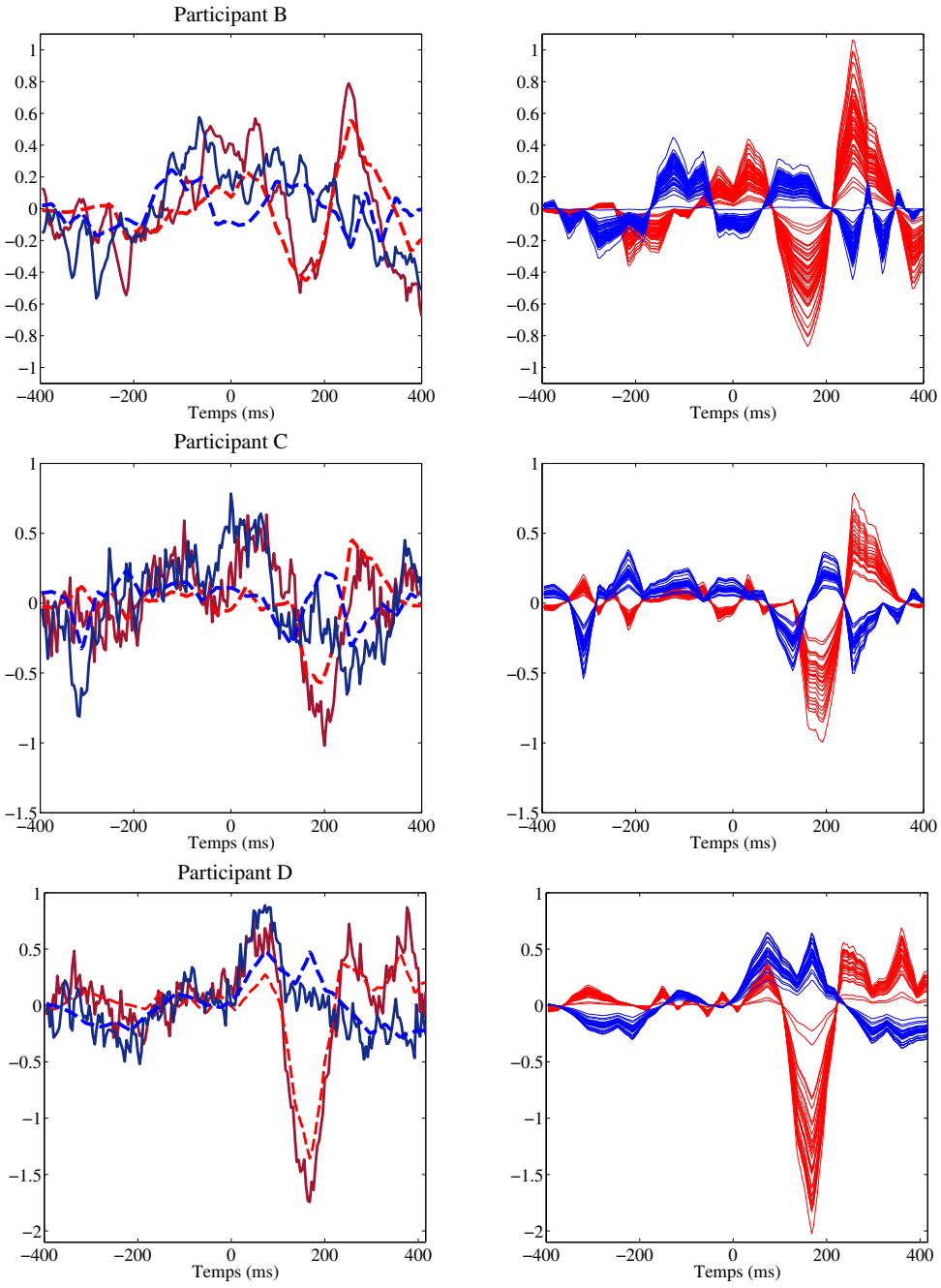


FIGURE C.5 – Ajustement des essais uniques

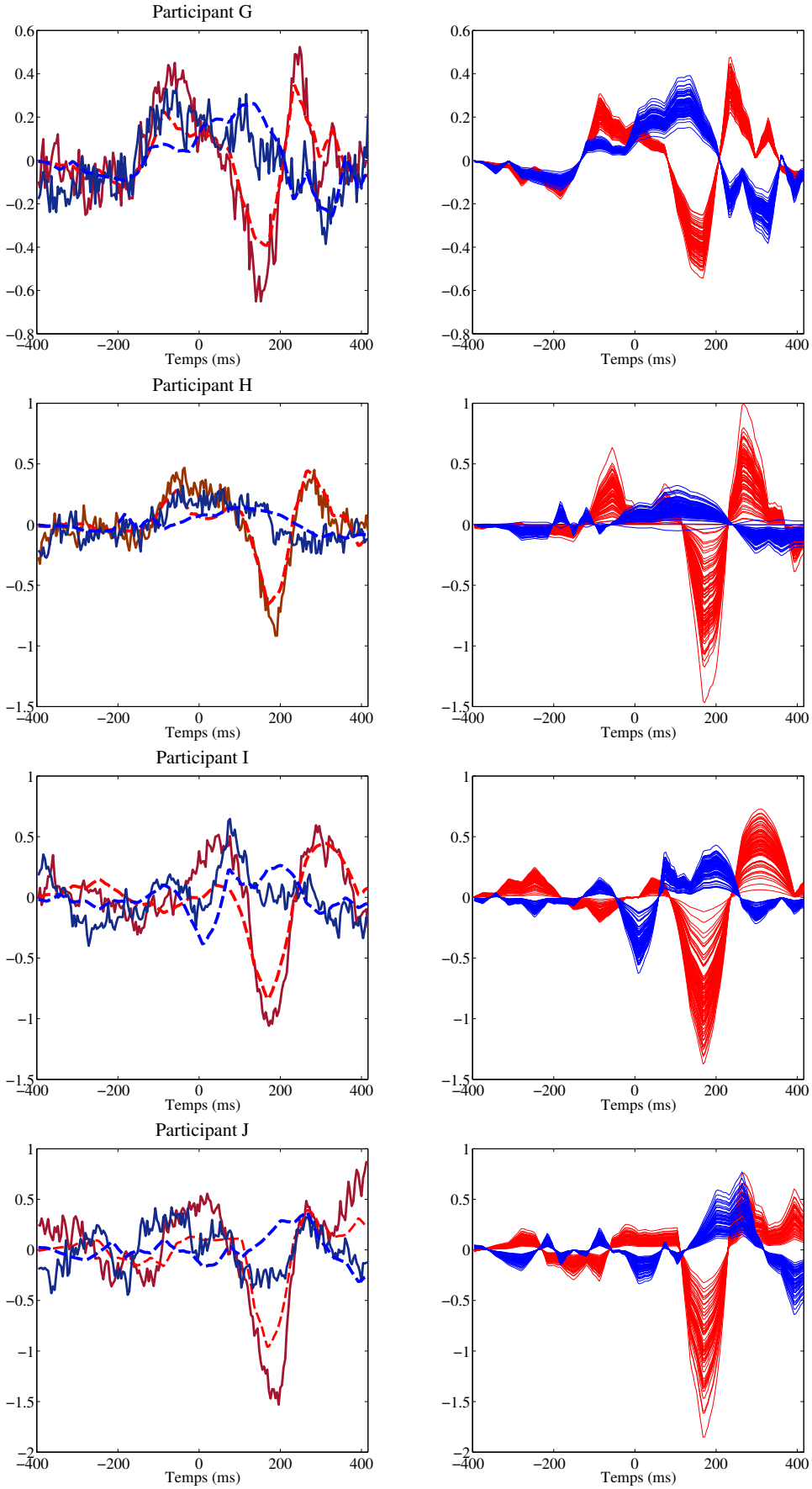


FIGURE C.6 – Ajustement des essais uniques

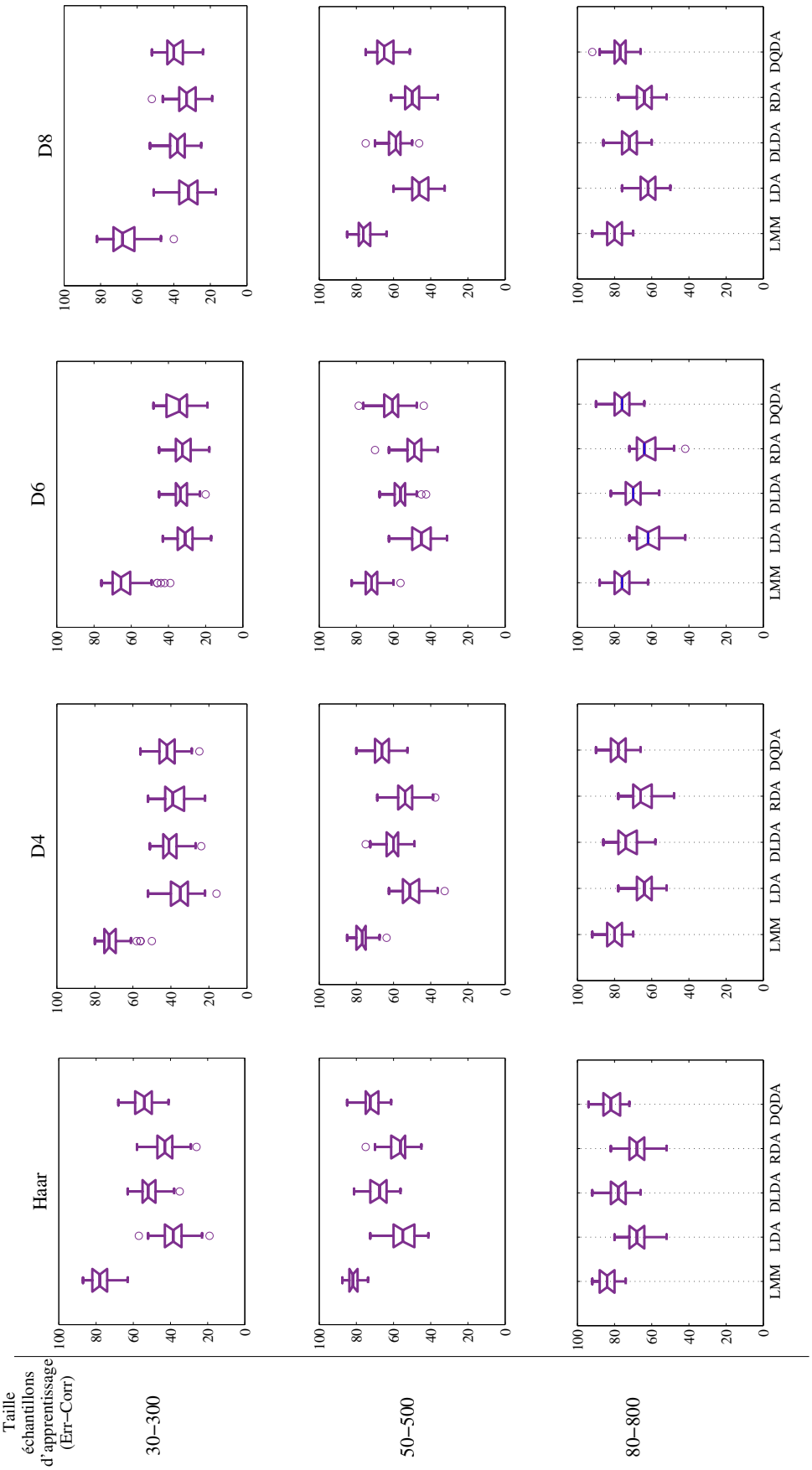


FIGURE C.7 – Participant A. Taux de bonne classification en fonction du choix de l'ondelette mère. Les résultats sont présentés pour différentes tailles d'échantillon d'apprentissage *Erreur-Correct*(30-300, 50-500 and 80- 800). Dans tous les cas, LMM donne les résultats les meilleurs.

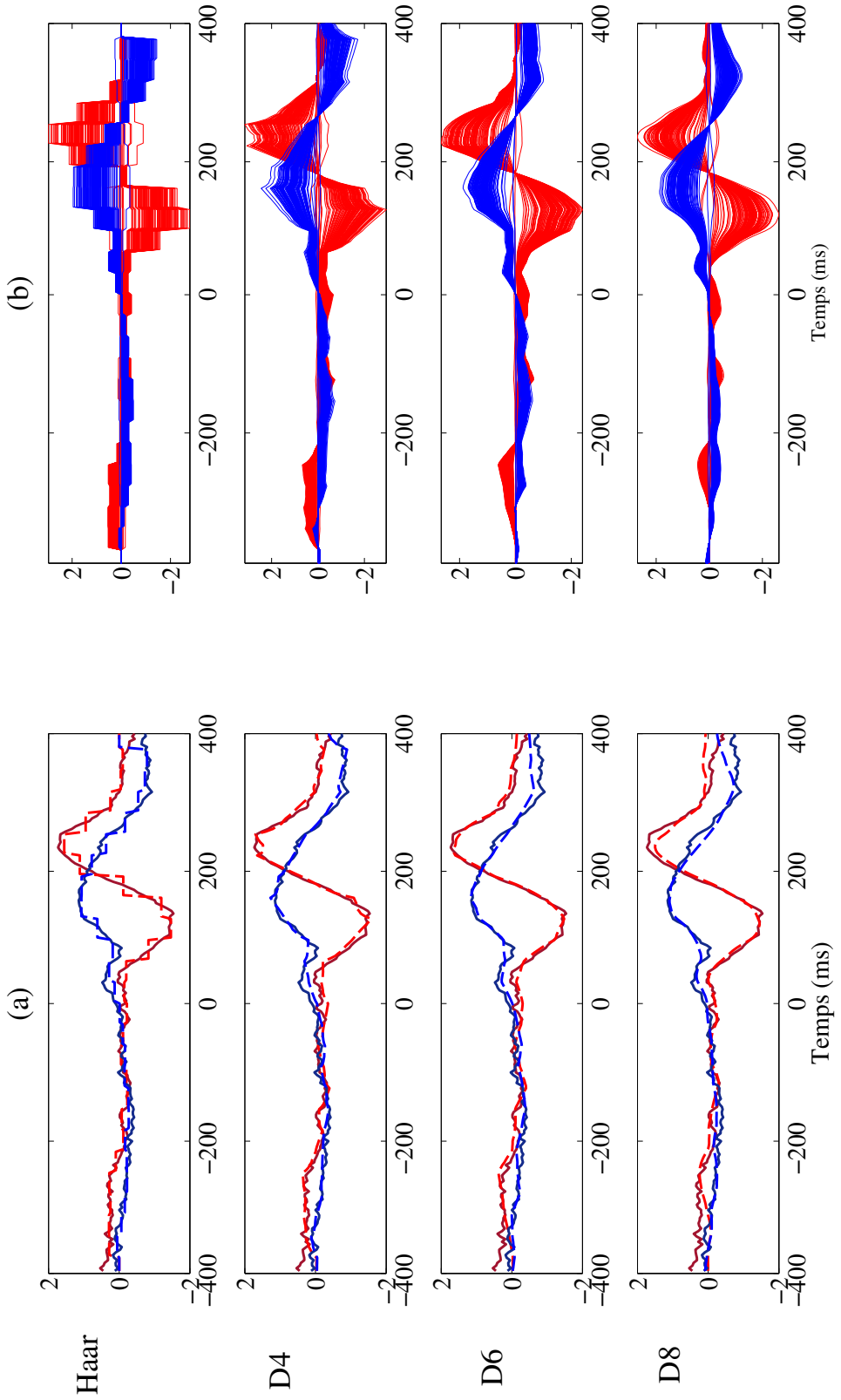


FIGURE C.8 – Participant A. Reconstrucion des signaux essai par essai par le modèle ajusté en fonction du choix de l'ondelette mère. (a) Moyennes des essais *Erreur* (rouge) et *Correct* (bleu) représentées sur l'électrode FCz. Traits pleins : moyennes des essais bruts dans chaque classe, tirets : moyenne des essais ajustés. (b) Représentation des essais uniques ajustés par le modèle dans les deux classes.

C.3.5 Détermination des vecteurs design des effets aléatoires

Dans le cas des potentiels d'erreur étudiés dans cette section, le choix du Γ^c défini comme la moyenne empirique dans la classe c semble pertinent. Nous allons néanmoins étudier ici l'apport de l'approche lasso pour la détermination des vecteurs *design* des effets aléatoires.

Mise en œuvre

Comme précisé dans la section 7.4.2, la procédure que nous proposons nécessite une initialisation des vecteurs Γ^c, $c = \{Erreur, Correct\}$. Nous choisissons ici de prendre les moyennes empiriques comme valeurs initiales :

$$\Gamma_{(0)}^{Erreur} = \bar{y}^{Erreur} \quad \text{et} \quad \Gamma_{(0)}^{Correct} = \bar{y}^{Correct}\,.$$

La détermination du paramètre λ optimal est définie à partir d'un jeu de données supplémentaires. Dans l'exemple présenté dans la suite, 30 essais erreur et 300 essais correct ont été considérés et pour chaque valeur de λ un taux moyen de bonne classification a été calculé par *leave-one-out*. λ_{opt} correspond à la valeur du paramètre qui minimise l'erreur de classement des essais de la classe minoritaire *Erreur* obtenue à partir de la règle de décision définie en (7.13), avec $\hat{\Gamma}_{\lambda}^{c}$.

Les résultats présentés dans la suite sont ceux obtenus pour le participant B. Notons qu'il s'agit d'une étude exploratoire de l'approche proposée et qu'il sera intéressant dans la suite de développer une étude comparative plus complète.

Résultats de classification

Nous comparons ici, à l'aide du score de Pierce, les résultats de classification obtenus par le modèle utilisant la moyenne empirique comme vecteur design de la partie aléatoire (LMM dans la suite) et le modèle utilisant la méthode lasso (LMM-Lasso dans la suite). Dans la table C.3.5, les taux moyens de classification des essais de la classe *Erreur* ainsi que les écarts-type associés (entre parenthèse) sont donnés pour les deux méthodes. Ils ont été calculés à partir d'une procédure de validation croisée avec 30 itérations. Notons que pour ce participant, la procédure lasso détecte moins bien la classe minoritaire. De plus, une comparaison des scores de Pierce montre qu'en moyenne les performances de LMM sont significativement meilleurs que celles de LMM-Lasso.

Taux classif. LMM	Taux classif. LMM Lasso	Score Pierce
62,6 (6,7)	51,1 (7,3)	0,12 (p-valeur $\ll 10^{-6}$)

Quelle(s) différence(s) pour la modélisation de la variabilité des essais ?

Nous nous intéressons ici à la différence introduite dans le modèle par le choix du vecteur Γ^c par lasso.

La figure C.9 représente les valeurs estimées de la composante de la variance des effets aléatoires τ^2 obtenues à partir des modèles LMM et LMM-Lasso pour 30 échantillons d'apprentissage différents. Notons que les estimations obtenues par la méthode LMM-Lasso semblent légèrement plus faibles mais toujours significativement différentes de zéro.

La figure C.10 représente l'estimation du signal d'intérêt pour chaque essais à partir des modèles LMM et LMM-Lasso estimés et de la prédiction des effets aléatoires. Cette dernière est directement liée à l'estimation de la variance des effets aléatoires τ^2 (7.15). Pour LMM, comme pour LMM-Lasso, les signaux ajustés *Erreur* ont une forme d'onde laissant apparaître une forte négativité pouvant s'apparenter à la négativité d'erreur. Les ajustements obtenus par LMM-Lasso sont moins variables que ceux obtenus par LMM sur certaines parties du signal. Ceci est cohérent par rapport à la différence observée sur l'estimation de τ^2 dans les deux cas (figure C.9).

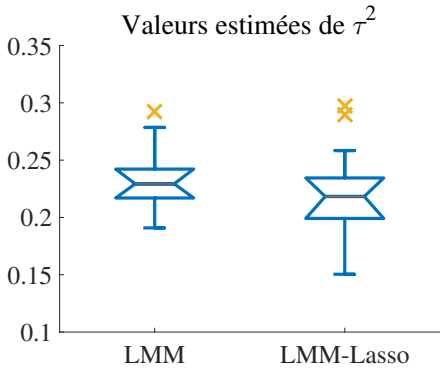


FIGURE C.9 – Estimations de la composante de la variance de l'effet aléatoire $uptau^2$ pour LMM et LMM-Lasso. Chaque boîte à moustaches représente une série de 30 estimations obtenues à partir de différents échantillons d'apprentissage de 40 essais *Erreur* et 400 essais *Correct*.

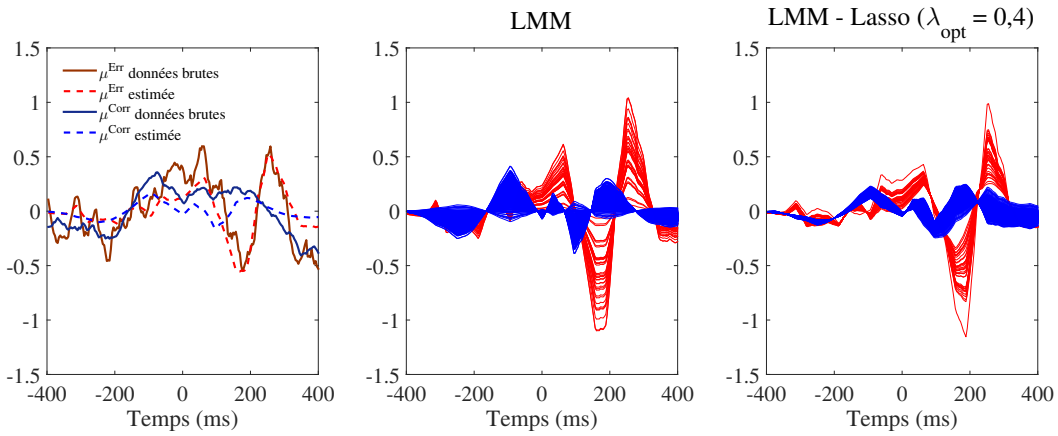


FIGURE C.10 – Illustration de l'ajustement des signaux par le modèle LMM et le modèle LMM-Lasso. La figure de gauche représente les moyennes des signaux *Erreur* (rouge) et *Correct* (bleu) sur l'électrode FCz pour le participant B. Traits pleins : moyennes des essais bruts dans chaque classe, tirets : moyennes estimées par le modèle (ici cette moyenne est identique pour LMM et LMM-Lasso). Figure centrale : représentation des essais ajustés par le modèle LMM dans les deux classes. Figure de droite : représentation des essais ajustés par le modèle LMM-Lasso (avec $\lambda_{opt} = 0,4$) dans les deux classes.

Cette première étude semble indiquer qu'en termes de modélisation la méthode lasso proposée

est pertinente. Cependant les premiers résultats obtenus en classification ne valident pas la méthode du point de vue de la discrimination. Nous ne pouvons donc tirer que des conclusions nuancées. Afin d'étudier l'apport de la méthode lasso pour le choix du vecteur design il faudrait en effet élaborer une étude systématique de comparaison en fonction : 1) des tailles d'échantillons d'apprentissage, 2) du choix d'initialisation $\Gamma^c_{(0)}$ et 3) des participants.

Bibliographie

Abramovich, F., Bailye, T. C., and Sapatinas, T. (2000). Wavelet analysis and its statistical applications. *Royal Statistical Society : Series D (The statistician)*, 49(1) :1–29. 32

Anderson, T. W. (2003). *An Introduction to Multivariate Statistical Analysis*. John Wiley and sons, 3rd edition. 45, 120

Atto, A. M. and Berthoumieu, Y. (2012). Wavelet packets of nonstationary random processes : Contributing factors for stationarity and decorrelation. *IEEE Transactions on Information Theory*, 58(1) :317–330. 26

Baayen, R. H., Davidson, D. J., and Bates, D. M. (2008). Mixed-effect modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59 :390–412. 106

Barachant, A. and Bonnet, S. (2011). Channel selection procedure using riemannian distance for BCI applications. In *5th International IEEE/EMBS Conference on Neural Engineering (NER)*. 87

Bartnik, A., Blinowska, K., and Durka, P. (1992). Single evoked potential reconstruction by means of wavelet transform. *Bioloical Cybernetics*, 67 :175–181. 25

Berger, H. (1929). Über das elektroenkephalogramm des menschen. *Archiv.für Psychiatrie und Nervenkrankheiten*, 87 :527–570. 18

Bickel, P. J. and Levina, E. (2004). Some theory for Fisher's linear discriminant function, 'naive Bayes', and some alternatives when there are many more variables than observations. *Bernoulli*, 10 :989–1010. 50, 54, 90

Bijma, F. and de Munck, J. C. (2008). A space-frequency analysis of MEG background processes. *NeuroImage*, 43 :478–488. 59

Bijma, F., de Munck, J. C., and Heethaar, R. M. (2005). The spatiotemporal MEG covariance matrix modeled as a sum of Kronecker products. *NeuroImage*, 27 :402–415. 59, 139

Bijma, F., de Munck, J. C., Huizenga, H. M., and Heethaar, R. M. (2003). A mathematical approach to the temporal stationarity of background noise in MEG/EEG measurements. *NeuroImage*, 20 :233–243. 59

Blankertz, B., Lemm, S., Treder, M., Haufe, S., and Muller, K. (2011). Single-trial analysis and classification of ERP components - a tutorial. *NeuroImage*, 56 :814–825. 39, 48, 85, 107, 109, 122

Blankertz, B., Tomioka, R.and Lemm, S., Kawanabe, M., and Muller, K. (2008). Optimizing spatial filters for robust EEG single-trial analysis. *IEEE Signal processing magazine*, 20. 29, 85

Buckheit, J., Chen, S., Donoho, D., Johnstone, I., and Scargle, J. (2005). About Wavelab. Technical report, Stanford University. 27, 32, 110

Burle, B., Roger, C., Vidal, F., and Hasbroucq, T. (2008). Spatio-temporal dynamics of information processing in the brain : recent advances, current limitations and future challenges. *International Journal of Bioelectromagnetism*, 10 :17–21. 11, 21

Callaway, E., Halliday, R., Naylor, H., and Thouvenin, D. (1984). The latency of the average is not the average of the latencies. *Psychophysiology*, 21 :571. 11, 21

Chaganty, H. M. and Naik, D. N. (2002). Analysis of multivariate longitudinal data using quasi-least squares. *Journal of Statistical Planning and Inference*, 103 :421–436. 63

Clemmensen, L., Hastie, T., Witten, D., and Ersboll, B. (2011). Sparse discriminant analysis. *Technometrics*, 58 :406–413. 54

Comon, P. and Jutten, C. (2007). *Séparation de sources 1 : concepts de base et analyse en composantes indépendantes*. Lavoisier, Paris, France. 24

Dal Seno, B., Matteucci, M., and Mainardi, L. (2009). Online detection of error-related potentials boosts the performance of mental typewriters. *Computational Intelligence and Neuroscience*, 2010 :499–519. 92

Daubechies, I. (1992). *Ten lectures on wavelets*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA. 26, 27, 110

Daubechies, I. (1993). Orthonormal bases of compactly supported wavelets : Variations on a theme. *SIAM Journal on Mathematical Analysis*, 24(2) :499–519. 26

Davidian, M. and Giltinan, D. M. (1995). *Nonlinear Models for Repeated Measurement Data*. Chapman and Hall, first edition. 106

Davidson, D. J. (2009). Functional mixed-effect models for electrophysiological responses. *Neurophysiology*, 41 :79–87. 106

Dawid, A. P. (1981). Some matrix-variate distribution theory : notational considerations and a Bayesian application. *Biometrika*, 68(1) :265–274. 64

Dayan, P. and Abbott, L. (2001). *Theoretical neuroscience*. MIT Press. 17

de Munck, J., Huizenga, H., Waldorp, L., and Heethaar, R. (2002). Estimating stationary dipoles from EEG/MEG data contaminated with spatially and temporally correlated background noise. *IEEE Transaction on Signal Processing*, 50(7) :1565–1572. 59

Delorme, A. and Makeig, S. (2004). EEGLAB : an open source toolbox for analysis of single-trial EEG dynamics. *Journal of Neuroscience Methods*, 134 :9–21. 25

Dudoit, S., Fridlyand, J., and Speed, T. P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association*, 97 :77–87. 49, 50, 54

Dutilleul, P. (1999). The MLE algorithm for the matrix normal distribution. *Journal of Statistical Computation and Simulation*, 64 :105–123. 64, 67

Dyrholm, M., Christoforou, C., and Parra, L. C. (2007). Bilinear discriminant component analysis. *Journal of Machine Learning Research*, 8 :1007–1021. 60

Dyrholm, M. and Parra, L. C. (2005). Smooth bilinear classification of EEG. In *28th annual international conference IEEE Engineering in medicine and biology society, Proceedings in*. 60

Eckart, C. and Young, G. (1936). The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3) :211–218. 80

Eriksen, B. A. and Eriksen, C. W. (1974). Effects of noise letters upon the identification of target letter in a non-search task. *Perception and Psychophysics*, 16 :143–149. 30

Falkenstein, M., Hoormann, J., Christ, S., and Hohnsbein, J. (2000). ERP components on reaction errors and their functional significance : a tutorial. *Biological Psychology. Special Issue : Error Processing and Adaptive Responding*, 51 :87–107. 88, 126

Farquhar, J. (2009). A linear feature space for simultaneous learning of spatio-spectral filters in BCI. *Neural Networks*, 22 :1278–1285. 59

Farwell, L. and Donchin, E. (1988). Talking off the top of tour head : toward a mental prosthesis utilizing event-related brain potentials. *Electroencephalography and clinical Neurophysiology*, 70 :510–523. 34

Fazli, S., Danóczy, M., Schelldorfer, J., and Müller, K.-R. (2011). $l1$-penalized linear mixed-effects models for high dimensional data with application to BCI. *NeuroImage*, 56 :2100–2108. 106, 140

Friedman, J. H. (1988). Regularized discriminant analysis. *Journal of the American Statistical Association*, 405 :165–175. 46, 47, 48, 70

Fruitet, J. (2012). *Interfaces Cerveau-Machines basées sur l'imagination de mouvements brefs : vers des boutons contrôlés par la pensée*. PhD thesis, Université de Nice - Sophia Antipolis. 19, 29

Fuentes, M. (2004). Testing for separability of spatial-temporal covariance functions. *Journal of Statistical Planning and Inference*, 136 :447–466. 63

Gasquet, C. and Witomski, P. (1999). *Fourier Analysis and Applications : Filtering, Numerical Computation, Wavelets*. Springer Science & Business Media. 24

Gençer, N., Williamson, S., Gueziec, A., and Hummel, R. (1996). Optimal reference electrode selection for electric source imaging. *Electroencephalogry and Clinical Neurophysiology.*, 99 :163–173. 18

Gibbons, D. and Chakraborti, S. (2010). *Nonparametric Statistical Inference*. Statistics : Textbooks and Monographs, fifth edition. 124

Glanz, H. and Carvalho, L. (2013). An expectation-maximization algorithm for the matrix normal distribution. *arXiv :1309.6609v1 [stat.ME]*. 64

Gupta, A. and Nagar, D. (2000). *Matrix variate distributions*. Chapman & Hall/CRC, Boca Raton, Florida, USA. 64, 74

Gutiérrez, D. and Escalona-Vargas, D. I. (2010). EEG data classification through signal spatial redistribution and optimized linear discriminants. *Computer methods and programs in biomedicine*, 97 :39–47. 29, 85, 86

Hasson-Meir, Y., Zhdanov, A., Hendler, T., and Intrator, N. (2011). Inference of brain mental states from spatiotemporal analysis of EEG single trials. *Proceedings of the international conference on Bio-inspired and signal processing*, pages 59–66. 59

Hastie, T., Buja, A., and Tibshirani, R. (1995). Penalized discriminant analysis. *The Annals of Statistics*, 23(1) :73–102. 43, 70, 86

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning*. Springer, second edition. 28, 39, 47, 54

Henderson, C. (1982). Analysis of covariance in the mixed model : Higher-level, non homogeneous, and random regressions. *Biometrics*, 38(3) :623–640. 150

Hollander, M., Wolfe, D. A., and Chicken, E. (2014). *Nonparametric Statistical Methods*. Wiley, third edition. 124

Huang, Y., Erdogmus, D., Hild II, K., Pavel, M., and Mathan, S. (2009). Mixed effects models for single-trial ERP detection in noninvasive brain computer interface design. In *Recent Advances in Biomedical Signal Processing*, pages 171–180. E-Book Preprint Bentham Science Publishers. 29, 86, 107

Huang, Y., Erdogmus, D., and Pavel, M. (2008). Mixed effects models for EEG evoked response detection. In *Machine Learning for Signal Processing, IEEE Workshop*. 86, 107

Huiskamp, G. (1991). Difference formulas for the surface laplacian on a triangulated surface. *Journal of Computational Physics*, 95 :477–496. 86

Huizenga, H. M., de Munck, J. C., Waldorp, L. J., and Grasman, R. P. (2002). Spatiotemporal EEG/MEG source analysis based on a parametric noise covariance model. *IEEE Transactions on biomedical engineering*, 49 :533–538. 59

Inoue, K. and Urahama, K. (2006). Non-iterative two-dimensional linear discriminant analysis. In *18th International Conference on Pattern Recognition (ICPR18)*, volume 2, pages 540–543. 73

Knuth, K. H., Shah, A. S., Truccolo, W. A., Ding, M., Bressler, S. L., and Schroeder, C. E. (2006). Differentially variable component analysis : Identifying multiple evoked components using trial-to-trial variability. *Journal of Neurophysiology*, 95 :3257–3276. 11, 21

Kolodziej, M., Majkowski, A., and Rak, R. J. (2012). Linear discriminant analysis as EEG features reduction technique for brain-computer interfaces. Technical report, Warsaw University of Technology. 39

Koopmans, L. H. (1995). *The spectral analysis of time series*. Probability and Mathematical Statistics. Academic Press. 108

Kotsiantis, S., Kanellopoulos, D., and Panayiotis, P. (2006). Handling imbalanced datasets : A review. *GESTS International Transactions on Computer Science and Engineering*, 30. 49

Krusienski, D. J., Sellers, E. W., Cabestaing, F., Bayoudh, S., McFarland, D., Theresa, V., and Wolpaw, J. R. (2006). A comparison of classification techniques for the P300 speller. *Journal of Neural Engineering*, 3(4) :299–305. 37

Kukleta, M. and Lamarche, M. (2001). Steep early negative slopes can be demonstrated in pre–movement bereitschaftpotential. *Clinical Neurophysiology*, 112 :1642–1649. 21

Lebedev, M. A. and Nicolelis, M. A. (2003). Brain–machine interfaces : past, present and future. *TRENDS in Neurosciences*, 29(9) :536–546. 29

Ledoit, O. and Wolf, M. (2003). Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of empirical finance*, 10 :603–621. 47, 48

Ledoit, O. and Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88 :365–411. 48

Li, M. and Yuan, B. (2005). 2D-LDA : A statistical linear discriminant analysis for image matrix. *Pattern Recognition Letters*, 26 :527–532. 73

Lotte, F., Congedo, M., Lécuyer, A., Lamarche, F., and Arnaldi, B. (2007a). A review of classification algorithms for EEG-based brain computer interfaces. *Journal of Neural Engineering*, 4. 28, 39

Lotte, F., Lécuyer, A., and Arnaldi, B. (2007b). Les interfaces cerveau-ordinateur : Utilisation en robotique et avancées récentes. In *Journées Nationales de le Recherche en Robotique, Obernai*. 29, 30

Lu, H., Plataniotis, K. N., and Venetsanopoulos, A. N. (2009). A taxonomy of emerging multilinear discriminant analysis solutions for biometric signal recognition. In *Biometrics : Theory, Methods and Applications*. John Wiley & Sons, Hoboken, NJ, USA. 73

Lu, N. and Zimmerman, D. L. (2005). The likelihood ratio test for a separable covariance matrix. *Statistics and probability letters*, 73 :449–457. 64, 67, 68

Luo, D., Ding, C., and Huan, H. (2009). Symmetric two dimensional linear discriminant analysis. In *IEEE Conference on Computer Vision and Pattern Recognition, 2009*, pages 2820 – 2827. 73

Mahanta, M., Aghaei, A., and Plataniotis, K. (2012). A Bayes optimal matrix-variate LDA for extraction of spatio-spectral features from EEG signals. In *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC 2012)*, pages 3955–3958. 25, 74, 77, 140

Mahanta, M. S., Aghaei, A. S., and Plataniotis, K. N. (2013). Regularized LDA based on separable scatter matrices for classification of spatio-spectral EEG patterns. In *IEEE International conference on acoustics, speech and signal processing (ICASSP)*. 70, 74

Mallat, S. (2008). *A Wavelet Tour of Signal Processing, Third Edition : The Sparse Way*. Academic Press, 3rd edition. 26, 27, 110

Manyakov, N. V., Chumerin, N., Combaz, A., and Van Hulle, M. M. (2011). Comparison of classification methods for P300 brain-computer interface on disabled subjects. *Computational Intelligence and Neuroscience*, 2011 :12 pages. 37

McCulloch, C. E., Searle, S. R., and Neuhaus, J. M. (2008). *Generalized, Linear and Mixed Models*. Wiley, second edition. 105, 106

McFarland, D. J., McCane, L. M., David, S. V., and Wolpaw, J. R. (1997). Spatial filter selection for EEG-based communication. *Electroencephalography and clinical Neurophysiology*, 103 :386–394. 85

Menardi, G. and Torelli, N. (2014). Training and assessing classification rules with unbalanced data. *Data Mining and Knowledge Discovery*, 28(1). 32

Meyer, D. E., Osman, A., Irwin, D. E., and Yantis, S. (1988). Modern mental chronometry. *Biological Psychology*, 26 :3–67. 21

Mitchell, M. W., Genton, M. G., and Gumpertz, M. L. (2006). A likelihood ratio test for separability of covariances. *Journal of multivariate analysis*, 97 :1025–1043. 68

Noushath, S., Kumar, G., and Shivakumara, P. (2006). $(2d)^2$ LDA : an efficient approach for face recognition. *Pattern Recognition*, 39(7) :1396–1400. 73

Nunez, P. and Srinivasan, R. (2006). *Electric fields of the brain*. Oxford University Press, 2nd edition. 17, 24

Parra, L. C., Christoforou, C., Gerson, A. D., Dyrholm, M., Luo, A., Wagner, M., Philiastides, M., and Sajda, P. (2008). Spatiotemporal linear decoding of brain state. *IEEE Signal processing magazine*, pages 107–115. 59

Parra, L. C., Spence, C. D., Gerson, A. D., and Sajda, P. (2005). Recipes for the linear analysis of EEG. *NeuroImage*, 28 :326–341. 39, 85

Perrin, M. (2012). *Coadaptation cerveau machine pour une interaction optimale : application au P300-Speller*. PhD thesis, Université Claude Bernard Lyon1. 29, 35

Perrin, M., Maby, E., Daligault, S., Olivier, B., and Mattout, J. (2012). Objective and subjective evaluation of online error correction during P300-based spelling. *Advances in Human-Computer Interaction*, 2012. 31

Perrin, M., Maby, E., and Mattout, J. (2011). Detecting and interpreting responses to feedback in BCI. In *Graz BCI International Workshop*, pages 116–119. 34

Quiroga, Q. R. (2000). Obtaining single stimulus evoked potentials with wavelet denoising. *Physica D*, 145 :278–292. 26

Quiroga, Q. R. and Garcia, H. (2003). Single-trial event-related potentials with wavelet denoising. *Clinical Neurophysiology*, 114 :376–390. 26

Quiroga, Q. R. and van Luijtelaar, E. L. J. M. (2002). Habituation and sensitization in rat auditory evoked potentials : a single-trial analysis with wavelet denoising. *International Journal of Psychophysiology*, 43(2) :141–153. 21

Ramirez, R. R. (2008). Source localization. *Scholarpedia*, 3(11). 24

Ramoser, H., Müller-Gerking, J., and Pfurtscheller, G. (2000). Optimal spatial filtering of single trial EEG during imagined hand movement. *IEEE Transactions on rehabilitation engineering*, 8(4) :441–446. 29

Rao, C. R. (1971a). Estimation of variance and covariance components : MINQUE theory. *Journal of Multivariate Analysis*, 1 :257–275. 147

Rao, C. R. (1971b). Minimum variance quadratic unbiaised estimation of variance components. *Journal of Multivariate Analysis*, 1 :445–456. 147

Rao, C. R. and Kleffe, J. (1988). *Estimation of variance components and applications*. Elsevier, Amsterdam. 105, 146

Rivet, B. and Souloumiac, A. (2013). Optimal linear spatial filters for event-related potentials based on spatio-temporal model : Asymptotical performance analysis. *Signal Processing*, 93 :387–398. 29

Rivet, B., Souloumiac, A., Attina, V., and Gibert, G. (2009). XDawn algorithm to enhance evoked potentials : application to brain computer interface. In *Biomedical engineering, IEEE Transactions*, volume 56, pages 2035–2043. 29, 35, 37

Roger, C., Bénar, C. G., Vidal, F., Hasbroucq, T., and Burle, B. (2010). Rostral cingulate zone and correct response monitoring : ICA and source localization evidences for the unicity of correct- and error-negativities. *NeuroImage*, 51 :391–403. 30, 31, 92

Roy, A. and Khattreeb, R. (2005). On discrimination and classification with multivariate repeated measures data. *Journal of Statistical Planning and Inference*, 134(2) :462–485. 63

Sanei, Saeid ans Chambers, J. (2007). *EEG signal processing*. Wiley. 24

Saporta, G. (2011). *Probabilités, Analyse des Données et Statistique*. Editions Technip, 3eme edition. 41, 43, 45, 51

Schafer, J. and Strimmer, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genetics. *Statistical applications in genetics and molecular biology*, 4(1). 47, 48

Searle, S. R., Casella, G., and McCulloch, C. E. (1992). *Variance components*. Wiley. 109, 111

Smulders, F. T., Kenemans, J. L., and Kok, A. (1996). Effects of task variables on measures of the mean onset latency of LRP depend on the scoring method. *Psychophysiology*, 33(2) :194–205. 21

Spinnato, J., Roubaud, M.-C., Burle, B., and Torrésani, B. (2014). Finding EEG space-time-scale localized features using matrix-based penalized discriminant analysis. In *International conference on Acoustics, Speech and Signal Processing (ICASSP 2014)*. 74, 78

Spinnato, J., Roubaud, M.-C., Burle, B., and Torrésani, B. (2015a). Detecting single-trial EEG evoked potential using a wavelet domain linear mixed model : Application to error potentials classification. *Journal of neural engineering*. 119

Spinnato, J., Roubaud, M.-C., Casini, L., Burle, B., and Torrésani, B. (2012). Une approche modèle mixte pour la classification supervisée de signaux électrophysiologiques. In *44èmes Journées de statistique de la SFDS*. 29, 86

Spinnato, J., Roubaud, M.-C., Perrin, M., Maby, E., Mattout, J., Burle, B., and Torrésani, B. (2015b). Analyse discriminante matricielle descriptive. application à l'étude de signaux EEG. In *47èmes Journées de statistique de la SFDS*. 74

Srivastava, M., Nahtman, T., and von Rosen, D. (2008). Models with a Kronecker product covariance structure : Estimation and testing. *Mathematical Methods of Statistics*, 17(4) :357–370. 67, 68, 69

Tandonnet, C., Burle, B., Hasbroucq, T., and Vidal, F. (2005). Spatial enhancement of EEG traces by surface laplacian estimation : comparison between local and global methods. *Clinical Neurophysiology*, 116 :18–24. 24

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 58(1) :267–288. 116

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005). Sparsity and smoothness via the fused Lasso. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 67 :91–108. 54

Tsiligkaridis, T. and Hero, A. O. (2013). Covariance estimation in high dimensions via Kronecker product expensions. *IEEE Transactions on Signal Processing*, 61 :5347–5360. 139

Tsiligkaridis, T., Hero, A. O., and Zhou, S. (2013). Convergence properties of Kronecker graphical Lasso algorithms. *IEEE Transactions on Signal Processing*, 61(7) :1743–1755. 139

Vetterli, M. and Kovačevic, J. (1995). *Wavelets and subband coding*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA. 27, 110

Vidaurre, C., Schlöogl, R. C., Scherer, R., and Pfurtscheller, G. (2006). A fully on-line adaptive BCI. *IEEE Transactions on Biomedical Engineering*, 53(6) :1214–1219. 39

Villaron, E. (2012). *Modèles aléatoires harmoniques pour les signaux électroencéphalographiques*. PhD thesis, Aix-Marseille Université. 25, 26, 59, 140

Villaron, E. and Torrésani, B. (2010). Harmonic hidden markov models for the study of EEG signals. In *European Signal Processing Conference (EUSIPCO 2010)*. 59

Wald, A. (1943). Tests of statistical hypothesis concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society*, 54 :426–482. 122

Wang, Z., Maier, A., Leopold, D. A., Logothetis, N. K., and Liang, H. (2007). Single-trial evoked potential estimation using wavelets. *Computers in Biology and Medecine*, 37 :463–473. 26

Weiss, G. M. (2004). Mining with rarity : A unifying framework. *ACM SIGKDD Explorations Newsletter*, 6(1) :7–19. 32, 49

Weiss, G. M. and Provost, F. (2001). The effect of class distribution on classifier learning : an empirical study. Technical report, Data, Inference, Analytics and Learning Lab. 32

Witkovský, V. (2002). Matlab algorithm mixed.m for solving Henderson's mixed model equations. Technical report, Institute of Measurement Sciences, Slovak Academy of Sciences. 112

Witten, D. and Tibshirani, T. (2011). Penalized classification using Fisher's linear discriminant. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 73(5) :753–772. 54

Wolpaw, J. R., Birbaumer, N., Heetderks, W. J., McFarland, D. J., Peckham, P. H., Schalk, G., Donchin, E., Quatrano, L. A., Robinson, C. J., and Vaughan, T. M. (2000). Brain-computer interface technology : a review of the first international meeting. *IEEE Transactions on Rehabilitation Engineering*, 8 :164–173. 29

Wolpaw, J. R., Birbaumer, N., McFarland, D. J., Pfurtscheller, G., and Vaughan, T. M. (2002). Brain-computer interfaces for communication and control. *Clinical Neurophysiology*, 113 :767–791. 29

Xie, J. and Qiu, Z. (2006). The effect of imbalanced data sets on LDA : A theoretical and empirical analysis. *Pattern Recognition*, 40 :557–562. 49

Xiong, H., Swamy, M., and Ahmad, M. (2006). Two-dimensional FLD for face recognition. *Pattern Recognition*, 38(7) :1121–1124. 73

Xue, J.-H. and Titterington, M. D. (2008). Do unbalanced data have a negative effect on LDA ? *Pattern Recognition*, 41 :1558–1571. 49

Ye, J., Janardan, R., and Li, Q. (2004). Two-dimensional linear discriminant analysis. In *Advances in Neural Information Processing Systems (NIPS 2004)*, volume 17, page 1569–1576. 73

Zhao, J., Yu, P. L., Shi, L., and Li, S. (2012). Separable linear discriminant analysis. *Computational Statistics and Data Analysis*, 56 :4290–4300. 73, 74, 77

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 67 :301–320. 54