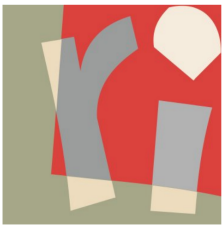
Représentations redondantes pour les signaux d'électroencéphalographie

Yoann Isaac



Université Paris-Sud

Ecole Doctorale d'Informatique de Paris-Sud

Laboratoire (d') Analyse de Données et Intelligence des Systèmes (CEA)

Laboratoire de Recherche en Informatique (Université Paris-Sud)

Discipline : Informatique

Thèse de doctorat

préparée sous la direction de Michèle Sebag et soutenue le 29 mai 2015 par

Yoann Isaac

Représentations redondantes pour les signaux d'électroencéphalographie

Composition du jury :

Président du jury :	M. Dominique Gouyou-Beauchamps	PR (Université Paris-Sud)
Rapporteurs :	M. Fabrizio De Vico Fallani	CR (INRIA)
	M. Alain Rakotomamonjy	PR (Université de Rouen)
Examinateur :	M. Gaël Varoquaux	CR (INRIA)
Encadrants :	M. Cédric Gouy-Pailler	Ingénieur Chercheur (CEA)
	M. Jamal Atif	PR (Université Paris-Dauphine)
Membre invitée :	Mme. Michèle Sebag	DR (Université Paris-Sud)

À ma famille,
À Lian, ma chérie,

Remerciements

Les trois années ayant mené à la rédaction de cette thèse ont été particulièrement riches aussi bien d'un point de vue scientifique qu'humain et je souhaite remercier chaleureusement l'ensemble des personnes ayant de près ou de loin contribué à ce travail.

En particulier, je souhaite exprimer ma gratitude à mon encadrement composé de Cédric Gouy-Pailler, Jamal Atif et de Michèle Sebag. Un grand merci pour m'avoir offert la chance de travailler sur ce projet, m'avoir guidé dans des domaines scientifiques que je connaissais peu, m'avoir aidé dans la rédaction de différentes communications et m'avoir fait découvrir le milieu de la recherche en général. J'ai énormément appris à votre contact et je vous remercie pour tout cela.
Je souhaite remercier également Quentin Barthélemy avec qui j'ai appris beaucoup sur les représentations redondantes et avec qui j'ai eu le plaisir de collaborer pour la représentation des signaux EEG à l'aide de ces approches. De même, je remercie Antoine Souloumiac pour l'ensemble des discussions que nous avons pu avoir sur les approches de filtrage spatial optimal pour la classification dans les interfaces cerveau-machine ainsi que pour son aide dans certaines dérivations mathématiques nécessaire à la mise en place d'algorithmes d'optimisation efficaces.
Par ailleurs, je tiens à exprimer ma gratitude à l'ensemble des membres de mon jury, à Alain Rakotomamonjy et Fabrizio De Vico Fallani pour avoir soigneusement relu et rapporté cette thèse, à Dominique Gouyou-Beauchamps pour avoir présidé ma soutenance et à Gaël Varoquaux pour ses propositions de pistes de recherche.

Ma gratitude va également à l'ensemble des membres de mes laboratoires d'accueil, le LADIS du CEA ainsi que l'équipe Tao du LRI, avec lesquels j'ai passé trois années très enrichissantes et agréables. En particulier, je remercie Maxime et Jérémy membres officiels de la geôle des thésards ainsi que, Flore, Émira, Paul, Antoine, Cécile, Néhémy, Crédo, Xavier, Lucile, Matthieu, Aurore, Aurélien et Anne Catherine pour votre constante bonne humeur et tous les bons moments passés dans le coin café.
Je souhaite enfin remercier ma famille et mes amis pour leur soutien. Je tiens à dire merci en particulier à mes amis thésards (hors laboratoires d'accueil), Dimitri, Thomas, Georges, Maud, qui m'ont beaucoup aidés dans les moments difficiles. Je tiens également à remercier ma mère et mon frère pour m'avoir aidé à tenir durant la rédaction ainsi que mes grands-parents pour tous les envois de fortifiants et les encouragements.
Pour finir, je remercie ma chérie, Lian, pour son amour, sa patience et son soutien tout au long de cette aventure.

Table des matières

Table des figures

Chapitre 1

Introduction

Le cerveau est l'organe le plus complexe du corps humain et probablement l'un des plus intéressants du monde vivant. Son fonctionnement a suscité l'intérêt des hommes depuis des temps très anciens aussi bien pour la compréhension de son rôle central au sein du système nerveux que pour son implication dans le processus cognitif. Ainsi, parmi les documents médicaux les plus anciens retrouvés à ce jour, les papyrus d'Edwin Smith datés du XIIIème siècle avant notre ère relatent des observations liant des dommages au cerveau à un déficit de la motricité, mettant en avant un lien entre celui-ci et les membres du corps [75]. Néanmoins, ce n'est que récemment que ce fonctionnement a pu être étudié avec plus de précision grâce à l'utilisation de méthodes permettant une mesure quantitative de l'activité cérébrale.
Parmi ces mesures, l'électroencéphalographie (EEG) est la plus ancienne mais aussi l'une des plus usitées. Elle permet la mesure de l'activité cérébrale d'un sujet en temps réel grâce à un ensemble d'électrodes placées sur son crâne capturant les modifications du champ électrique de son cerveau. Enregistrés pour la première fois sur un être humain par Hans Berger [20] en 1929, les signaux EEG sont toujours d'une grande utilité de nos jours du fait de leurs résolutions temporelles élevées et de la facilité avec laquelle ils peuvent être mesurés. Ils sont utilisés aussi bien pour l'étude du fonctionnement [49] cérébral que pour le diagnostic médical [227] ou les interfaces cerveau-machine (ICM) [230].
Ces dernières sont des systèmes permettant à une personne de contrôler une machine via la modulation de son activité cérébrale sans aucun concours de ses muscles. Ces systèmes sont apparus dans les années 70 [224] et sont toujours très étudiés de nos jours. Ils sont principalement utilisés pour des applications médicales et notamment pour les patients atteints du syndrome d'enfermement [127] mais leur utilisation dans les domaines de recherche du jeu vidéo ou bien de la réalité virtuelle a également fait l'objet de travaux [143].

1.1 Contexte et problématique

Du fait de leur utilisation dans ces divers contextes, l'analyse des signaux d'électroencéphalographie reste un domaine de recherche très actif. Ce sont des signaux complexes dont il peut être difficile d'extraire les composantes d'intérêts. De nombreuses méthodes ont été conçues dans ce but mais aucune n'y arrive à ce jour de façon suffisamment fiable et précise. Ces signaux sont composés d'un mélange d'une multitude d'activités cérébrales différentes et sont, de plus, très variables aussi bien d'un sujet à l'autre qu'entre deux enregistrements réalisés sur un même sujet ce qui les rend particulièrement difficiles à traiter. Caractériser de tels signaux nécessite la prise en compte de leurs variations temporelles, spatiales ainsi que fréquentielles, toutes trois importantes pour leur analyse. De nombreuses méthodes

d'analyse de ces signaux ne prennent en considération qu'un ou deux de ces aspects et sont ainsi incapables de considérer l'ensemble des informations qu'ils contiennent. Par ailleurs, ces approches manquent aussi fréquemment de flexibilité, capturant seulement leurs composantes ayant les plus grandes amplitudes.
Afin d'obtenir cette flexibilité ainsi que l'intégration des différents aspects de ces signaux dans une même approche, nous considérons dans cette thèse l'utilisation de représentations redondantes. Le principe de ces représentations est de décrire des signaux à l'aide de combinaisons linéaires de signaux basiques nommés atomes appartenant à une famille appelée dictionnaire pouvant être composée d'un nombre d'éléments supérieur à la taille des signaux [139]. Ces représentations se sont montrées particulièrement intéressantes pour la réalisation de diverses tâches telles que le débruitage [68], la compression [12] ou bien la séparation de sources [30] sur des classes de signaux aussi diverses que les images [68], les enregistrements audio [187] ou les électrocardiogrammes [73]. Qui plus est, sous certaines conditions, les représentations obtenues en plus de décrire précisément les signaux, peuvent être interprétables permettant ainsi une utilisation plus aisée de celles-ci.
La grande flexibilité de ce type de représentation n'est tout de même pas sans contreparties. Du fait de la redondance du dictionnaire, il existe pour chaque signal une infinité de décompositions possibles sur celui-ci. Par conséquent, une contrainte est considérée sur les solutions de ces décompositions afin qu'elles soient uniques. Une contrainte de parcimonie est le plus souvent choisie pour cela. Cette dernière suppose que les signaux peuvent être représentés par une combinaison linéaire d'un faible nombre d'atomes du dictionnaire, on parle alors de décomposition parcimonieuse des signaux. Ce type de décomposition bien qu'efficace pour certaines applications (voir les références ci-dessus) présente néanmoins des inconvénients lorsque le rapport signal à bruit (RSB) est faible et/ou le dictionnaire mal conditionné. Les décompositions parcimonieuses obtenues dans ces cas sont en effet instables (décompositions différentes de signaux représentant un même phénomène) et les composantes extraites des signaux peuvent ne pas refléter les « réelles » composantes sous-jacentes de ceux-ci.

Dans l'optique de la mise en place de représentations redondantes pour les signaux EEG qui ont effectivement un RSB faible (≈ -20 db), notre étude va s'intéresser à la conception de contraintes permettant de limiter ces inconvénients en utilisant certaines connaissances *a priori* sur ces signaux pour guider les décompositions vers des solutions plausibles physiologiquement. La construction de telles contraintes appelées aussi régularisations et des algorithmes de décompositions associés sera ainsi le point central de ce travail, l'objectif final étant l'obtention de représentions flexibles et interprétables des signaux EEG à l'aide de modèles redondants.
Par ailleurs, bien que la conception de représentations pour les ICM ne soit pas notre but, nous souhaitons que les approches proposées puissent être utilisées dans un tel contexte. Nous avons donc décidé d'une part, de travailler directement sur les signaux EEG et non sur les sources cérébrales, afin de conserver des coûts de calcul raisonnables et d'autre part de concevoir des régularisations pouvant s'appliquer sur chaque signal séparément afin notamment de pouvoir réaliser des traitements en ligne (avec des signaux arrivants les uns après les autres).

1.2 Contributions et organisation du document

Les contributions de cette thèse sont de plusieurs natures :

1. des modèles de décomposition régularisée sont proposés pour les signaux EEG,
2. des algorithmes de décomposition sont construits afin de résoudre les problèmes d'optimisation associés à ces décompositions,
3. ces modèles sont évalués sur des signaux synthétiques réalistes ainsi que sur des signaux EEG réels.

Afin de présenter ces contributions, ce document est organisé comme suit :

— Les chapitres 2 et 3 sont des chapitres d'état de l'art. Le chapitre 2 introduit les signaux EEG, les propriétés de ces signaux y sont décrites, de même que les méthodes communément utilisées pour leur traitement. Ce chapitre fait apparaître les faiblesses de ces méthodes d'analyse et les motivations qui nous ont poussé à étudier des modèles redondants. Le chapitre 3 s'intéresse ensuite à ces représentations de manière approfondie. Les aspects théoriques de ces modèles y sont abordés ainsi que les algorithmes permettant de les mettre en œuvre. Ces chapitres visent à permettre la compréhension des chapitres de contribution suivants, les choix effectués par la suite pour la modélisation des signaux EEG n'ayant de sens qu'à la lumière de ces informations.

— Le chapitre 4 est dédié à la construction de régularisations pour la décomposition de signaux EEG sur des dictionnaires. Ces régularisations sont fondées sur des propriétés connues de ces signaux afin de guider les décompositions vers des solutions plausibles physiologiquement. Deux modèles de décompositions régularisées sont proposés : une décomposition sur un dictionnaire d'atomes temporels régularisée spatialement et une décomposition sur un dictionnaire d'atomes spatiaux régularisée en temps.

— Le chapitre 5 traite de la mise en place du modèle de décomposition régularisée spatialement introduit dans le chapitre précédent. Des algorithmes sont proposés afin de réaliser cette décomposition. Dans un contexte d'analyse temps-fréquence, ces algorithmes sont ensuite évalués sur des signaux artificiels avant d'être appliqués à la détection du potentiel évoqué P300.

— Le chapitres 6 s'attache ensuite à la conception de l'algorithme de décomposition avec régularisation en analyse nécessaire au modèle spatial régularisé temporellement. Cet algorithme est évalué sur des signaux synthétiques ainsi que pour l'identification de potentiels évoqués P300.

— Enfin, le chapitre 7 est consacré à l'extension d'un modèle classique EEG (modèle des micro-états) à l'aide d'une approche d'apprentissage de dictionnaire utilisant l'algorithme de décomposition conçu dans le chapitre précédent.

1.3 Notations

Vecteurs et matrices :

- $\mathbf{a}$, les vecteurs sont notés à l'aide de lettres minuscules en gras,
- A, les matrices sont notées à l'aide de lettres majuscules,
- $\mathbf{a}(i)$, représente le i-ième élément de $\mathbf{a}$,
- $A(.,i) = A(i)$, représente la i-ième colonne de A,
- $A(i,.) = (A^T(i))^T$, représente la i-ième ligne de A,
- $A(j,i)$, représente l'élément de la matrice A présent dans la ligne j et dans la colonne i,
- I_n, représente la matrice identité de dimension n.

Opérateurs et fonctions :

- $\langle .,. \rangle$ opérateur de produit scalaire,
- $(.)^t$, opérateur de transposition,
- $(.)^\dagger$, pseudo-inverse, $X^\dagger = (X^T X)^{-1} X^T$,
- $|.|$, valeur absolue d'un scalaire,
- $[.]$, opérateur de concaténation de matrices,
- $\#\{.\}$, cardinal d'un ensemble,
- *span*., espace engendré par une famille de vecteurs,
- *supp*(.), ensemble des indices des éléments non-nuls d'un vecteur (support).

Normes :

- $\|\mathbf{x}\|_p = (\sum_i |\mathbf{x}_i|^p)^{\frac{1}{p}}$, norme p du vecteur $\mathbf{x}$,
- $\|X\|_{p,q} = (\sum_i \|X(i)\|_p^q)^{\frac{1}{q}}$, norme $\ell_{p,q}$ de la matrice X,
- $\|X\|_p = \|X\|_{p,p}$, raccourci pour l'application d'une norme $\ell_{p,p}$ sur la matrice X,
- $\|X\|_F = \|X\|_{2,2}$, norme de Frobénius de la matrice X,
- $\|\mathbf{x}\|_0 = \#\{\mathbf{x}(i) = 0\}$ pseudo-norme ℓ_0 du vecteur $\mathbf{x}$.

1.4 Abréviations

1.4.1 Abréviations françaises

ACP	Analyse en Composantes Principales
ACI	Analyse en Composantes Indépendantes
ICM	Interface Cerveau-Machine
PE	Potentiels Évoqués
t.q.	tel que

1.4.2 Abréviations anglaises

AAR	Adaptive AutoRegressive
BLDA	Bayesian Linear Discriminant Analysis
ERS	Event Related Synchronisation
ERD	Event Related Desynchronisation
FDA	Fisher Discriminant Analysis
LDA	Linear Discriminant Analysis
MP	Matching Pursuit
MVAR	Multivariate Autoregressive
MVARMA	Multivariate Autoregressive Moving Average
OMP	Orthogonal Matching Pursuit
SSVEP	Steady State Visual Evoked Potential
SVM	Support Vector Machine

Chapitre 2

Électroencéphalographie

Sommaire

Ce chapitre décrit comment l'activité cérébrale peut être mesurée par électroencéphalographie ainsi que les approches utilisées sur les signaux obtenus pour les analyser dans un cadre médical et dans le contexte des interfaces cerveau-machine. La section 2.1 présente les origines de cette mesure, son principe, ainsi que les protocoles utilisés pour l'acquisition des signaux EEG. Les propriétés de ces signaux ainsi que les processus neurophysiologiques sous-jacents sont ensuite présentés dans la section 2.2 de même que les principales applications. Enfin, les méthodes classiques d'analyse de ces signaux sont décrites dans la section 2.3. Nous verrons en particulier les raisons ayant motivé notre étude et l'intérêt que peuvent avoir les représentations redondantes pour l'étude des signaux EEG.

2.1 Mesure de l'activité cérébrale par électroencéphalographie

L'électroencéphalographie est une méthode permettant de mesurer l'activité électrique du cerveau grâce à un ensemble d'électrodes placées à la surface du crâne du sujet. Cette mesure présente l'avantage d'être non invasive, relativement peu chère et facile à réaliser.

2.1.1 Les débuts de l'électroencéphalographie

C'est en 1875 que Richard Caton rapporta pour la première fois la mesure d'une activité électrique à la surface du crâne d'un animal réalisée à l'aide d'un galvanomètre. Les premières mesures réalisées sur un être humain furent effectuées par Hans Berger en 1924 et publiées en 1929 [20]. Ce premier enregistrement est présenté dans la Fig. 2.1. Ces travaux permirent la découverte du rythme *alpha* appelé aussi rythme de Berger (voir section 2.2 pour les différents rythmes cérébraux en EEG). Ils furent poursuivis par Edgar Douglas Adrian qui obtint le prix Nobel de médecine en 1932.

FIGURE 2.1 – Premier enregistrement d'un signal EEG humain par Hans Berger en 1924 [20].

L'utilisation pratique de cette méthode dans un cadre médical ne démarra qu'à partir des années cinquante, principalement pour le diagnostic de l'épilepsie dont les caractéristiques sont aisément repérables sur un enregistrement EEG [227]. Cette mesure fut plus tard utilisée pour l'analyse des cycles du sommeil ou la détection de maladies psychiatriques [78] (voir section 2.2.4). Les premiers travaux proposant l'utilisation d'une telle mesure pour le contrôle d'une machine furent eux réalisés en 1971 [224].

2.1.2 Origine des signaux mesurés par électroencéphalographie

Le cerveau est l'organe central du système nerveux. Il assure la régulation de la plupart des fonctions vitales et permet la réalisation des tâches cognitives. Il est situé dans la boîte crânienne et comporte environ 10^{11} cellules neuronales. C'est l'organe le plus complexe du corps humain dont le fonctionnement , même s'il est étudié depuis de nombreuses années, n'est pas encore complètement compris. Différents travaux ont tout de même permis de le cartographier en identifiant des zones directement liées à certaines fonctions corporelles. Un exemple d'une telle cartographie est présenté dans la figure 2.2.

L'activité électrique du cerveau mesurée en surface du crâne est due à des transferts ioniques réalisés au niveau neuronal. Un neurone est une cellule composée d'un corps cellulaire aussi appelé Soma et des prolongements. Ces prolongements sont de deux types, un axone unique, permettant le passage d'un potentiel d'action transmettant les informations

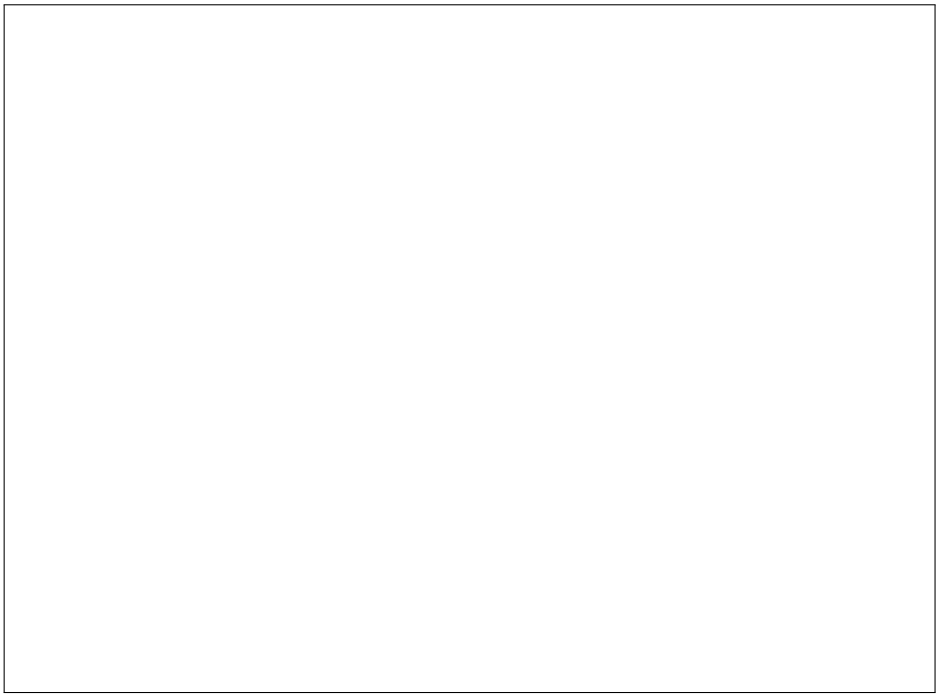


FIGURE 2.2 – Cartes fonctionnelles du cerveau : association de certaines aires du cerveau avec des fonctions corporelles. Source : [207].

et de nombreuses dendrites recevant des informations des axones d'autres neurones par l'intermédiaire de synapses. Cette structure est présentée ci-dessous dans la figure 2.3.

Lorsqu'une information transite entre plusieurs neurones, les équilibres ioniques de ceux-ci sont modifiés. D'une part, les potentiels postsynaptiques provoquent une diminution du potentiel autour des cellules et déclenchent ainsi des mouvements de charge dans les corps cellulaires appelés courants primaires. D'autre part, des courants secondaires à l'extérieur des cellules sont créés et compensent les courants primaires.
Les champs électriques induits par ces courants, pris un par un, sont trop faibles pour être distingués à la surface du crâne. Les signaux observés par électroencéphalographie sont en fait le résultat de l'excitation de neurones particuliers localisés dans la couche corticale et appelés neurones pyramidaux. Ces neurones sont orientés perpendiculairement au crâne. Les neurones pyramidaux situés dans une zone spécifique du cortex reçoivent des signaux identiques provenants de neurones voisins d'une manière telle que leurs potentiels postsynaptiques sont synchronisés temporellement. C'est cette cohérence temporelle dans des régions particulières du cortex qui permet la mesure par électroencéphalographie d'une partie de l'activité électrique du cerveau : principalement l'activité cérébrale présente à la surface du cortex. Ces cellules pyramidales sont schématisées dans la figure 2.4.

Les ondes électromagnétiques produites par ces assemblées de neurones se propagent jusqu'aux électrodes à la surface du crâne en traversant plusieurs milieux : cerveau, liquide

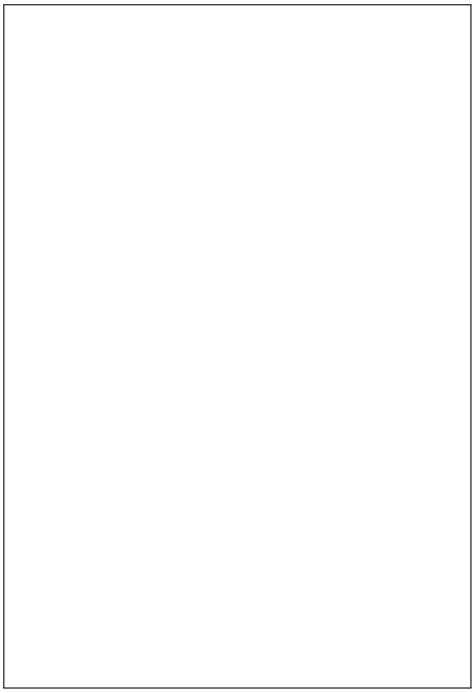


FIGURE 2.3 – Schéma d'un neurone. Source : [193]

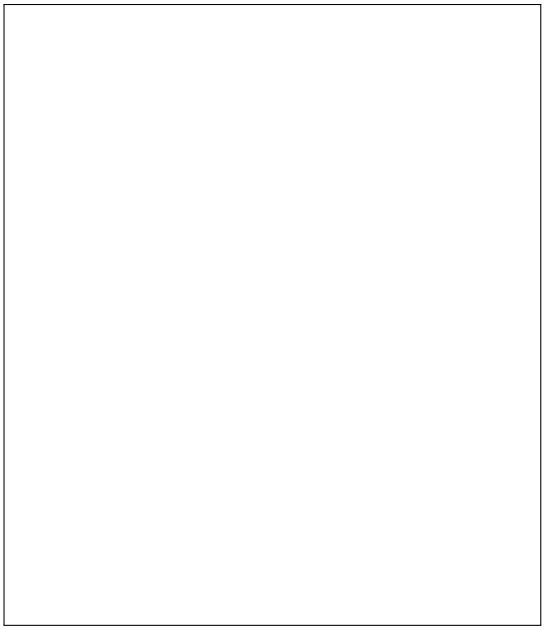


FIGURE 2.4 – Dessin de la couche corticale et des neurones pyramidaux. Source : [235].

céphalo-rachidien, os du crâne et peau. Cette propagation entraîne une atténuation du signal suivant une loi en $\frac{1}{r^2}$ (r étant la distance entre la source et les électrodes). Par ailleurs, étant donné le faible temps de propagation de l'onde comparé à la période du signal (fréquences

faibles des signaux EEG), cette propagation est généralement considérée instantanée. Cette dernière considération est appelée hypothèse quasi-statique.

2.1.3 Protocole de mesure EEG

Une mesure électroencéphalographique est réalisée via un casque d'électrodes placé sur le crâne du sujet. Les positions des électrodes sont standardisées selon le système 10-20 [112] dont un schéma est présenté dans la figure 2.4. Le nombre d'électrodes peut varier d'une expérience à l'autre de seulement une électrode à plus d'une centaine (jusqu'à 256 pour des montages à grande densité) en fonction du phénomène à étudier. Ces électrodes sont généralement en argent (Ag/AgCl) et utilisées avec un gel permettant d'améliorer la conduction. Des électrodes dites sèches (sans gel) sont aussi de plus en plus utilisées pour leur facilité de mise en place.

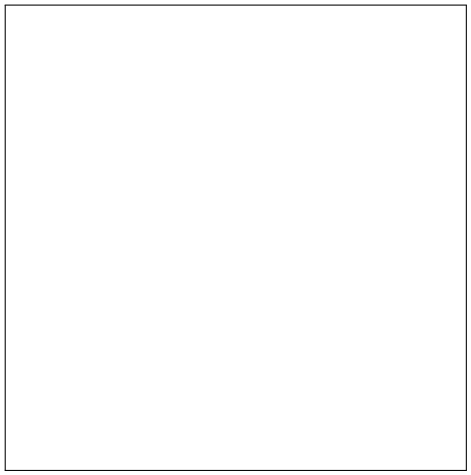


FIGURE 2.5 – Schéma de positionnement des électrodes sur le crâne pour les mesures d'électroencéphalographie : système 10-20. Source : [105].

Les différences de potentiel que l'on souhaite mesurer étant très faibles, un amplificateur différentiel est utilisé. Ces différences sont calculées entre les électrodes et une électrode de référence souvent placée sur le haut du crâne (Cz) l'oreille ou parfois le nez.

2.2 Propriétés des signaux EEG, processus neurophysiologiques et applications

Les signaux EEG se démarquent d'autres mesures de l'activité cérébrale par une résolution temporelle élevée limitée uniquement par la vitesse de mesure des capteurs. La fréquence d'échantillonnage de ces mesures se situe ainsi communément entre 100Hz et 5000Hz en fonction du matériel utilisé. Cette propriété permet leur utilisation dans l'étude de phénomènes transitoires courts de l'ordre de la centaine de millisecondes et explique la popularité de cette mesure dans les systèmes d'interfaces cerveau-machines (en plus de sa praticité). En revanche, la résolution spatiale de cette approche est assez faible. D'une part,

car ces signaux sont souvent mesurés avec des montages possédant seulement quelques dizaines d'électrodes même si les montages à grande densité d'électrodes ($\geq$ 150 électrodes) limitant cet inconvénient sont de plus en plus utilisés. D'autre part, car la faible conduction du crâne entraîne un phénomène de diffusion spatiale des ondes électromagnétiques, un phénomène dont nous discuterons plus en détails dans le chapitre 4.
Les principales difficultés apparaissant dans l'étude de phénomènes physiologiques à partir de ces signaux résident dans le rapport signal à bruit (RSB) souvent très faible de ceux-ci (de l'ordre de -20dB) et leurs grandes variabilités aussi bien entre plusieurs sujets qu'entre différentes sessions de mesures. Le signal d'intérêt est noyé dans les autres activités cérébrales et est ainsi difficile à extraire. Cette extraction peut néanmoins être facilitée par la connaissance des caractéristiques de ces signaux et l'utilisation de plusieurs enregistrements d'un phénomène (plusieurs essais). Les signaux EEG étant multidimensionnels, les caractéristiques utiles pour cela peuvent être spatiales, temporelles ou bien fréquentielles.

2.2.1 Potentiels évoqués

Les potentiels évoqués (ou PE) sont des modifications de l'activité cérébrale engendrée par des événements externes ou internes. Deux catégories de PE apparaissent en pratique : les PE stationnaires (SSEP en anglais pour Steady State Evoked Potential) et les PE transitoires (ou simplement non-stationnaires). Ce sont des réponses verrouillées en temps et en phase avec des événements [1].

2.2.1.1 Potentiels évoqués stationnaires

Ces PE apparaissent lorsqu'un sujet perçoit un stimulus périodique de manière répétée. Ces stimuli provoquent une augmentation de la puissance spectrale associée à leur fréquence de présentation et des harmoniques associées [191]. La détection de tels potentiels est donc réalisée via une analyse fréquentielle.
Spatialement, ils apparaissent dans la zone du cortex associée au sens stimulé. Les stimulus utilisés communément pour provoquer ce type de PE sont visuels, auditifs ou sensitifs. Un exemple d'analyse fréquentielle pour un SSEP visuel avec une fréquence de 7Hz est donné dans la figure 2.6.

SSEP dans les ICM : Les ICM basées sur les potentiels évoqués utilisent plusieurs excitateurs envoyant des stimuli à des fréquences différentes. Lorsque le sujet se concentre sur l'un d'entre eux en particulier, une analyse harmonique permet l'identification de celui-ci et l'exécution de la commande associée. Les stimuli les plus fréquemment utilisés dans ce contexte sont visuels [165, 241], mais des études ont montré que des stimuli somatosensoriels [164] (stimulations tactiles) ou auditifs [24] pouvait être également envisagés.
Ce sont des systèmes très populaires car en plus de ne nécessiter aucun entraînement, ils permettent au sujet de réaliser un choix entre de nombreuses possibilités ($\geq$ 5 fréquences

1. Ces réponses possèdent des instants d'arrivé et des phases décalés de valeurs fixes par rapport à celles des stimuli

FIGURE 2.6 – Spectre d'un signal EEG mesuré en position occipital pour un potentiel évoqué stationnaire visuel provoqué par des stimuli de fréquence 7 Hz. Source : [80].

différentes, jusqu'à 48 dans [84]) en gardant un taux d'identification élevé (généralement $\geq 90\%$ [59]) et possèdent ainsi un taux de transfert d'information élevé (20-30 bits/min [59] au minimum, pouvant aller jusqu'à 68 bits/min [84]) comparé à d'autres systèmes d'ICM. L'inconvénient majeur de ce type d'ICM réside dans le nombre élevé de stimuli reçu par le sujet ; ce nombre important de stimuli pouvant être assez inconfortable et fatiguant pour celui-ci, voire même dangereux pour des sujets souffrant d'épilepsie.

2.2.1.2 Potentiels évoqués transitoires

Ces potentiels sont asservis en temps et en phase à l'événement dont ils sont la réponse. Un moyennage temporel sur un grand nombre de leurs enregistrements permet leur extraction. Parmi les PE transitoires étudiés on trouve le P300 apparaissant lorsqu'un stimulus rare survient au milieu d'une suite d'autres stimuli, le N400 apparaissant dans des tâches cognitives liées aux représentations sémantiques ou bien le N170 observé lors de l'identification de visages. Le P300 est très utilisé dans le contexte des ICM.
Prenons ce dernier comme exemple pour illustrer les PE transitoires. Il survient 300 ms après un stimulus rare et est situé soit dans la zone fronto-temporale pour le P3a soit dans la zone pariétale pour le P3b (pour plus de détails voir [188]). Les profils temporels du P3a et P3b obtenus par moyennage ainsi que les topographies associées sont présentés dans la figure 2.7.

Utlisation du P300 pour les ICM : Ce PE peut être provoqué par l'apparition d'un stimulus rare cible au milieu de nombreux stimuli non-cibles. L'attention du sujet envers le stimulus cible est tout aussi importante pour son apparition que le fait qu'il n'apparaisse que rarement. Un tel potentiel a été utilisé pour la première fois dans le contexte des ICM en 1988 [74] afin de permettre à un sujet d'épeler des mots pour une application nommée « P300 speller ». Pour cela, le sujet est mis devant une grille de lettres 6×6 dont les colonnes et les lignes s'illuminent de façon aléatoire. En se concentrant sur la lettre qu'il veut utiliser et en comptant mentalement le nombre d'illuminations de celle-ci, des P300 sont induits dans son activité cérébrale à chacune des illuminations de cette lettre. La détection de ces P300 permet le repérage des lignes et des colonnes dans lesquelles la lettre apparaît et

FIGURE 2.7 – Moyenne de l'évolution temporelle des potentiels des électrodes Pz et Fz sur des enregistrements contenant le potentiel évoqué P300. Mise en évidence des potentiels P3a et P3b ainsi que de leurs topographies respectives. Source : [31]

l'identification de celle-ci. L'identification de la lettre correcte nécessite souvent plusieurs répétitions de la série d'illuminations. Le P300 est utilisé de nos jours dans des systèmes d'ICM variés avec différents types de stimuli [127], le « P300 speller » étant tout de même le plus courant avec parfois des grilles différentes [216].

2.2.2 Variation des rythmes cérébraux

Contrairement aux potentiels évoqués verrouillés à la fois en temps et en phase par rapport à un événement, d'autres types d'activités cérébrales peuvent se manifester par une réponse uniquement verrouillée en phase. Ces activités sont induites par la synchronisation ou la désynchronisation d'assemblées de neurones spécifiques et ont donc été nommées ERS et ERD en anglais pour « Event Related Synchronisation » et « Event Related Desynchronisation ». Leur extraction ne peut pas être réalisée par moyennage comme les PE mais elles présentent des caractéristiques fréquentielles pouvant être détectées. Ainsi, elles peuvent être repérées par l'observation des modifications de puissance spectrale dans certaines bandes de fréquences [184].
Le premier processus neurophysiologique de ce type à avoir été étudié est le « blocage » de la bande *alpha* repéré par Hans Berger [20] en 1929 qui apparaît lorsque le sujet ouvre ses yeux. Dans les années qui suivirent de nombreuses observations similaires mettant en évidence ce type d'activités ont conduit à leurs caractérisations à travers six bandes de fréquences dont la nomenclature est présentée ci-dessous.

Rythme *delta* δ **: 0.1-3.5 Hz** — Oscillations très lentes de l'activité électrique cérébrale. Observées principalement durant le sommeil profond de l'adulte et chez les très jeunes enfants.

Rythme *theta* θ **: 4-7 Hz** — Ces oscillations se rencontrent chez les enfants, les adolescents et les jeunes adultes. Elles sont aussi caractéristiques des états de somnolence, de certains processus émotionnels et des phases de mémorisation.

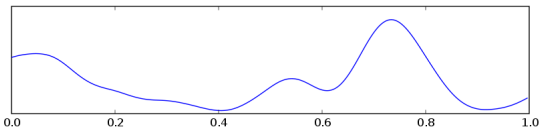
Rythme *alpha* α **: 8-12 Hz**	Ce sont les oscillations ayant les plus grandes amplitudes. Elles sont principalement localisées dans les parties occipitales du cortex. Ce rythme est associé aux états de relaxation et aux moments où les yeux sont fermés.
Rythme *mu* μ **: 7.5-15 Hz**	Ce rythme est caractéristique des activités sensorimotrices. Il est localisé au niveau du cortex moteur (ce qui le différencie du rythme alpha).
Rythme *beta* β **: 13-30 Hz**	Observé chez l'adulte éveillé, il est caractéristique des périodes d'activité intense et de concentration. Ces oscillations sont aussi liées à l'exécution de mouvements.
Rythme *gamma* γ **:** $>$ **30 Hz**	Oscillations les plus rapides pouvant aller jusqu'à 80 ou même 100 hz elles sont associées aux tâches cognitives supérieures comme le liage perceptif.

Un exemple de ces différents rythmes est donné dans la figure 2.8.

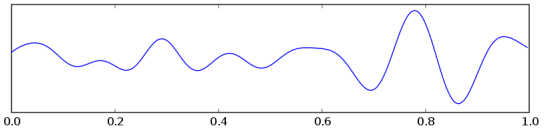
Différentes ICM ont été conçues pour exploiter la capacité d'un sujet à modifier volontairement ces rythmes cérébraux, soit directement après un entrainement, soit en exécutant une tâche mentale précise.

Rythmes corticaux lents : Des études ont montré qu'un sujet pouvait apprendre à modifier le voltage associé à des rythmes cérébraux très lents ($<$1Hz) grâce à des séances d'entrainement durant lesquelles un retour sensoriel lui permet d'évaluer celui-ci [186]. La mise en place d'une ICM fondée sur ce rythme est ainsi possible, le temps d'entrainement représentant son principal inconvénient.

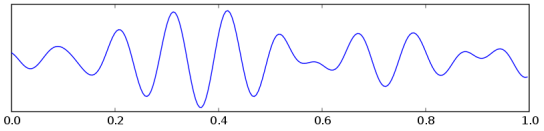
Rythmes sensori-moteurs : Des ERS/ERD peuvent être provoqués par l'exécution d'une tâche cognitive particulière sans intervention de stimulus externe. La détection de ceux-ci permet alors la détermination de l'intention du sujet. Les rythmes sensori-moteurs sont très utilisés dans ce contexte. Les différentes zones du cortex sensorimoteur sont en effet organisées de manière à ce que chacune corresponde à une zone du corps. Cette disposition est qualifiée de somatotopique. Grâce à celle-ci, certaines activités cérébrales associées à cette zone du cortex sont séparables. Cette séparation est possible lorsque ces activités présentent des localisations spatiales relativement distantes. Ces activités étant caractérisées par des ERS/ERD se répercutant sur les rythmes *mu* et *bêta*, un filtrage spatial dans ces bandes de fréquences est généralement considéré.
Les tâches cognitives associées aux mouvements des bras sont par exemple séparables. Ces tâches sont en effet situées de manière symétrique par rapport au centre du cortex sensorimoteur à une distance suffisante pour permettre une séparation. De plus, l'imagination de ces mouvements induit des activités cognitives situées dans les mêmes régions [189]. Des ICM dites d'imagerie motrice exploitant ce processus ont été conçues [185]. Celles-ci permettent à l'utilisateur d'effectuer un choix dans un programme en imaginant le mouvement



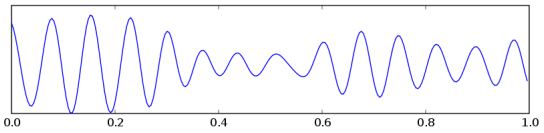
(a) Rythme *delta*.



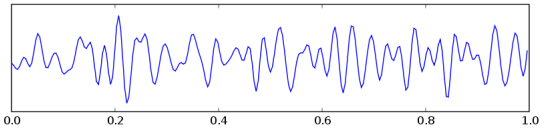
(b) Rythme *theta*.



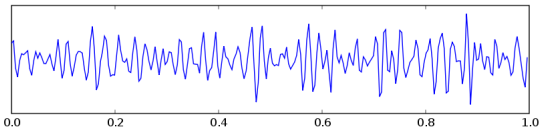
(c) Rythme *alpha*.



(d) Rythme *mu*.



(e) Rythme *beta*.



(f) Rythme gamma.

FIGURE 2.8 – Exemples de décours temporels (sur 1 seconde) pour différents rythmes cérébraux. Source : [83].

d'un des deux bras. D'autres tâches motrices peuvent également être considérées comme le mouvement de la langue ou des pieds [34].

2.2.3 Micro-états cérébraux

Les deux sections précédentes ont présenté les caractéristiques des signaux EEG lorsque des tâches cognitives particulières sont réalisées. De façon plus générale, certaines études ont montré que les signaux EEG peuvent être modélisés par une suite d'états cérébraux, chacun caractérisé par une topographie (forme spatiale) spécifique [136, 180]. Les signaux EEG présentent en effet des phases durant lesquelles une topographie reste stable et se maintient pour une durée allant de 60 à 120 ms avant de changer brutalement pour évoluer

vers une autre topographie stable. Ces états ont été associés à des étapes de réalisation de tâches cognitives.
Des études statistiques réalisées sur des suites temporelles de tels états ont lié les modifications de leurs ordres d'apparition ou de leur durée à des maladies mentales comme la schizophrénie [135] ou la dépression [205]. Ils ont également permis la caractérisation des signaux EEG associés aux états de repos d'un sujet [238]. À notre connaissance, cette propriété des signaux EEG n'a pas encore été considérée pour la création d'ICM.
Nous reviendrons sur cet aspect des signaux EEG dans les chapitres 4 et 7 dans lesquels nous considèrerons des modélisations prenant en compte cette propriété.

2.2.4 Applications

Les signaux EEG ont de nombreuses applications en supplément de leurs utilisations pour la compréhension du fonctionnement cérébral.

2.2.4.1 Diagnostic médical

Les signaux EEG permettent le diagnostic de problèmes neurologiques et de maladies mentales. La détection de l'épilepsie peut par exemple être réalisée par le repérage de changements brusques du contenu fréquentiel de ces signaux [199]. De même, les problèmes de dégénérescence cérébrale associés à la maladie d'Alzheimer ou à la maladie de Creutzfeldt-Jakob [172] peuvent être détectés par des anomalies des puissances spectrales associées à certaines bandes de fréquences. Certains problèmes mentaux comme la schizophrénie ou la dépression peuvent aussi être diagnostiqués par l'analyse d'anomalies de certains potentiels évoqués [36] de même que les effets de certaines médications [78].

2.2.4.2 ICM

Les interfaces cerveau-machine sont des systèmes permettant à un être humain de contrôler une machine via la modulation de son activité cérébrale uniquement, sans aucune aide de ses muscles [230]. Ces systèmes sont principalement utilisés dans un cadre médical, notamment pour les malades atteints du syndrome d'enfermement [128] pour lesquels ce type de systèmes offre un moyen de communication avec d'autres personnes. Ces dernières années, d'autres applications de ces systèmes ont été envisagées comme l'exploration d'environnements en réalité virtuelle ou le contrôle de jeux vidéos [131]. Toutefois, leurs performances devront être améliorées avant de réaliser ce genre de contrôles complexes. Les ICM peuvent etre modèlisées via le schéma donné dans la figure 2.9.

Les ICM sont constituées d'une boucle de rétroaction permettant le contrôle de la machine par l'humain en temps-réel. L'activité cérébrale est premièrement capturée. Cette mesure peut être réalisée via un ensemble de méthodes (EEG, MEG, IRMf, ...), l'EEG étant l'approche la plus utilisée en raison des multiples avantages évoqués en début de ce chapitre. Une fois la mesure effectuée, les signaux capturés sont pré-traités afin de pouvoir en extraire des caractéristiques permettant leur classification. La sortie du classifieur est ensuite transformée en contrôle pour la machine et un retour sensoriel (la plupart du temps

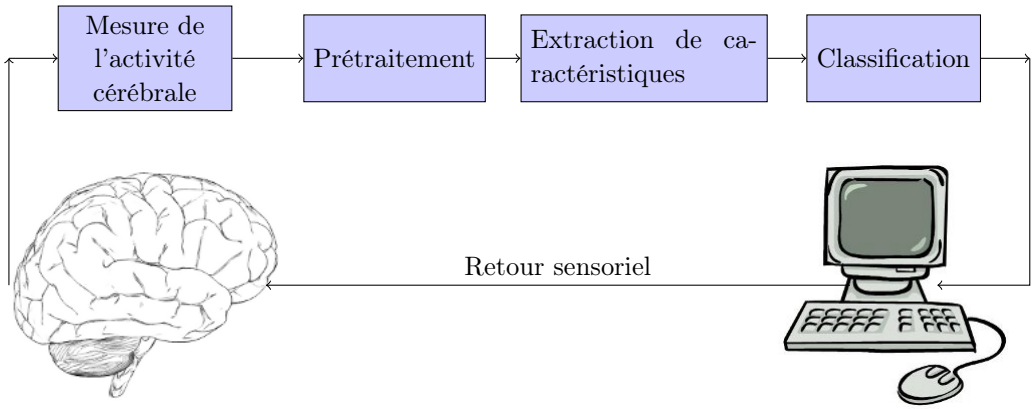


FIGURE 2.9 – Schéma de fonctionnement d'une interface cerveau-machine.

visuel) est envoyé au sujet afin qu'il puisse mettre à jour son activité cérébrale en fonction de ce que le système a interprété de ses intentions.

2.3 Méthodes d'analyse des signaux EEG

Depuis la découverte de ces signaux, de nombreux travaux ont eu pour objet la conception de méthodes permettant de les analyser et de les comprendre. Nous nous attachons ici à la description des principales approches développées. Les méthodes liées à la classification en ICM seront aussi brièvement décrites.
Dans toute cette section nous considérons l'analyse de signaux EEG mesurés sur T pas temporels et C électrodes. Ces signaux sont représentés par des matrices de mesure : $Y \in \mathbb{R}^{T \times C}$.

2.3.1 Modélisations temporelles

Dans une modélisation dynamique du signal, sa valeur à l'instant t est définie en fonction de ses valeurs aux instants précédents. De manière générale, ces modèles peuvent être décrits graphiquement par le schéma présenté dans la figure 2.10.
La suite des $X(t,.)$ représente la suite des états cachés du système étudié [2] au cours du temps tandis que les $Y(t,.)$ sont les variables observées et donc les seules utilisables directement pour la compréhension du comportement de celui-ci. Pour les signaux EEG, les états $X(t,.)$ du système sont les états mentaux du sujet et les observations $Y(t,.)$ sont les mesures de l'activité électrique par les électrodes.
Ces modèles sont caractérisés dans un formalisme probabiliste markovien par :

— une ditribution *a priori* du premier état $p(X(0,.))$
— une distribution de transition entre états $p(X(t,.)|X(t-1,.),\ldots,X(0,.))$,

2. Ces états sont codés ici par des vecteurs lignes concaténés dans la matrice X.

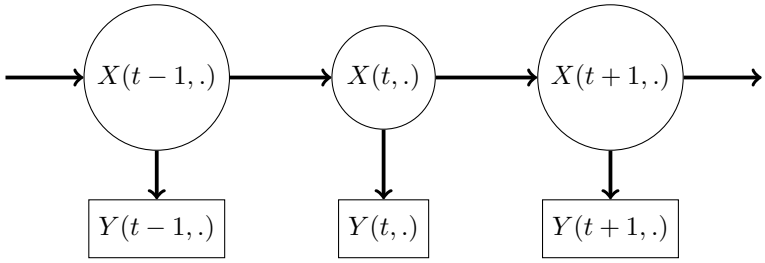


FIGURE 2.10 – Modélisation temporelle générale des signaux EEG, l'évolution des états (cachés) cérébraux est décrite par la matrice X tandis que les mesures associées obtenues sur les électrodes sont représentées par la matrice Y.

— un modèle d'observation décrit par la distribution $p((Y(t,.)|X(t,.))$, ces observations étant indépendantes entre elles.

Toutes les approches d'analyse dynamique n'utilisent pas ce type de formalisme mais la plupart d'entres elles peuvent être exprimées dans ce cadre. Pour le traitement des signaux EEG, les modèles temporels les plus couramment utilisés sont les modèles MVAR (« Multivariate autoregressive » en anglais) et MVARMA (pour « MVAR moving average ») [232]. Ils peuvent s'écrire pour le canal i d'un signal Y comme suit :

AR :

$$Y(t,i) = -\sum_{k=1}^{p} A(i,k)Y(t-k,i) - \sum_{j=1,j\neq k}^{C}\sum_{k=1}^{p} A(j,k)Y(t-k,j) + N(t,i)$$

ARMA :

$$Y(t,i) = -\sum_{k=1}^{p} A(i,k)Y(t-k,i) - \sum_{j=1,j\neq k}^{C}\sum_{k=1}^{p} A(j,k)Y(t-k,j) + \sum_{k=1}^{q} B(k,i)N(t-k,i)$$

où p et q sont les ordres de prédiction et N le bruit. Ce sont des modèles linéaires à paramètres constants dans lesquels les observations et les états sont fusionnés et appartiennent à un espace continu.
Ces modèles ont été introduits en EEG au début des années 70 [146, 86] et ont notamment permis la réalisation d'analyses spectrales de ces signaux de manière efficace avec une grande précision (estimateur spectral avec entropie maximum) et peu de paramètres. Ils ont été utilisés pour l'analyse de phénomènes transitoires [142], la segmentation des signaux EEG [29] et l'élimination d'artefacts oculaires et musculaires [200].
La limite principale de ces modèles réside dans leurs paramètres fixés pour toute la durée du signal, imposant une étape de segmentation parfois difficile (non-stationnarité des signaux EEG) avant le traitement des signaux. Ainsi, des modèles AAR (« adaptive » AR aussi appelé TVAR pour « time varying » AR) permettant des variations de ces paramètres au cours du temps ont été mis en place. Ils ont permis notamment l'étude des différentes phases du sommeil [200] et se sont montrés particulièrement intéressants pour la caractérisation de

phénomènes transitoires [9]. Cette approche a également été utilisée dans le cadre des ICM pour la classification de tâches motrices [185].
Récemment, une modélisation par chaines de Markov cachées utilisant toutes les possibilités du schéma présenté plus haut a été considérée pour les signaux EEG et appliquée à la classification de signaux d'imagerie motrice [174]. Cette modélisation, présentée dans la figure 2.11, s'est montrée plus efficace pour la séparation du mouvement des bras que des méthodes de classification linéaire.

FIGURE 2.11 – Automate de Markov avec états cachés pour la représentation de signaux EEG. Source : [174]

Par ailleurs, d'autres types de modélisation ont également été envisagés pour l'étude de ces signaux. Des modélisations chaotiques de leurs comportements dynamiques ont par exemple été mises en place et se sont montrées efficaces dans l'étude de l'épilepsie. Le lecteur pourra se référer à [114] pour plus de détails sur ces modélisations. De même, la modélisation causale dynamique (DCM en anglais pour « Dynamic causal modelling ») appliquée récemment aux signaux EEG a fourni une nouvelle approche permettant par exemple une caractérisation intéressante de réponses évoquées [53].

Ces modèles peuvent tous être exprimé sous la forme d'une représentation d'état s'écrivant de manière générique :

$$X(t+1,.) = AX(t,.) + BN_e(t,.),$$
$$Y(t+1,.) = CX(t,.) + DN_o(t,.),$$

permettant par exemple une résolution via la méthode du filtre de Kalman comme dans [85]. Cette dernière réalise pour chaque pas de temps une prédiction de l'état courant en fonction de l'état précédent puis corrige la prédiction à partir des observations.

À noter que ces modélisations dynamiques présentent un intérêt à la fois génératif étant donné qu'elles permettent d'expliciter la construction des signaux EEG et discriminatif, permettant dans certains cas la séparation de différentes classes d'activités cérébrales.

2.3.2 Transformées et analyses temps-fréquence

Les propriétés fréquentielles des signaux EEG sont très étudiées du fait de leurs liens directs avec les processus neurophysiologiques. Les analyses harmoniques sont généralement effectuées grâce à une transformée de Fourier discrète des signaux. Pour le canal j du signal Y elle s'écrit :

$$Y_f(k,j) = \sum_{t=1}^{T-1} Y(t,j) \exp(-\frac{2i\pi kt}{T})$$

Cette transformée a permis de nombreuses avancées dans la compréhension de l'activité cérébrale [121, 61] et notamment l'identification d'associations entre certaines tâches cognitives et des bandes de fréquences particulières [104]. De la même façon, les transformées en cosinus et sinus discrètes ont permis la caractérisation de rythmes lents tels que ceux observés lorsqu'un sujet est sous l'effet de drogue ou durant certaines phases du sommeil [3]. Elles s'écrivent (même forme pour le sinus) :

$$Y_c(k,j) = \sum_{t=1}^{T-1} Y(t,j) \cos(\frac{\pi}{T}(t+\frac{1}{2})k)$$

Le problème principal de ces transformées réside dans leur traitement purement fréquentiel des signaux qui est efficace pour la description de phénomènes stationnaires mais échoue dans la caractérisation de processus transitoires. Par conséquent, des représentations temps-fréquence ont été considérées afin de décrire ces derniers. La transformée de Fourier discrète à court terme fut notamment utilisée [63]. Celle-ci réalise la transformée de Fourier du signal multiplié auparavant par une fonction fenêtre w (souvent gaussienne) sélectionnant un interval temporel spécifique :

$$Y_{sf}(k,m,j) = \sum_{t=1}^{T-1} Y(t,j)\ w(t-m)\ \exp(-\frac{2i\pi kt}{T})\ .$$

Des transformées en ondelettes furent ensuite considérées. De la même manière que les transformées précédentes, leurs composantes correspondent à des produits scalaires entre le signal à analyser et un ensemble de signaux basiques localisés en temps et en fréquence. Ces signaux basiques sont issus de la discrétisation de fonctions issues d'une unique fonction mère par dilatation et décalage temporel afin de couvrir la partie du plan temps-fréquence étudiée. Comparée à la transformée de Fourier à court terme, elles permettent une meilleure caractérisation de certaines singularités des signaux. Pour le traitement des EEG, elles se sont notamment montrées efficaces pour la prévision des crises d'épilepsie [87]. Une visualisation des représentations temps-fréquence obtenues avec ces transformées est présentée dans la figure 2.12.

Enfin, l'analyse temps-fréquence des signaux EEG a été considérée d'un point de vue génératif, *i.e.* à travers la recherche de composantes temps-fréquences permettant la reconstruction des signaux dans un ensemble de signaux basiques nommé dictionnaire. Pour un dictionnaire constitué d'une base orthogonale, la décomposition d'un signal sur celui-ci

FIGURE 2.12 – Exemple d'une représentation temps-fréquence de mesures EEG enregistrées pendant le sommeil d'un adulte obtenue à l'aide d'un dictionnaire de Gabor. Les lettres représentent des structures significatives du sommeil identifiées par un expert. Source : [63].

peut être réalisée de la même façon que dans les transformées précédentes (produit scalaire avec les éléments). Dans le cas d'un dictionnaire possédant plus d'éléments que la taille des signaux, la décomposition sélectionne quelques composantes permettant une reconstruction suffisante du signal décomposé. Ces décompositions dites parcimonieuses seront étudiées en détail dans la suite de ce document. Dans l'analyse de signaux EEG, une telle décomposition sur un dictionnaire de Gabor a permis l'obtention de représentations temps-fréquences intéressantes [62, 64], le dictionnaire de Gabor permettant un pavage optimal du plan temps-fréquence.
Pour plus de détails sur les approches temps-fréquence, le lecteur pourra se référer à [152].

2.3.3 Filtrage spatial

La dimension spatiale des signaux EEG revêt une grande importance étant donné que la plupart des processus neurophysiologiques sont localisés dans des régions spécifiques du cortex cérébral. De nombreuses stratégies de filtrage spatial ont été étudiées afin de mieux extraire certaines activités cérébrales ou de permettre une meilleure séparation de plusieurs d'entre elles. Ces filtrages spatiaux sont réalisés de manière linéaire :

$$F = YU$$

avec U la matrice regroupant les filtres spatiaux. Ce modèle linéaire se base sur l'hypothèse d'une mesure au niveau des électrodes d'un mélange instantané des signaux provenant de sources à l'intérieur du cerveau (approximation quasi-statique).

La sélection de certaines électrodes pour l'étude de tâches cognitives est la manière la plus simple de filtrer spatialement des signaux EEG. Un expert est nécessaire pour réaliser

ce choix. Il peut ne pas être idéal pour tous les sujets, mais cette sélection est suffisante pour l'analyse de certaines activités cérébrales très localisées. Pour cela, soit $\mathbf{e} \in \mathbb{R}^N$ le vecteur des indices des N électrodes sélectionnées, la i-ième colonne de $U \in \mathbb{R}^{C\times N}$ s'écrit :

$$U(j,i) = \begin{cases} 1 & \text{si } j = \mathbf{e}(i) \\ 0 & \text{sinon.} \end{cases}$$

L'utilisation d'une électrode de référence dont le signal est soustrait aux signaux de toutes les autres est aussi une manière de filtrer ces signaux. Le choix de celle-ci est un sujet de discussion au sein de la communauté et influence de façon importante les résultats de certaines méthodes [103]. Soit e l'électrode choisie, la i-ième colonne de $U \in \mathbb{R}^{C\times C-1}$ s'écrit :

$$U(j,i) = \begin{cases} 1 & \text{si } j = i \\ -1 & \text{si } j = e \text{ et } i \neq j \\ 0 & \text{sinon.} \end{cases}$$

Il est également possible de prendre comme électrode de référence une électrode virtuelle ayant pour signal la moyenne de ceux obtenus sur les autres électrodes. Cette approche nommée CAR pour « Common Average Reference » est assez populaire et peut-être vue comme un filtre spatial passe haut éliminant la composante spatiale constante des canaux (de fréquence spatiale nulle). Une justification théorique de cette méthode est donnée dans [21]. Quelle que soit la référence choisie, ce choix vise en général à éliminer un biais spatial permettant l'obtention de meilleures performances des approches utilisées par la suite. $U \in \mathbb{R}^{C\times C}$ s'écrit dans ce cas :

$$U(j,i) = \begin{cases} \frac{C-1}{C} & \text{si } j = i \\ -\frac{1}{C} & \text{sinon.} \end{cases}$$

Un filtrage laplacien est aussi couramment considéré pour le traitement des données EEG. La justification d'un tel filtrage tient à la forme lisse des profils spatiaux de ces signaux. Leurs composantes sont, après avoir été émises par une zone du cortex, diffusées par le crâne avant la mesure sur les électrodes. Cette diffusion entraîne une localisation difficile des activitées cérébrales, qui apparaissent « floutées » spatialement dans les mesures EEG, d'où l'utilisation d'un filtrage laplacien pour améliorer cette mesure. Soit $V(i)$ l'ensemble des électrodes se situant dans le voisinage de l'electrode i, la e-iéme colonne de la matrice $U \in \mathbb{R}^{C\times C}$ s'écrit :

$$U(j,i) = \begin{cases} \#V(i) & \text{si } j = i \\ -1 & \text{si } j \in V(i) \\ 0 & \text{sinon.} \end{cases}$$

La comparaison de ce filtrage avec l'utilisation d'une électrode de référence est présentée

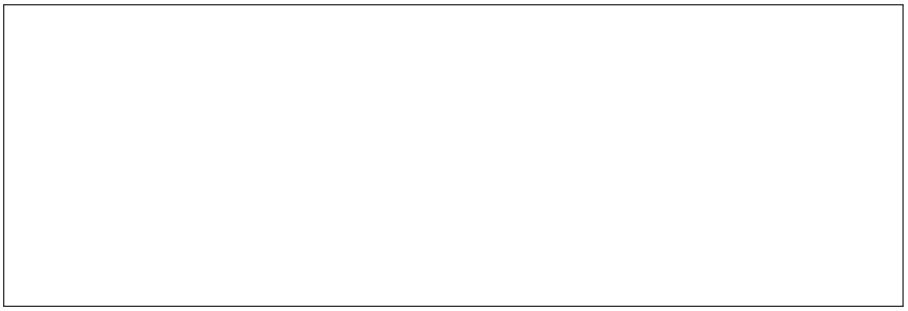


FIGURE 2.13 – Visualisation de différents filtres spatiaux : électrode de référence (oreille), CAR, laplacien étroit et laplacien large. La moyenne des activités des électrodes en noir est soustrait de l'activité de l'électrode C3 en rouge. Source : [157].

dans [157]. Ces filtres sont schématisés dans la figure 2.13. À noter que des filtres laplaciens à base de splines ont aussi été développés [204].

Même s'ils sont relativement efficaces, ces filtres ne sont ni adaptés aux tâches cognitives à extraire ni aux sujets. Par conséquent, des approches par composantes ont été utilisées pour ce filtrage, permettant la conception de filtres spatiaux optimaux pour des tâches spécifiques et à des sujets particuliers. Parmi celles-ci, l'analyse en composantes principales (ACP) ainsi que l'analyse en composantes indépendantes (ACI) sont très populaires. N'étant pas seulement utilisées pour le filtrage spatial, nous discuterons de celles-ci plus longuement dans la section suivante qui traite spécifiquement de ce type d'approches.

Reconstruction de sources cérébrales. L'ensemble de ces approches de filtrage tente de se rapprocher des « vraies » composantes des signaux EEG. Ces composantes correspondent aux activités des assemblées de neurones sources des signaux mesurés à la surface du crâne. La recherche et la séparation de ces activités a donné également lieu à de nombreux travaux (voir [99] pour une revue). Le nombre de sources étant bien plus important que le nombre d'électrodes, ce problème ne présente pas de solution unique (problème mal conditionné). Ainsi, un filtrage spatial linéaire classique comme ceux présentés plus haut ne permet pas la séparation de ces sources. Ce problème nommé aussi problème inverse EEG s'exprime naturellement à travers un modèle linéaire redondant de la forme :

$$Y = GX + N$$

pour un signal $Y \in \mathbb{R}^{C \times T}$, une matrice de gain $G \in \mathbb{R}^{C \times S}$ ($C \ll S$), une matrice de coefficient $X \in \mathbb{R}^{S \times T}$ et N une matrice de bruit.
La matrice de gain (ou « lead field matrix ») lie les sources cérébrales aux mesures sur les électrodes. Cette matrice peut être estimée via la résolution du problème direct EEG, consistant à modèliser la propagation des ondes électromagnétiques dans le crâne de manière à obtenir la valeur du signal mesuré sur les électrodes lorsqu'une source est active [95, 163]. La sélection d'une solution pour ce problème inverse est réalisée en fonction de connaissances *a priori* sur les sources. La résolution de ce type de problème étant directement liée aux représentations étudiées dans ce document, nous détaillerons les approches proposées une fois les représentations redondantes introduites formellement dans la section 3.5.2.3.

2.3.4 Analyse en composantes

Les approches par composantes cherchent à extraire de ces signaux un ensemble de composantes les constituants. Ces composantes peuvent être spatiales, temporelles ou fréquentielles, elles sont donc transverses à ces différents aspects. Ce sont donc des approches principalement génératives. Elles peuvent être catégorisées en fonction des dimensions des composantes extraites. Celles utilisées pour l'EEG considèrent le plus souvent un modèle linéaire de type :

$$Y = \Phi X \ .$$

2.3.4.1 Analyse concernant une seule dimension

L'analyse en composantes principales (ACP) est la première de ces méthodes à avoir été utilisée pour l'analyse de signaux EEG [129, 33, 208]. C'est une méthode d'analyse statistique permettant de décorréler les données. Pour cela, les signaux sont projetés dans une nouvelle base $X = UY$ de tel sorte que leurs matrices de covariance après projection soient diagonales. Pour un signal Y possédant une matrice de covariance empirique $\Sigma = Y^T Y$, le calcul de U peut être réalisé via la diagonalisation de Σ (ou de manière équivalente la décomposition en valeurs singulières de Y) :

$$\Sigma = Y^T Y = W D W^T \ .$$

Il est alors possible de choisir $U = W^T$ ou bien $U = D^{\frac{-1}{2}} W^T$ afin de décorréler Y, le second choix permettant un blanchiment du signal de manière à obtenir l'identité comme matrice de covariance. Pour des signaux distribués de manière gaussienne ayant des moments d'ordre supérieur à deux nuls, cette transformation permet l'obtention de composantes indépendantes de ceux-ci.
D'un point de vue géométrique, cette méthode calcule les axes principaux des données considérées. Une réduction de dimension peut être réalisée via cette approche en choisissant les vecteurs propres de la diagonalisation de Σ associés aux plus grandes valeurs propres.
De nombreux travaux ont analysé, appliqué et étendu cette méthode. Pour une revue récente le lecteur pourra se référer à [1].
Cette méthode a été particulièrement étudiée pour l'analyse spatiale des signaux EEG et des variantes adaptées à un cadre discriminatif ont été développée. Ces méthodes permettent la séparation de plusieurs activités cérébrales spécifiques ou bien le rehaussement d'une de ces activités par rapport aux autres classes. Leur point commun est la recherche des extrema de la fonction suivante :

$$J(\mathbf{u}) = \frac{\mathbf{u}^T \Sigma_1 \mathbf{u}}{\mathbf{u}^T \Sigma_2 \mathbf{u}}. \tag{2.1}$$

$\mathbf{u}$ représente ici l'un des filtres spatiaux de U, Σ_1 la matrice de covariance d'une classe de signaux et Σ_2 la matrice de covariance d'une autre classe (ou du bruit). J est une fonction connue sous le nom de quotient de Rayleigh [82]. L'invariance d'échelle de cette fonction $J(\alpha\mathbf{u}) = J(\mathbf{u})$ permet d'écrire ce problème sous la forme :

$$\underset{\mathbf{u}}{\text{argopt}} \quad \mathbf{u}^T \Sigma_1 \mathbf{u} \quad \text{t.q} \quad \mathbf{u}^T \Sigma_2 \mathbf{u} = 1. \tag{2.2}$$

Sa solution aprés dérivation du lagrangien est : $\Sigma_1 u = \lambda \Sigma_2 u$. Ainsi, les *extrema* de J sont les vecteurs propres généralisés des matrices Σ_1 et Σ_2.
Dans un contexte de classification à deux classes, les filtres spatiaux retenus sont les vecteurs propres associés aux valeurs propres les plus faibles et les plus élevées. Cette approche porte alors le nom de CSP pour « Common Spatial Pattern » [189] et est efficace pour la classification de tâches motrices. De nombreuses variantes de cette approche ont été developpées, le lecteur intéressé pourra se référer à [27] pour une revue. Dans le cas du rehaussement d'un potentiel évoqué par rapport au bruit, une approche nommée xDAWN [192] a été créée. Seuls les filtres associés aux plus grandes valeurs propres sont alors sélectionnés. Les performances de ce type d'approche dépendent principalement de la qualité de l'estimation des matrices de covariances [145].

L'analyse en composantes indépendantes (ACI) a été également considérée pour les signaux EEG [150]. Elle s'est notamment montrée particulièrement efficace pour l'élimination d'artefacts occulaires [225]. Cette approche, permet l'extraction de composantes statistiquement indépendantes des signaux par projection linéaire comme précédemment : $X = UY$. Contrairement à l'ACP réalisant uniquement une décorrelation, celle-ci cherche une transformation rendant nuls tous les moments d'ordre 2 et supérieurs. Différentes approches ont été développées pour l'obtention de la matrice U de projection. Celles-ci considèrent pour cela différents critères sur les composantes de sortie X :

— maximum de vraisemblance $\Leftrightarrow$ minimum de la distribution de Kullback-Leibler,
— minimisation de l'entropie
— minimisation de l'information mutuelle.

Contrairement à l'ACP ces filtres ne sont pas nécessairement orthogonaux.

Le lecteur pourra se référer à [113] pour plus de détails.

Des études EEG ont également développé des méthodes de clustering. Celles-ci associent à chaque élément d'un signal un des centres des clusters et peuvent ainsi être vues comme des approches par composantes où les éléments ne sont pas associés à une combinaison de composantes, mais à une seule. Une telle approche est utilisée avec des composantes spatiales pour le calcul des microétats présentée dans la section 2.2.3 [180].

2.3.4.2 Analyse concernant plusieurs dimensions

D'autres méthodes tentent d'extraire des composantes sur plusieurs dimensions simultanément. L'approche PARAFAC qui est une généralisation de la ACP permet par exemple l'extraction de composantes sur deux [160] ou trois [159] dimensions. Les composantes des différentes dimensions sont liées de manière à ce que la méthode soit équivalente à l'extraction de composantes multidimensionnelles de rang 1. Elle a notamment été utilisée pour la caractérisation de certains PE.
L'extraction de composantes spatio-temporelles plus générales a aussi été étudiée à travers l'apprentissage d'un dictionnaire multivarié redondant [16]. Cette approche réalisée de manière invariante dans le temps s'est montrée efficace pour l'évaluation du P300. Cette étude fut réalisée en parallèle des travaux présentés dans ce document (voir section 3.5.2.4).

2.3.5 Sélection de caractéristiques et classification

Dans un contexte discriminatif, généralement pour les ICM, la classification des signaux EEG est réalisée via la chaine de traitement présentée dans la figure 2.9. Le choix des caractéristiques utilisées pour la classification dépend des tâches cognitives considérées. Cette extraction peut être effectuée par les différentes approches présentées dans les sections précédentes. En ce qui concerne l'étape de discrimination, elle est généralement composée d'une étape de sélection de caractéristiques, puis de l'étape de classification proprement dite.
Deux approches peuvent être considérées pour la sélection des caractéristiques. Celle-ci est effectuée, soit, par le classifieur lui-même (voir par exemple [214] pour une intégration des différentes étapes discriminatives des ICM), soit, indépendamment du classifieur avant l'étape de classification [125](« wrapper vs filter »). Les approches gloutonnes ainsi que celles utilisant une mesure de corrélation (ou d'information mutuelle) entre les caractéristiques et les labels sont les plus courantes pour les ICM [206, 6].
En ce qui concerne l'étape de classification, différentes approches peuvent également être adoptées. Le choix du classifieur dépend de plusieurs critères tels que la taille des signaux, la valeur du rapport signal à bruit, la variance des données ou encore la taille de l'ensemble d'entrainement [144].
Nous notons ici $F \in \mathbb{R}^{N\times K}$ la matrice de caractéristiques associée aux signaux $Y^k, k \in \{1,\dots,K\}$ et $\mathbf{z} \in \mathbb{R}^K$ le vecteur des étiquettes de ces signaux. De manière générale, les algorithmes d'apprentissage supervisé tentent d'apprendre une fonction f permettant d'obtenir l'étiquette d'un signal en fonction de ses caractéristiques. Dans le cadre probabiliste de la théorie de la décision [106] cette fonction est évaluée via la minimisation du *risque* s'exprimant comme suit :

$$min_f \;\; \mathbb{E}(l(f(F),\mathbf{z}))$$

avec $\mathbb{E}$ l'espérance et l une fonction de perte mesurant la proximité entre les vraies étiquettes et celles obtenues avec f. Ce risque est estimé en considérant la distribution empirique des données, permettant d'écrire la minimisation précédente :

$$\min_f \;\; \frac{1}{K}\sum_{k=1}^{K} l(f(F(k)),\mathbf{z}(k))) \;\; .$$

Afin d'évaluer un tel apprentissage de f, les données disponibles sont généralement divisées en deux ensembles, un ensemble d'entraînement sur lequel f est apprise et un ensemble de tests permettant d'évaluer l'efficacité de f sur de nouvelles données. L'un des inconvénients de ce type d'apprentissage concerne la capacité de généralisation de la fonction apprise. Son apprentissage peut en effet la rendre très spécialisée pour les données d'entrainement mais peu performante pour la classification de signaux non-présent dans l'ensemble d'entrainement. Ce problème, nommé sur-apprentissage est illustré dans la figure 2.14.
L'ajout d'une régularisation R contraignant cette fonction peut permettre de limiter ce sur-apprentissage. L'apprentissage de f s'écrit alors :

$$\min_f \;\; \frac{1}{K}\sum_{k=1}^{K} l(f(F(k)),\mathbf{z}(k))) + R(f)$$

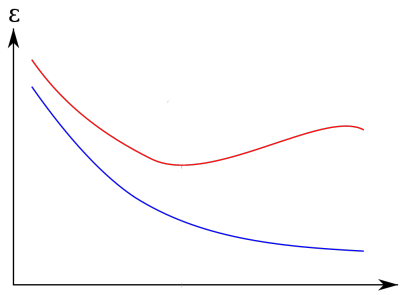


FIGURE 2.14 – Taux d'erreur de classification en fonction de la taille de la base d'entrainement. En bleu, la classification sur l'ensemble d'entrainement et en rouge sur l'ensemble de test.

Parmi les fonctions de perte classique, on trouve par exemple la fonction de perte quadratique $l(f(F(k)), \mathbf{z}(k))) = (f(F(k)) - \mathbf{z}(k))^2$ ou la « hinge loss » $l(f(F(k)), \mathbf{z}(k))) = \max(0, 1 - \mathbf{z}(k) f(F(k)))$. En ce qui concerne les régularisations, elles dépendent de la structure de f et pénalisent fréquemment la norme (ℓ_1, ℓ_2, ..., etc.) des paramètres de f en fonction d'*a priori* sur la solution du problème.

Les signaux EEG étant très bruités et souvent de grandes dimensions, les algorithmes les plus couramment utilisés sont linéaires. Ceux-ci sont robustes vis-à-vis du bruit et peuvent classifier des signaux de grandes dimensions dans des temps raisonnables.
Le modèle linéaire pour le signal k s'écrit :

$$\mathbf{z}(k) = \sum_{n=1}^{N} \mathbf{w}(n+1) F(n,i) + \mathbf{w}(1), \;\Leftrightarrow\; \mathbf{z}(k) = \sum_{n=1}^{N+1} \mathbf{w}(n) G(n,i),$$

ou $\mathbf{w}$ est le vecteur de coefficients du classifieur et G la matrice des vecteurs de caractéristiques augmentées par une rangée de 1. Géométriquement, le vecteur de coefficients w est un hyperplan de l'espace de caractéristiques séparant les classes dont un exemple en dimension 2 est donné dans la figure 2.15.

Les algorithmes de classification linéaire les plus courants pour les ICM sont le LDA (Li-

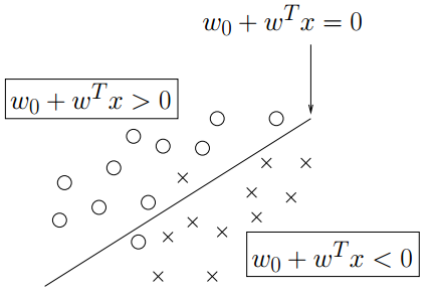


FIGURE 2.15 – Séparation linéaire de deux classes de vecteurs par une droite.

near Discriminant Analysis), le FDA (Fisher Discriminant Analysis) et le SVM (Séparateur

à Vaste Marge). La principale différence entre le LDA et le FDA réside dans l'hypothèse d'une distribution gaussienne des données que fait le LDA, le FDA ne s'appuyant que sur les moyennes et variances des données. En ce qui concerne le SVM, il permet une séparation maximisant les marges entre les données et l'hyperplan au prix du réglage d'un paramètre. La popularité de ce dernier provient également de la possibilité de séparations non-linéaires qu'elle permet à l'aide de fonctions noyaux permettant des séparations linéaires dans des espaces de plus grandes dimensions sans surcoût de calcul. De nombreux classifieurs ont été étudiés dans le contexte des ICM, le lecteur intéressé pourra se référer à [144] pour une comparaison de ceux-ci ainsi qu'à [60] pour une vision générale de la reconnaissance de motifs.

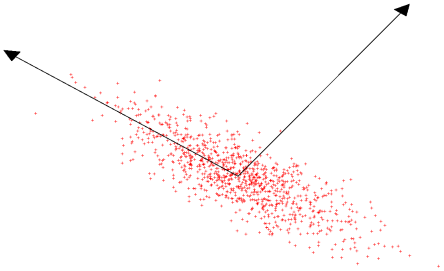
2.4 Représentations redondantes pour les EEG

2.4.1 Principes des approches redondantes

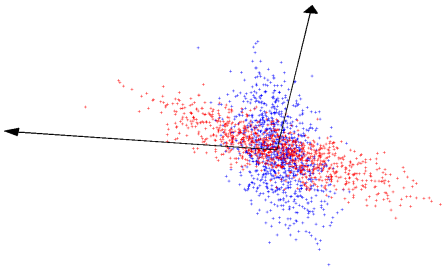
Dans ce travail, nous nous attachons à la mise en place de représentations redondantes pour les signaux EEG. Ces représentations sont des approches par composantes dont le nombre de composantes excède la dimension des signaux étudiés. Ces familles de composantes génèrent donc l'espace de nos signaux, mais ne sont pas contraintes à être orthogonales comme les composantes de l'ACP et peuvent comporter un nombre d'éléments plus grand que la taille des signaux contrairement à l'ACP ou l'ACI. Elles offrent ainsi une souplesse de représentation beaucoup plus grande que les autres approches.
Cette souplesse est mise en avant pour des signaux synthétiques ci-dessous (inspirée de [139]). Ces signaux sont générés à partir de plusieurs distributions gaussiennes multimodales 2D (1000 points par distribution). Les différentes directions de variance des distributions sont mises en avant par des couleurs bien que tous les points appartiennent à une seule classe.
La figure 2.16a présente une distribution gaussienne 2D simple ainsi que les directions obtenues avec une ACP. L'ACP est ici suffisante pour décrire les données étant donné que la méthode capture correctement la direction de variance maximale. Elle échoue par contre sur le second exemple présenté figure 2.16b où la distribution possède 2 directions non-orthogonales. L'utilisation d'une ACI (figure 2.16c) permet par contre une description efficace de ces données en retrouvant les deux directions principales de variance grâce à des composantes moins contraintes. L'exemple final donné dans la figure 2.16d et 2.16e propose une distribution à trois directions que l'ACI n'arrive pas à capturer contrairement à une approche redondante qui retrouve correctement celles-ci.
Les approches redondantes seront abordées en détail dans le chapitre suivant.

2.4.2 Intérêt des représentations redondantes pour la description des signaux EEG

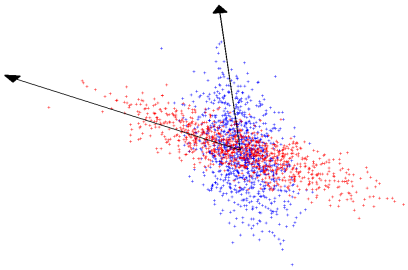
Comme nous l'avons vu dans ce chapitre les signaux EEG sont complexes et de nombreuses méthodes ont été développées pour leur analyse. Parmi les difficultés rencontrées, les principales sont l'intégration de l'ensemble des aspects de ces signaux dans une même approche, le rapport signal à bruit très faible, la variabilité des signaux aussi bien entre



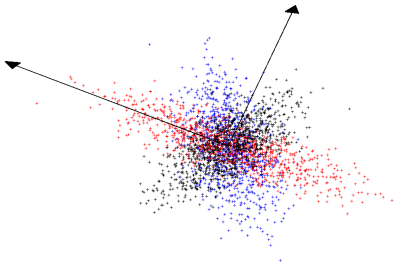
(a) ACP pour une distribution simple.



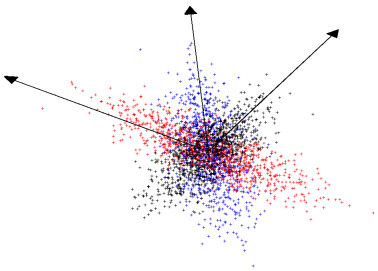
(b) ACP pour une distribution à deux directions non orthogonales.



(c) ACI pour une distribution à deux directions non orthogonales.



(d) ACI pour une distribution à trois directions.



(e) Représentation redondante pour une distribution à trois directions.

plusieurs sujets qu'entre différentes sessions de mesure et l'obtention de représentations interprétables.

Les modélisations dynamiques permettent une description fine de l'aspect temporel des EEG, néanmoins, elles nécessitent de nombreux signaux pour l'estimation de leurs paramètres et ne permettent pas l'intégration facile de l'aspect fréquentiel. En ce qui concerne les méthodes temps-fréquences que nous avons passées en revue, elles peuvent être vues comme des approches par composantes avec un ensemble de composantes fixes. Elles sont interprétables, permettent l'intégration des différents aspects des EEG mais pas d'adaptation aux variations des signaux. Enfin, les approches de filtrage spatial de la littérature permettent au contraire une adaptation à la variabilité des signaux en utilisant des approches d'extraction des composantes principales mais sont limitées à la taille des signaux

pour le nombre de composantes considérées et sont souvent peu interprétables.
Les représentations redondantes brièvement introduites plus haut pourraient permettre :

- d'intégrer les aspects spatiaux, temporels et fréquentiels de ces signaux dans un cadre de travail unifié,
- de pouvoir décrire efficacement les signaux grâce à une grande flexibilité,
- d'apprendre des composantes adaptées aux sujets,
- et d'intégrer aisément des connaissances *a priori* dans le modèle afin de guider la méthode vers des composantes plausibles et interprétables.

De plus, dans un contexte discriminatif comme celui des ICM, les approches génératives comme celle-ci peuvent se révéler intéressantes [130] afin d'éviter le surapprentissage des classifieurs lorsque les signaux sont très bruités ou lorsque la base d'entrainement contient de nombreuses données aberrantes (« outliers » en anglais).

Chapitre 3

Décomposition de signaux sur un dictionnaire redondant

Sommaire

Nous allons nous intéresser dans ce chapitre aux représentations redondantes de manière plus approfondie. Le but de ce chapitre est de présenter celles-ci pour un modèle multicanal classique de façon à permettre au lecteur de mieux les comprendre avant de passer à leur étude dans le cadre des signaux EEG.
La section 3.1 introduit ces représentations ainsi que leurs formalisations à travers la théorie des repères. La section 3.2 s'attache ensuite à formaliser le problème de décomposition parcimonieuse de signaux sur un dictionnaire tandis que la section 3.3 présente des algorithmes permettant de réaliser cette décomposition. La section 3.4 expose les principes de l'apprentissage de dictionnaires ainsi que les algorithmes classiques associés. Enfin, la section 3.5 passe en revue les modèles redondants déja mis en place pour les signaux EEG afin de permettre une meilleure compréhension des contributions de cette thèse décrites dans les chapitres suivants.
La littérature associée à ce domaine étant très importante, nous ne présentons ici que les

éléments permettant une bonne compréhension de la suite de ce texte. Pour aller plus loin, le lecteur pourra s'intéresser au livre de Ole Christensen [46] pour une introduction à la théorie des repères (« frames » en anglais) ainsi qu'à celui de Michael Elad [67] concernant la parcimonie et les représentations redondantes dans le cadre du traitement de signal. L'histoire des transformations en traitement du signal est également traitée dans [194] avec comme évolution l'apprentissage de représentations redondantes.

3.1 Représentations redondantes

3.1.1 Dictionnaires pour la représentation de signaux

Ces dernières années, de nombreuses approches considérant des modèles redondants ont été développées pour le traitement de divers types de signaux. Le principe de ces modèles est de représenter un signal par une combinaison linéaire de signaux élémentaires pris dans une famille dite redondante possédant un nombre d'éléments plus élevé que la dimension des signaux. Ces signaux élémentaires sont appelés atomes et sont réunis dans un dictionnaire. De manière formelle, dans un cadre multidimensionnel ce type de modèle linéaire peut être défini simplement comme suit.

Modèle redondant multidimensionnel (ou multicanal) :
Soit $Y \in \mathbb{R}^{C\times T}$ un signal multidimensionnel et $\Phi = [\phi_1, \ldots, \phi_{N_\Phi}] \in \mathbb{R}^{C\times N_\Phi}$ un dictionnaire ($C \leq N_\Phi$), ce type de modèle linéaire en composantes peut être écrit comme suit :

$$\begin{aligned} \forall i \in \{1, \ldots, T\}, \quad & Y(i) = \Phi X(i) + N(i) \\ & Y = \Phi X + N \end{aligned} \tag{3.1}$$

avec $X \in \mathbb{R}^{N_\Phi \times T}$ une matrice de coefficients et $N \in \mathbb{R}^{C\times T}$ une matrice de bruit. Chaque signal C-dimensionnel $Y(i)$ est approché par une combinaison linéaire des éléments de Φ dont les coefficients sont les $X(i)$.

Comme nous l'avons vu dans le chapitre précédent, l'un des intérêts majeurs de ces modèles en comparaison avec d'autres approches en composante comme l'ACP ou l'ACI réside dans leur flexibilité. Les familles constituant les dictionnaires sont génératrices des espaces qu'elles décrivent mais le nombre d'éléments les constituants n'est pas limité et aucune contrainte d'orthogonalité n'est imposée. Ainsi, ces familles sont soumises à peu de contraintes et elles permettent une description flexible d'un espace.
Pour le traitement de signal, une représentation redondante semble relativement naturelle étant donné qu'en règle générale il n'existe aucune raison de limiter la taille de la famille utilisée pour décrire un phénomène au nombre de capteurs utilisés pour le mesurer. Pour dire cela autrement, le nombre de composantes élémentaires pouvant apparaître dans un phénomène est en général beaucoup plus grand que le nombre de capteurs utilisés pour le mesurer ; d'où une meilleure capacité de description des approches redondantes [215, 13].
L'utilisation de ce type de représentations présente néanmoins certaines difficultés. La famille des éléments de Φ n'étant pas libre, la décomposition des signaux sur celle-ci n'est

pas unique. Deux problèmes principaux apparaissent donc dans la mise en place de ces représentations : le choix du critère permettant de sélectionner des coefficients de décomposition pour un signal donné ainsi que le choix du dictionnaire. Les réponses dépendent sans surprise de la classe des signaux considérés ainsi que de la tâche à effectuer. Lorsque le dictionnaire est bien choisi et la décomposition réalisée de façon à respecter la structure du signal, la représentation obtenue peut grandement faciliter la réalisation de certaines tâches (voir section 3.5).

3.1.2 Les repères

Mathématiquement, les repères (« frames ») étendent la notion de base [46] et permettent la formalisation des approches utilisant des dictionnaires.
Une base est une famille de vecteurs qui est à la fois libre et génératrice :

$$\{\Phi(k)\}_{k=1}^{C} \text{ est une base de } \mathbb{R}^C \quad \Leftrightarrow \begin{cases} span\{\Phi(k)\}_{k=1}^{C} = \mathbb{R}^C \\ \nexists \{\lambda_k\}_{k=1}^{C} \neq \mathbf{0} \text{ t.q } \sum_{k=1}^{C} \lambda_k \Phi(k) = 0. \end{cases}$$

Lorsque un espace $\mathbb{R}^C$ possède une base $\{\Phi(k)\}_{k=1}^{C}$ alors les vecteurs de cet espace peuvent s'écrire de manière unique comme combinaison linéaire des éléments de celle-ci :

$$\forall \mathbf{x} \in \mathbb{R}^C, \text{ il existe un unique ensemble de scalaires } \{\lambda_k\}_{k=1}^{C} \text{ t.q } \mathbf{x} = \sum_{k=1}^{C} \lambda_k \Phi(k) \ .$$

Un repère est une famille de vecteur de l'espace qui est génératrice de celui-ci comme une base mais qui n'est pas nécessairement libre. La propriété ci-dessus n'est alors pas toujours vérifiée. Le concept de repère est plus général que celui de base et l'englobe. De manière formelle les repères sont définis comme suit :

Définition 1. *Une famille de vecteurs $\{\Phi(k)\}_{k=1}^{N_\Phi}$ est un repère de l'espace $\mathbb{R}^C$ s'il existe des constantes $0 < A \leq B < \infty$ tel que :*

$$\forall \mathbf{x} \in \mathbb{R}^C, A\|\mathbf{x}\|_2^2 \ \leq \ \sum_{i=1}^{N_\Phi} |\langle \mathbf{x}, \Phi(i) \rangle|^2 \ \leq \ B\|\mathbf{x}\|_2^2$$

Les scalaires A et B sont appelés bornes du repère et ne sont pas uniques sauf si l'on considère pour A le supremum des scalaires vérifiant la propriété précédente et de façon équivalente l'infimum pour B. Un repère est dit ajusté si $A = B$ (« tight » en anglais), dans ce cas il préserve l'angle entre les vecteurs et la géométrie de l'espace.
Pour un repère quelconque $\{\Phi(k)\}_{k=1}^{N_\Phi}$ de $\mathbb{R}^C$, il n'existe pas de décomposition unique d'un élément $\mathbf{x} \in \mathbb{R}^C$ de cet espace sur cette famille. Toutefois, la théorie des repères fournit un théorème de représentation généralisant la décomposition obtenue pour les bases à l'aide de l'opérateur de repère.

Définition 2. *Soit $\{\Phi(k)\}_{k=1}^{N_\Phi}$ un repère de $\mathbb{R}^C$, $\mathbf{a} \in \mathbb{R}^{N_\Phi}$ un vecteur de coefficients et $\mathbf{x}$ un élément de $\mathbb{R}^C$. L'opérateur de repère S est défini par :*

$$S : \mathbb{R}^C \to \mathbb{R}^C, Sx = UU^*\mathbf{x} = \sum_{i=1}^{N_\Phi} \langle \mathbf{x}, \Phi(i) \rangle \Phi(i) \ .$$

L'opérateur S lie un vecteur de $\mathbb{R}^C$ à la combinaison linéaire des éléments du repère obtenue en prenant pour coefficients les produits scalaires du vecteur avec ces éléments. Dans le cas d'une base orthogonale $\forall \mathbf{x} \in \mathbb{R}^C \quad S\mathbf{x} = \mathbf{x}$, toutefois cette propriété n'est pas valable de manière générale pour les repères. La décomposition d'un élément de $\mathbb{R}^C$ sur un repère peut être tout de même obtenue via le théorème de représentation suivant.

Théorème 1. *Soit $\{\Phi(k)\}_{k=1}^{N_\Phi}$ un repère de $\mathbb{R}^C$ associé à un opérateur de repère S. S est inversible, auto-adjoint et*

$$\forall \mathbf{x} \in \mathbb{R}^C, \quad \mathbf{x} = \sum_{i=1}^{N_\Phi} \langle \mathbf{x}, S^{-1}\Phi(i)\rangle \Phi(i) = \sum_{i=1}^{N_\Phi} \langle \mathbf{x}, \Phi(i)\rangle S^{-1}\Phi(i)$$

La famille $\{S^{-1}\Phi(i)\}_{i=1}^{N_\Phi}$ est appelée repère dual canonique du repère $\{\Phi(k)\}_{k=1}^{N_\Phi}$. Lorsque cette famille est regroupée dans une matrice nommée Φ^*, nous avons $\Phi^* = \Phi^\dagger$ avec $\Phi^\dagger$ la pseudo-inverse de la matrice Φ du repère *i.e.* $\Phi^\dagger = (\Phi^T\Phi)^{-1}\Phi^T$.
Parmi les coefficients permettant de représenter un vecteur $\mathbf{x} \in \mathbb{R}^C$ avec les éléments d'un repère $\{\Phi(k)\}_{k=1}^{N_\Phi}$, ceux obtenus par produit scalaire avec les éléments du repère dual canonique $\{\langle \mathbf{x}, S^{-1}\Phi(i)\rangle\}_{i=1}^{N_\Phi}$ sont les coefficients de norme ℓ_2 la plus faible. Les décompositions de signaux sur des repères grâce à ces coefficients se sont notamment montrées efficaces pour des tâches de débruitage [57].

3.1.3 Représentations parcimonieuses

Dans ce chapitre nous allons nous intéresser à des ensembles de coefficients différents : ceux possédant un grand nombre d'éléments nuls, *i.e* les coefficients permettant la reconstruction d'un signal avec peu d'éléments du repère. La décomposition est alors qualifiée de parcimonieuse. Ce choix pour le critère de sélection des coefficients répond au principe du rasoir d'Occam, suggérant le choix de la solution la plus simple pour ces décompositions : ici, la solution impliquant un minimum de composantes simultanément actives dans un signal. Il est intéressant de noter que des travaux sur le cortex visuel des mammifères [76, 176] ont mis en évidence un fonctionnement de l'aire V1 utilisant ce type de représentations. Lorsque le signal d'une image perçu par l'œil est reçu dans cette aire, les cellules qui la composent se comportent comme des filtres spatio-temporels et fournissent une représentation parcimonieuse pouvant être utilisée par des neurones des couches suivantes pour l'interprétation du signal reçu. Ainsi, le type de codage parcimonieux considéré dans ce chapitre est apparu naturellement chez les mammifères au cours de leur évolution comme une approche efficace de codage de l'information.

3.2 Décomposition parcimonieuse sur un dictionnaire

Représenter un signal par décomposition parcimonieuse sur un dictionnaire est un problème non-trivial. Avant de présenter les différents algorithmes développés pour cela nous allons discuter ici des propriétés de ce problème.
Dans toute cette partie $Y \in \mathbb{R}^{C\times T}$ est un signal multidimensionnel et $\Phi = [\phi_1, \ldots, \phi_{N_\Phi}] \in \mathbb{R}^{C\times N_\Phi}$ un dictionnaire ($C \leq N_\Phi$).

3.2.1 Formalisation du problème de décomposition parcimonieuse

La décomposition parcimonieuse de Y sur Φ est formalisée à travers le problème d'optimisation suivant :

$$\underset{X\in\mathbb{R}^{N_\Phi\times T}}{\arg\min} \|Y-\Phi X\|_F^2 \quad \text{t.q.} \quad \|X\|_0 < J \tag{3.2}$$

où la quasi-norme ℓ_0 est définie comme suit :

$$\|X\|_0 = \#\{(i,j) \mid \forall i \in \{1,\dots,N_\Phi\},\ \forall j \in \{1,\dots,T\},\ X(i,j)=0\}.$$

Pour une matrice X donnée, l'ensemble des indices des éléments non nuls est appelé support de X. Cette minimisation est un problème combinatoire NP-difficile [168](voir la réduction polynomiale de ce problème à 3-SAT dans [217]). Ainsi, quand les dimensions de celui-ci augmentent, il devient difficile de le résoudre dans un temps raisonnable. La solution est en général approchée par des méthodes gloutonnes (voir section 3.3.1). Ces méthodes peuvent toutefois se révéler peu stables dans les cas où le rapport signal à bruit est faible, le dictionnaire mal conditionné ou les signaux peu parcimonieux sur le dictionnaire choisi.

Une autre possibilité a été développée pour la réalisation d'une décomposition parcimonieuse : la relaxation de la contrainte ℓ_0 en contrainte ℓ_1 [43, 210]. Ce problème de décomposition parcimonieuse avec contrainte relaxée se formalise traditionnellement comme suit :

$$\underset{X\in\mathbb{R}^{N_\Phi\times T}}{\arg\min} \|Y-\Phi X\|_F^2 + \lambda\|X\|_1. \tag{3.3}$$

Ce relâchement de la norme ℓ_0 en norme ℓ_1 permet l'obtention d'un problème convexe qui peut-être résolu via des algorithmes d'optimisation convexe classiques (voir section 3.3.2). Lorsque le dictionnaire est de rang plein, ce problème est strictement convexe et sa solution est unique. L'utilisation de cette norme provoque par contre un biais dans les décompositions du fait de la prise en compte de la valeur absolue des coefficients.
Il peut être difficile de voir en quoi la norme ℓ_1 permet l'obtention d'une décomposition parcimonieuse. Une illustration géométrique classique de cet effet peut être obtenue via le problème monodimentionnel suivant :

$$\underset{\mathbf{x}}{\arg\min} \|\mathbf{y}-\Phi\mathbf{x}\|_2^2 \quad \text{t.q} \quad \|\mathbf{x}\|_1 < J, \tag{3.4}$$

dans lequel le dictionnaire n'est composé que de deux atomes. La figure 3.1 présente les boules unités ℓ_1 et ℓ_2, ainsi que les lignes de niveau de la fonction $f : \mathbf{x} \to \|\mathbf{y}-\Phi\mathbf{x}\|_2^2$. À l'optimum, la ligne de niveau est tangente à la boule. Contrairement à la norme ℓ_2, la norme ℓ_1 encourage les solutions obtenues sur les coins de cette boule et donc les solutions parcimonieuses (voir [148] pour plus de détails).
Des normes intermédiaires $\|.\|_p, \quad p < 1$ ont également été étudiées dans ce contexte [93] ainsi que des versions lisses de la norme ℓ_0 [161], toutefois nous ne détaillerons pas celles-ci dans ce document étant donné qu'elles n'ont pas fait l'objet de travaux durant cette étude.

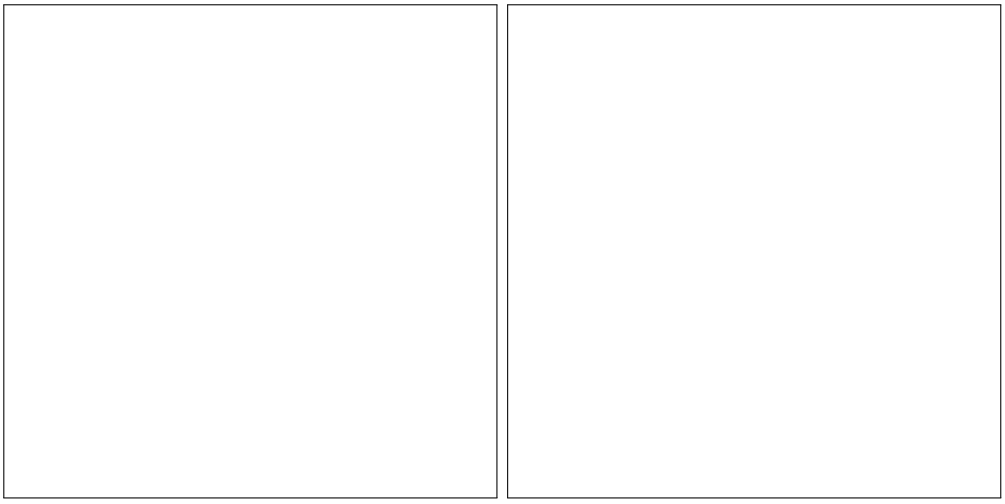


FIGURE 3.1 – Visualisation en 2D de l'effet d'une contrainte ℓ_1 sur la solution d'un problème d'optimisation. Le point rouge représente la solution du problème pour : à gauche un problème contraint par une terme ℓ_1 et à droite un problème contraint par un terme ℓ_2. Source : thèse de Julien Mairal [148].

Quelque soit la méthode utilisée, la solution du problème présenté Eq.(3.2) peut être difficile à approcher et dépend fortement du dictionnaire considéré et du degré de parcimonie des signaux. De nombreux travaux ont analysé ce problème et fourni des indices permettant d'évaluer la difficulté de celui-ci en fonction de ces deux aspects. Ces considérations théoriques sont importantes dans la mise en place de représentations redondantes et nous allons donc nous y intéresser dans la section suivante.
Avant d'aborder ces résultats il est important de noter qu'une autre formulation du problème de décomposition parcimonieuse est possible.

Analyse versus Synthèse Le problème décrit plus haut cherche à trouver une matrice de décomposition parcimonieuse et est nommé problème en synthèse. Ces dernières années un problème proche a également été étudié [69, 167], on parle de formulation en analyse (forme stricte ℓ_0 ici) :

$$\underset{X \in \mathbb{R}^{N_\Phi \times T}}{\arg\min} \|Y - X\|_F^2 \quad \text{t.q} \quad \|\Phi^* X\|_0 < J. \tag{3.5}$$

Le but de celui-ci est d'obtenir une approximation X de Y dont la projection sur Φ^* est parcimonieuse. Lorsque le dictionnaire est une base non-singulière, les problèmes formulés en synthèse et en analyse sont équivalents si l'on choisit $\Phi^* = \Phi^{-1}$. Dans un cadre redondant, cette équivalence n'est plus de mise lorsque $\Phi^* = \Phi^\dagger$ mais les problèmes obtenus sont proches (voire [69] pour une comparaison). L'équivalence peut être obtenue en construisant le dictionnaire dual décrit dans [141]. L'approche en analyse s'est montrée particulièrement efficace pour des tâches de débruitage [183]. Dans cette thèse nous considérons exclusivement des approches en synthèse, néanmoins des régularisations en analyse (de type $\|XP\|_0$) sont étudiées, voir section 3.2.3.

3.2.2 Considérations théoriques à propos des décompositions parcimonieuses

L'une des questions majeures apparaissant avec ce type de décomposition concerne l'efficacité des algorithmes résolvant les problèmes définis plus haut à retrouver la structure parcimonieuse sous-jacente des signaux. Cette efficacité dépend de certaines propriétés du dictionnaire, de la structure parcimonieuse des signaux ainsi que du niveau de bruit. Nous présentons ici quelques résultats importants.

3.2.2.1 Caractérisation des dictionnaires

Lors d'une décomposition parcimonieuse, les atomes du dictionnaire peuvent être vus comme des directions dans l'espace des signaux et sont généralement normalisés $\|\Phi(i)\|_2 = 1, \forall i \in \{1, \dots, N_\Phi\}$. La distance entre ces directions conditionne fortement la décomposition et peut être mesurée par la corrélation entre les atomes. Ainsi, l'indicateur le plus simple permettant d'évaluer le comportement d'un dictionnaire est la corrélation maximale entre deux atomes. Ce nombre correspond à la distance minimale entre deux atomes et est appelé la cohérence μ du dictionnaire :

Définition 3. *La cohérence d'un dictionnaire Φ est définie par :*

$$\mu = \max_{i,j,\ t.q\ i \neq j} |\langle \Phi(i), \Phi(j) \rangle|$$

Intuitivement, nous pouvons comprendre son importance en considérant un signal jouet $\mathbf{x}$ colinéaire à un atome du dictionnaire $\mathbf{x} = \lambda\phi$. Même lorsque celui-ci est bruité $\mathbf{x}_\varepsilon = \mathbf{x} + \varepsilon$, nous souhaiterions pouvoir retrouver ϕ grâce à une décomposition parcimonieuse. Les atomes choisis durant celle-ci sont ceux les plus corrélés avec le signal. Si un atome φ appartenant au dictionnaire est trés corrélé avec ϕ, alors sa corrélation avec $\mathbf{x}_\varepsilon$ pourrait être plus élevée que celle avec ϕ et il pourrait être choisi à sa place par erreur.
Ce n'est pas une mesure très fine du comportement d'un dictionnaire mais la cohérence est facile à calculer et permet une caractérisation rapide de celui-ci. La théorie des repères fournit une borne inférieure de cet indice appelée borne de Welch d'ordre 1 :

$$\mu \geq \sqrt{\frac{N_\Phi - C}{C(N_\Phi - 1)}}$$

Cette borne est atteinte pour des « *equiangular tight frames* »(ETF) [46], les atomes sont alors peu corrélés et la récupération des structures parcimonieuses plus aisée.
Une caractérisation plus fine des dictionnaires peut être obtenue via les valeurs des mesures de cohérence cumulée généralisant la notion de cohérence. Ceux-ci mesurent la distance minimale entre les sous-espaces engendrés par un nombre particulier d'atome [220, 102] (m ici) :

Définition 4. *La cohérence cumulée μ_1 d'ordre m (fonction de Babel) d'un dictionnaire Φ est définie comme suit :*

$$\mu_1(m) = \max_{\Delta\ t.q\ \#\Delta = m} \max_{w \in \Delta} \sum_{\lambda \in \Delta} |\langle \Phi(w), \Phi(\lambda) \rangle|$$

Le calcul de ces valeurs devient néanmoins long dès que la taille du dictionnaire dépasse plusieurs dizaines d'éléments. Un dictionnaire est dit quasi-incohérent lorsque μ_1 augmente lentement par rapport à la taille des sous-ensembles d'atomes considérés.

Une autre caractérisation des dictionnaires a également été développée dans l'étude du problème relaché [40, 38] à travers un ensemble de constantes d'isométrie (« *restricted isometric properties* ») définies comme suit :

Définition 5. *La constante d'isométrie δ_s d'un dictionnaire Φ est définie comme le plus petit nombre tel que :*

$$(1-\delta_s)\|\mathbf{x}\|_2^2 \;\leq\; \|\Phi\mathbf{x}\|_2^2 \;\leq\; (1+\delta_s)\|\mathbf{x}\|_2^2$$

pour tout $\mathbf{x}$ possédant un support de taille s.

Ces constantes peuvent être interprétées comme des mesures du « degré d'orthogonalité » du dictionnaire pour des sous-espaces de différentes dimensions.

Les mesures obtenues grâce à ces caractérisations des dictionnaires permettent de dériver des conditions sur la capacité des algorithmes de décompositions parcimonieuses à retrouver la structure sous-jacente de signaux ainsi que sur l'unicité des solutions des problèmes présentés plus haut.
Avant de s'intéresser à ces résultats, notons que dans l'optique de la mise en place d'un modèle redondant, le choix du dictionnaire est important d'un point de vue modélisation afin d'obtenir les propriétés souhaitées sur cette représentation (interprétabilité, adaptation aux signaux, ...) mais les propriétés mentionnées ici doivent également être prises en compte. Ainsi, un compromis entre la capacité descriptive du dictionnaire et son conditionnement pour les décompositions parcimonieuses est souvent nécessaire avec le modèle de décomposition présenté plus haut. Une autre possibilité proposée dernièrement considère l'utilisation d'un dictionnaire de modélisation et d'un autre dictionnaire utilisé uniquement pour le choix des atomes lors de la décomposition [201]. Les valeurs des cohérences cumulées ont d'ailleurs été étendues à cette dernière décomposition [102].

3.2.2.2 Récupération de la structure parcimonieuse

De nombreux résultats ont été obtenus concernant la capacité d'algorithmes de décomposition parcimonieuse à retrouver la décomposition la plus parcimonieuse d'un signal. Nous présentons ici quelques-uns des résultats les plus importants dans un cadre monodimensionel.
Considérons tout d'abord un cas sans bruit ou les signaux peuvent être exprimés via une combinaison parcimonieuse des atomes du dictionnaire. Le problème traité est alors :

$$\arg\min_{\mathbf{x}\in\mathbb{R}^{N_\Phi}} \|\mathbf{x}\|_* \quad \text{t.q.} \quad \mathbf{y} = \Phi\mathbf{x}. \tag{3.6}$$

ou $\|.\|_*$, peut être soit la norme ℓ_0 soit la norme ℓ_1.
Lorsque une décomposition suffisamment parcimonieuse d'un signal existe alors les problèmes ℓ_0 et ℓ_1 sont équivalents et présentent une unique solution [55].

Théorème 2. *Soit* $\mathbf{x}$ *un vecteur de coefficients tel que* $\mathbf{y} = \Phi\mathbf{x}$.
Si $\|\mathbf{x}\|_0 < \frac{1+\mu^{-1}}{2}$ *alors* $\mathbf{x}$ *est la solution du problème de décomposition* ℓ_0 *et du problème* ℓ_1.

Cette condition est néanmoins rarement vérifiée pour des signaux réels car elle nécessite des signaux très parcimonieux et/ou un dictionnaire possédant des atomes très peu corrélés.

Les travaux de Tropp et al. [217, 220] ont introduit une autre condition assurant la résolution du problème ℓ_0 par les algorithmes de décomposition. Celle-ci est liée au coefficient de récupération exacte (« *Exact Recovery Coefficient* » (ERC)) défini comme suit :

Définition 6. *Soit* Δ *un ensemble d'indices des atomes du dictionnaire* Φ,

$$ERC(\Phi, \Delta) = 1 - \max_{w \neq \Delta} \|\Phi_\Delta^\dagger \Phi(w)\| \ .$$

Le théorème associé à cette condition s'écrit alors :

Théorème 3. *Soit* $\mathbf{x} \in \mathbb{R}^C$ *un signal dont la représentation la plus parcimonieuse sur le dictionnaire* Φ *a pour support* Δ_{opt}, *une condition suffisante pour que l'Orthogonal Matching Pursuit (OMP définit section 3.3) et les méthodes résolvant le problème* ℓ_1 *retrouve ce support est :*

$$ERC(\Phi, \Delta_{opt}) \geq 0 \ .$$

Cette condition est notamment vérifiée pour tous les signaux possédant une décomposition dont la taille du support vaut m lorsque [220] :

$$\mu_1(m-1) + \mu_1(m) \leq 1 \ .$$

Lorsque les valeurs de cohérence cumulée sont calculables en temps raisonnable, pour des dictionnaires paramétriques par exemple, ce résultat permet de caractériser efficacement les décompositions sur ce dictionnaire.
D'autres résultats importants s'expriment en fonction des constantes d'isometrie [38]. Dans un cas non bruité :

- si $\delta_{2s} \leq 1$, le problème ℓ_0 possède une solution de degré de parcimonie s,
- si $\delta_{2s} \leq \sqrt{2} - 1$, la solution du problème ℓ_1 est la même que celle du probléme ℓ_0 (de parcimonie s).

Dans le cas général bruité, dont le problème s'écrit :

$$\arg\min_{x \in \mathbb{R}^{N_\Phi}} \|\mathbf{x}\|_* \quad \text{t.q.} \quad \|\mathbf{y} - \Phi\mathbf{x}\|_2^2 \leq \varepsilon,$$

elles permettent également la démonstration du théorème suivant.

Théorème 4. *Lorsque* $\delta_{2s} \leq \sqrt{2} - 1$ *alors la solution du problème* ℓ_1 *notée* $\hat{\mathbf{x}}$ *respecte :*

$$\|\hat{\mathbf{x}} - \bar{\mathbf{x}}\|_2^2 \leq C_0 s^{-\frac{1}{2}} \|\bar{\mathbf{x}} - \bar{\mathbf{x}}_s\|_1 + C_1 \varepsilon \tag{3.7}$$

avec C_0 *et* C_1 *des constantes,* $\mathbf{y} = \Phi\bar{\mathbf{x}}$ *et* $\bar{\mathbf{x}}_s$ *un vecteur composé des* s *éléments les plus grands de* $\bar{\mathbf{x}}$.

Les constantes d'isométrie δ_s de dictionnaires classiques ayant été dérivées dans divers travaux ce théorème est particulièrement intéressant pour des décompositions sur ceux-ci.

3.2.3 Régularisations et décompositions structurées

Les considérations précédentes assurent l'efficacité des décompositions parcimonieuses sous certaines conditions concernant les signaux et le dictionnaire. En pratique, elles peuvent se révéler efficaces dans un cadre plus large mais ne permettent pas l'obtention de représentations satisfaisantes lorsque le bruit est trop fort ou le dictionnaire choisi composé d'atomes très corrélés. Dans ces situations, il est possible d'utiliser certaines connaissances *a priori* sur les signaux afin d'obtenir de meilleures décompositions en contraignant celle-ci grâce à des termes de régularisations prenant en compte ces connaissances. De manière générale le problème de décomposition s'écrit alors :

$$\underset{X \in \mathbb{R}^{N_\Phi \times T}}{\arg\min} \|Y - \Phi X\|_F^2 \quad \text{t.q} \quad R(X) < J$$

Dans un contexte de représentation redondante, les régularisations induisant une parcimonie ont été particulièrement étudiées. De façon générale, elles cherchent à imposer une structure de zéros dans les coefficients. Nous les présentons ici dans un formalisme ℓ_1.

3.2.3.1 Décomposition simultanée

La régularisation la plus naturelle pour une décomposition de signaux multicanaux (d'un ensemble de signaux monodimensionnels en général) est celle imposant une structure parcimonieuse commune aux différents canaux. Elle est formalisée comme suit :

$$R(X) = \|X\|_{2,1}$$

où la norme $\ell_{2,1}$ est défini par $\|X\|_{2,1} = \sum_i \|X(i,.)\|_2$.
Divers travaux ont étudié les propriétés de cette régularisation [41, 71, 101, 222, 221]. Ils montrent que l'utilisation des informations venant de plusieurs signaux possédant une structure parcimonieuse commune permet l'obtention d'une meilleure décomposition de ceux-ci. Les conditions sur la récupération de la structure parcimonieuse des signaux présentées plus haut en monodimensionnel peuvent en général être étendues de manière assez simple. Le theorème 2 et la condition obtenue grâce à l'ERC sont notamment étendu à ce contexte dans [41].

3.2.3.2 Décomposition structurée

La décomposition précédente peut être vue comme un cas particulier d'une régularisation étudiée dans un cadre monodimensionnel sous le nom de « Group-LASSO » [239]. Cette dernière permet la mise en place de structures parcimonieuses complexes en forçant la solution d'une décomposition à être parcimonieuse par groupe, *i.e.* à mettre à zéros les variables par groupe. Dans un cadre multidimensionnel ces groupes peuvent être choisis soit pour forcer des ensembles de canaux à se décomposer d'une manière identique soit dans le but de former des ensembles d'atomes du dictionnaire qui seront utilisés ensembles.
Formellement, soit $g^i, i \in \{1, \ldots, N_g\}$ des groupes d'indices de canaux (respectivement,

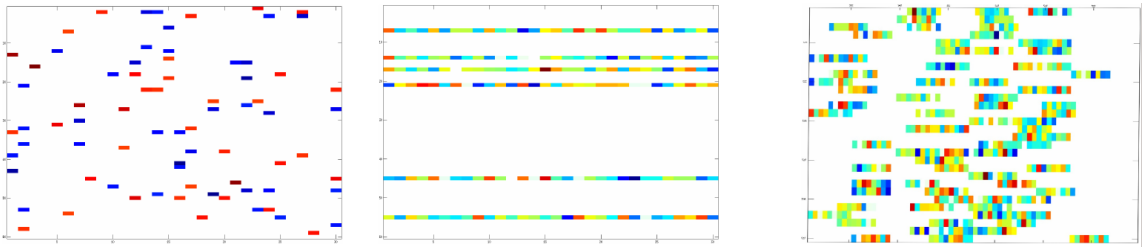


FIGURE 3.2 – Gauche : décomposition parcimonieuse simple. Centre : décomposition simultanée. Droite : décomposition structurée avec groupes de canaux

d'atomes du dictionnaire), elle s'écrit :

$$R(X) = \sum_{j=1}^{N_\Phi} \sum_{i=1}^{N_g} \|X(j, g^i)\|_2 \quad \text{resp} \quad R(X) = \sum_{j=1}^{T} \sum_{i=1}^{N_g} \|X(g^i, j)\|_2$$

Les deux possibilités peuvent être utilisées simultanément et des groupes 2D créés. Pour une étude approfondie des propriétés de ce type de régularisations voir [116, 70].
La figure 3.2 illustre la différence entre les matrices de coefficient obtenues pour une décomposition simple, simultanée et structurée par groupe.

3.2.3.3 Régularisations en analyse

Les régularisations en analyse ont également attiré l'attention de la communauté ces dernières années [223]. Pour une décomposition de signaux multicanaux, elles peuvent s'exprimer de deux façons différentes :

$$R_P^{Analyse1}(X) = \|XP\|_1 \quad \text{ou} \quad R_P^{Analyse2}(X) = \|PX\|_1 \ .$$

Dans la première formulation, la régularisation impose aux lignes de la matrice de décomposition X une projection parcimonieuse sur les filtres composants P (colonnes de P), permettant d'assurer une structure particulière entre les vecteurs de décomposition des canaux du signal Y. Dans la seconde, les filtres sont les lignes de P et impose une structure dans les décompositions de chaque canal pour les coefficients des différents atomes.
Dans un cadre monodimensionnel, ce type de régularisation a notamment été étudié pour des problèmes où un signal $\mathbf{x}$ est généré par une combinaison linéaire d'éléments d'un dictionnaire $\mathbf{y} = D\mathbf{x}$ mais n'est accessible qu'à travers un ensemble de mesures liées linéairement à lui par $\mathbf{z} = \Phi\mathbf{y} + \mathbf{n}$ (avec Φ une matrice de mélange pouvant être un dictionnaire et $\mathbf{n}$ un vecteur de bruit). Dans ce dernier cas, la résolution du problème :

$$\arg\min_{\mathbf{v}} \|\mathbf{z} - \Phi\mathbf{v}\|_2^2 + \lambda\|D^\dagger\mathbf{v}\|_1$$

permet une meilleure approximation de $\mathbf{y}$ qu'une simple décomposition parcimonieuse (pour un paramètre de régularisation λ adapté). Dans un cas monodimensionnel, un théorème équivalent à celui présenté dans l'eq. (3.7) a été démontré dans [39] et borne la distance

entre la solution du problème ci-dessus et le vrai signal. Ce théorème se fonde sur l'introduction de variantes des constantes d'isométrie (RIP) dépendantes du dictionnaire D(D-RIP). Dans d'autres contextes, des régularisations du même type ont été étudiées avec par exemple des contraintes de Variation Totale (VT) [198, 44] ou encore de parcimonie sur des dictionnaires d'ondelettes [190].

Autres régularisations Parmi les autres régularisations envisagées pour des décompositions sur dictionnaire, nous pouvons citer les régularisations de type Tikhonov et celles imposant des coefficients de décomposition non-négatifs [132].
Comme pour la régularisation en analyse, dans un contexte multidimensionnel une régularisation de Tikhonov peut être appliquée soit sur les canaux :

$$R_P^{Tikhonov1}(X) = \|XP\|_F^2,$$

avec une matrice P encodant le lien supposé entre les canaux, soit sur les coefficients associés aux différents atomes :

$$R_P^{Tikhonov2}(X) = \|PX\|_F^2,$$

pour une matrice P encodant une information sur les coefficients associés aux différents atomes. C'est notamment une régularisation populaire pour imposer des structures de coefficients lisses dans la matrice de décomposition [242, 88].

3.3 Algorithmes de décomposition

Nous allons maintenant nous intéresser aux algorithmes permettant la réalisation de décompositions parcimonieuses. Ces algorithmes se divisent en deux catégories : des approches gloutonnes essayant d'approcher la solution du problème strict ℓ_0 (Eq. 3.2) et des méthodes d'optimisation numérique convexe permettant la résolution du problème relaché ℓ_1 (Eq. 3.3).

3.3.1 Méthodes gloutonnes

Ces approches tentent d'approcher la solution du problème ℓ_0 en sélectionnant les atomes itérativement. Ce sont des méthodes rapides mais ne garantissent pas l'obtention du minimum global du problème. Les algorithmes les plus courants sont des variantes du Matching Pursuit (MP) décrit dans [153], auparavant connu sous le nom de Projection Pursuit [79].
Le principe du MP est de sélectionner à chaque itération l'atome permettant une décroissance maximale de l'erreur de reconstruction courante et de soustraire sa contribution au résidu courant, la valeur de ce résidu ayant été initialisé avec le signal. De manière formelle, pour un signal monodimensionnel $\mathbf{y}$, soit le vecteur des coefficients de décomposition $\mathbf{x}^j$ et le résidu $\mathbf{r}^j$ à l'itération j, les étapes du MP peuvent être décrites comme suit :

Sélection

$$index^{j+1} = \arg\min_i \|\mathbf{r}^j - \Phi(i)\langle \mathbf{y}, \Phi(i)\rangle\|_F^2$$
$$\Leftrightarrow \quad index^{j+1} = \arg\max_i \quad |\langle \mathbf{y}, \Phi(i)\rangle|$$

Mise à jour

$$\mathbf{r}^{j+1} = \mathbf{r}^j - \Phi(index^{j+1})\langle \mathbf{y}, \Phi(index^{j+1})\rangle,$$
$$\mathbf{x}^{j+1} = \mathbf{x}^j, \quad \mathbf{x}^{j+1}(index^{j+1}) = \langle \mathbf{y}, \Phi(index^{j+1})\rangle$$

avec $\mathbf{r}^0 = \mathbf{y}$. Cette approche est sous optimale étant donné que la projection du signal sur l'atome à l'iteration j ne prend pas en compte les projections précédentes. Ainsi, une variante nommée Orthogonal Matching Pursuit (OMP) a été conçue par Pati et al. [181]. Celle-ci réalise l'étape de mise à jour grâce à une projection orthogonale du signal sur l'ensemble des atomes choisis aux cours des itérations. Soit Δ^j les indices des atomes sélectionnés à l'itération j, l'étape de mise à jour est réalisée comme suit :

$$\mathbf{x}^{j+1} = \arg\min_{\mathbf{x}} \|\mathbf{y} - \Phi(\Delta^j)\mathbf{x}\|_2^2$$
$$\mathbf{r}^{j+1} = \mathbf{r}^j - \Phi(\Delta^j)\mathbf{x}^{j+1}$$

Les itérations sont stoppées (MP et OMP) lorsque le nombre d'atomes sélectionnés atteint J ou bien lorsque la norme relative du résidu par rapport à celle du signal $\frac{\|\mathbf{r}^j\|_2^2}{\|\mathbf{y}^j\|_2^2}$ est suffisamment faible.
De nombreuses variantes de ces méthodes ont été développées, parmi celles-ci on trouve l'OLS [42] (Orthogonal Least Square) sélectionnant les atomes de manière à faire décroitre au maximum la norme du résidu après mise à jour ainsi que des approches sélectionnant plusieurs atomes à chaque itération tel que le Stagewise OMP [58] (StOMP), le Regularized OMP [169] (ROMP) ou le Two Stage Thresholding [151].
Dans un contexte multicanal, le Simultaneous OMP (SOMP) [222] permet d'approcher la solution du problème obtenue pour une régularisation $\ell_{2,0}$ abordée un peu plus haut. Son principe est le même que celui de l'OMP en dehors du fait que les atomes sont sélectionnés pour tous les canaux simultanément. De même, le Block OMP(BOMP) [70] a été développé pour des problèmes de type « Group-LASSO ».
À noter que l'OMP a également été adapté à une formulation en analyse du problème avec le GAP [167].

3.3.2 Optimisation convexe

Le problème relaché de décomposition parcimonieuse est convexe et non différentiable. Dans le cas d'un dictionnaire de rang plein, il est même strictement convexe et présente l'avantage de n'avoir qu'une unique solution. Sa résolution necessite néanmoins un temps supérieur à l'approximation de la solution du problème ℓ_0 par des méthodes gloutonnes (en général).
Introduit dans [43], le problème de décomposition ℓ_1 fut résolu via sa reformulation en un

problème de programmation linéaire (LP) équivalent à l'aide d'une méthode *primal-dual*. D'autres reformulations sous formes de problèmes quadratiques ou coniques du second ordre ont par la suite été également étudiées [37]. Ces approches s'étant révélées stables mais lentes d'autres approches ont été développées.
Parmi celles-ci, les algorithmes proximaux sont particulièrement efficaces [47] et permettent la résolution de problèmes de grandes dimensions dans des temps raisonnables, notamment grâce à l'algorithme FISTA [171] inspiré des travaux de Nesterov [18] et ayant une convergence quadratique. Il s'agit d'algorithmes itératifs comprenant deux étapes principales (par itération) : un pas de descente de gradient sur la partie différentiable de la fonction à minimiser et une projection du résultat sur la contrainte induite par la régularisation parcimonieuse réalisée grâce à un opérateur proximal. Cette méthode fait suite aux (Tw)IST [52, 25] et spliting operator [48] ainsi qu'à leurs nombreuses variantes. Nous décrirons en détail FISTA dans le chapitre 5.
L'utilisation d'algorithmes *primal-dual* de type lagrangien augmenté (ADMM et split Bregman) est une autre approche importante pour la résolution de problèmes ℓ_1. Dernièrement, elles se sont montrées particulièrement efficaces pour la résolution de ce type de problèmes [90] grâce à une identification rapide des zéros d'une décomposition. Elles seront également détaillées dans le chapitre 6.
Les approches homotopiques sont aussi très utilisées pour la résolution de ces problèmes. Leur principe réside dans le calcul d'un chemin de régularisation, *i.e.* les solutions pour différentes valeurs du paramètre de régularisation. À partir d'une valeur pour laquelle la solution est simple à calculer, le chemin est créé en calculant les solutions de proche en proche grâce aux propriétés du problème considéré. Le LARS [66] résout les problèmes ℓ_1 sur ce principe en calculant les conditions pour lesquelles certaines variables deviennent non nulles. Elle fait partie des approches dites d'« active set » résolvant des sous-problèmes successivement en incorporant les différentes variables petit à petit. De manière générale, ce type d'approches (ou plus généralement : les approches par « working set ») est utilisé pour les problèmes ℓ_1 à cause de l'avantage que leur procure la parcimonie des solutions [133].
Enfin, des algorithmes de descente par blocs ont aussi été développés pour ces problèmes [81], notamment pour des régularisations induisant des blocs de non-zéros ainsi que des approches dites de ℓ_2 repondérées [8] utilisant des approximations différentiables du problème. Le lecteur intéressé pourra se référer à [11] pour plus de détails sur la plupart de ces algorithmes (et une comparaison) et [90] pour les approches de type split Bregman.

3.4 Apprentissage de dictionnaire

Le choix du dictionnaire dans la mise en place d'une représentation redondante est primordial étant donné que les structures parcimonieuses obtenues dans la décomposition de signaux sur un dictionnaire sont spécifiques à celui-ci. Dans de nombreux cas, des dictionnaires conçus de manière paramétrique sont efficaces pour la représentation des signaux considérés [229, 194] : dictionnaire de Gabor, de Harr, de différentes ondelettes, concaténations de bases orthogonales, etc. Néanmoins, pour certaines classes de signaux, une meilleure représentation peut être obtenue grâce à l'apprentissage d'un dictionnaire contenant leurs

principales composantes [139].

3.4.1 Formalisation du problème

Nous présentons ici l'apprentissage du dictionnaire Φ à partir d'une base de signaux $\mathbf{y}_k, k \in \{1, \ldots, K\}$ monodimensionnels dont les décompositions sont notées $\mathbf{x}_k, \in \{1, \ldots, K\}$. L'adaptation des approches présentées à des signaux multidimensionnels sera présentée séparément par la suite.

Les premiers travaux sur l'apprentissage de dictionnaires l'ont formalisé de manière probabiliste [176, 139, 126] comme maximisation d'une fonction de vraisemblance :

$$\hat{\Phi} = \arg\max_{\Phi} \quad \log(\prod_{k=1}^{K} P(\mathbf{y}_k, \Phi)) = \arg\max_{\Phi} \quad \sum_{k=1}^{K} \log(\int P(\mathbf{y}_k|\mathbf{x}_k, \Phi) P(\mathbf{x}_k) d\mathbf{x}_k)$$

Le terme $P(\mathbf{y}_k|\mathbf{x}_k, \Phi)$ est un terme d'attache aux données (proximité entre $\mathbf{y}_k$ et $\Phi\mathbf{x}_k$), tandis que $P(x_k)$ dépend de la distribution des coefficients de décomposition. En prenant comme hypothèse une distribution laplacienne ou de Cauchy des coefficients, l'intégrale peut être remplacée par son maximum. En supposant de plus une distribution gaussienne centrée du bruit sur le terme d'attache aux données, le problème est équivalent aux minimisations suivantes :

$$\hat{\Phi} = \arg\min_{\Phi, \mathbf{x}_1, \ldots, \mathbf{x}_K} \sum_{k=1}^{K} \|\mathbf{y}_k - \Phi\mathbf{x}_k\|_2^2 \text{ t.q } \|\mathbf{x}_k\|_0 < J \quad \text{pour} \quad P(\mathbf{x}_k) \sim Cauchy \tag{3.8}$$

$$\hat{\Phi} = \arg\min_{\Phi, \mathbf{x}_1, \ldots, \mathbf{x}_K} \sum_{k=1}^{K} \|\mathbf{y}_k - \Phi\mathbf{x}_k\|_2^2 + \lambda\|\mathbf{x}_k\|_1 \quad \text{pour} \quad P(\mathbf{x}_k) \sim Laplace$$

Les deux mêmes termes d'attache aux données et de distribution des coefficients apparaissent ici. Afin d'éviter des problèmes d'échelles, la norme des atomes du dictionnaire est bornée sans perte de généralité de l'approche : $\|\Phi(i)\|_2 \leq 1, i \in \{1, \ldots, N_\Phi\}$. Cette formulation est la plus utilisée actuellement.
Ce problème d'optimisation est non convexe et l'obtention du minimum globale reste difficile. Toutefois, différents algorithmes ont été développés pour s'en approcher. Ils seront décrits dans la section suivante.

Lien avec le problème du packing Conceptuellement, l'apprentissage d'un dictionnaire est équivalent à l'extraction des directions principales des signaux considérés. Nous avons déja vu que la distance entre les atomes (au sens de la corrélation) du dictionnaire doit être maximale pour un bon conditionnement des décompositions et l'obtention d'une représentation stable (section 3.2.2.1). Ainsi, lorsque les éléments du dictionnaire sont de normes unités, cet apprentissage peut être vu comme un problème de pavage de la boule unité de l'espace de description des données. Ce pavage ne doit pas être réalisé sur l'ensemble de la boule mais sur les directions permettant la description des données [217]. Une illustration 2D du pavage d'un cercle par les atomes d'un dictionnaire est présentée dans la figure 3.3. Par conséquent, plusieurs méthodes d'apprentissage de dictionnaires ont été

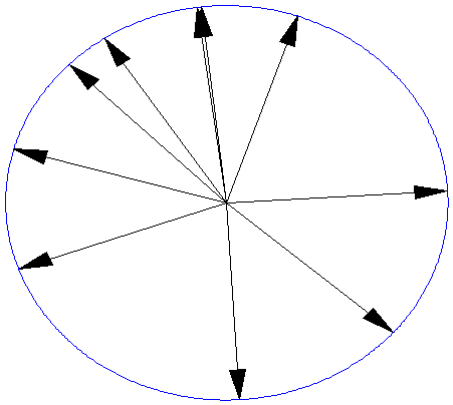


FIGURE 3.3 – Pavage (non-optimal) d'un cercle par les atomes d'un dictionnaire 2D.

développées avec pour optique une représentation précise des données et des corrélations entre atomes minimales, voir par exemple l'INK-SVD [147].

3.4.2 Algorithmes

Le problème (eq. (3.8)) qui nous intéresse ici est non convexe lorsque la minimisation concerne à la fois le dictionnaire Φ et les coefficients de décomposition $\{x_1, \ldots, x_K\}$. Néanmoins, les sous-problèmes minimisant le coût par rapport aux coefficients ou au dictionnaire le sont. Ainsi, les algorithmes d'apprentissage de dictionnaires résolvent généralement ces sous-problèmes de façon alternés afin de s'approcher de la solution du problème complet. Ils se divisent en deux catégories suivant qu'ils effectuent le traitement des données par blocs (« batch ») ou au fur et à mesure (« online »).
Dans la description des algorithmes présentés ci-dessous, Y représente la matrice regroupant les K signaux monodimensionnels et X la matrice regroupant les vecteurs de décompositions. Ces algorithmes se différencient principalement par la manière dont est réalisée la mise à jour du dictionnaire, la méthode de décomposition pouvant être choisie en fonction de l'application.

Approches « batch » À chaque itération, les approches « batch » décomposent l'ensemble des signaux sur le dictionnaire puis mettent à jour le dictionnaire en fonction des décompositions obtenues par résolution du problème suivant :

$$\Phi^{i+1} = \arg\min_{\Phi} \|Y - \Phi X^{i+1}\|_F^2 \quad \text{t.q } \|\Phi(j)\|_2^2 \leq 1, \forall j \in \{1, \ldots, N_\Phi\}$$

Les algorithmes sont stoppés lorsque les erreurs de reconstruction sont suffisamment faibles. Nous présentons ici quelques-unes des principales mises à jour :

MV (Maximum de vraisemblance) [176, 139]	Mise à jour du dictionnaire par descente de gradient. $\Phi^{i+1} = \Phi^i + \eta(Y - \Phi X)X^T$ avec $\eta > 0$ le pas du gradient.

MOD (Method of Optimal Directions) [72] Mise à jour par résolution directe du problème : $\Phi = YX^T(XX^T)^{-1}$.

MAP (Maximum à posteriori) [126] Mise à jour bayésienne.
$\delta\Phi^i = \frac{1}{K}\Phi^i(X - Y)X^T$
$\Phi^{i+1} = \Phi^i - \eta(\delta\Phi^i - tr(\Phi^i\ \delta\Phi^i)\Phi^i)$, $\eta > 0$.

K-SVD [5] Mise à jour successive par SVD des atomes et des coefficients simultanément. Pour chaque atome : calcul de la matrice de rang 1 la plus proche de la matrice représentant la contribution de cet atome aux signaux.

Méthodes « online » Les approches « online » quant à elles mettent à jour le dictionnaire après la décomposition de chaque signal. Elles peuvent donc être choisies lorsque les signaux arrivent petit à petit et permettent un redémarrage à « chaud » [1]. Elles présentent également l'avantage d'être moins sensibles aux minima locaux dans le cas où tous les signaux sont disponibles. Dans ce dernier cas les signaux sont choisis de manière aléatoire et traités successivement. La distance entre le dictionnaire obtenu pour le minimum global du problème et celui obtenu pour le minimum local atteint par des algorithmes online peut par ailleurs être borné sous certaines conditions [117].

La mémoire des décompositions déja effectuées est importante pour ces algorithmes. Certains algorithmes utilisent uniquement la dernière décomposition effectuée pour la mise à jour du dictionnaire [17, 4]. Ces algorithmes doivent alors posséder un paramètre diminuant au cours des itérations permettant de donner un poids de plus en plus important aux dictionnaires de l'itération précédente. Pour [17] et [4], la mise à jour est réalisée par descente de gradient. À l'itération j, après la décomposition $\mathbf{x}_k$ du signal $\mathbf{y}_k$ sur les atomes Δ_k, elle s'écrit :

$$\Phi(\Delta_k)^{j+1} = \Phi(\Delta_k)^j - \eta^j(\mathbf{y}_k - \Phi(\Delta_k)^j\mathbf{x}_k)\mathbf{x}_k^T$$

C'est alors le pas η de cette descente qui est diminué au cours des itérations afin de gérer cette mémoire et d'assurer la convergence.
D'autres méthodes [149, 203] construisent des structures de données enregistrant les décompositions antérieures afin de les prendre en compte dans l'étape de mise à jour. Une stratégie est en général proposée afin d'oublier dans ces structures les décompositions trop anciennes. Ces dernières approches se sont montrées particulièrement efficaces et rapides.
Pour l'algorithme décrit dans [149] auquel nous nous réfèrerons avec l'abréviation ODL (« Online Dictionary Learning »), la mise à jour s'écrit :

$$\begin{aligned} A^{j+1} &= A^j + \mathbf{x}_k\mathbf{x}_k^T \\ B^{j+1} &= B^j + \mathbf{y}_k\mathbf{x}_k^T \\ D^{j+1} &= \arg\min_D \frac{1}{j+1}(Tr(D^TDA^{j+1}) - Tr(D^TB^{j+1})) \end{aligned}$$

1. Démarrage de l'algorithme utilisant les informations obtenues dans les exécutions précédentes.

Les structures de mémoire sont alors les matrices A et B et il est possible de supprimer de ces matrices les contributions des décompositions les plus anciennes aisément.

Adaptations aux signaux multidimensionnels : Pour les algorithmes « batch », l'utilisation de signaux multidimensionnels revient à les décomposer de manière séparée puis à concaténer les décompositions pour la mise à jour du dictionnaire. Pour les approches « online », les décompositions sont naturellement séparées et la mise à jour nécessite, lorsque des structures de mémoire sont présentes, de pouvoir les mettre à jour pour l'ensemble des canaux décomposés (mode mini-batch ODL [149]), dans le cas sans mémoire la loi de progression du paramètre d'importance du dictionnaire précédent doit être ajustée.

Régularisations L'apprentissage des atomes du dictionnaire peut être régularisé comme la décomposition. Il est possible par exemple de forcer une parcimonie sur les atomes [197] ou bien une contrainte TV [148].

Analyse L'apprentissage de dictionnaire pour une formulation en analyse a dernièrement été l'objet de plusieurs travaux. L'algorithme du K-SVD a notamment été adapté pour cette formulation [196].

3.5 État de l'art des modèles redondants pour les EEG

3.5.1 Quelques applications des modèles redondants

Les modèles redondants se sont montrés efficaces pour de nombreuses tâches.

Débruitage Ces représentations permettent un débruitage efficace de certains types de signaux lorsque le dictionnaire contient les composantes importantes de ceux-ci. Elles ont notamment été considérées pour le débruitage d'images [68] de signaux audio [187] ou de signaux d'électrocardiographie [73].

Séparation de sources Elles sont également efficaces pour la séparation de sources et la résolution de problèmes inverses de par la mise en place naturelle de problème mal-posé dans un cadre redondant [30, 52].

Compression et décodage Dans le domaine de la transmission de signal ces représentations sont également intéressantes, aussi bien pour la compression des signaux [12] que pour leur décodage [40].

Classification Des tâches de classifications ont également considéré ce type de représentation, par exemple pour la reconnaissance de visages [233].

3.5.2 Modèles redondants pour les signaux EEG

Des modèles redondants ont également été envisagés et mis en place pour le traitement des signaux EEG. Nous allons les passer en revue dans cette section en les catégorisant en fonction des applications considérées.

3.5.2.1 Modèles de décomposition

Avant de décrire ces différents travaux, il est nécessaire d'introduire les différents modèles linéaires de décomposition qui ont été étudiés.

Modèle multicanal Nous avons déja introduit le modèle multicanal dans ce chapitre, deux décompositions peuvent être envisagées pour celui-ci, une pour chaque dimension des signaux. Dans le cas des signaux EEG, les décompositions peuvent être réalisées d'un point de vue spatial ou bien temporel. Ainsi, soient $\Phi_t \in \mathbb{R}^{T \times N_{\Phi_t}}$ un dictionnaire d'atomes temporels et $\Phi_s \in \mathbb{R}^{C \times N_{\Phi_s}}$ un dictionnaire d'atomes spatiaux, les deux modèles peuvent s'écrire pour le signal Y comme suit :

$$\begin{aligned} \text{Spatial :} \quad & Y = \Phi_s X_s + N_s \\ \text{Temporel :} \quad & Y^T = \Phi_t X_t + N_t \end{aligned}$$

avec $X_t \in \mathbb{R}^{N_{\Phi_t} \times C}$, $X_s \in \mathbb{R}^{N_{\Phi_s} \times T}$ les matrices de décomposition et $N_s \in \mathbb{R}^{C \times T}$, $N_t \in \mathbb{R}^{T \times C}$ les matrices de bruit.
La figure 3.4 présente le schéma d'un modèle multicanal avec dictionnaire temporel.

Modèle multivarié Le modèle multivarié considère un dictionnaire d'atomes spatio-temporels. Soit $\Phi \in \mathbb{R}^{C \times T \times N_\Phi}$ un tenseur d'ordre trois regroupant une famille $\Phi^i \in \mathbb{R}^{C \times T}, i \in \{1, \ldots, N_\Phi\}$ d'atomes spatio-temporels, il s'écrit comme suit pour le signal Y :

$$Y = \sum_{i=1}^{N_\Phi} \Phi^i \mathbf{x}(i) + N$$

ou $\mathbf{x}$ est le vecteur de coefficients, N une matrice de bruit.
Une visualisation de ce modèle est présentée dans la figure 3.5

Modèles à deux dictionnaires Différentes études, dont [192], sur les signaux EEG et notamment sur les potentiels évoqués rapportent que ceux-ci peuvent être correctement approchés par des matrices de rang 1 obtenues par le produit dyadique [2] d'une topographie et d'un signal temporel. Un autre modèle utilisant un dictionnaire spatial et un dictionnaire temporel peut donc être envisagé. Soit $\Phi_s \in \mathbb{R}^{C \times N_{\Phi_s}}$ et $\Phi_t \in \mathbb{R}^{T \times N_{\Phi_t}}$ ces deux dictionnaires, il s'écrit pour le signal Y :

$$Y = \Phi_s X \Phi_t^T + N$$

2. Le produit dyadique A de deux vecteurs $\mathbf{x}$ et $\mathbf{y}$ s'écrit : $A(i,j) = \mathbf{x}(j)\mathbf{y}(i)$ ou $A = \mathbf{x}\mathbf{y}^T$.

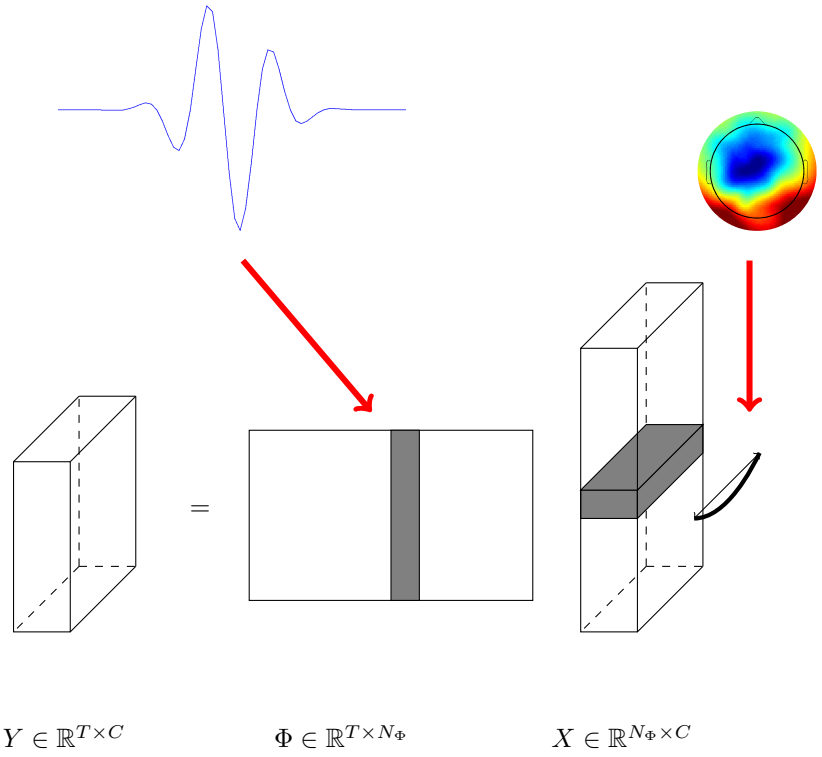


FIGURE 3.4 – Modèle de décomposition multicanal EEG lorsqu'un dictionnaire temporel est considéré. Chaque canal du signal Y est décomposé à l'aide des atomes temporels concaténés dans Φ. Chaque ligne de la matrice des coefficients X est alors une topographie associée à un atome du dictionnaire.

avec $X \in \mathbb{R}^{N_{\Phi_s} \times N_{\Phi_t}}$ la matrice de décomposition et $N \in \mathbb{R}^{C \times T}$ une matrice de bruit.
Pour une décomposition parcimonieuse non-structurée, ce modèle est un cas particulier du précédent pour des atomes spatio-temporel de rang 1 avec toutefois l'avantage de ne pas avoir à construire complètement le dictionnaire pour réaliser la décomposition.
Ce modèle est schématisé dans la figure 3.6.

3.5.2.2 Analyse temps-fréquence

Plusieurs approches de décomposition parcimonieuse sur des dictionnaires de Gabor ont été développées pour l'analyse temps-fréquence de signaux EEG [63]. Ces méthodes considèrent un modèle multicanal et effectuent la décomposition de manière gloutonne à l'aide de l'algorithme du « Matching Pursuit » (MP) ou de variantes de celui-ci. Ces variantes sont caractérisées par les critères utilisés pour le choix des atomes à chaque itération. Nous exprimons ici ces critères pour un dictionnaire $\Phi \in \mathbb{R}^{T \times N_\Phi}$ et un ensemble de K signaux EEG $\{Y^k \in \mathbb{R}^{T \times C}, \forall k \in \{1, \ldots, K\}\}$.
Le premier modèle étudié décompose les canaux indépendamment les uns des autres [62].

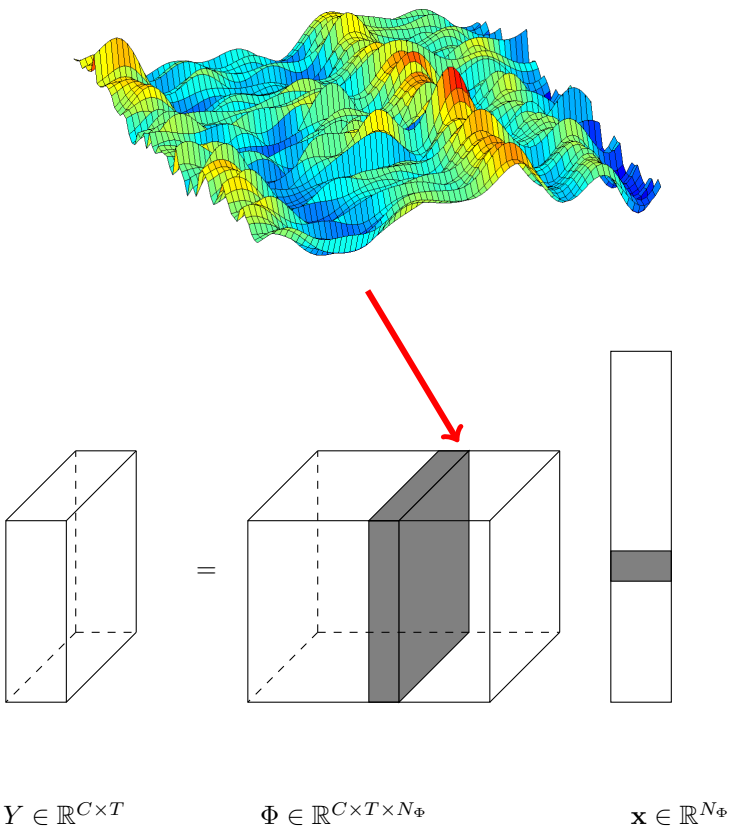


FIGURE 3.5 – Modèle de décomposition multivarié EEG, le dictionnaire Φ est composé d'atomes spatio-temporels, le signal Y est représenté par une combinaison de ces éléments pondérés par les coefficients de $\mathbf{x}$.

Pour le canal c l'indice du meilleur atome à l'itération j est calculé par :

$$\mathbf{index}^{j+1}(c) = \arg\max_i |\langle R^j(c), \Phi(i)\rangle|$$

avec $R^j(c)$ le résidu associé au canal c à l'itération j.
Le « Multichannel MP » (MMP) a ensuite été rapidement considéré avec différents critères. Le MMP originel [100] propose un critère naturel d'extension du modèle monocanal :

$$index^{j+1} = \arg\max_i \sum_{c=1}^{C} |\langle R^j(c), \Phi(i)\rangle|^2,$$

correspondant à une régularisation de type $\|X\|_{2,0}$.
Dans [65] il est proposé de choisir les atomes en se basant sur la moyenne des corrélations des différents canaux :

$$index^{j+1} = \arg\max_i |\sum_{c=1}^{C} \langle R^j(c), \Phi(i)\rangle|$$

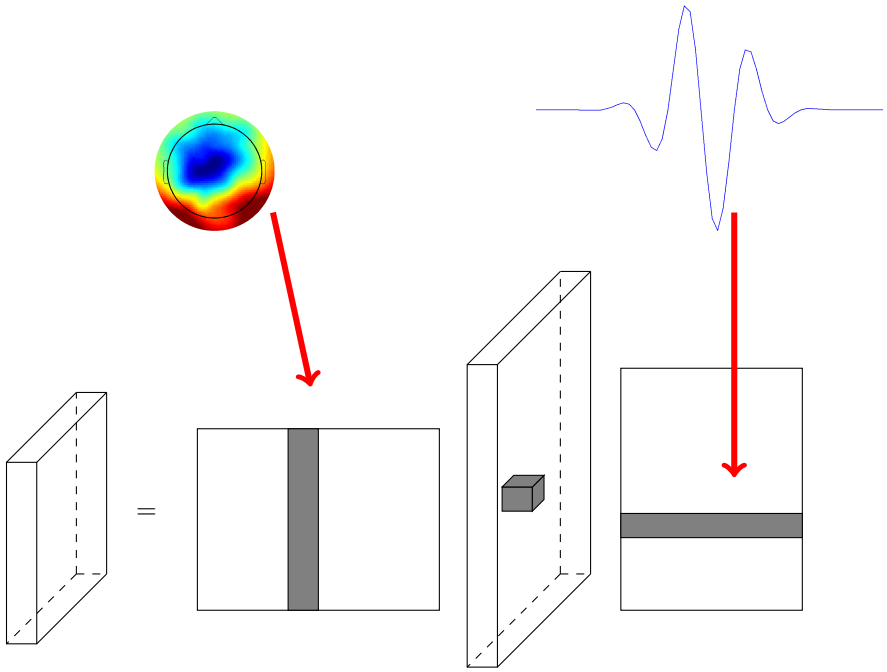


FIGURE 3.6 – Modèle de décomposition EEG utilisant deux dictionnaires : un dictionnaire spatial Φ_s et un dictionnaire temporel Φ_t. Le signal Y est représenté par une combinaison linéaire pondérée d'atomes spatio-temporels de rang 1 obtenus par produit d'atomes de Φ_s et d'atomes de Φ_t. Les coefficients de pondération sont regroupés dans la matrice X.

Contrairement au critère précèdent, les atomes possédant une corrélation positive avec le résidu sont ici favorisés, ce qui en fait un critère plus restrictif. Ce critère se rapproche de la régularisation imposant la positivité des coefficients introduit dans la section 3.2.3 bien que la positivité ne soit pas réellement forcée (les coefficients pouvant être négatifs). Dans le cadre de séries temporelles comme les EEG, ce critère favorise le choix d'atome en phase avec le résidu. Ces deux critères ont de même été utilisés avec des coefficients complexes de décomposition [155, 98, 202], autorisant des phases différentes pour tous les canaux.
Plus récemment, une variante nommée Consensus MP [19] a été introduite pour un modèle monocanal utilisant plusieurs essais. Celle-ci ne fixe pas une structure parcimonieuse commune aux essais. À chaque itération, la décomposition est effectuée en suivant les étapes suivantes :

— sélection pour chaque essai de l'ensemble des atomes les plus corrélés à l'essai,
— choix par vote d'un atome dit de consensus, représentant au mieu tous les essais,
— soustraction pour chaque essai de la contribution de l'atome de l'ensemble calculé à l'étape 1 le plus proche de l'atome de consensus.

En plus des contraintes sur la structure parcimonieuse, des méthodes ont été également conçues de manière à contraindre les topographies des coefficients associés aux différents atomes à respecter certains *a priori*. L'un des premiers travaux considérant une telle approche est décrit dans [124], les décompositions temps-fréquence des signaux multicanaux sont effectuées de manière à obtenir des topographies régulières dans les coefficients. Cette décomposition est néanmoins différente du modèle considéré dans ce chapitre car les coefficients de décomposition sont obtenus par minimisation d'une norme ℓ_2 (unique d'aprés la théorie des repères) et donc non parcimonieux. Plus précisément, pour un signal Y elle résout le problème suivant :

$$\hat{X} = \arg\min_X \|Y - \Phi X\|_F^2 + \lambda_1^2 \|X\|_2^2 + \lambda_2^2 \|XS\|_F^2$$
$$\hat{X} = \arg\min_X \|Y - \Phi X\|_F^2 + \|X(\lambda_1 I + \lambda_2 S)\|_F^2$$

Un algorithme de clustering est alors employé sur les topographies obtenues dans X pour obtenir des micros-états temps-fréquence.
Une étude plus récente s'intéresse à une décomposition similaire, mais cette fois-ci avec une décomposition parcimonieuse réalisée à l'aide d'une variante du MP appelé le Dependency MMP. Le critère de cette dernière s'exprime comme suit :

$$index^{j+1} = \arg\max_i \frac{\sum_{c=1}^{C} |\langle R^j(c), \Phi(i)\rangle|^2}{1 + W_j(X^j, \langle R^j, \Phi(i)\rangle)},$$

avec $W_j(X^j, \mathbf{c}^T) = S(D_j(X^j, \mathbf{c}^T))$ ou S est une fonction de $\mathbb{R}^+$ bornée et D une fonction permettant de mesurer la présence de la propriété souhaitée sur la décomposition, X^j donne accès à la décomposition de l'itération précédente et $\mathbf{c}^T$ contient les coefficients obtenus (vecteur ligne ici) si l'atome i est choisi. Cette étude propose alors une fonction W permettant d'effectuer le choix des atomes de façon à obtenir des topographies régulières. Celle-ci s'écrit :

$$D_j(X^j, \mathbf{c}^T) = p_1 \frac{(\|\mathbf{c}^T P\|_2^2)^{p_3}}{(\|\mathbf{c}^T\|_2^2)^{p_2}} p_4^{\ln(j+1)}$$

avec P la matrice de l'opérateur laplacien pour l'ensemble d'électrodes considéré et p_1, p_2, p_3, p_4, des paramètres scalaires. Comme auparavant, une fois la décomposition effectuée un algorithme de clustering est utilisé afin d'extraire les micro-états des topographies obtenues dans la matrice de décomposition, permettant l'obtention d'une décomposition encore plus parcimonieuse.
Enfin, on peut noter également l'étude décrite dans [97] réalisant une décomposition temps-fréquence dont les topographies sont prises dans un ensemble de modes spatiaux définis par des fonctions de Bessel. Le modèle utilisé est alors un modèle à deux dictionnaires. Un MP adapté au modèle multivarié (utilisant un produit scalaire matriciel) est considéré dans ce cas :

$$index_i^{j+1}, index_k^{j+1} = \arg\max_{i,k} \langle R^j, \Phi_t(i)\Phi_s^T(k)\rangle$$
$$\text{avec : } \langle A, B\rangle = \sum_{k=1}^{i}\sum_{l=1}^{j} A(k,l)B(k,l)$$

Ce dernier produit scalaire est celui induisant la norme de Frobenius.

3.5.2.3 Problème inverse et localisation de sources

Le problème inverse en EEG s'exprime naturellement comme une décomposition multicanal sur un dictionnaire spatial redondant. Il s'écrit pour une signal Y comme suit (voir la section 2.3.3) :

$$Y = GX + N$$

avec $G \in \mathbb{R}^{C \times 3S}$ ($C \ll 3S$) la matrice de gain liant les S sources cérébrales aux C électrodes. Pour chaque source, 3 atomes sont présents dans G, représentant des orientations différentes : x, y et z. Nous supposons dans la suite que ces atomes sont contigus dans la matrice de gain.
Afin de sélectionner une solution à ce problème, différentes contraintes fondées sur des connaissances *a priori* sur les sources ont été proposées. Les premiers travaux ont considéré des contraintes ℓ_2 avec notamment l'approche LORETA [179] proposant la contrainte suivante :

$$R^{LORETA}(X) = \|BWX\|_F^2$$

pour W une matrice de poids permettant de gérer différents biais et B un opérateur laplacien. Ces approches sont généralement regroupées sous l'abréviation MNE pour « Minimum Norm Estimate » et présentent l'avantage d'un calcul rapide de la solution.
Néanmoins, ces méthodes ne permettant l'obtention que de solutions peu précises (« floutées »), des régularisations parcimonieuses ont été envisagées sous l'hypothèse d'un nombre faible de sources actives simultanément. Cette famille d'approches regroupées sous l'abréviation MCE pour « Minimum Current Estimate » comprend notamment l'utilisation de régularisations ℓ_1 [154] et ℓ_p ($0 \leq p \leq 1$) [92] permettant l'obtention de solutions parcimonieuses plus précises. La résolution du problème est alors réalisée en le réécrivant sous forme de problème de programmation linéaire ou en utilisant l'algorithme du simplex pour les problèmes ℓ_1 et grâce à l'algorithme FOCUSS pour les problèmes ℓ_p.
Le principal inconvénient de ces dernières approches provient du fait qu'elles ne prennent pas en compte les orientations des sources ainsi que l'aspect temporel du signal, effectuant des décompositions indépendantes pour chaque instant. De nombreux travaux récents ont donc proposé des modèles pour remédier à cet inconvénient. Concernant l'orientation, les études décrites dans [107, 54, 178] proposent l'utilisation d'une norme $\ell_{1,2}$ qui peut s'écrire comme une norme $\ell_{1,2}$ groupée ici :

$$R(X) = \sum_{j=1}^{T} \sum_{i=1}^{S} \|X(g^i, j)\|_2$$

avec $g^i, i \in \{1, \ldots, S\}$ les groupes correspondants aux atomes représentants la même source pour les trois orientations. Avec cette régularisation, les triplets d'atomes des sources sont sélectionnés simultanément, permettant une gestion efficace des orientations. Le modèle

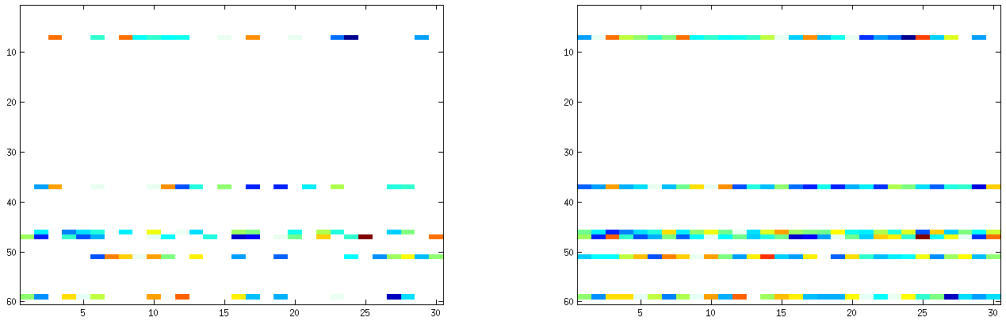


FIGURE 3.7 – Comparaison de décompositions contraintes par une régularisation $\ell_1 + \ell_{2,1}$ à gauche et une régularisation $\ell_{2,1}$ à droite.

proposé dans [108] considère également cette régularisation avec une décomposition supplémentaire sur des sphères 3D dans l'espace des sources.
Concernant la dimension temporelle, une régularisation $\ell_{1,2}$ est également proposée dans [178]. Celle-ci est généralisée dans [94] grâce à des normes mixtes pondérées sur deux niveaux ℓ_{pq} :

$$R_{w;pq}(X) = \|X\|_{w;pq} = \left(\sum_{s=1}^{S} (\sum_{t=1}^{T} w_{s,t} |X(s,t)|^p)^{\frac{q}{p}} \right)^{\frac{1}{q}},$$

permettant de rendre actives des sources pour toute la durée du signal et améliorant leur choix grâce à l'information obtenue pour chaque pas de temps. De plus, des normes mixtes sur trois niveaux ℓ_{pqr} sont également définis dans [94] :

$$R_{w;pqr}(X) = \|X\|_{w;pqr} = \left(\sum_{s=1}^{S} \left(\sum_{t=1}^{T} (\sum_{k=1}^{K} w_{s,t,k} |X(s,t,k)|^p)^{\frac{q}{p}} \right)^{\frac{r}{q}} \right)^{\frac{1}{r}},$$

avec X un tenseur d'ordre 3 composé de K essais. Cette dernière permet de prendre en compte K essais représentants un phénomène à l'étude et ainsi pouvoir obtenir un choix des sources plus efficace car basé sur plus d'informations.
Les signaux EEG n'étant toutefois pas stationnaires, rendre une source active durant tout le signal n'est pas une solution idéale. Pour surmonter ce problème, il est possible comme proposé dans [162] de considérer une norme $\ell_1 + \ell_{2,1}$:

$$R_{1+2,1}(X) = \rho \|X\|_1 + (1-\rho) \sum_{s=1}^{S} \|X(s)\|_F$$

permettant la sélection d'une ligne dans la matrice de décomposition mais forçant les coefficients de cette ligne à être parcimonieux. L'effet de cette régularisation est illustré dans la figure 3.7.
Différentes études ont également proposé l'utilisation d'un dictionnaire temps-fréquence Φ pour décrire l'évolution temporelle des sources à travers un modèle à deux dictionnaires :

$$Y = GX\Phi^T + N \ .$$

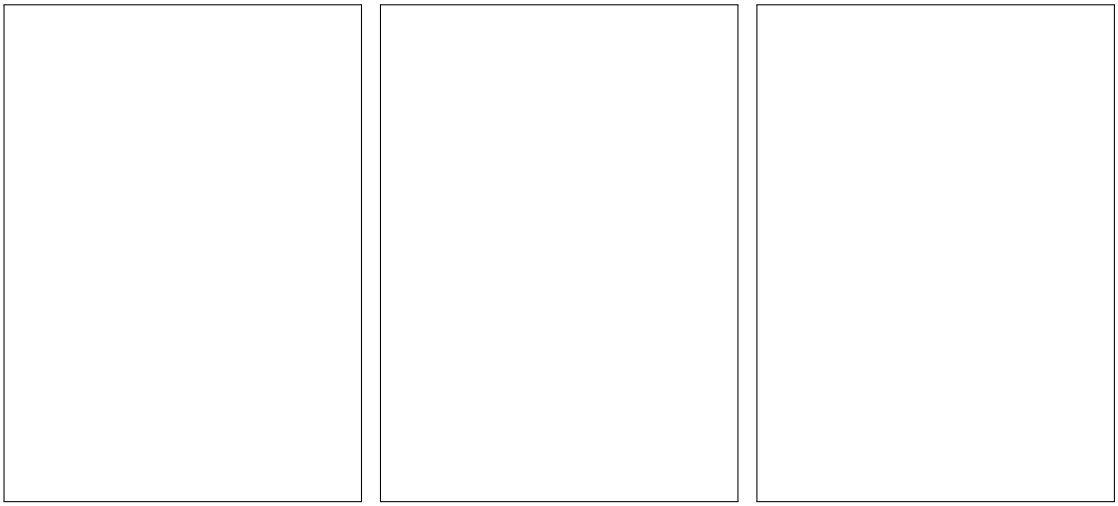
Bolstad et al. [32] proposent avec ce modèle de grouper les atomes dans chacun des dictionnaires par proximité spatiale et temporelle. Les groupes 2D composés des intersections de ces groupes spatiaux et temporels dans la matrice de coefficients sont alors considérés pour une régularisation $\ell_{p,1}$ par groupe. Formellement, soit $\{g_s^i, i \in \{1, \ldots, N_{g_s}\}\}$ les groupes spatiaux, $\{g_t^i, i \in \{1, \ldots, N_{g_t}\}\}$ les groupes temporels, elle s'écrit comme suit :

$$R_{2,1}^{g_s,g_t}(X) = \sum_{i,j} \|vect(X(g_s^i, g_t^j))\|_p$$

$X(g_s^i, g_t^j)$ la sous-matrice de X correspondante au groupe 2D d'intersection de g_s^i et g_t^j et $vect(.)$ l'opérateur de vectorisation. L'hypothèse sous-jacente à cette régularisation est la localisation spatiale et temporelle des activités cérébrales qui ainsi doivent s'exprimer sur des atomes proches spatialement et temporellement.
Une régularisation $\ell_1 + \ell_{2,2}$ (régularisation classique nommée « elastic net » [242]) est proposée dans [209] qui considère une version vectorielle du modèle double. Celle-ci suppose une parcimonie totale sur la matrice de coefficient et une régularité des coefficients. Des hypothèses similaires sont réalisées dans [96] qui propose d'utiliser la régularisation $R_{1+2,1}$ introduite plus tôt, toutefois dans ce dernier cas une sélection parcimonieuse des sources est réalisée en plus de la sélection parcimonieuse sur tous les coefficients.

3.5.2.4 Autres études

Certaines études ont proposé l'apprentissage d'un dictionnaire pour la représentation des signaux EEG. Parmi celles-ci, l'étude décrite dans [120] propose l'apprentissage d'un dictionnaire invariant par translation pour un seul canal grâce à l'algorithme MoTIF [119]. Certains des atomes appris présentent un contenu spectral localisé correspondant à des bandes de fréquence importantes des signaux EEG et peuvent donc être plausibles physiologiquement.
Dans un autre contexte, le K-SVD a été appliqué aux signaux EEG pour les BCI [158] dans le but d'apprendre des composantes spatiales ainsi que des composantes temporelles pour l'identification de potentiels d'erreur et du PE P300. Les signaux sont débruités à l'aide des composantes apprises grâce à un OMP modifié : certains atomes représentant la moyenne des signaux sont intégrés automatiquement dans les atomes choisis. Le débruitage à l'aide des composantes spatiales permet l'amélioration de la classification qui est réalisée avec un FDA.
Pour le débruitage de signaux neuroélectriques intracrâniens, un apprentissage de dictionnaires invariant par translation a également été développé récemment [109]. Celui-ci considère un modèle monocanal avec plusieurs essais. La décomposition parcimonieuse des signaux (indépendamment les uns des autres) est réalisée via l'algorithme LARS modifié et la mise à jour du dictionnaire est effectuée à l'aide d'un algorithme de descente par blocs.
Enfin, un algorithme d'apprentissage de dictionnaires invariant par translation multivarié a été proposé dans [16]. La décomposition des signaux est réalisée via l'algorithme OMP adapté en multivarié où le choix des atomes est effectué à l'aide d'une inter-corrélation et non d'un produit scalaire de manière à identifier les décalages temporels optimaux (comme la décomposition du MoTIF). Une descente de gradient est proposée pour la mise à jour



(a) Moyenne (b) Moindres carrés (c) Dictionnaire multivarié

FIGURE 3.8 – Comparaison d'estimation du P300. À gauche, estimé à l'aide d'une simple moyenne des essais, au centre par résolution du problème des moindres carrés et enfin à droite par apprentissage d'un dictionnaire multivarié. Source : [16]

du dictionnaire. Deux résultats intéressants sont obtenus avec cette approche, d'une part, l'apprentissage d'un dictionnaire temporel permet de décrire la variance des signaux EEG avec beaucoup moins d'atomes qu'un dictionnaire paramètrique temps fréquences, d'autre part, un tel apprentissage peut être efficace pour la caractérisation de potentiels évoqués comme le P300. La figure 3.8 présente l'estimation du P300 avec une telle approche comparée à celles obtenues par moyennage et résolution simple du problème des moindres carrés associés.

Des études pour la compression de signaux EEG ont également été effectuées à l'aide de modèles redondants dans le cadre du « compressed sensing », voir par exemple [10] proposant une décomposition sur un dictionnaire de Gabor avec régularisation $\ell_{2,1}$ ou [2] passant en revue différentes approches pour le choix du dictionnaire et la méthode de résolution du problème de décomposition parcimonieuse.

Enfin, différentes régularisations parcimonieuses ont été proposées dans des modèles non nécessairement redondants construits pour la classification de signaux EEG. Ces régularisations permettent le plus souvent la sélection de variables assurant une classification optimale des signaux. Des termes de régularisation simple ℓ_1 ont été notamment étudiés pour des variantes d'approches classiques comme le « Sparse CSP » [237] et plus récemment une régularisation plus élaborée a été développée pour un classifieur vaste marge [77]. Dans cette dernière étude, la régularisation proposée est conçue de manière à sélectionner à partir de signaux de plusieurs sujets, les électrodes permettant de maximiser la détection d'un PE donné. La sélection de ces canaux est réalisée via une norme mixte $\ell_{1,q}$ avec $1 \geq q \geq 2$.

Chapitre 4

Régularisations pour la décomposition de signaux EEG

Sommaire

Le point central de cette thèse consiste en la mise en place de régularisations pour la décomposition de signaux EEG sur des familles redondantes. Les signaux d'électroencéphalographie ayant été présentés et les modèles redondants décrits, nous pouvons maintenant nous intéresser plus en détail à ces régularisations.
Celles-ci ont pour objectif :

- l'obtention de décompositions moins sensibles au bruit et plus stables même lorsque le dictionnaire est très cohérent,
- l'apprentissage de dictionnaires plausibles physiologiquement pour la représentation de signaux EEG.

De plus, comme nous l'avons déjà évoqué précédemment, nous souhaitons analyser les signaux EEG dans l'espace des électrodes (et non des sources), essai par essai, indépendamment les uns des autres, c'est-à-dire sans utiliser les informations de plusieurs enregistrements simultanément. Ces contraintes réduisent les types de modèles et de régularisations possibles.

Les régularisations appliquées sur les coefficients de décomposition peuvent affecter deux aspects différents de la structure de ces coefficients :

La structure de choix des atomes Ces régularisations dépendent directement de la structure du dictionnaire. Lorsque celui-ci est conçu de manière paramétrique, ce type de régularisations peut par exemple grouper les coefficients correspondants à des atomes possédant des propriétés similaires. Dans le cas

des EEG, il s'agit d'une proximité spatiale, temporelle ou fréquentielle. Ces régularisations sont par contre plus difficiles à mettre en place lorsque les atomes sont appris sauf si les propriétés des atomes sont fortement contraintes durant l'apprentissage.

La décomposition de plusieurs signaux

Ces régularisations contraignent la manière dont la décomposition simultanée de plusieurs signaux (monodimensionnels ou multidimensionnels) est effectuée, l'information obtenue sur plusieurs signaux permettant un choix plus robuste des atomes. Elles sont indépendantes du dictionnaire et prennent en compte le lien existant entre ces signaux. Lorsqu'ils correspondent par exemple aux signaux des différentes électrodes de l'EEG, la relation entre ceux-ci dépend de la configuration spatiale de ces électrodes.

Dans notre cas (essai par essai), les régularisations possibles pour les modèles multivariés et à deux dictionnaires ne peuvent concerner que la structure de choix des atomes alors que pour le modèle multicanal, elles peuvent affecter sur une dimension (les lignes de la matrice de décomposition) le choix des atomes et sur l'autre (les colonnes) la façon dont sont décomposés les différents signaux monodimensionnels. Le schéma présenté dans la figure 4.1 illustre les groupements d'atomes ou de canaux pouvant être obtenus avec ces deux types de régularisations dans la matrice de décomposition pour un modèle multicanal.

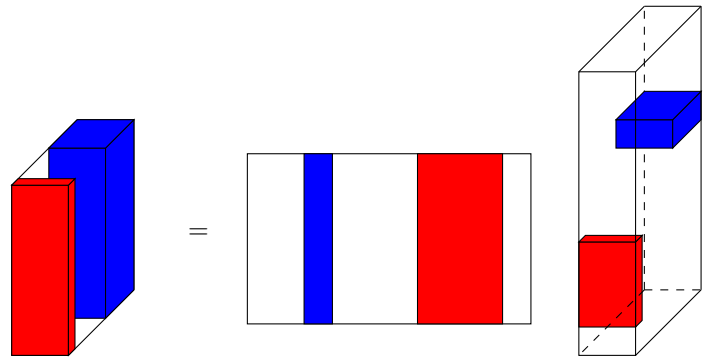


FIGURE 4.1 – Régularisations par groupes des coefficients d'une décomposition multicanale : en rouge, le regroupement de plusieurs atomes pour la décomposition d'un canal et en bleu, le regroupement de plusieurs canaux décomposés sur le même atome.

Les objectifs évoqués plus haut nous poussant à nous intéresser à ce second type de régularisations, c'est donc le modèle multicanal que nous allons étudier et pour lequel nous allons mettre en place ces régularisations. Notre étude se concentre donc sur la décomposition simultanée de signaux monodimensionnels.

Pour des signaux spatio-temporels comme les signaux EEG, nous avons vu que le modèle multicanal peut considérer une décomposition sur un dictionnaire spatial ou sur un dictionnaire temporel. Dans la suite de ce chapitre, nous présenterons des régularisations étudiées pour ces deux contextes. Nous nous intéresserons dans un premier temps à la régularisation spatiale de décompositions temporelles afin de pouvoir par exemple guider des décompositions temps-fréquence vers des solutions plausibles physiologiquement. Nous proposerons ensuite une régularisation temporelle fondée sur le modèle des micro-états pour une décomposition spatiale. Cette dernière a été conçue dans l'optique d'une extension de ce modèle que nous présenterons dans le chapître 7.

Remarque. Les développements mathématiques associés à ces régularisations font apparaitre une équation particulière nommée équation de Sylvester. La résolution de celle-ci dans notre contexte est présentée dans l'annexe A par souci de simplicité de lecture. Elle ne sera donc pas explicitée dans les sections et chapitres suivants.

4.1 Régularisations spatiales

4.1.1 Hypothèses et *a priori* neurophysiologiques

L'activité électrique mesurée lors de la réalisation d'une tâche mentale présente une cohérence spatiale. Celle-ci peut être utilisée pour guider les décompositions sur des dictionnaires temporels. Les topographies particulières associées à ces activités sont souvent difficiles à obtenir, néanmoins certaines des propriétés générales de celles-ci peuvent être prises en compte dans les décompositions à travers des termes de régularisations spatiales. Ici, deux hypothèses particulières sont réalisées à propos des topographies associées à l'activité des sources cérébrales :

- **elles sont lisses**, ne présentant pas de variations spatiales trop brusques.
- **elles sont localisées spatialement**, *i.e.* elles possèdent des *extrema* étalés sur un faible nombre d'électrodes.

Ces hypothèses s'appuient sur les études du problème direct EEG [173], *i.e.* les études cherchant à comprendre la relation entre les variations d'activités d'ensembles de neurones (sources cérébrales) et les variations des mesures obtenues sur les électrodes. Celles-ci s'attachent à la modélisation de la propagation d'ondes électromagnétiques à l'intérieur de la tête, permettant l'obtention pour une source donnée de la topographie mesurée au niveau des électrodes. Elles considèrent un modèle de la tête composé de plusieurs couches ayant chacune des conductivités uniformes et constantes. Ce modèle peut être composé simplement de plusieurs sphères concentriques ou fondé sur une forme de tête réaliste. Dans chacune de ces couches, la propagation des ondes est modélisée à partir des équations de Maxwell sous hypothèse de quasi-stationnarité [1]. Dans une telle modélisation, l'activité électrique des sources cérébrales est liée linéairement aux potentiels mesurés au niveau des électrodes [163, 23] et permet d'approcher la forme des topographies associées à l'activité

1. Hypothèse valable ici étant donné le faible temps de propagation de l'onde comparé à la période du signal (fréquences faibles des signaux EEG)

de différentes sources cérébrales. Une approche classique permettant de mettre en lumière cette relation linéaire est décrite dans l'annexe B.
Le potentiel mesuré sur une électrode est ainsi une moyenne spatiale de la contribution des activités électriques de différentes sources cérébrales. Ce moyennage dépend de la distance avec ces sources, de leurs orientations ainsi que l'amplitude de leurs activités [173]. Ainsi, pour un ensemble de sources cérébrales contiguës les unes aux autres, il semble raisonnable de supposer une localisation spatiale de la mesure de leur activité au niveau des électrodes. Cette localisation est en réalité perturbée par la propagation des ondes dans le crâne du fait de la faible conductivité électrique de ce dernier. Les activités sont par conséquent relativement étalées spatialement et présentent un aspect « flou », le crâne agissant comme un filtre spatial passe-bas du signal.
Les hypothèses effectuées plus haut découlent directement de ces observations : d'une part les topographies des activités recherchées sont régulières et lisses spatialement, et d'autre part elles présentent des extrema localisés sur quelques électrodes malgré un étalement assez important.

Grâce aux outils développés récemment, il est possible de visualiser les topographies correspondantes à des sources cérébrales spécifiques et ainsi de mieux comprendre les hypothèses effectuées précédemment. Pour cela, nous avons construit un modèle de tête réaliste d'un sujet à partir d'enregistrements d'imagerie par résonance magnétique (IRM) pour ensuite calculer de manière précise la matrice de gain pour des sources réparties uniformément dans le cerveau [2]. La figure 4.2 présente un tel modèle de tête [177] à 3 couches : peau, crâne, cortex.
La matrice de gain est calculée à partir d'une discrétisation du cerveau en voxels (petits éléments de volume) pour un ensemble de 64 électrodes placées selon le standard 10-20 (voir la figure 2.4) à la surface du crâne. Pour chacun des voxels, 3 topographies sont calculées pour 3 orientations différentes (x, y et z) du dipôle représentant son activité électrique, permettant ainsi d'obtenir par combinaison linéaire l'ensemble des topographies pouvant être générées par des sources situées dans ce voxel. Nous avons effectué ici ce calcul grâce au solver direct OpenMEEG [95]. Quelques-unes des topographies obtenues sont présentées dans la figure 4.3.

Ces topographies permettent une visualisation des phénomènes décrits plus haut, elles présentent des extrema localisés et sont régulières.
Nous allons maintenant nous attacher à la mise en place de régularisations permettant d'assurer le respect de ces *a priori* dans les matrices de décomposition.

4.1.2 Formalisation mathématique

Comme nous l'avons vu précédemment, pour un signal EEG $Y \in \mathbb{R}^{T \times C}$ et un dictionnaire temporel $\Phi \in \mathbb{R}^{T \times N_\Phi}$, le modèle de décomposition multicanal peut s'exprimer comme

2. Les données utilisées proviennent de la librairie FieldTrip : http ://fieldtrip.fcdonders.nl/tutorial/headmodel_eeg

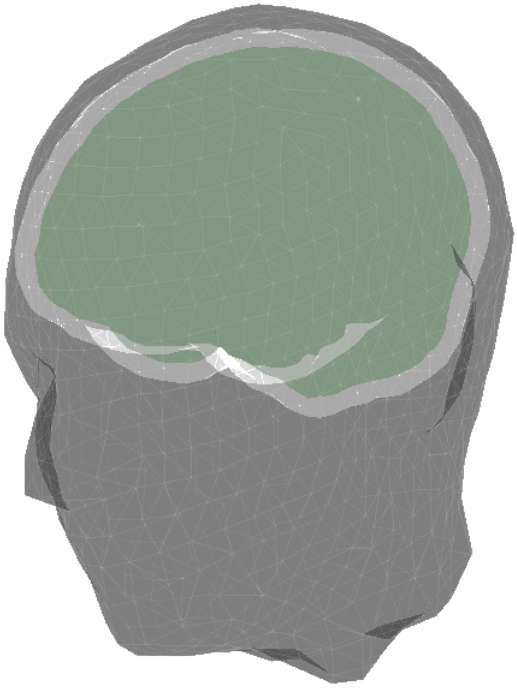


FIGURE 4.2 – Représentation 3D d'un modèle de tête réaliste à trois couches obtenu à l'aide d'enregistrements MRI.

suit :

$$Y = \Phi X + N$$

avec $X \in \mathbb{R}^{N_\Phi \times C}$ une matrice de décomposition et $N \in \mathbb{R}^{T \times C}$ une matrice de bruit.
Le problème de décomposition associé s'écrit alors :

$$\hat{X} = \underset{X \in \mathbb{R}^{N_\Phi \times C}}{\arg\min} \quad \|Y - \Phi X\|_F^2 + R(X).$$

avec $R(X)$ un terme de régularisation encodant les *a priori* sur la solution.
Afin de guider la décomposition de signaux EEG vers des solutions plausibles physiologiquement respectant les propriétés décrites ci-dessus, nous nous intéressons maintenant à la création de termes de régularisations spatiales.

Ces régularisations utilisent la structure spatiale des électrodes. Des voisinages à 4 et 8 électrodes sont envisagés ici et illustrés dans la figure 4.4. Pour une électrode donnée, un voisinage de 4 éléments ne prend en compte que les électrodes présentent au-dessus, en dessous, à droite et à gauche de celle-ci tandis qu'un voisinage de 8 éléments considère également les électrodes adjacentes sur les diagonales.

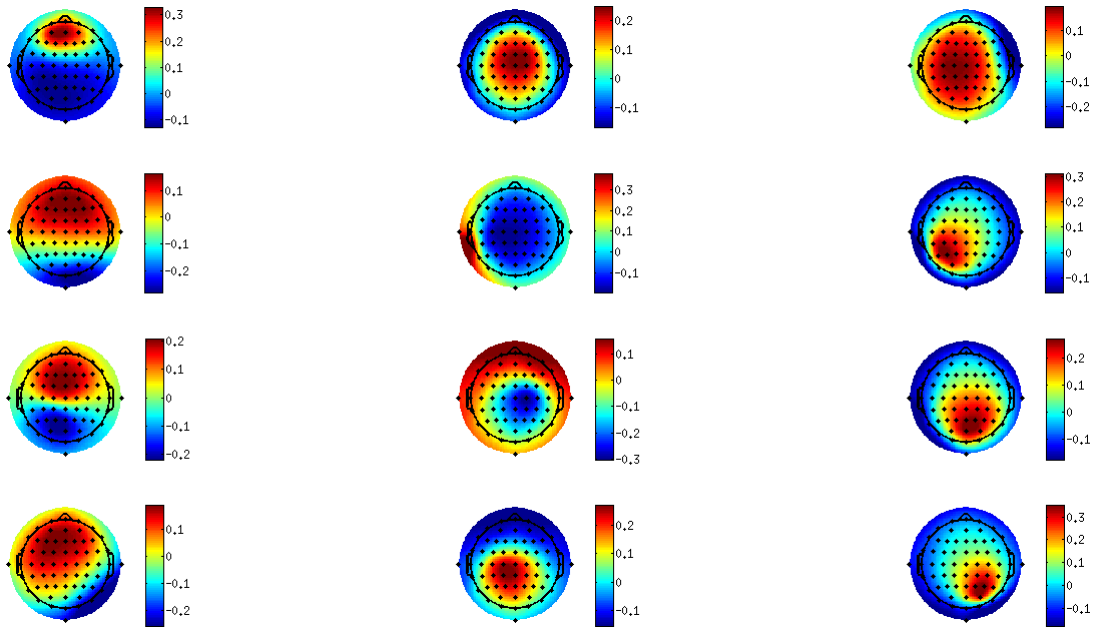


FIGURE 4.3 – Exemple de topographies (normalisées) de la matrice de gain obtenue par résolution du problème direct EEG pour un modèle de tête réaliste.

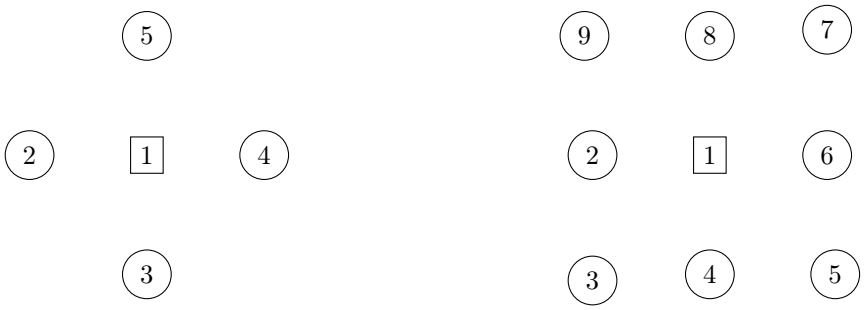


FIGURE 4.4 – Voisinages 2D de taille 4 (à gauche) et 8 (à droite).

4.1.2.1 Régularisations parcimonieuses structurées

Dans un premier temps, nous considérons l'hypothèse de localisation des activités cérébrales. Comme nous l'avons vu précédemment, cette localisation est perturbée par la diffusion des ondes électromagnétiques par le crâne et donc les activités cérébrales affectent de nombreuses électrodes simultanément. Ainsi, il semble naturel de considérer premièrement une régularisation imposant une structure commune à l'ensemble des canaux. Nous rappelons alors le terme de régularisation associé (formalisme ℓ_1) :

$$R_{2,1}(X) = \sum_{n=1}^{N_\Phi} \|X(n,.)\|_2 = \|X\|_{2,1}$$

Néanmoins, afin de mieux identifier ces activités, nous souhaiterions qu'elles s'expriment dans la matrice de décomposition seulement sur quelques électrodes (celles correspondantes aux extrema), améliorant ainsi la résolution spatiale de la représentation. Cette préoccupation est classique dans l'étude des signaux EEG et un filtrage spatial passe-haut est souvent considéré pour cela avec par exemple le filtrage laplacien présenté dans le chapitre 2. Ici, nous proposons de considérer une régularisation parcimonieuse par groupe, avec des groupes composés d'électrodes voisines spatialement. Ce type de régularisation peut être obtenu via la régularisation du « Group Lasso » introduite précédemment :

$$R^g_{2,1}(X) = \sum_{j=1}^{N_\Phi} \sum_{i=1}^{C} d_i \|X(j,\ g_i)\|_2,$$

avec $\{g_i,\ i \in \{1,\dots,C\}\}$ les groupes d'électrodes voisines spatialement et $\{d_i, i \in \{1,\dots,C\}\}$ les coefficients de pondération associés aux différents groupes.
Toutefois, ces groupes étant non disjoints, le support de décomposition obtenu avec ce terme n'est pas l'union d'un faible nombre de groupes d'électrodes voisines mais le complémentaire de l'union de groupes mis à zéro (voir [116] pour plus de détails). Cette régularisation sélectionne en effet des groupes de coefficients qui sont mis à zéro, or nous voulons ici une régularisation choisissant des groupes de coefficients non-nuls. Récemment [115, 175], une régularisation considérant des variables latentes a été proposée pour cela, le problème de régression associé est nommé « Latent Group Lasso ». Soit $\{\mathbf{v}^i_j \in \mathbb{R}^C \mid \forall i \in \{1,\dots,C\}\}$ l'ensemble des variables latentes associées à chacun des groupes d'électrodes pour l'atome j, cette régularisation s'écrit :

$$R^{lg}_{2,1}(X) = \sum_{j}^{N_\Phi} \min_{\mathbf{v}^i_j \in \mathbb{R}^C} \sum_{i=1}^{C} d_i \|\mathbf{v}^i_j\|_2 \quad \text{t.q.} \quad \begin{cases} \sum_{i=1}^{C} \mathbf{v}^i_j = X(j,.) \\ \forall i \in \{1,\dots C\},\ supp(\mathbf{v}^i_j) \subset g_i \end{cases}$$

avec $supp(\mathbf{x})$ l'ensemble des coefficients non-nuls du vecteur $\mathbf{x}$ (support).
Les supports de décomposition obtenus avec cette régularisation sont les complémentaires de ceux obtenus avec la régularisation du « Group Lasso ». Une illustration de la différence entre ces deux régularisations ($R^{lg}_{2,1}$ et $R^g_{2,1}$) en terme de groupes possibles de variables non-nulles est présentée dans la figure 4.5. À noter qu'elles sont équivalentes lorsque les groupes sont disjoints.

Concernant les coefficients de pondération associés aux différents groupes, ils dépendent de la taille de ces groupes. Cette taille peut varier en fonction du nombre d'électrodes voisines considérée ainsi que de l'emplacement de l'électrode centrale du groupe (faible nombre de voisins sur les bords de la grille de capteurs notamment). Il est possible de les fixer ici comme suit :

$$\forall i \in \{1,\dots,C\},\quad d_i \ = \frac{1}{\#\{g_i\}}\ .$$

4.1.2.2 Régularisation de lissage

En ce qui concerne l'*a priori* de régularité des topographies associées à l'activité électrique des sources cérébrales, nous avons choisi une régularisation de type Tikhonov [91].

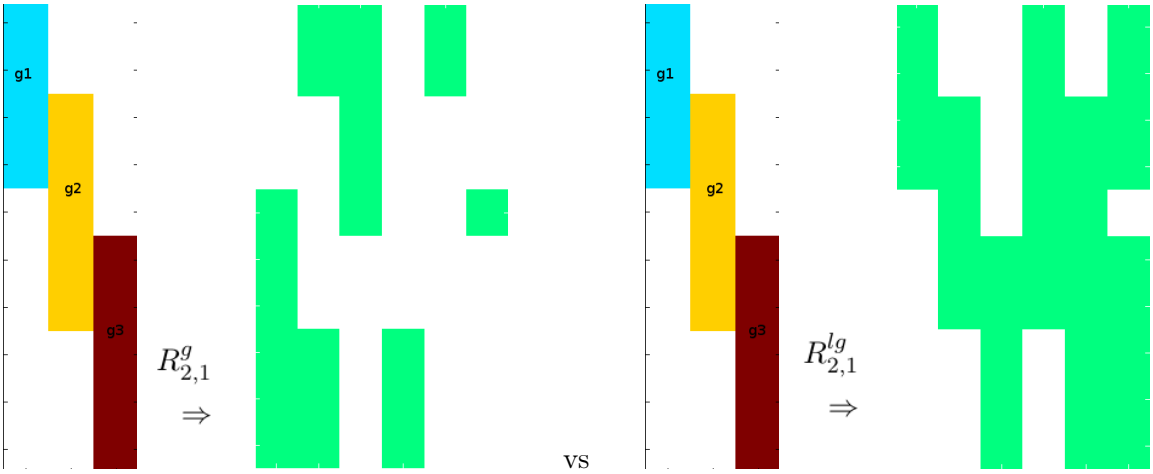


FIGURE 4.5 – Comparaison des groupes possibles pour les régularisations du « Group-Lasso » et du « Latent Group-Lasso » avec 3 groupes : g_1, g_2 et g_3 (inspirée de [175]). Ces groupes sont complémentaires.

Elle s'exprime ici sous la forme suivante :

$$R(X) = \|XL\|_F^2,$$

avec L une matrice encodant l'*a priori* spatial.
Le caractère lisse des coefficients peut être obtenu de manière classique en considérant une régularisation laplacienne qui peut être formalisée ici comme suit :

$$R_{lap}(X) = \sum_{j}^{N_\Phi} \sum_{i=1}^{C} \left(\#\{g_i\}\ X(j,i) - \sum_{s \in g_i} X(j,s) \right)^2 = \|XL_2\|_F^2,$$

avec $\{g_i,\ i \in \{1,\dots,C\}\}$ les groupes d'électrodes voisines introduites précédemment, d la taille du voisinage et L_2 la matrice de l'opérateur laplacien. Contrairement à ce dont nous avons discuté plus haut, le laplacien est utilisé ici en tant qu'opérateur de lissage et non pour un filtrage passe haut, étant donné le fait que l'on considère la minimisation de la norme du vecteur obtenu après son application.
Une régularisation forçant plus simplement les coefficients de décomposition des électrodes proches spatialement à avoir des coefficients proches est également envisageable. Pour cela, nous considérons la régularisation suivante :

$$R_{spa}(X) = \sum_{j}^{N_\Phi} \sum_{i=1}^{C} \sum_{s \in g_i} \frac{1}{2} (X(j,i) - X(j,s))^2 = \|XL_1\|_F^2,$$

avec L_1 la matrice de régularisation associée.
Pour un voisinage de taille 4, si l'électrode 1 à pour voisines les électrodes 2, 3, 4 et 5 alors

les matrices L_1 et L_2 associées sont écrites :

$$L_1 = \begin{pmatrix} 1 & 1 & 1 & 1 \\ -1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix}, \quad L_2 = \begin{pmatrix} 4 \\ -1 \\ -1 \\ -1 \\ -1 \end{pmatrix}$$

Lorsque les distances entre les électrodes sont connues et fixes, ces régularisations peuvent être pondérées. Il est par exemple possible de considérer les poids $w_{i,j} = \frac{1}{d_{i,j}}, i, j \in \{1, \ldots, C\}$ où $d_{i,j}$ représente la distance entre les électrodes i et j. Les régularisations s'écrivent alors :

$$R_{w,lap}(X) = \sum_{j}^{N_\Phi} \sum_{i=1}^{C} \left(\sum_{s \in g_i} w_{i,s}\ X(j,i) - \sum_{s \in g_i} w_{i,s} X(j,s) \right)^2 = \|X L_{w,2}\|_F^2,$$

$$R_{w,spa}(X) = \sum_{j}^{N_\Phi} \sum_{i=1}^{C} \sum_{s \in g_i} \frac{1}{2} \left(w_{i,s} \left(X(j,i) - X(j,s) \right) \right)^2 = \|X L_{w,1}\|_F^2$$

avec $L_{w,1}$ et $L_{w,2}$ les matrices de régularisation pondérées.
Ces régularisations sont des filtres spatiaux passe-bas. Ils permettent donc d'éliminer tout bruit de fréquence spatiale élevée. Leur efficacité dans l'élimination d'un bruit blanc pour les topographies observées plus haut peut être illustrée via une expérience simple.

Pour cela, nous créons un ensemble de K signaux à partir d'un dictionnaire temps-fréquence de Gabor Φ_t et d'un dictionnaire de topographies Φ_s obtenu via la résolution du problème direct introduit plus haut. Chaque signal Y_k est construit comme la somme de N signaux spatio-temporels créés en associant des atomes de Φ_t et des atomes de Φ_s :

$$Y^k = \sum_{n=1}^{N} \Phi_t(\kappa_{k,t}(n)) \mathbf{c}_k(n) \Phi_s(\kappa_{k,s}(n))^T,$$
$$Y_k = \Phi_t(\kappa_{k,t}) X^k$$

avec $\kappa_{k,t}$ et $\kappa_{k,s}$ les vecteurs d'indices des atomes temporels et spatiaux du signal k et $\mathbf{c}_k$ le vecteur de coefficients associé.
Nous souhaitons évaluer la capacité des régularisations définies plus haut à permettre la récupération de la matrice de coefficients X^k lorsque Y^k est bruité et les indices des atomes temporels $\kappa_{k,t}$ connus. Ainsi pour $\bar{Y}^k = Y^k + N$ la version bruitée de Y_k pour une matrice de bruit blanc gaussien N, nous considérons la résolution du problème suivant :

$$\hat{X}^k = \min_X \|\bar{Y}^k - \Phi_t(\kappa_{k,t}) X\|_F^2 + \mu \|X L_*\|_F^2$$

avec $\mu \geq 0$ et L_* l'une des matrices de régularisation spatiale introduite plus haut. C'est un problème convexe différentiable dont la résolution fait intervenir l'équation de Sylvester

suivante :

$$\Phi_t(\kappa_{k,t})^T \Phi_t(\kappa_{k,t}) \hat{X}^k + \mu \hat{X}^k L_* L_*^T = \Phi_t(\kappa_{k,t})^T \bar{Y}^k \quad .$$

Celle-ci rentre dans le cadre de la proposition 1 et est résolue comme expliquée dans l'annexe A.

Pour différents niveaux du rapport signal à bruit (RSB), toutes les régularisations et des voisinages de 4 et de 8 éléments, ces solutions sont calculées pour l'ensemble des K signaux et l'évaluation est effectuée via le critère suivant : $\varepsilon(k) = \frac{\|\hat{X}^k - X^k\|_F^2}{\|X^k\|_F^2}$. Le paramètre μ est déterminé sur un ensemble d'entrainement par validation croisée pour chacun des cas testés. La valeur optimale de ce paramètre obtenue pour une régularisation laplacienne lorsque le critère est minimal augmente avec la valeur du RSB comme l'illustre la figure 4.6.

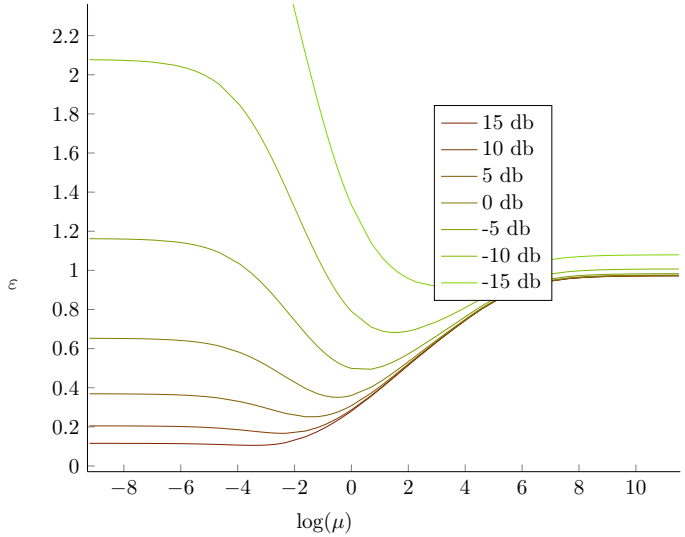


FIGURE 4.6 – Valeurs optimales du paramètre de régularisation μ pour la résolution d'un problème des moindres carrés régularisé par un terme de lissage. Les courbes présentées correspondent à différentes valeurs du rapport signal à bruit (bruit blanc).

Les résultats de l'expérience pour $K = 150$ sont présentés dans le tableau de la figure 4.7. Les différents voisinages évalués sont indiqués en exposant. Ces résultats montrent que les régularisations spatiales permettent une récupération plus efficace de la structure sous-jacente des signaux (débruitage). Un voisinage étendu à 8 électrodes est plus efficace que celui à 4 électrodes. De plus, pour un tel débruitage, la régularisation laplacienne est plus efficace que la régularisation locale.

Apprentissage de la régularisation de lissage La matrice $P = LL^T$ joue un rôle important dans ces régularisations. Le terme $\|X^k L_*\|_F^2$ peut en effet être vu comme une

	15 db	**10 db**	**5 db**	**0 db**	**-5 db**	**-10 db**	**-15 db**	**Moy**
Brut	0.11445	0.20342	0.36188	0.64507	1.1407	2.0425	3.6243	1.1618
R^4_{spa}	0.10503	0.17004	0.26052	0.38232	0.53025	0.70159	0.90939	0.43702
R^8_{spa}	0.10652	0.17296	0.26829	0.39203	0.53728	0.71146	0.90964	0.4426
$R^4_{w,spa}$	0.10387	0.16721	0.25577	0.37771	0.52427	0.6979	0.90835	0.43358
$R^8_{w,spa}$	0.10462	0.16876	0.25829	0.37761	0.52297	0.6922	0.89726	0.43167
R^4_{lap}	0.10511	0.16821	0.25127	0.35838	0.49404	0.67861	0.9064	0.42315
R^8_{lap}	0.10402	0.16499	0.24655	0.35201	0.48801	0.66833	0.89731	0.41732
$R^4_{w,lap}$	0.10459	0.16686	0.24833	0.35436	0.4903	0.67715	0.90766	0.42132
$R^8_{w,lap}$	**0.10212**	**0.16083**	**0.23875**	**0.3423**	**0.4787**	**0.66076**	**0.89275**	**0.41089**

FIGURE 4.7 – Erreurs quadratiques de reconstruction obtenues pour différentes régularisations de lissage lorsque les signaux sont contaminés par un bruit blanc. Les résultats sont présentés pour différentes valeurs du rapport signal à bruit.

pénalisation fondée sur la norme quadratique $\|\mathbf{x}\|_P = \mathbf{x}^T P \mathbf{x}$ s'écrivant :

$$\|X^k\|^2_{P,2} = \sum_{i=1}^{N_\Phi} X^k(i,.) P X^k(i,.)^T = \|X^k L_*\|^2_F,$$

avec P une matrice symétrique réelle définie positive. L'apprentissage d'une telle norme peut être envisagé pour une meilleure adaptation à un certain ensemble de topographies comme celles associées par exemple à la résolution du problème direct pour un sujet particulier.

Soit $X = [X(1), \ldots, X(T)] \in \mathbb{R}^{C \times T}$ un ensemble de T topographies, nous souhaitons apprendre une norme quadratique permettant de réaliser le débruitage de telles topographies. Formellement, soit $Y = [Y(1), \ldots, Y(T)]$ des topographies obtenues par bruitage des topographies concaténées dans X, notre objectif est de pouvoir débruiter de tels signaux via la résolution du problème suivant :

$$\hat{X} = \underset{X \in \mathbb{R}^{C \times T}}{\arg\min} \|Y - X\|^2_F + \mu \|X^T\|^2_{P,2} \ ,$$

avec P une matrice symétrique réelle définie positive. La solution d'un tel problème s'écrit :

$$\hat{X} = (I_C + \mu P)^{-1} Y$$

avec I_C la matrice identité de taille C.
Ainsi, pour un modèle de bruit donné selon lequel un ensemble de signaux bruités Y est créé à partir de l'ensemble des topographies disponibles X, l'apprentissage de P peut être effectué via la résolution du problème suivant :

$$\hat{P} = \underset{P \in \mathcal{S}_\mathcal{C}}{\arg\min} \|X - (I_C + \mu P)^{-1} Y\|^2_F$$

avec $\mathcal{S}_C$ l'ensemble des matrices symétriques réelles définies positives de taille C.
Les matrices symétriques réelles définies positives étant inversibles (avec des inverses possédant les mêmes propriétés), ce problème peut être reformulé comme suit :

$$\hat{Q} = \underset{Q \in \mathcal{S}_C}{\arg\min} \|X - QY\|_F^2 \ ,$$

avec $Q = (I_C + \mu P)^{-1}$.
La résolution de ce problème peut être réalisée comme décrit dans [231], elle aboutit à :

$$QYY^T + YY^TQ = XY^T + YX^T \ .$$

Cette dernière équation est une équation de Sylvester (en Q) particulière dont la résolution peut être effectuée comme décrit dans l'annexe A.

Afin d'évaluer l'efficacité d'un tel apprentissage, nous avons créé à partir des topographies obtenues dans le modèle direct précédemment décrit un ensemble de 1000 topographies bruitées (bruit blanc gaussien). La mesure des performances de l'apprentissage a ensuite été réalisée par validation croisée sur cet ensemble en utilisant $\frac{9}{10}$ des signaux pour l'entrainement et $\frac{1}{10}$ pour l'évaluation. Le critère retenu pour cette évaluation est simplement l'erreur quadratique ($\frac{\|\hat{X}-X\|_F^2}{\|X\|_F^2}$), la moyenne des résultats pour différents RSB est présentée dans la figure 4.8 et comparée avec celle obtenue pour la régularisation $R^8_{w,lap}$.

	15 db	**10 db**	**5 db**	**0 db**	**-5 db**	**-10 db**	**-15 db**	**Moy**
$R^8_{w,lap}$	**0.022147**	0.039151	0.068162	0.11444	0.18218	0.27325	0.37694	0.15375
Appris	0.022466	**0.038544**	**0.064338**	**0.10083**	**0.15167**	**0.21577**	**0.30181**	**0.12792**

FIGURE 4.8 – Comparaison des erreurs quadratiques de reconstruction obtenues avec la régularisation de lissage apprise et la régularisation laplacienne. Les résultats sont présentés pour différentes valeurs du rapport signal à bruit.

4.1.3 Décomposition temporelle régularisée spatialement

Les régularisations spatiales que nous avons choisies d'étudier ayant été présentées, nous pouvons écrire maintenant notre problème de décomposition régularisée spatialement :

$$\hat{X} = \underset{X \in \mathbb{X}^{C \times T}}{\arg\min} \|Y - \Phi X\|_F^2 + \lambda R^{lg}_{2,1}(X) + \mu \|XL_*\|_F^2$$

avec L_* une des matrices de régularisations spatiales considérées et $R^{lg}_{2,1}(X)$ la régularisation du « Latent Group LASSO » présentée plus haut pour des groupes d'électrodes voisines.
Nous étudierons ce problème d'optimisation sous sa forme convexe et stricte (formalisme ℓ_0) dans le chapitre 5 avant de s'intéresser à son application dans le cadre de l'analyse temps-fréquence de signaux EEG.

4.2 Régularisation temporelle

4.2.1 Hypothèses et *a priori* neurophysiologiques

Nous considérons maintenant la régularisation temporelle d'une décomposition sur un dictionnaire spatial. Le comportement dynamique des signaux EEG étant complexe et fortement dépendant des tâches mentales effectuées, une régularisation utilisable dans un cadre général semble difficile à mettre en place. Toutefois, nous avons déjà évoqué dans le premier chapitre que certaines études [134, 138, 136] ont observé que **ces signaux pouvaient être modélisés par des suites de topographies restant stables quelques dizaines de millisecondes chacunes et présentant des transitions brusques entre elles**.
C'est sur cette dernière hypothèse que nous allons créer notre régularisation, *i.e.* nous supposons dans la suite que les topographies apparaîssant dans un signal EEG comportent des transitions brusques et des zones de stabilité.

L'observation des suites de topographies EEG fait apparaître des topographies qui possèdent des *extrema* prononcés entrecoupés par des topographies « plates ». Afin d'évaluer l'évolution temporelle de ce comportement, deux indices ont été proposés [136] :

- le GFP (« Global Field Power ») mesurant la déviation d'une topographie $\mathbf{y} \in \mathbb{R}^C$ par rapport à la moyenne des électrodes $\bar{\mathbf{y}}$:

$$GFP(\mathbf{y}) = \sqrt{\frac{\sum_{i=1}^{C}(\mathbf{y}(i) - \bar{\mathbf{y}})^2}{C}}$$

- le GMD (« Global Map Dissimilarity ») évaluant la différence entre deux topographies $\mathbf{y}, \mathbf{z} \in \mathbb{R}^C$:

$$GMD(\mathbf{y}, \mathbf{z}) = \sqrt{\frac{1}{C}\sum_{i=1}^{C}\left(\frac{\mathbf{y}(i) - \bar{\mathbf{y}}}{GFP(\mathbf{y})} - \frac{\mathbf{z}(i) - \bar{\mathbf{z}}}{GFP(\mathbf{z})}\right)^2}$$

Un exemple des courbes obtenues avec ces indices pour un signal EEG réel est donné dans la figure 4.9 extraite de [136].

Le GFP mesure la déviation standard spatiale et permet donc d'évaluer à quel point une topographie est « marquée » avec des *extrema* élevés (fort GFP) ou bien « plate » et ne possèdent que de faibles *extrema* (faible GFP). En ce qui concerne le GMD, il caractérise la différence de forme de deux topographies indépendamment de leurs GFP (et donc de leurs normes). Les courbes présentées ci-dessus permettent d'observer que le GMD est caractérisé par un ensemble de pics correspondant à des changements brusques de topographies. Entre ces pics, des maxima du GFP associés à des topographies plus marquées apparaîssent.
Les phases de relative stabilité apparaîssant entre les modifications brusques des suites de topographies sont nommées micro-états et sont associées à des topographies moyennes.

4.2.2 Formalisation mathématique

Le modèle de décomposition spatiale multicanal pour un signal EEG $Y \in \mathbb{R}^{C\times T}$ (transposé du signal de la section précédente) et un dictionnaire spatial $\Phi \in \mathbb{R}^{C\times N_\Phi}$ peut s'expri-

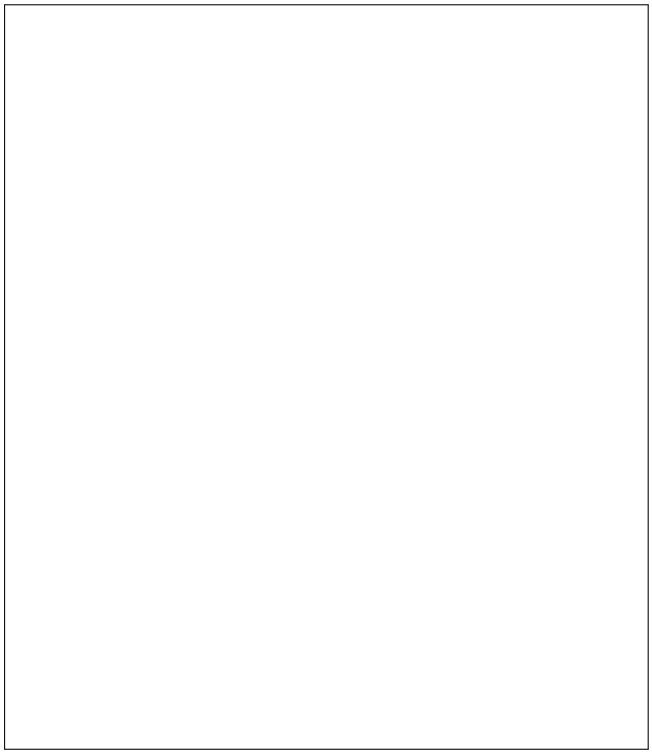


FIGURE 4.9 – Exemple des évolutions temporelles du GFP et du GMD pour un enregistrement EEG sur une durée de 211 ms. Les maxima du GFP sont notés par des astérisques. Source : [136].

mer comme suit :

$$Y = \Phi X + N$$

avec $X \in \mathbb{R}^{N_\Phi \times T}$ une matrice de décomposition et $N \in \mathbb{R}^{C \times T}$ une matrice de bruit.
Le problème de décomposition associé s'écrit alors :

$$\hat{X} = \underset{X \in \mathbb{R}^{N_\Phi \times T}}{\arg\min} \quad \|Y - \Phi X\|_F^2 + R(X). \tag{4.1}$$

avec $R(X)$ un terme de régularisation encodant les *a priori* temporels.

Afin de respecter l'*a priori* de stabilité des micro-états, nous souhaitons que la matrice de coefficients obtenue avec une telle décomposition soit parcimonieuse par bloc. Les blocs souhaités ne sont pas connus à l'avance comme pour la régularisation spatiale considérée précédemment.
Formellement, une telle structure pour X peut être écrite comme suit :

$$\forall c \in \{1, \cdots, C\}, \quad X^T(c) = \sum_{i=1}^{\mathbf{n_c}} \alpha^c(i) \mathbf{1}_{\kappa_{i,c}}.$$

avec $\{\kappa_{i,c},\ \forall i \in \{1,\cdots,n_c\}\}$ une partition de $[1,\cdots,T]$ correspondant aux blocs du c-ième canal de X et $\{\alpha^c(i) \in \mathbb{R},\ \forall i \in \{1,\cdots,n_c\}\}$ les coefficients associés. Une illustration de cette structure est présentée dans la figure 6.1 ; les coefficients associés à l'un des micro-états sont affichés en gris.

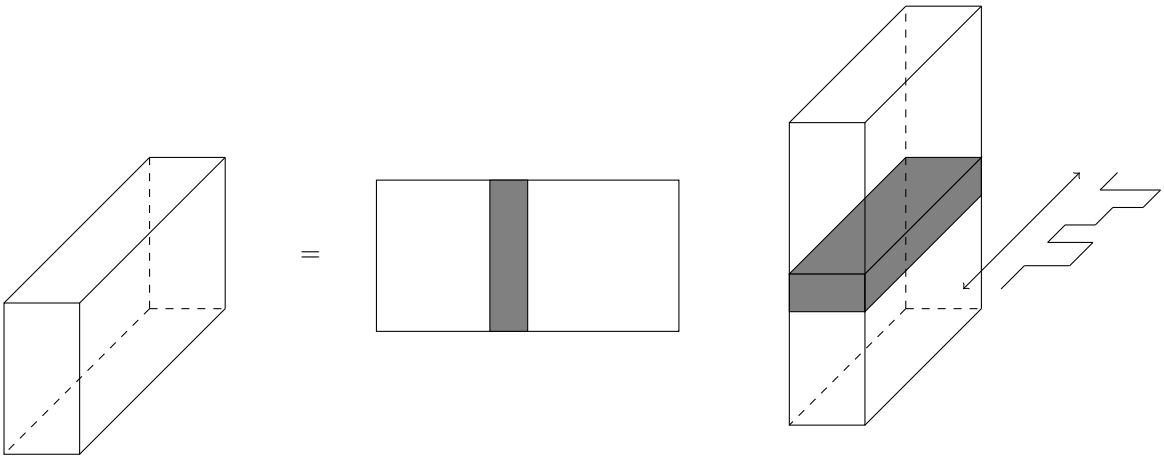


FIGURE 4.10 – Décomposition par bloc parcimonieux d'une série temporelle multidimensionnelle.

Dans le but d'obtenir de telles structures pour les matrices de décompositions du problème de l'Eq.(4.1) la régularisation $R(X)$ doit permettre une détection des ruptures du signal. Une régularisation de type Variation Totale (VT) permet la détection de ces ruptures. Cette régularisation a été introduite dans le domaine de l'image pour des tâches de débruitage [198] et a été étudiée de façon intensive [51]. Afin de détecter les ruptures d'un signal, cette régularisation pénalise le gradient de celui-ci en valeur absolue. Dans notre formalisme discret elle s'exprime comme suit :

$$R_{VT}(X) = \sum_{t=2}^{T} \|X(t) - X(t-1)\|_1 = \|XP\|_1,$$

$$\text{avec } P = \begin{pmatrix} -1 & & & \\ 1 & -1 & & \\ & 1 & \ddots & \\ & & \ddots & -1 \\ & & & 1 \end{pmatrix}$$

Dans notre problème d'apprentissage, cette régularisation est combinée avec une régularisation parcimonieuse afin d'imposer un nombre faible d'états actifs en même temps.

La régularisation VT s'exprimant naturellement avec une norme ℓ_1, nous étudierons cette combinaison de régularisations pour une pénalisation parcimonieuse ℓ_1. Notre régularisation s'écrit donc :

$$R_{FL}(X) = \lambda\|X\|_1 + \mu\|XP\|_1.$$

avec λ et μ les paramètres de régularisation gérant l'importance de ces termes.

Le principal atout de cette régularisation réside dans sa capacité à identifier les ruptures du signal et donc les phases de stabilité des suites temporelles de topographies. Son inconvénient majeur provient du fait qu'elle ne prend pas en compte les variations régulières des topographies entre les ruptures.

4.2.3 Décomposition spatiale régularisée temporellement

Le problème de décomposition régularisée temporellement associé à notre hypothèse peut maintenant s'écrire comme suit :

$$(\hat{X}^1, \dots, \hat{X}^K, \hat{\Phi}) = \underset{X^1,\dots,X^K,\Phi}{\arg\min} \sum_{k=1}^{K} \|Y^k - \Phi X^k\|_F^2 + \lambda\|X^k\|_1 + \mu\|X^k P\|_1 \tag{4.2}$$

Dans un contexte monodimensionnel (et non-redondant), ce problème a été introduit sous le nom de « Fused-LASSO » dans [211]. Ce type de régularisation a été étudié ensuite dans différents contextes comme l'étalement de spectre par saut de fréquence [7], certaines études en géophysique [89], l'apprentissage multi-tâches [240], l'analyse de tendance [123], l'estimation de matrice de covariance [50], l'analyse de marqueurs génétiques [122] ou encore la détection de ruptures [28].
Malgré sa convexité, ses termes non-différentiables le rendent difficile à résoudre, en particulier dans un cas multicanal et redondant comme celui qui nous occupe. Par conséquent, avant de revenir au modèle des micro-états pour la modélisation des signaux EEG (chapitre 7) nous allons nous intéresser à ce problème et proposer un schéma d'optimisation pour sa résolution dans le chapitre 6.
La conception d'un tel algorithme est également intéressante pour d'autres types de régularisations utilisant la combinaison d'une régularisation ℓ_1 et d'une régularisation en analyse. Nous avons notamment vu dans le précédent chapitre que ce type de régularisation permettait de retrouver plus facilement la structure parcimonieuse d'un signal sur un dictionnaire lorsque cette structure pouvait s'exprimer par une combinaison linéaire des atomes d'un autre dictionnaire dont le dual était utilisé dans la régularisation en analyse (condition D-RIP).

4.3 Conclusion

Ce chapitre a présenté la conception de régularisations pour la décomposition de signaux EEG sur des dictionnaires redondants. Ces régularisations ont été conçues de manière à contraindre ces décompositions à respecter des *a priori* physiologiques.

Nous nous sommes intéressés dans un premier temps à une décomposition temporelle pour laquelle la combinaison d'une régularisation parcimonieuse et d'une régularisation de lissage a été proposée. Nous avons ensuite conçu une régularisation temporelle pour le problème transposé permettant de réaliser des décompositions spatiales. Cette régularisation est fondée sur le modèle des micro-états et se compose d'un terme de parcimonie et d'un terme permettant de conserver les ruptures temporelles des signaux.

Ces régularisations ayant été construites, nous allons maintenant nous concentrer sur la mise en place d'algorithmes de décomposition utilisant celles-ci. Le chapitre 5 est ainsi dédié à la décomposition temps-fréquence régularisée spatialement de signaux EEG et le chapitre 6 à leur décomposition spatiale régularisée temporellement. Une extension du modèle des micro-états à l'aide de cette dernière décomposition sera par la suite mise en place dans le chapitre 7.

Chapitre 5

Décomposition temps-fréquence régularisée spatialement

Sommaire

Ce chapitre est dédié à la mise en place d'algorithmes permettant de réaliser les décompositions temps-fréquence régularisées spatialement proposées dans le chapitre 4. Après un rapide rappel du modèle considéré et des problèmes d'optimisation associés, des approches gloutonnes sont proposées pour la résolution de la formulation ℓ_0 (strict) des problèmes et des méthodes d'optimisation convexe sont considérées pour la résolution des problèmes de décomposition ℓ_1 (relachés). Ces algorithmes sont ensuite évalués sur des signaux synthétiques réalistes avant d'être appliqués sur des signaux réels pour la détection du PE P300 dans un contexte discriminatif.

5.1 Modèle de décomposition temps-fréquence

5.1.1 Modèle linéaire multicanal

Soit $Y \in \mathbb{R}^{T \times C}$ un signal EEG enregistré sur C électrodes durant T pas temporels et $\Phi = [\phi(1), \ldots, \phi(N_\Phi)] \in \mathbb{R}^{T \times N_\Phi}$ un dictionnaire temps-fréquence redondant de taille N_Φ ($N_\Phi \gg T$). Nous étudions dans ce chapitre la décomposition $X \in \mathbb{R}^{N_\Phi \times C}$ de signaux

multicanaux Y sur le dictionnaire Φ s'exprimant comme suit :

$$\begin{aligned} \mathbf{Y}(c) &= \Phi\mathbf{X}(c) + \mathbf{n}(c), \quad c \in \{1, \ldots, C\} \ , \\ Y &= \Phi X + N \end{aligned}$$

avec $N = [\mathbf{n}(1), \ldots, \mathbf{n}(C)] \in \mathbb{R}^{T \times C}$ une matrice de bruit gaussien. Une visualisation est proposée dans la figure 3.4.
Nous considérons ici différentes combinaisons de régularisations spatiales (régularisations parcimonieuses et régularisations de lissage) introduites dans le chapitre précédent afin de guider ce type de décompositions vers des solutions plausibles physiologiquement.

5.1.2 Dictionnaires temps-fréquence

Un dictionnaire temps-fréquence est composé d'atomes localisés en temps et en fréquence décrivant la partie du plan temps-fréquence correspondant aux phénomènes étudiés. Un atome temps-fréquence $\Phi(\gamma), \quad 1 \leq \gamma \leq N_{\Phi_t}$, est ainsi caractérisé par une localisation temporelle u_γ, une localisation fréquentielle ξ_γ ainsi que des étalements $\sigma_t(\gamma)$, $\sigma_w(\gamma)$ dans ces deux dimensions. Ces quantités sont définies comme suit [152] :

$$u_\gamma = \sum_t t \ |\Phi(t,\gamma)|^2, \qquad \xi_\gamma = \frac{1}{2\pi} \sum_w w \ |\dot{\Phi}(w,\gamma)|^2$$

$$\sigma_t^2(\gamma) = \sum_t (t - u_\gamma)^2 \ |\Phi(t,\gamma)|^2, \qquad \sigma_w^2(\gamma) = \frac{1}{2\pi} \sum_w (w - \xi_\gamma)^2 \ |\dot{\Phi}(w,\gamma)|^2$$

où $\dot{\Phi}(w,\gamma)$ est la transformée de Fourier de $\Phi(t,\gamma)$. Cette localisation peut être visualisée dans la plan temps-fréquence par une boîte d'Heisenberg, illustrée dans la figure 5.1.
Le principe d'incertitude d'Heisenberg [152] limite la surface de cette boîte : $\sigma_t \sigma_w \geq \frac{1}{2}$. Ainsi, des atomes possédants une localisation temporelle étroite permettront une description temporelle fine du signal mais seront étalés en fréquence et ne donneront qu'une description fréquentielle de résolution faible.
Parmi les dictionnaires temps-fréquence classiques, le dictionnaire de Gabor est connu pour offrir une concentration optimale de l'énergie dans le plan temps-fréquence. Il s'est avéré particulièrement efficace pour la représentation des signaux EEG étant donnée sa capacité à capturer à la fois leurs composantes transitoires et oscillatoires (voir section 2.2). C'est donc ce dictionnaire que nous étudierons pour ce modèle.

Un dictionnaire de Gabor invariant en temps et en fréquence est construit à partir d'une fenêtre gaussienne [152]. Cette fenêtre est dilatée, modulée et translatée de manière à former les différents atomes du dictionnaire. Pour une échelle (dilatation) spécifique $s = 2^j$, une fenêtre gaussienne discrète peut être écrite :

$$g_j(k) = 2^{-\frac{j}{2}+\frac{1}{4}} \ \exp\left(-\pi(2^{-j}k)^2\right).$$

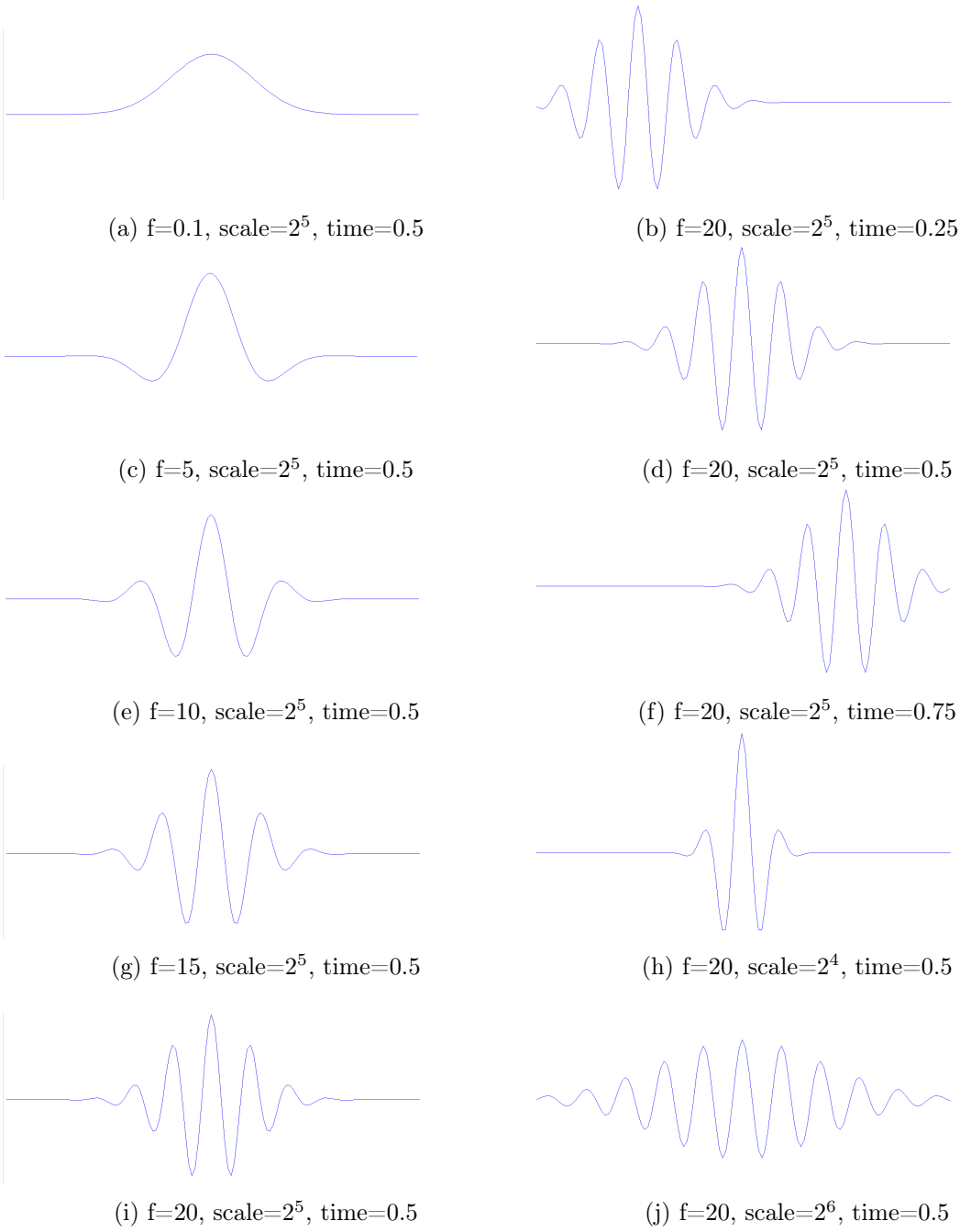
Le dictionnaire de Gabor associé à cette échelle peut alors être construit comme suit :

$$\Phi_j = \{ \ \phi_j(k) = K_j \ g_j(k - qu_j) \ \cos(k\xi_j p) \ \}_{0 \leq q \leq N_q, 0 \leq p \leq N_p}$$

FIGURE 5.1 – Boîte d'Heisenberg d'un atome temps-fréquence Φ_γ, les dimensions de la boîte correspondent aux étalements en temps (abscisse) et en fréquence (ordonné) de l'atome. Source : [152].

où K_j est un facteur de normalisation, u_j et ξ_j les intervalles temporels et fréquentiels de la discrétisation, et N_q, N_p les valeurs maximales des translations (en temps et en fréquence). Un dictionnaire complètement invariant est construit avec $u_j = 1$ et $\xi_j = 2\pi/C$. Toutefois ce dictionnaire possédant un nombre important d'atomes et une grande cohérence, la paramétrisation suivante est souvent choisie : $u_j = 2^j\eta^{-1}$, $\xi_j = 2\pi 2^{-j}\eta^{-1}$ et $N_q = \eta C 2^{-j}$, $N_p = \eta 2^j$. Dans cette dernière discrétisation, le paramètre η permet la gestion des intervalles temporels et fréquentiels et donc du pavage temps-fréquence. Lorsque $\eta \geq 1$ alors le dictionnaire est un repère.
Un dictionnaire de Gabor multi-échelles est l'union de tels dictionnaires ayant des échelles différentes. Une visualisation de différents exemples d'atomes de Gabor est présentée dans la figure 5.2.

Le choix des intervalles de temps et de fréquence est primordial pour assurer de bonnes décompositions, car il est directement lié à la proximité (au sens de la corrélation ici) entre les atomes et donc à la cohérence du dictionnaire. Ainsi, en choisissant des intervalles de petites tailles, la représentation des signaux est plus fine, mais les décompositions sont moins stables.
L'objectif des régularisations que nous allons considérer ici est de guider la décomposition vers des solutions plausibles en utilisant des connaissances neurophysiologiques et cela même lorsque le signal est bruité ou le dictionnaire très cohérent. Une telle régularisation vise à permettre l'obtention de décompositions stables même lorsque l'on souhaite aboutir à une description plus fine des signaux, en diminuant par exemple les intervalles temporels et fréquentiels de la discrétisation du dictionnaire temps-fréquence sur lequel sont réalisées les décompositions.



(a) f=0.1, scale=2^5, time=0.5

(b) f=20, scale=2^5, time=0.25

(c) f=5, scale=2^5, time=0.5

(d) f=20, scale=2^5, time=0.5

(e) f=10, scale=2^5, time=0.5

(f) f=20, scale=2^5, time=0.75

(g) f=15, scale=2^5, time=0.5

(h) f=20, scale=2^4, time=0.5

(i) f=20, scale=2^5, time=0.5

(j) f=20, scale=2^6, time=0.5

FIGURE 5.2 – Atomes d'un dictionnaire de Gabor pour différents paramètres

5.1.3 Régularisations parcimonieuses

Deux régularisations parcimonieuses sont étudiées pour ce modèle. D'une part, elles permettent d'assurer une parcimonie dans le choix des atomes (entre les lignes de la matrice de décomposition), répondant ainsi à l'hypothèse de localisation des activités cérébrales dans

le plan temps-fréquence qui a été abordée dans le premier chapitre. Le signal de chaque canal est alors représenté par seulement quelques atomes.
D'autre part, elles imposent une cohérence spatiale des coefficients de décomposition (agissant sur les colonnes de la matrice de décomposition).

Pour cela, une régularisation imposant une structure parcimonieuse commune aux canaux est premièrement considérée :

$$R_{2,1}(X) = \|X\|_{2,1} = \sum_{j=1}^{N_\Phi} \|X(j,.)\|_2 \ .$$

l'utilisation de l'ensemble des canaux pour le choix des atomes permettant un choix plus cohérent de ceux-ci.

Dans un second temps, afin de prendre en compte de manière plus fine la structure spatiale des signaux EEG, une régularisation imposant une parcimonie sur les groupes de canaux correspondant à des électrodes spatialement proches est étudiée. Pour cela, la régularisation du « Latent Group LASSO » est considérée. Soit $\{g_i \mid i \in \{1,\dots,C\}\}$ les groupes d'électrodes voisines et $\{\mathbf{v}_j^i \in \mathbb{R}^C \mid \forall i \in \{1,\dots,C\}\}$ les variables latentes associées à chacun de ces groupes pour l'atome j, celle-ci s'écrit :

$$R_{2,1}^{lg}(X) = \sum_{j=1}^{N_\Phi} \min_{\mathbf{v}_j^i \in \mathbb{R}^C} \sum_{i=1}^{C} d_i \|\mathbf{v}_j^i\|_2 \ \text{ t.q. } \begin{cases} \sum_{i=1}^{C} \mathbf{v}_j^i = X(j,.) \\ \forall i \in \{1,\dots C\}, supp(\mathbf{v}_j^i) \subset g_i \end{cases}$$

Un voisinage de 8 est choisi ici pour la création de ces groupes. Les coefficients de pondération $\{d_c \mid \forall c \in \{1,\dots,C\}\}$ introduits dans le chapitre précédent ne sont pas considérés (fixés à 1) afin de ne pas augmenter le nombre de paramètres à régler mais les algorithmes proposés dans les sections suivantes peuvent être adaptés pour les prendre en compte.
Ces régularisations sont exprimées ici sous leurs formes relachées mais nous considèrerons également leurs formes ℓ_0 dans la suite. Elles sont nommées respectivement $R_{2,0}$ et $R_{2,0}^{lg}(X)$.

5.1.4 Régularisations spatiales de lissage

En supplément de ces régularisations parcimonieuses, nous considérons ici une régularisation de lissage. Celle-ci prend en compte l'effet de diffusion des signaux électriques cérébraux par le crâne et répond donc à l'*a priori* de régularité spatiale des composantes que nous souhaitons extraire. Cette régularité est attendue des coefficients obtenus dans les lignes de la matrice de décomposition.
Parmi les régularisations de lissage envisagées dans le chapitre précédent, l'expérience de débruitage réalisée (section 4.1.2.2) indique que la plus efficace est la régularisation laplacienne pondérée s'écrivant comme suit :

$$R_{w,lap}^{8}(X) = \sum_{j} \sum_{i=1}^{C} \left(\sum_{s \in g_i} w_{i,s} \ X(j,i) - \sum_{s \in g_i} w_{i,s} X(j,s) \right)^2 = \|X L_2^8\|_F^2$$

dans lequel le chiffre 8 correspond à la taille du voisinage considéré (même notations que dans le chapitre précédent) et les poids w calculés comme inverses des distances entre les électrodes.
L'utilisation d'une régularisation apprise n'est pas considérée ici, mais elle sera envisagée dans de futurs travaux.

5.1.5 Problèmes d'optimisation associés

Le problème d'optimisation associé à ces décompositions s'écrit de manière générale comme la minimisation d'une combinaison d'un terme d'attache aux données et de termes de régularisations pouvant être écrites sous forme de contraintes ou de pénalisations du terme principal. Dans cette étude, nous considérons la résolution du problème à la fois sous sa forme stricte ℓ_0 et sous sa forme relachée ℓ_1.
Forme stricte :

$$\hat{X} = \underset{X \in \mathbb{R}^{N_\Phi \times C}}{\arg\min} \quad \|Y - \Phi X\|_F^2 + \mu \|X L_*\|_F^2 \quad t.q \quad R^0(X) < J \tag{5.1}$$

Forme relachée :

$$\hat{X} = \underset{X \in \mathbb{R}^{N_\Phi \times C}}{\arg\min} \quad \|Y - \Phi X\|_F^2 + \mu \|X L_*\|_F^2 + \lambda R^1(X) \tag{5.2}$$

où $\lambda \geq 0$, $J \geq 0$ et $\mu \geq 0$ sont les paramètres de régularisation, L_* l'une des matrices encodant la régularisation spatiale et R^0 (resp. R^1) l'une des régularisations parcimonieuses étudiées sous forme stricte (resp. relachée).

Dans le cas d'une décomposition sur un dictionnaire de Gabor, la grande cohérence de ce dernier n'assure pas l'équivalence entre les problèmes ℓ_0 et ℓ_1.
Lorsque la puissance du bruit n'est pas trop élevée et la parcimonie des signaux suffisamment forte, ces deux approches présentent des performances similaires [219, 218]. Le critère ERC (voir section 3.2.2) permet d'ailleurs d'assurer des décompositions exactes avec ces deux approches. Les méthodes gloutonnes permettant l'approximation de la solution du problème ℓ_0 étant en général plus rapide que les méthodes résolvant le problème ℓ_1, elles sont préférées dans ce cas. Au contraire, lorsque le bruit est fort et/ou les signaux peu parcimonieux, les approches ℓ_1 offrent des garanties de stabilité pour les décompositions les rendant plus intéressantes [56]. C'est pourquoi, nous considérons ici ces deux approches en vue de les comparer pour le traitement des signaux EEG.

5.2 Stratégies d'optimisation

Afin de réaliser ces décompositions, nous nous attachons maintenant à la mise en place d'algorithmes résolvant les problèmes d'optimisation associés sous les deux formes étudiées.

5.2.1 Optimisation convexe pour le formulation ℓ_1

La formulation relâchée du problème est strictement convexe et différentes approches d'optimisation classiques peuvent être considérées pour sa résolution. Une approche proximale a été choisie pour la minimisation de notre fonction de coût. Celle-ci présente une convergence quadratique permettant la résolution de notre problème dans des temps raisonnables et ne nécessite pas de réglage de paramètres.

5.2.1.1 Approches proximales

Les méthodes proximales [48, 47, 11] permettent la minimisation de fonctions de coût composées de la somme de deux fonctions convexes présentant des régularités particulières. Ces problèmes sont écrits pour un problème matriciel comme suit :

$$\hat{X} = \arg\min_X \quad f_1(X) + f_2(X)$$

avec f_1 une fonction convexe differentiable possédant un gradient S-Lipschitzien [1] et f_2 une fonction convexe. Ce type de problème apparaît pour la réalisation de nombreuses tâches en traitement du signal et les approches proximales sont fréquemment considérées lorsque la fonction f_2 est non-différentiable comme dans le cas qui nous occupe. Nous ne rentrerons pas ici dans les détails de ces approches que le lecteur intéressé pourra trouver dans les références mentionnées plus haut, mais le schéma d'optimisation général est présenté de manière à comprendre son utilisation pour notre problème.
Les méthodes proximales se rapprochent itérativement de la solution du problème traité à partir de la linéarisation de la fonction de coût autour du point courant. Ainsi, à chaque itération i le problème suivant est résolu :

$$\hat{X} = \arg\min_X f_1(X^i) + \langle \nabla f_1(X^i), X - X^i \rangle + f_2(X) + \frac{S}{2}\|X - X^i\|_F^2$$

avec X^i le point courant et S le coefficient de Lipschitz du gradient de f_1. Le terme quadratique permet de limiter la recherche du point suivant à une distance bornée (par le coefficient de Lipschitz) du point courant. Ce problème est équivalent à la minimisation suivante :

$$\hat{X} = \arg\min_X \frac{1}{2}\|X - (X^i - \frac{1}{S}\nabla f_1(X^i))\|_F^2 + \frac{1}{S} f_2(X).$$

Le terme $X^i - \frac{1}{S}\nabla f_1(X^i)$ est le résultat d'une descente de gradient au point courant par rapport à f_1 avec pour pas de descente $\frac{1}{S}$. La minimisation précédente réalise donc une projection du résultat de cette descente sur la contrainte représentée par la fonction f_2. Lorsque cette contrainte est nulle, c'est une simple descente de gradient qui est réalisée tandis que pour une fonction indicatrice d'un ensemble, cette approche correspond à celle du gradient projeté.

1. C'est à dire : $\forall (X^1, X^2), \quad |\nabla f_1(X^1) - \nabla f_1(X^2)| \leq S\|X^1 - X^2\|$.

La caractérisation de cette projection dépend de l'opérateur proximal défini pour une fonction f_2 comme suit :

$$Prox_{f_2}(Z) = \arg\min_X \quad \frac{1}{2}\|X - Z\|_F^2 + f_2(X) \ ,$$

permettant d'écrire la minimisation de l'approche proximale :

$$Prox_{\frac{1}{S}f_2}\left(X^i - \frac{1}{S}\nabla f_1(X^i)\right) \ .$$

Lorsque la fonction f_2 est une norme pondérée par un coefficient $f_2 = c\Omega$ alors cet opérateur proximal peut être exprimé en fonction de l'opérateur de projection $(Proj)$ sur la boule de rayon c de la norme dual associée Ω^* [11] :

$$Prox_{c\Omega} = Id - Proj_{\Omega^* \leq c},$$

avec Id l'opérateur identité.
Cette norme duale Ω^* est définie par :

$$\Omega^*(X) = \max_{\Omega(Z)\leq 1} \langle Z, X\rangle$$

Le dual de la norme duale est la norme initiale. Parmi les couples classiques de normes duales, il est bien connu que la norme ℓ_∞ est couplée avec la norme ℓ_1, que la norme ℓ_2 est son propre dual et qu'une norme ℓ_p est le dual de la norme ℓ_q si $\frac{1}{p} + \frac{1}{q} = 1$.

Les approches proximales comportent généralement les étapes suivantes à l'itération i :

$$X^i = Prox_{\frac{1}{S}f_2}(X^{i-\frac{1}{2}} - \frac{1}{S}\nabla f_1(X^{i-\frac{1}{2}}))$$
$$X^{i+\frac{1}{2}} = h(X^1, \ldots, X^i) \ .$$

La fonction h permet à partir des itérations précédentes de calculer une valeur intermédiaire permettant parfois d'augmenter la vitesse de convergence.
L'algorithme de base parfois appelé ISTA (Iterative Soft Thresholding Algorithm) n'utilise pas de valeurs intermédiaires : $h(X^i, \ldots, X^1) = X^i$. Pour le problème mentionné plus haut, l'algorithme FISTA (Fast ISTA) a été choisi. Ce dernier présenté dans [18] est basé sur les travaux de Nesterov [171] et présente une convergence quadratique en $O(\frac{1}{i^2})$. À l'itération i il suit les étapes présentées plus haut avec :

$$t^{i+1} = \frac{1 + \sqrt{1 + 4(t^i)^2}}{2}$$
$$h(X^i, \ldots, X^1) = X^i + \frac{t^i - 1}{t^{i+1}}(X^i - X^{i-1}).$$

La valeur du coefficient de Lipschitz S du gradient de f_1 peut être parfois calculée analytiquement. Dans le cas contraire, un majorant local de ce coefficient est calculable via une recherche linéaire à chaque itération. Cette approche peut également accélérer la convergence de l'algorithme lorsque S est connu. En effet, la valeur de S trouvée analytiquement

est liée à la plus grande variation du gradient de f_1 dans l'espace de recherche. Or, localement, les variations de celui-ci peuvent être bien plus faibles et autoriser ainsi de plus grands pas de descente.
Dans [18], Beck et al. propose de calculer un majorant de ce coefficient en le faisant évoluer selon une progression géométrique jusqu'à ce qu'il respecte un critère évalué localement. Plus précisément, pour une évolution géométrique caractérisée par η, à l'itération i le coefficient de Lipschitz est calculé par $S_i = \eta^{k_i} S_{i-1}$ avec k_i l'indice le plus petit permettant de respecter le critère suivant :

$$\bar{S} = \eta^k S_{i-1}, \quad Z = Prox_{\frac{1}{\bar{S}} f_2}(X^{i-\frac{1}{2}} - \frac{1}{\bar{S}} \nabla f_1(X^{i-\frac{1}{2}})),$$
$$f_1(Z) + f_2(Z) \leq f_1(X^{i-\frac{1}{2}}) + \langle \nabla f_1(X^{i-\frac{1}{2}}), Z - X^{i-\frac{1}{2}} \rangle + f_2(Z) + \frac{\bar{S}}{2} \|Z - X^{i-\frac{1}{2}}\|_2^F.$$

5.2.1.2 Application à notre problème et choix d'implémentation

Dans le contexte de la décomposition qui nous intéresse, l'algorithme FISTA peut être appliqué en prenant $f_1 : X \to \|Y - \Phi X\|_F^2 + \mu \|X L_*\|_F^2$ et f_2 l'une des régularisations parcimonieuses étudiées.
Concernant la première régularisation $R_{2,1}$, l'opérateur proximal est bien connu. Ainsi, pour chaque ligne $\mathbf{x} = X(i,.)$ de la matrice de décomposition l'opérateur proximal de $R_{2,1}$ s'écrit :

$$Prox_{\lambda R_{2,1}}(\mathbf{x}) = \mathbf{x} \max\left(0,\ 1 - \frac{\lambda}{\|\mathbf{x}\|_2}\right).$$

Concernant la régularisation par groupe avec variables latentes, cet opérateur est plus complexe à calculer. Nous avons choisi d'utiliser la même reformulation que celle proposée dans [175]. Cette dernière consiste à transformer le problème de manière à obtenir un problème équivalent régularisé également pour obtenir une parcimonie par groupe mais avec des groupes disjoints. Cette transformation n'est tout de même pas gratuite car la dimension de l'espace d'optimisation est alors plus grande et donc la résolution plus coûteuse en temps de calcul.
Dans le cas du problème étudié dans [175] qui considère le modèle $\mathbf{y} = X\mathbf{w}$ avec une régularisation par groupe sur $\mathbf{w} \in \mathbb{R}^K$, cette reformulation implique une duplication des colonnes de X effectuée comme suit :

$$\bar{X} = [X(g_1), \ldots, X(g_N)],$$

pour N groupes avec [.] un opérateur de concaténation. Cette transformation peut être écrite $\bar{X} = XP$ pour $P \in \mathbb{R}^{K \times \sum_{i=1}^{N} \#\{g_i\}}$ une matrice définie par :

$$P = [P(g_1), \ldots, P(g_N)],$$
$$\text{avec pour le groupe j :} \quad \forall k \in \{1, \ldots, \#\{g_j\}\} \quad P_{g_j}(g_j(k), k) = 1.$$

Pour notre problème, lorsque la régularisation $R_{2,1}^{lg}(X)$ est considérée, cette reformulation aboutit à :

$$\hat{X} = \underset{X \in \mathbb{R}^{N_\Phi \times C}}{\arg\min} \|Y - \Phi X P^T\|_F^2 + \mu \|X P^T L_*\|_F^2 + \lambda \sum_{j}^{N_\Phi} \sum_{c=1}^{C} \|X(j, \tilde{g}_c)\|_2,$$

avec cette fois-ci des groupes $\{\tilde{g}_j, j \in \{1,\dots,C\}\}$ disjoints :

$$\tilde{g}_1 = \{1,\dots,\#\{g_1\}\},$$
$$\forall c \in \{2,\dots,C\},\ \ \tilde{g}_c = \left\{\sum_{k=1}^{c-1}\#\{g_k\}+1,\dots,\sum_{k=1}^{c}\#\{g_k\}\right\}\ .$$

L'augmentation de la dimension de X entraîne comme pour la régression étudiée dans [175] un surcoût de calcul pour notre minimisation mais permet la résolution du problème à partir du même opérateur proximal que celui utilisé pour une régularisation par groupe considérant des groupes disjoints. Cet opérateur est similaire à celui utilisé pour la norme $\ell_{2,1}$ [2], l'opération réalisée sur les lignes étant effectuée sur les groupes de chaque ligne. Ainsi, pour chacune des lignes et chacun des groupes $\mathbf{x} = X(j,\tilde{g}_c)$ de cette matrice, l'opérateur proximal s'écrit :

$$Prox_{\lambda R_{2,1}}(\mathbf{x}) = \mathbf{x}\max\left(0,\ 1-\frac{\lambda}{\|\mathbf{x}\|_2}\right)\ .$$

En ce qui concerne le coefficient de Lipshitz S, nous avons fait le choix d'en calculer localement un majorant à chaque itération via la méthode proposée dans [18] afin d'accélérer la convergence et d'effectuer les descentes de gradient les plus grandes possibles à chaque itération.
L'algorithme complet est résumé dans la table 5.1.
Les gradients des parties différentiables des fonctions de coût sont aisément calculables. Pour $f_1(X) = \|Y-\Phi X\|_F^2 + \mu\|XL_*\|_F^2$:

$$\nabla_X(f_1)(X) = -2\Phi^T(Y-\Phi X) + \mu X L_* {L_*}^T,$$

alors que pour $f_1(X) = \|Y-\Phi XP^T\|_F^2 + \mu\|XP^TL_*\|_F^2$:

$$\nabla_X(f_1)(X) = -2\Phi^T(Y-\Phi XP^T)P + \mu XP^TL_*{L_*}^TP.$$

5.2.2 Approches gloutonnes pour la formulation ℓ_0

Comme nous l'avons vu dans le chapitre 3, la forme stricte du problème de décomposition parcimonieuse est NP-difficile. Il n'est donc pas envisageable de vouloir le résoudre de façon exacte pour des problèmes de moyenne/grande dimension. Par conséquent, nous allons développer ici des algorithmes gloutons permettant d'approcher la solution du problème (5.1) pour les différentes régularisations parcimonieuses étudiées.
Les algorithmes que nous considérons suivent des étapes similaires à celles de l'Orthogonal Matching Pursuit (OMP), *i.e.* chaque itération est composée d'une étape de recherche du meilleur atome pour la représentation du résidu courant et d'une étape de mise à jour du résidu par soustraction de la contribution des atomes choisis jusque-là au signal décomposé. Pour l'itération i nous noterons ici le résidu U^i (et non pas R afin d'éviter toute confusion avec les régularisations).

2. La régularisation $\ell_{2,1}$ est en fait un cas particulier de la régularisation par groupe lorsqu'un seul groupe contenant l'ensemble des atomes est considéré.

Paramètres du modèle : λ, μ.
Paramètres de la méthode : $\eta \geq 1$, $L^0 \geq 0$.

procedure FISTA(Y, Φ, L_*)
  Initialiser X^0, $X^{\frac{1}{2}} = X^0$, $t^0 = 1$ et $i = 1$
  while $i \leq iterMax$ ou $\frac{\|X^i - X^{i-1}\|_2}{\|X^i\|_2} \leq \varepsilon$ **do**
    k_i = Trouver le plus petit k tel que :
      $\bar{S} = \eta^k S^{i-1}$, $Z = Prox_{\frac{1}{\bar{S}} f_2}(X^{i-\frac{1}{2}} - \frac{1}{\bar{S}} \nabla f_1(X^{i-\frac{1}{2}}))$
      $f_1(Z) + f_2(Z) \leq f_1(X^{i-\frac{1}{2}}) + \langle \nabla f_1(X^{i-\frac{1}{2}}), Z - X^{i-\frac{1}{2}} \rangle + f_2(Z) + \frac{\bar{S}}{2} \|Z - X^{i-\frac{1}{2}}\|_F^2$
    $S_i = \eta^{k_i} S_{i-1}$
    $X^i = Prox_{\frac{1}{S^i} f_2}(X^{i-\frac{1}{2}} - \frac{1}{S^i} \nabla f_1(X^{i-\frac{1}{2}}))$
    $t^{i+1} = \frac{1+\sqrt{1+4(t^i)^2}}{2}$
    $X^{i+\frac{1}{2}} = X^i + \frac{t^i - 1}{t^{i+1}}(X^i - X^{i-1})$
    $i = i + 1$
  end while
  return X^i
end procedure

TABLE 5.1 – FISTA - Algorithme détaillé

5.2.2.1 MP multidimensionnel régularisé spatialement (« MSCMP : Multidimensional Spatially Constrained MP »)

Dans un premier temps, nous considérons le cas de la régularisation $R_{2,0}$. Sans régularisation spatiale, le SOMP [220] (« Simultaneous OMP ») permet de réaliser une telle décomposition. Nous allons adapter ce dernier pour notre problème afin de prendre en compte la régularisation spatiale.

Principe

Le chapitre 3 a présenté les variantes du MMP considérant des régularisations spatiales. Celles-ci modifient généralement l'étape de choix des atomes de manière à sélectionner des atomes sur lesquels la projection du signal respecte certaines contraintes. Ces approches présentent l'inconvénient de considérer cette projection de manière non contrainte en utilisant le produit scalaire pour réaliser celle-ci. Nous proposons pour notre problème de réaliser le choix des atomes en prenant en compte une projection contrainte spatialement. De façon plus formelle, nous recherchons l'atome et les coefficients de décompositions résolvants pour le résidu U^i de l'itération courante i la minimisation suivante :

$$\hat{\mathbf{x}}, \hat{j} = \underset{\mathbf{x} \in \mathbb{R}^C, j}{\arg\min} \quad \|U^i - \Phi(j)\mathbf{x}^T\|_F^2 + \mu \|\mathbf{x}^T L_*\|_F^2$$

Pour un atome j fixé du dictionnaire, la projection s'écrit [3] :

$$\begin{aligned}\hat{\mathbf{x}} &= \arg\min_{\mathbf{x}\in\mathbb{R}^C} \|U^i - \Phi(j)\mathbf{x}^T\|_F^2 + \mu\|\mathbf{x}^T L_*\|_F^2,\\ \hat{\mathbf{x}} &= (I_C + L_* L_*^T)^{-1}\ (U^i)^T\Phi(j)\ \ ,\end{aligned}$$

avec I_C la matrice identité de taille C.
En notant $Q = (I_C + L_* L_*^T)^{-1}$, la sélection de l'atome est effectuée alors en résolvant :

$$\begin{aligned}&\arg\min_j \quad \|U^i - \Phi(j)\Phi(j)^T U^i Q^T\|_F^2 + \mu\|\Phi(j)^T U^i Q^T L_*\|_F^2,\\ =\ &\arg\min_j \quad -Tr\left(\Phi(j)^T (U^i) Q^T (U^i)^T \Phi(j)\right)\\ =\ &\arg\max_j \quad \langle\ \Phi(j)^T U^i Q^T,\ \ \Phi(j)^T U^i\ \rangle\\ =\ &\arg\max_j \quad \|\Phi(j)^T U^i\|_{Q^T}^2\end{aligned}$$

avec $\|.\|_{Q^T}$ la norme quadratique associée à la matrice Q^T.

Une fois la sélection de l'atome réalisée pour l'itération courante i, le résidu est recalculé en fonction de tous les atomes choisis jusque-là. Soit Δ^i l'ensemble des indices des atomes choisis à l'itération i, les coefficients de décomposition sont calculés par résolution du problème suivant :

$$\hat{X}(\Delta^i, .) = \arg\min_{X\in\mathbb{R}^{\#\{\Delta^i\}\times C}} \|Y - \Phi(\Delta^i)X\|_F^2 + \mu\|XL_*\|_F^2, \tag{5.3}$$

puis le résidu est calculé par : $U^{i+1} = Y - \Phi(\Delta^i)\hat{X}(\Delta^i, .)$.
Le problème défini dans l'equation (5.3) peut être résolu de manière exacte à chaque itération. Soit $h : X\ \to \|Y - \Phi(\Delta^i)X\|_F^2 + \mu\|XL_*\|_F^2$, nous avons :

$$\frac{d}{dX}h = -2\Phi(\Delta^i)^T(Y - \Phi(\Delta^i)X) + \mu X L_* L_*^T$$

et donc :

$$\begin{aligned}&\frac{d}{dX}h(\hat{X}(\Delta^i, .)) = 0\\ \Leftrightarrow\ &\Phi(\Delta^i)^T\Phi(\Delta^i)X(\Delta^i, .) + \mu X(\Delta^i, .)L_* L_*^T = \Phi(\Delta^i)^T Y\end{aligned}$$

Cette dernière expression est une équation de Sylvester de forme $A^i X + XB = C$ avec $A^i = \Phi(\Delta^i)^T\Phi(\Delta^i)$ et $B = \mu L_* L_*^T$ des matrices symétriques réelles. La résolution d'un tel problème est présentée dans l'annexe A. La proposition 1 montre notamment l'équivalence de celui-ci avec un système diagonal pouvant être résolu directement.

3. Solution classique d'un problème des moindres carrées régularisé à l'aide d'une régularisation de Tikhonov. Obtenue directement par dérivation.

Remarques et détails d'implémentation : L'étape de projection sur les atomes précédemment choisis étant contrainte spatialement, le résidu calculé pour l'étape suivante n'est pas orthogonal à l'espace engendré par ces atomes. Ainsi, il est possible qu'un atome soit choisi plusieurs fois, notamment lorsque le paramètre μ est grand. Pour éviter cela, notre implémentation ne considère que la liste des atomes encore non utilisés pour la sélection d'un nouvel atome.
Concernant la seconde étape, la matrice B ne dépendant pas de l'itération, il est possible de pré-calculer sa diagonalisation. Au contraire, pour la matrice A^i, la diagonalisation doit être réalisée à chaque itération. A^i étant une matrice de dimension i le coût de calcul grandit au fur et à mesure des itérations mais reste raisonnable étant donné la nature parcimonieuse de la décomposition effectuée (faible nombre d'itérations *i.e.* $J \ll T$).
En fonction de l'application considérée, différents critères d'arrêt de l'algorithme peuvent être envisagés. Il est par exemple possible d'arrêter celui-ci après qu'un nombre fixé d'atomes ait été choisi ou bien lorsque la norme du résidu devient suffisamment faible.
L'algorithme complet du MSCMP est résumé dans la table 5.2.

Paramètres du modèle : J, μ.
Paramètres d'arrêt : ε

procedure MSCMP(Y, Φ, L_*)
  Initialiser $X = 0$, $U^0 = Y$, $\Delta^0 = \{\}$ et $i = 1$
  Calculer $Q = (I_C + \mu L_* L_*^T)^{-1}$
  Diagonaliser $B = \mu L_* L_*^T = F D_B F^T$
  while $i \leq J$ ou $\|U^i\|_2^2 \leq \varepsilon$ **do**
    Choix d'un atome :
    $id^i = \max_j \ \langle\, \Phi(j)^T U^t Q^T, \ \ (U^i)^T \Phi(j) \rangle$
    $\Delta^i = \{\Delta^{i-1}, id^i\}$
    Calcul du résidu :
      Diagonaliser $A = \Phi(\Delta^i)^T \Phi(\Delta^i) = G D_A G^T$
      Construire O à partir de D_A et D_B (Eqn. A.5)
      $X^i = X^{i-1}$
      $(X^i)(\Delta^i, .)) = F(F^T \Phi(\Delta^i)^T Y G \oslash O) G^T$
      $U^i = Y - \Phi X^i$
    $i = i + 1$
  end while
  return X^{i-1}
end procedure

TABLE 5.2 – MSCMP - Algorithme détaillé

5.2.2.2 MP multidimensionnel structuré et régularisé spatialement (« MSCSMP : Multidimensional Spatially Constrained Structured MP »)

Nous nous intéressons maintenant à la régularisation du « Latent Group LASSO » noté $R_{2,0}^{lg}$.

Principe :
De la même manière que dans l'algorithme précédent, nous considérons un algorithme en deux étapes réalisant à chaque itération le choix d'un atome et d'un groupe de canaux avant de mettre à jour le résidu en fonction des choix réalisés précédemment.
La première étape s'inspire de celle définie plus haut et nécessite l'introduction de nouvelles notations. Ainsi, nous notons ici $\left\{\theta_j^i, \forall j \in \{1,\dots,N_\Phi\}\right\}$ les supports sélectionnés pour chaque atome à l'itération i et $U^i_{-\theta_j}$ le résidu à l'itération i dont la contribution des coefficients θ_j^i a été supprimée. $U^i_{-\theta_j}$ est ainsi obtenu par :

$$U^i_{-\theta_j^i} = U^i + \Phi(j)X^i(j,\theta_j^i).$$

L'objectif à chaque itération est de choisir un groupe de canaux et un atome de manière à réduire au maximum la norme du résidu tout en respectant la contrainte spatiale induite par la régularisation. Nous proposons de réaliser cette étape en résolvant le problème suivant :

$$\begin{aligned}&\hat{\mathbf{x}}, \hat{j}, \hat{c} = \underset{\mathbf{x}\in\mathbb{R}^C, j, c}{\arg\min} \quad \|U^i_{-\theta_j^i}(\bar{\theta}^i_{jc}) - \Phi(j)\mathbf{x}^T\|_F^2 + \mu(\|\mathbf{x}^T L_*\|_F^2 - RU_j^i),\\ &\text{avec} \quad \bar{\theta}^i_{jc} = \theta_j^i \cup g_c \quad \text{et} \quad RU_j^i = \|(X^i)^T(j)L_*\|_2^2.\end{aligned}$$

Pour chaque atome j et chaque groupe c, $\bar{\theta}^i_{jc}$ représente le support de décomposition obtenu si ce choix est réalisé tandis que RU_j^i correspond à la valeur de la contrainte spatiale obtenue à l'itération précédente pour l'atome j. Ainsi, la minimisation précédente permet d'effectuer le choix d'un atome et d'un groupe de manière à minimiser le résidu en améliorant le lissage des coefficients de décomposition.
Comme précédemment pour j et c fixés, il est possible de calculer le vecteur $\hat{\mathbf{x}}$ comme suit :

$$\begin{aligned}&\hat{\mathbf{x}} = \underset{\mathbf{x}\in\mathbb{R}^C}{\arg\min} \|U^i_{-\theta_j^i}(\bar{\theta}^i_{jc}) - \Phi(j)\mathbf{x}^T\|_F^2 + \mu\|\mathbf{x}^T L_*^T(\bar{\theta}^i_{jc})\|_F^2,\\ \Leftrightarrow\ &\hat{\mathbf{x}} = \left(I_{\#\{\bar{\theta}^i_{jc}\}} + L_*^T(\bar{\theta}^i_{jc})L_*^T(\bar{\theta}^i_{jc})^T\right)^{-1} (U^i(\bar{\theta}^i_{jc}))^T\Phi(j).\end{aligned}$$

Le choix de l'atome et du groupe est donc fait simplement en calculant pour tous les couples (j,c) la valeur du coût présenté plus haut de façon à choisir le couple lui donnant une valeur minimale :

$$\begin{aligned}&\hat{j}, \hat{c} = \underset{j,c}{\arg\min} \quad \|U^i_{-\theta_j^i}(\bar{\theta}^i_{jc}) - \Phi(j)(\bar{\mathbf{x}}^i_{jc})^T\|_2^2 + \mu(\|(\bar{\mathbf{x}}^i_{jc})^T L_*\|_2^2 - RU_j^i),\\ &\text{avec} \quad \bar{\mathbf{x}}^i_{jc} = \left(I_{\#\{\bar{\theta}^i_{jc}\}} + L_*^T(\bar{\theta}^i_{jc})L_*^T(\bar{\theta}^i_{jc})^T\right)^{-1} (U^i(\bar{\theta}^i_{jc}))^T\Phi(j).\end{aligned}$$

Intéressons-nous maintenant à la mise à jour du résidu. La résolution exacte directe présentée dans la section précédente n'est plus applicable ici et nous proposons donc d'effectuer cette étape à l'aide de l'algorithme FISTA que nous avons déjà présenté. Le problème à résoudre peut être écrit :

$$\hat{X} = \underset{X\in\mathbb{R}^{N_\Phi\times C}}{\arg\min} \quad \|Y - \Phi X\|_F^2 + \mu\|XL_*\|_F^2 \quad \text{t.q.} \quad \forall j \in \{1,\dots,N_\Phi\}, \quad X(j,(\theta_j^i)_{comp}) = \mathbf{0}$$

avec $\{1,\dots,C\} = \theta_j^i \cup (\theta_j^i)_{comp}$. C'est un problème de décomposition sur un dictionnaire redondant régularisée spatialement dont la structure parcimonieuse des coefficients est fixée strictement. Pour se replacer dans le cadre du coût à minimiser avec l'algorithme FISTA la contrainte stricte doit être transformée en un terme de pénalité. Pour cela, il suffit de l'écrire à l'aide d'une fonction indicatrice $\mathcal{X}_{\{\theta_1^i,\dots,\theta_{N_\Phi}^i\}}$. $\mathcal{X}_{\{\theta_1^i,\dots,\theta_{N_\Phi}^i\}}(X)$ est nulle lorsque le support de $X^T(j)$ est inclus dans θ_j^i pour tous les atomes j et vaut $+\infty$ dans le cas contraire.
L'opérateur proximal de la fonction $\mathcal{X}$ s'écrit alors simplement pour la ligne $\mathbf{x} = X^T(j)$ comme suit :

$$\forall c \in \{1,\dots,C\}$$
$$Prox_{\mathcal{X}}(\mathbf{x})(c) = \begin{cases} \mathbf{x}(c) & \text{si } c \in \theta_j^i, \\ 0 & \text{sinon.} \end{cases}$$

En pratique, moins le support comporte d'éléments, plus le nombre d'itérations permettant la convergence est faible. Ainsi, au fûr et à mesure des itérations de l'algorithme global (MSCSMP) le coût de cette mise à jour va augmenter. Pour des représentations suffisamment parcimonieuses, ce coût est négligeable comparé au coût de calcul de la recherche des meilleurs atomes et des meilleurs groupes.
L'algorithme complet du MSCSMP est résumé dans la table 5.3.

5.3 Évaluation expérimentale

Nous nous attachons maintenant à l'évaluation des algorithmes présentés plus haut. Dans un premier temps, nous considérons la décomposition de signaux synthétiques afin de mesurer la capacité des méthodes à retrouver les structures parcimonieuses sous-jacentes des signaux lorsque ceux-ci respectent les caractéristiques supposées des composantes EEG. Nous examinons ensuite l'intérêt de ces algorithmes pour la décomposition temps-fréquence de signaux réels EEG dans un contexte discriminatif.

5.3.1 Récupération de la structure sous-jacente de signaux synthétiques

Afin d'évaluer la capacité des algorithmes proposés à extraire les différentes composantes d'un signal lorsque celles-ci respectent nos hypothèses, nous allons construire un ensemble de signaux synthétiques à partir de telles composantes, puis décomposer des versions bruitées de ceux-ci avec les algorithmes présentés plus haut.

5.3.1.1 Création des signaux

Pour cette expérience, nous construisons un ensemble de K signaux $\{Y^k, k \in \{1,\dots,K\}\}$ à partir d'un dictionnaire temporel fixé Φ. Chaque signal Y^k est synthétisé via l'équation suivante $Y^k = \Phi X^k$ après construction d'une matrice de décomposition X^k respectant les *a priori* de notre modèle, *i.e.* construite à partir de quelques composantes régulières et localisées spatialement. Plus précisément, pour chaque signal, N_a atomes sont choisis de manière

Paramètres du modèle : J, μ.
Paramètres d'arrêt : ε

procedure MSCSMP(Y, Φ, L_*)
  Initialiser $X = 0$, $U^0 = Y$ et $i = 1$
  Initialiser $\theta_j^0 = \{\}$ pour $j \in \{1, \dots, N_\Phi\}$
  Calculer $Q = (I_C + \mu L_* L_*^T)^{-1}$
  while $i \leq J$ ou $\|U^i\|_2^2 \leq \varepsilon$ **do**
    Choix d'un atome et d'un groupe :
      $\forall j \in \{1, \dots, N_\Phi\}$ calculer :
        $RU_j^i = \|(X^i)^T(j) L_*\|_2^2,\ U^i_{-\theta_j^i} = U^i + \Phi(j) X^i(j, \theta_j^i)$
      $\forall j \in \{1, \dots, N_\Phi\},\ \forall c \in \{1, \dots, C\}$ calculer :
        $\bar{\theta}^i_{jc} = \theta_j^i \cup g_c,\ \bar{\mathbf{x}}^i_{jc} = \left(I_{\#\{\bar{\theta}^i_{jc}\}} + L_*^T(\bar{\theta}^i_{jc}) L_*^T(\bar{\theta}^i_{jc})^T\right)^{-1} \ (U^i(\bar{\theta}^i_{jc}))^T \Phi(j)$
      Résoudre :
      $\hat{j}, \hat{c} = \arg\min_{j,c} \ \|U^i_{-\theta_j^i}(\bar{\theta}^i_{jc}) - \Phi(j)(\bar{\mathbf{x}}^i_{jc})^T\|_2^2 + \mu(\|(\bar{\mathbf{x}}^i_{jc})^T L_*\|_2^2 - RU_j^i)$
    Calcul du résidu :
      Résoudre à l'aide FISTA :
        $X^{i+1} = \arg\min_{X \in \mathbb{R}^{N_\Phi \times C}} \ \|Y - \Phi X\|_2^2 + \mu \|X L_*\|_2^2$
                t.q. $\forall j \in \{1, \dots, N_\Phi\},\ X(j, (\theta_j^i)^C) = \mathbf{0}$
      $U^{i+1} = Y - \Phi X^{i+1}$
    $i = i + 1$
  end while
  return X^i
end procedure

TABLE 5.3 – MSCSMP - Algorithme détaillé

aléatoire dans le dictionnaire et des vecteurs de coefficients leur sont associés. Ces vecteurs de coefficients sont sélectionnés dans un dictionnaire spatial Φ_s créé à partir de la résolution du problème direct vu dans le chapitre 4 afin d'obtenir des signaux réalistes.
La construction d'un signal est donc résumée par :

$$Y^k = \sum_{i=1}^{N_a} \mathbf{a}^k(i)\Phi(j_i^k)\Phi_s^T(l_i^k),$$

avec $\{j_i^k, i \in \{1,\dots,N_a\}\}$ les indices des atomes temporels choisis pour le signal k, $\{l_i^k, i \in \{1,\dots,N_a\}\}$ les indices des atomes spatiaux associés. et $\mathbf{a^k}$ le vecteur des coefficients de pondération.
Une fois les signaux créés, ils sont bruités, avant d'être décomposés sur le dictionnaire avec les algorithmes correspondants aux régularisations étudiées. Afin de se rapprocher du bruit présent dans les signaux EEG, le bruit choisi pour chaque canal est un bruit blanc filtré à l'aide d'un filtre passe bas de manière à obtenir une distribution des fréquences en $\frac{1}{f^\alpha}, \alpha \in \mathbb{N}$. Aucune structure spatiale n'est, par contre, supposée pour ce bruit. Nous notons les signaux bruités $\{\tilde{Y}^k, k \in \{1,\dots,K\}\}$.

5.3.1.2 Experience 1

Nous souhaitons évaluer dans un premier temps la capacité des différentes approches à retrouver la structure parcimonieuse sous-jacente des signaux lorsque tous les canaux sont utilisés pour le choix des atomes, l'avantage obtenu en ne conservant pour ce choix que peu de groupes de canaux sera discuté dans une deuxième expérience. Nous comparons pour cela l'ensemble des régularisations étudiées aussi bien pour le problème relâché où elles sont notées $R_{2,1}, R_{2,1} + R^8_{w,lap}, R^{lg}_{2,1}, R^{lg}_{2,1} + R^8_{w,lap}$ que pour le problème strict où nous les notons $R_{2,0}, R_{2,0} + R^8_{w,lap}, R^{lg}_{2,0}, R^{lg}_{2,0} + R^8_{w,lap}$. Ces tests sont effectués pour différents niveaux de bruit (RSB) et différentes valeurs du nombre de composantes utilisées pour la création des signaux.

Protocole de test et paramètres. Afin de pouvoir évaluer de la même manière les décompositions obtenues avec les approches gloutonnes et celles obtenues par résolution du problème convexe, nous sélectionnons pour chacune des approches N_a atomes (lignes) et projetons le signal sur ces atomes en conservant la contrainte spatiale (même paramètre de régularisation spatiale que pour la décomposition). Ainsi, pour chaque régularisation et chaque signal, la décomposition est réalisée via une des approches présentées plus haut, puis les signaux sont projetés sur les N_a principaux atomes à l'aide de l'algorithme FISTA avec une structure de zéro fixée, *i.e.* de manière similaire à l'algorithme du gradient projeté. Pour les régularisations convexes, cette projection présente également l'avantage d'éviter le biais inhérent aux approches ℓ_1 dû à la valeur du terme de la régularisation parcimonieuse.
En ce qui concerne la sélection des atomes avant projection, les algorithmes gloutons sont simplement stoppés lorsque N_a atomes différents ont été choisis tandis que pour les approches convexes, les atomes sélectionnés sont ceux correspondant aux N_a lignes de la matrice de décomposition possédant les plus grandes normes. Pour les problèmes relâchés, le

choix du paramètre de régularisation parcimonieuse λ est crucial. Après avoir observé que le paramètre optimal dépendait fortement du signal considéré, nous avons choisi de prendre la valeur permettant d'approximer le signal avec un taux de reconstruction donné : $\frac{\|Y-\Phi X\|_F}{\|Y\|_F}$, la valeur de ce taux étant déterminée sur un ensemble d'entraînement pour chaque valeur de N_a et chaque valeur du RSB au regard des critères définis dans le paragraphe suivant. De la même façon, pour toutes les régularisations, le paramètre de régularisation spatiale optimal μ est également déterminé sur un ensemble d'entraînement.

L'expérience est réalisée pour deux dictionnaires différents, le premier Φ^1 de faible cohérence $\big(\mu_1(\Phi^1) \leq 0.2\big)$ est généré par des atomes dont les coefficients sont issus d'une distribution gaussienne $\mathcal{N}(0, I)$ et le second Φ^2 est un dictionnaire de Gabor de cohérence plus élevée $\big(\mu_1(\Phi^2) \geq 0.9\big)$.
En ce qui concerne les autres paramètres, les choix effectués sont résumés dans la figure 5.3.

Modèle		Composantes
$C = 64$	$T = 64$	$\mathbf{a} \sim \mathcal{N}(\mathbf{0}, I)$
$N_{\Phi^1} = 78$	$N_{\Phi^2} = 83$	$j \sim \mathcal{U}(1, N_\Phi)$
$K = 100$		$l \sim \mathcal{U}(1, N_{\Phi_s})$

FIGURE 5.3 – Paramètres de l'expérience permettant l'évaluation des régularisations proposées dans la récupération de structures sous-jacentes de signaux artificiels.

Critères Une évaluation simple des décompositions obtenues $\{\hat{X}^k, \{1, \ldots, K\}\}$ peut être effectuée via le critère suivant :

$$\varepsilon_1(\hat{X}^k, X^k) = \frac{\|X^k - \hat{X}^k\|_F}{\|X^k\|_F} \ .$$

Ce critère prend en compte à la fois le choix des atomes temporels et celui des coefficients de décomposition des différents canaux. Il semble donc pertinent pour mesurer le bon recouvrement des composantes. Un problème apparaît néanmoins pour un dictionnaire temps-fréquence de Gabor Φ^2 possédant une cohérence forte. En effet, pour une décomposition temps-fréquence des signaux EEG, nous souhaitons pouvoir retrouver les activités cérébrales constituantes d'un signal et donc pas nécessairement les indices exacts des atomes. Or, avec le critère précédent, le choix de mauvais atomes est très pénalisant et la cohérence forte du dictionnaire favorise ces erreurs.
Ainsi, nous introduisons un autre critère, nous nous fixons comme objectif la récupération des composantes spatio-temporelles présentes dans les signaux. Soit Y^k un signal construit

comme suit :

$$Y^k = \sum_{i=1}^{N_a} \mathbf{a}^k(i)\Phi(j_i^k)\Phi_s^T(l_i^k) = \sum_{i=1}^{N_a} A_i^k,$$

avec $\{A_i^k, i \in \{1,\dots,N_a\}\}$ les composantes spatio-temporelles du signal, et soit $\hat{Y}^k = D\hat{X}^k = \sum_{i=1}^{N_a} \hat{A}_i^k$ le signal reconstruit après avoir obtenu une matrice de décomposition $\hat{X}^k$ avec l'un des algorithmes. Nous souhaitons savoir si les composantes de Y^k ont été retrouvées durant la décomposition. Le critère est alors calculé itérativement en effectuant les étapes suivantes à chaque itération :

1. choix des deux composantes les plus proches entre les deux ensembles $\{A_i^k, i \in \{1,\dots,N_a\}\}$ et $\{\hat{A}_i^k, i \in \{1,\dots,N_a\}\}$:

$$\hat{i},\hat{j} = \arg\max_{i,j} \frac{\langle \hat{A}_i^k, \hat{A}_j^k \rangle}{\|\hat{A}_i^k\|_F \|\hat{A}_j^k\|_F} \ ,$$

2. mise à jour du critère :

$$\varepsilon_2(\hat{X}^k, X^k) = \varepsilon_2(\hat{X}^k, X^k) + \frac{\langle \hat{A}_{\hat{i}}^k, \hat{A}_{\hat{j}}^k \rangle}{\|\hat{A}_{\hat{i}}^k\|_F \|\hat{A}_{\hat{j}}^k\|_F} \ ,$$

3. suppression des composantes $\hat{i}$ et $\hat{j}$ des ensembles.

Le critère est initialisé à 0 à la première itération. Les distances entre les composantes sont calculées indépendamment des normes de celles-ci, étant donné que l'on souhaite retrouver toutes les composantes et pas seulement celles ayant les plus grandes amplitudes (contrairement au critère précédent). Plus la valeur du critère est élevée, meilleure est la performance de la méthode pour le recouvrement des composantes spatio-temporelles.
Afin d'obtenir des résultats comparables avec les deux dictionnaires, nous considérons ce critère dans les deux cas, bien que le premier critère soit suffisant pour le dictionnaire ayant une faible cohérence.

Résultats et discussion. Les résultats obtenus avec le dictionnaire Φ^1 sont présentés dans la figure 5.4 et les résultats obtenus avec le dictionnaire Φ^2 dans la figure 5.5. Ces résultats doivent être lus en prenant en compte que le critère doit être maximisé.

Plusieurs observations peuvent être effectuées à partir de ces résultats :

— Comme attendu, le recouvrement des composantes des signaux synthétiques devient plus difficile lorsque le nombre de composantes augmente ou bien lorsque le rapport signal à bruit diminue. De même, la difficulté de ce problème augmente avec la cohérence du dictionnaire.

— Concernant la régularisation de lissage, il apparaît pour l'ensemble des régularisations parcimonieuses qu'elle permet d'améliorer de façon significative les performances des méthodes dans la récupération des composantes sous-jacentes des signaux. Cette régularisation permet d'une part d'obtenir des coefficients réguliers spatialement et d'autre part de mieux choisir les atomes temporels.

	5 db	**0 db**	**-5 db**	**-10 db**	**Moy**
$R_{2,0}$	8.8004	8.3033	7.0684	5.5244	7.4241
$R_{2,0} + R^8_{w,lap}$	8.9898	8.5589	7.6531	7.0669	8.0671
$R^{lg}_{2,0}$	8.9655	8.3596	7.3155	5.8497	7.6226
$R^{lg}_{2,0} + R^8_{w,lap}$	9.037	**8.6719**	7.7753	6.8248	8.0772
$R_{2,1}$	8.8284	8.2376	6.7296	5.0699	7.2164
$R_{2,1} + R^8_{w,lap}$	9.0226	8.6258	7.9329	7.2263	8.2019
$R^{lg}_{2,1}$	8.8522	8.25	6.6576	4.4584	7.0545
$R^{lg}_{2,1} + R^8_{w,lap}$	**9.0563**	8.6256	**7.9344**	**7.2701**	**8.2216**

(a) $N_a = 10$

	5 db	**0 db**	**-5 db**	**-10 db**	**Moy**
$R_{2,0}$	13.0378	11.7326	9.8353	7.4028	10.5021
$R_{2,0} + R^8_{w,lap}$	13.3851	12.4479	11.1553	9.7286	11.6792
$R^{lg}_{2,0}$	13.1216	12.1204	10.3478	7.9381	10.882
$R^{lg}_{2,0} + R^8_{w,lap}$	**13.4411**	12.4537	11.2302	9.7486	11.7184
$R_{2,1}$	12.9342	11.3308	9.1757	6.8515	10.0731
$R_{2,1} + R^8_{w,lap}$	13.3027	**12.5295**	**11.3786**	**10.1201**	**11.8327**
$R^{lg}_{2,1}$	12.8348	10.9687	9.0614	6.6095	9.8686
$R^{lg}_{2,1} + R^8_{w,lap}$	13.2787	12.4569	11.3591	10.0088	11.7759

(b) $N_a = 15$

	5 db	**0 db**	**-5 db**	**-10 db**	**Moy**
$R_{2,0}$	16.9679	15.3123	12.2269	9.267	13.4435
$R_{2,0} + R^8_{w,lap}$	17.4656	16.3023	14.8295	12.9449	15.3856
$R^{lg}_{2,0}$	17.2083	15.3002	13.0916	9.9701	13.8925
$R^{lg}_{2,0} + R^8_{w,lap}$	17.5259	16.3474	14.5898	12.7756	15.3097
$R_{2,1}$	17.0372	15.0722	12.0869	8.552	13.1871
$R_{2,1} + R^8_{w,lap}$	**17.5487**	**16.553**	**15.1224**	**13.3204**	**15.6361**
$R^{lg}_{2,1}$	16.8843	14.4582	11.5858	8.1745	12.7757
$R^{lg}_{2,1} + R^8_{w,lap}$	17.2846	16.3988	14.9863	13.2749	15.4862

(c) $N_a = 20$

	5 db	**0 db**	**-5 db**	**-10 db**	**Moy**
$R_{2,0}$	21.1372	18.4037	14.7511	10.8672	16.2898
$R_{2,0} + R^8_{w,lap}$	21.6991	20.1001	18.2767	15.405	18.8702
$R^{lg}_{2,0}$	21.1769	18.8906	15.714	11.8134	16.8987
$R^{lg}_{2,0} + R^8_{w,lap}$	**21.7774**	**20.3269**	17.733	15.3518	18.7973
$R_{2,1}$	20.7814	17.8498	14.2223	10.3394	15.7982
$R_{2,1} + R^8_{w,lap}$	21.4595	20.077	18.4737	**16.334**	**19.0861**
$R^{lg}_{2,1}$	20.4857	17.177	13.6767	9.7703	15.2774
$R^{lg}_{2,1} + R^8_{w,lap}$	21.2632	19.9699	**18.5233**	16.2601	19.0041

(d) $N_a = 25$

FIGURE 5.4 – Évaluation des méthodes de décomposition parcimonieuse régularisée spatialement dans la récupération de structures sous-jacentes de signaux artificiels lorsque le dictionnaire est peu cohérent.

	5 db	**0 db**	**-5 db**	**-10 db**	**Moy**
$R_{2,0}$	4.7128	4.2596	3.6987	2.887	3.8895
$R_{2,0} + R^8_{w,lap}$	5.2405	5.1324	4.825	4.2271	4.8563
$R^{lg}_{2,0}$	4.715	3.997	3.2234	2.5479	3.6208
$R^{lg}_{2,0} + R^8_{w,lap}$	5.5563	5.3614	4.99	4.2876	5.0488
$R_{2,1}$	5.1631	4.3962	3.5079	2.5877	3.9137
$R_{2,1} + R^8_{w,lap}$	5.7715	**5.4163**	4.8475	4.2652	5.0751
$R^{lg}_{2,1}$	4.9285	4.1731	3.3754	2.584	3.7652
$R^{lg}_{2,1} + R^8_{w,lap}$	**5.7891**	5.3584	**5.0607**	**4.414**	**5.1556**

(a) $N_a = 10$

	5 db	**0 db**	**-5 db**	**-10 db**	**Moy**
$R_{2,0}$	6.1433	5.5058	4.7642	3.5939	5.0018
$R_{2,0} + R^8_{w,lap}$	7.5571	7.3658	**7.0315**	6.0804	7.0087
$R^{lg}_{2,0}$	6.4062	5.3204	4.6679	3.692	5.0217
$R^{lg}_{2,0} + R^8_{w,lap}$	7.6261	**7.543**	6.9092	6.0149	7.0233
$R_{2,1}$	6.164	5.3158	4.8076	4.2671	5.1386
$R_{2,1} + R^8_{w,lap}$	**7.906**	7.3003	6.7346	**6.0869**	7.007
$R^{lg}_{2,1}$	5.9163	5.056	4.6041	4.0993	4.9189
$R^{lg}_{2,1} + R^8_{w,lap}$	7.8753	7.4081	6.9251	5.9321	**7.0351**

(b) $N_a = 15$

	5 db	**0 db**	**-5 db**	**-10 db**	**Moy**
$R_{2,0}$	8.5739	7.9054	6.6063	5.294	7.0949
$R_{2,0} + R^8_{w,lap}$	10.3151	10.0525	9.5595	8.441	9.592
$R^{lg}_{2,0}$	8.6276	7.4882	6.2181	5.3951	6.9322
$R^{lg}_{2,0} + R^8_{w,lap}$	10.3857	9.9143	9.5155	7.9629	9.4446
$R_{2,1}$	9.0074	7.7391	6.6459	5.9436	7.334
$R_{2,1} + R^8_{w,lap}$	10.502	10.2103	9.584	**8.8548**	**9.7878**
$R^{lg}_{2,1}$	8.6464	7.7966	6.4397	5.9069	7.1974
$R^{lg}_{2,1} + R^8_{w,lap}$	**10.6216**	**10.2163**	**9.6432**	8.6257	9.7767

(c) $N_a = 20$

	5 db	**0 db**	**-5 db**	**-10 db**	**Moy**
$R_{2,0}$	11.1454	10.1695	8.5902	6.4789	9.096
$R_{2,0} + R^8_{w,lap}$	12.9444	12.5698	12.0784	10.6738	12.0666
$R^{lg}_{2,0}$	11.247	9.978	8.6268	6.8985	9.1876
$R^{lg}_{2,0} + R^8_{w,lap}$	13.0369	12.3452	11.7559	10.5474	11.9213
$R_{2,1}$	11.0439	9.7483	8.363	6.3432	8.8746
$R_{2,1} + R^8_{w,lap}$	**13.2904**	**12.8304**	12.0534	**11.3513**	**12.3814**
$R^{lg}_{2,1}$	11.3554	9.6791	8.5546	7.0292	9.1546
$R^{lg}_{2,1} + R^8_{w,lap}$	13.0736	12.7401	**12.1007**	11.1244	12.2597

(d) $N_a = 25$

FIGURE 5.5 – Évaluation des méthodes de décomposition parcimonieuse régularisée spatialement dans la récupération de structures sous-jacentes de signaux artificiels lorsque le dictionnaire est un dictionnaire de Gabor de cohérence élevée.

— À propos de la comparaison des approches d'optimisation convexe et des approches gloutonnes : elles présentent des performances similaires lorsque le rapport signal à bruit n'est pas trop faible (5db, 0db). Dans ces cas, les performances des approches convexes semblent légèrement supérieures mais nécessitent l'optimisation du paramètre λ pour les différents signaux (plusieurs décompositions pour obtenir le bon taux de reconstruction). Pour cette raison, les approches gloutonnes semblent dans ce contexte plus intéressantes du fait d'un coût de calcul moindre (lorsque la parcimonie des signaux est supposée forte). Au contraire pour des RSB plus faibles, les approches convexes présentent des performances significativement plus élevées pouvant justifier cette optimisation du paramètre de régularisation.
— Enfin, concernant les régularisations par groupes, nous pouvons observer qu'elles présentent des performances très similaires avec les régularisations imposant une structure parcimonieuse commune. Cette expérience n'est néanmoins pas adaptée pour évaluer complètement ces régularisations par groupes étant donné qu'elle n'évalue que le choix des atomes par ces approches et ne s'intéresse pas au choix des groupes. Nous pouvons seulement remarquer ici qu'elles permettent de sélectionner les atomes aussi efficacement que les approches imposant une structure parcimonieuse commune lorsque tous les canaux sont utilisés.

5.3.1.3 Expérience 2

Nous allons maintenant nous intéresser de manière plus approfondie aux régularisations par groupes et à la manière dont la sélection des atomes peut être améliorée grâce à celles-ci lorsque le choix des groupes est pris en compte. Étant donné les résultats de l'expérience précédente montrant l'efficacité des approches ℓ_1 pour un RSB faible que nous supposons pour les signaux EEG ainsi que le coût calculatoire élevé du MSCSMP lorsque le nombre de groupes à sélectionner augmente, nous étudions uniquement des approches convexes ici.

Protocole de test et paramètres. L'expérience que nous considérons ici est très similaire à celle effectuée dans la section précédente. Les mêmes signaux sont décomposés à l'aide de l'algorithme FISTA pour les régularisations $R_{2,1}$, $R_{2,1}+R^8_{w,lap}$, $R^{lg}_{2,1}$ et $R^{lg}_{2,1}+R^8_{w,lap}$ pour différentes valeurs du rapport signal à bruit et du nombre de composantes (N_a) sous-jacentes de ceux-ci.
Comme précédemment, pour chaque régularisation et chaque signal, après décompositions du signal, une sélection de N_a atomes est effectuée puis une projection contrainte spatialement est appliquée sur ceux-ci. Pour la régularisation $R_{2,1}$, les N_a principaux atomes sont choisis comme ceux ayant les vecteurs de coefficients de plus grandes normes ℓ_2 comme dans l'expérience 1. En revanche, pour la régularisation par groupes, ce choix est réalisé différemment. Pour chaque atome, les normes ℓ_2 des vecteurs de coefficients des différents groupes sont calculées. La somme des N_g valeurs les plus élevées parmi les normes calculées pour chaque atome est alors utilisée comme critère à maximiser pour choisir les N_a principaux atomes. L'expérience est réalisée pour plusieurs valeurs de N_g : 8, 25, 50.
Cette expérience est effectuée pour les deux dictionnaires conçus précédemment et le critère défini plus haut est conservé. Les paramètres λ et μ sont également gérés comme dans la

première expérience.

Résultats et discussion Les résultats obtenus avec le dictionnaire Φ^1 sont présentés dans la figure 5.6 et les résultats obtenus avec le dictionnaire Φ^2 dans la figure 5.7.

Ces résultats permettent de comprendre l'intérêt des régularisations par groupes pour l'identification correcte des atomes et donc des composantes spatio-temporelles lorsqu'elles sont combinées avec une régularisation spatiale. En particulier nous pouvons faire les observations suivantes :

- Comme précédemment, la régularisation spatiale permet d'améliorer significativement les résultats.
- La régularisation par groupes permet l'obtention de meilleures performances que la régularisation imposant une structure parcimonieuse commune aux différents canaux dans la plupart des cas. Il est d'ailleurs intéressant de noter que cette remarque est également valable pour les décompositions sans régularisations spatiales pour lesquelles l'augmentation des performances est encore plus importante. Cet effet peut être expliqué par l'utilisation de seulement quelques groupes de canaux pour le choix des atomes, les groupes sélectionnés participent fortement à la reconstruction et correspondent donc aux canaux sur lesquels les topographies des activités sont maximales. Ces *maxima* sont moins fortement impactés par le bruit et permettent un meilleur choix des atomes.
- De plus, même si les différences sont faibles, il semble que plus le nombre d'activités à retrouver est grand et plus les performances sont grandes pour une sélection avec peu de groupes alors que pour peu d'activités, un choix réalisé avec un grand nombre de groupes est plus efficace. Concernant ce phénomène, il peut être expliqué en remarquant que plus le nombre d'activités composant les signaux grandit et plus l'identification de l'une d'entre elles ne peut se faire aisément que sur un nombre de canaux faible : les canaux les plus caractéristiques de celle-ci sont les moins caractéristiques des autres.

Ces observations sont particulièrement marquées avec le dictionnaire de faible cohérence, mais restent également valables pour le dictionnaire de Gabor.

5.3.2 Détection de P300

L'intérêt des méthodes proposées plus haut ayant été montré sur des données artificielles, nous les évaluons maintenant sur données réelles. En particulier, nous nous intéressons à la détection de potentiels évoqués P300, introduit dans le chapitre 2.
Pour rappel, c'est un potentiel évoqué très utilisé dans les systèmes ICM, généralement observé lorsqu'un stimulus rare cible apparaît dans une suite de stimuli non-cibles. Plus précisément, il apparaît entre 250 et 450 ms après le stimulus cible et présente une amplitude maximale dans la région pariétale. Son amplitude et son délai d'apparition dépendent de différents facteurs comme la durée séparant deux stimuli cibles (pour une revue voir [188]).

	5 db	**0 db**	**-5 db**	**-10 db**	**Moy**
$R_{2,1}$	6.3229	5.9697	4.7486	3.7572	5.1996
$R_{2,1} + R^8_{w,lap}$	9.269	8.7539	8.1662	7.5459	8.4337
$R^{lg}_{2,1}$, $N_g = 8$	6.4578	6.5692	5.4936	4.557	5.7694
$R^{lg}_{2,1} + R^8_{w,lap}$, $N_g = 8$	9.295	8.8726	8.2916	7.4913	8.4876
$R^{lg}_{2,1}$, $N_g = 25$	6.3115	6.567	5.3522	4.2645	5.6238
$R^{lg}_{2,1} + R^8_{w,lap}$, $N_g = 25$	9.2921	**8.8852**	8.3061	7.5932	8.5192
$R^{lg}_{2,1}$, $N_g = 8$	6.2385	6.4217	5.1084	4.1858	5.4886
$R^{lg}_{2,1} + R^8_{w,lap}$, $N_g = 50$	**9.3588**	8.8125	**8.3065**	**7.6673**	**8.5362**

(a) $N_a = 10$

	5 db	**0 db**	**-5 db**	**-10 db**	**Moy**
$R_{2,1}$	11.8355	9.5978	8.0949	6.1411	8.9173
$R_{2,1} + R^8_{w,lap}$	13.3341	**12.6961**	11.9478	11.0199	12.2495
$R^{lg}_{2,1}$, $N_g = 8$	11.3529	11.2361	7.9273	6.9112	9.3569
$R^{lg}_{2,1} + R^8_{w,lap}$, $N_g = 8$	13.4334	12.5838	**11.961**	11.044	12.2555
$R^{lg}_{2,1}$, $N_g = 25$	11.2607	11.2514	7.6014	6.546	9.1649
$R^{lg}_{2,1} + R^8_{w,lap}$, $N_g = 25$	**13.4788**	12.5539	11.9335	**11.1031**	**12.2673**
$R^{lg}_{2,1}$, $N_g = 8$	11.1726	11.0549	7.3496	6.2879	8.9663
$R^{lg}_{2,1} + R^8_{w,lap}$, $N_g = 50$	13.1753	12.6808	11.8465	10.9813	12.171

(b) $N_a = 15$

	5 db	**0 db**	**-5 db**	**-10 db**	**Moy**
$R_{2,1}$	10.9126	12.2475	8.0712	5.407	9.1596
$R_{2,1} + R^8_{w,lap}$	17.2188	**16.3907**	14.9762	**13.4746**	15.5151
$R^{lg}_{2,1}$, $N_g = 8$	15.9677	14.0582	9.8131	7.4155	11.8136
$R^{lg}_{2,1} + R^8_{w,lap}$, $N_g = 8$	**17.4777**	16.3824	**15.2158**	13.3699	**15.6115**
$R^{lg}_{2,1}$, $N_g = 25$	15.7805	13.9146	9.8465	7.2603	11.7005
$R^{lg}_{2,1} + R^8_{w,lap}$, $N_g = 25$	17.2725	16.226	15.1484	13.4098	15.5142
$R^{lg}_{2,1}$, $N_g = 8$	15.6981	13.5867	9.5835	6.9734	11.4604
$R^{lg}_{2,1} + R^8_{w,lap}$, $N_g = 50$	17.2607	16.2198	14.8467	13.3995	15.4317

(c) $N_a = 20$

	5 db	**0 db**	**-5 db**	**-10 db**	**Moy**
$R_{2,1}$	17.2473	12.2407	10.457	7.4958	11.8602
$R_{2,1} + R^8_{w,lap}$	21.5596	20.1237	18.0322	15.7209	18.8591
$R^{lg}_{2,1}$, $N_g = 8$	20.71	17.5028	13.8843	8.5675	15.1662
$R^{lg}_{2,1} + R^8_{w,lap}$, $N_g = 8$	**21.6917**	20.0597	**18.1083**	**15.8708**	**18.9326**
$R^{lg}_{2,1}$, $N_g = 25$	20.7279	17.2392	13.5463	8.3535	14.9667
$R^{lg}_{2,1} + R^8_{w,lap}$, $N_g = 25$	21.5094	**20.127**	17.9724	15.8358	18.8612
$R^{lg}_{2,1}$, $N_g = 8$	20.6764	16.8771	13.0992	8.2613	14.7285
$R^{lg}_{2,1} + R^8_{w,lap}$, $N_g = 50$	21.391	19.9624	17.8387	15.6955	18.7219

(d) $N_a = 25$

FIGURE 5.6 – Évaluation des méthodes de décomposition parcimonieuse régularisée par groupe spatialement dans la récupération de structures sous-jacentes de signaux artificiels lorsque le dictionnaire est peu cohérent.

	5 db	**0 db**	**-5 db**	**-10 db**	**Moy**
$R_{2,1}$	5.0364	3.8903	3.5622	2.5886	3.7694
$R_{2,1} + R^8_{w,lap}$	**6.3231**	**5.6957**	5.2053	4.3115	5.3839
$R^{lg}_{2,1}$, $N_g = 8$	5.0351	4.0303	3.294	2.5016	3.7153
$R^{lg}_{2,1} + R^8_{w,lap}$, $N_g = 8$	6.2195	5.5535	5.1261	4.4818	5.3452
$R^{lg}_{2,1}$, $N_g = 25$	4.7115	3.9107	3.3554	2.6128	3.6476
$R^{lg}_{2,1} + R^8_{w,lap}$, $N_g = 25$	6.2936	5.6433	**5.309**	**4.5199**	**5.4415**
$R^{lg}_{2,1}$, $N_g = 8$	4.6802	4.0561	3.3867	2.6395	3.6906
$R^{lg}_{2,1} + R^8_{w,lap}$, $N_g = 50$	6.2046	5.646	5.1623	4.4854	5.3746

(a) $N_a = 10$

	5 db	**0 db**	**-5 db**	**-10 db**	**Moy**
$R_{2,1}$	5.0767	4.2159	4.0189	3.7863	4.2744
$R_{2,1} + R^8_{w,lap}$	8.1338	**7.6222**	6.792	6.3036	7.2129
$R^{lg}_{2,1}$, $N_g = 8$	5.7095	4.9711	4.6254	3.7795	4.7714
$R^{lg}_{2,1} + R^8_{w,lap}$, $N_g = 8$	8.0319	7.6093	7.0304	**6.5755**	**7.3118**
$R^{lg}_{2,1}$, $N_g = 25$	5.6876	4.7124	4.5516	3.7227	4.6686
$R^{lg}_{2,1} + R^8_{w,lap}$, $N_g = 25$	8.1719	7.4969	7.0189	6.4351	7.2807
$R^{lg}_{2,1}$, $N_g = 8$	5.5372	4.6728	4.5268	3.7446	4.6203
$R^{lg}_{2,1} + R^8_{w,lap}$, $N_g = 50$	**8.2619**	7.5966	**7.0313**	6.0986	7.2471

(b) $N_a = 15$

	5 db	**0 db**	**-5 db**	**-10 db**	**Moy**
$R_{2,1}$	6.3222	6.5442	5.6769	5.0153	5.8896
$R_{2,1} + R^8_{w,lap}$	10.7989	9.8981	9.2849	8.7643	9.6865
$R^{lg}_{2,1}$, $N_g = 8$	8.484	7.6153	6.6847	5.7665	7.1376
$R^{lg}_{2,1} + R^8_{w,lap}$, $N_g = 8$	10.7553	**10.2718**	9.5376	**9.0583**	**9.9058**
$R^{lg}_{2,1}$, $N_g = 25$	8.2223	7.4897	6.2537	5.7256	6.9228
$R^{lg}_{2,1} + R^8_{w,lap}$, $N_g = 25$	**10.8025**	10.218	**9.6672**	8.8797	9.8919
$R^{lg}_{2,1}$, $N_g = 8$	8.2863	7.4017	6.5256	5.5606	6.9436
$R^{lg}_{2,1} + R^8_{w,lap}$, $N_g = 50$	10.6109	10.1974	9.5777	8.8021	9.797

(c) $N_a = 20$

	5 db	**0 db**	**-5 db**	**-10 db**	**Moy**
$R_{2,1}$	10.4816	9.1711	6.9991	6.4939	8.2864
$R_{2,1} + R^8_{w,lap}$	13.3904	13.0953	**12.5415**	**11.5687**	**12.6489**
$R^{lg}_{2,1}$, $N_g = 8$	11.2342	9.7739	8.6594	6.5428	9.0526
$R^{lg}_{2,1} + R^8_{w,lap}$, $N_g = 8$	13.534	12.9698	12.4961	11.4306	12.6076
$R^{lg}_{2,1}$, $N_g = 25$	11.3487	9.8931	8.5609	6.2346	9.0093
$R^{lg}_{2,1} + R^8_{w,lap}$, $N_g = 25$	**13.5357**	12.9912	12.5347	11.3231	12.5962
$R^{lg}_{2,1}$, $N_g = 8$	11.5692	9.8042	8.4924	6.1673	9.0083
$R^{lg}_{2,1} + R^8_{w,lap}$, $N_g = 50$	13.4392	**13.1025**	12.4783	11.3522	12.5931

(d) $N_a = 25$

FIGURE 5.7 – Évaluation des méthodes de décomposition parcimonieuse régularisée spatialement dans la récupération de structures sous-jacentes de signaux artificiels lorsque le dictionnaire est un dictionnaire de Gabor de cohérence élevée.

5.3.2.1 Données

Pour cette expérience, nous utilisons le jeu de données IIb de la deuxième compétition ICM [26]. Ces signaux ont été enregistrés avec un montage à 64 électrodes et une fréquence d'échantillonnage de 240Hz. Les essais que nous considérons sont extraits entre 150 et 450 ms après les stimuli lumineux et filtrés pour ne conserver que les fréquences entre 0.3 et 20 Hz avec un filtre de Butterworth d'ordre 4. Le jeu de données comporte plusieurs sessions de mesure, nous considérons dans cette expérience pour les résultats de classification la première session du sujet 1 depuis laquelle ont été extraits 500 signaux contenant un P300 et 500 n'en contenant pas.

5.3.2.2 Protocole de test et paramètres

Nous souhaitons ici évaluer les régularisations et méthodes étudiées dans ce chapitre pour la décomposition de potentiels évoqués P300 sur un dictionnaire temps-fréquence. Pour cela, les signaux sont décomposés sur le dictionnaire pour les régularisations : $R_{2,1}$, $R_{2,1} + R^8_{w,lap}$, $R^{lg}_{2,1}$ et $R^{lg}_{2,1} + R^8_{w,lap}$ à l'aide de l'algorithme FISTA. La forme relâchée du problème a été choisie pour les mêmes raisons que dans l'expérience 2 (hypothèse d'un RSB faible).
Pour chaque signal, après décomposition, de la même façon que pour les signaux synthétiques, des atomes sont sélectionnés puis le signal projeté sur ces atomes avec la même valeur du paramètre de régularisation spatiale que pour la décomposition. Cette projection est réalisée via l'algorithme FISTA avec une structure de zéros fixe et permet de supprimer le biais du terme de régularisation parcimonieuse introduit dans la décomposition. La sélection des atomes est effectuée de la même manière que pour l'expérience 2 pour plusieurs valeurs de N_a et N_g.
Une fois cette étape réalisée, l'évaluation de ces décompositions est effectuée grâce à une étape de classification. Nous supposons ici que les régularisations proposées permettent un choix plus cohérent des atomes malgré le bruit et les variabilités inter-essais. Plus précisément, soit deux ensembles d'atomes non nécessairement disjoints correspondants l'un à la présence d'un P300 l'autre à l'absence de celui-ci, nous supposons que les essais se décomposent plus fréquemment sur l'ensemble des atomes associés à leurs classes lorsque les régularisations spatiales proposées sont employées. Les composantes obtenues sur des atomes n'appartenant pas à ces ensembles sont alors considérées comme du bruit.
Ainsi dans un premier temps, pour chaque cas considéré (nombre d'atomes sélectionnés, régularisations, et paramètres de régularisation), les principaux atomes de décomposition des deux classes sont identifiés par décomposition de signaux représentatifs obtenus par moyennage d'essais de ces deux classes (ici obtenues sur d'autres sessions du jeu de données). Ensuite, à partir des matrices de décompositions obtenues précédemment, les essais sont reconstruits en utilisant uniquement les atomes identifiés comme permettant de représenter l'une des deux classes dans le cas considéré. La classification est enfin réalisée pour ces essais reconstruits à l'aide du BLDA (« Bayesian Linear Discriminant Analysis ») qui s'est montré particulièrement efficace pour la classification de ce type de signaux [111].

Pour chaque régularisation, les paramètres de régularisations sont déterminés sur le

score de classification à l'aide d'une validation croisée. Concernant le dictionnaire, un dictionnaire de Gabor multi-échelles Φ a été conçu pour cette expérience. Plus précisément, il est construit comme la concaténation des atomes de dictionnaires de Gabor d'échelle $s = 16$, $s = 32$ et $s = 64$ pour des fréquences comprises entre 0 et 20hz. Il possède une cohérence élevée : $\mu_1(\Phi, 1) \geq 0.9$ et une taille de $N_\Phi = 185$ pour des signaux possédant 73 échantillons temporels.

5.3.2.3 Résultats de classification et discussion

La figure 5.8 présente les scores de classifications obtenus avec les signaux reconstruits après décompositions régularisées pour différents nombres de composantes. Nous comparons (dans l'ordre) les régularisations : $R_{2,1}$, $R_{2,1} + R^8_{w,lap}$ puis $R^{lg}_{2,1}$ et $R^{lg}_{2,1} + R^8_{w,lap}$ avec un choix d'atomes basé sur 10 groupes et enfin ces mêmes régularisations pour un choix basé sur 50 groupes.

Ces résultats mettent en évidence l'intérêt des régularisations proposées pour l'identification de composantes dans les signaux EEG et l'obtention de décompositions plus stables. En particulier nous pouvons faire les observations suivantes :

— Comme attendu, le score de classification augmente avec le nombre de composantes utilisées pour les décompositions. L'approche utilisée étant générative et non-discriminative, les composantes sont choisies de manière à reconstruire les signaux, or les composantes de plus grandes amplitudes choisies en premier ne sont pas obligatoirement les plus discriminatives. De plus, du fait de la cohérence du dictionnaire rendant les décompositions plus sensibles aux bruits et de notre choix de n'utiliser que les atomes sélectionnés par la décomposition des signaux représentatifs des deux classes, l'augmentation de N_a entraîne tout simplement une prise en compte de plus d'atomes (et donc de plus de composantes possibles).

— Un test de type « Wilcoxon signed-rank » ($p<0.05$) montre que pour toutes les valeurs de N_a, les régularisations spatiale permettent une amélioration des scores de classification à l'exception du cas de la régularisation $R^{lg}_{2,1}10$ pour N_a. Cette amélioration est plus importante lorsque le nombre de composantes est faible. Ce phénomène peut s'expliquer par la plus grande importance du choix des composantes lorsque peu de celles-ci sont utilisées pour la classification et montre que ces régularisations de lissage permettent d'améliorer le choix des atomes de manière significative.

— La régularisation par groupe est significativement (test de type « Wilcoxon signed-rank ») plus efficace que celle imposant une structure parcimonieuse commune à l'ensemble des canaux lorsque le nombre de composantes est grand *i.e.* pour $N_a = 70$. Ainsi, le choix des atomes est plus efficace lorsqu'il s'effectue sur quelques groupes de canaux contiguës qui correspondent aux *maxima* des profils spatiaux des composantes comme nous l'avions déjà observé dans l'expérience 2. Le nombre de groupes utilisés pour la sélection des atomes semble peu faire varier les résultats même si la variance des résultats est légèrement plus faible lorsque le nombre de groupes pris en compte est grand.

Par ailleurs, même si l'approche proposée est purement générative, lorsque le nombre de composantes prises en compte est grand ($N_a = 70$), les scores de classification sont

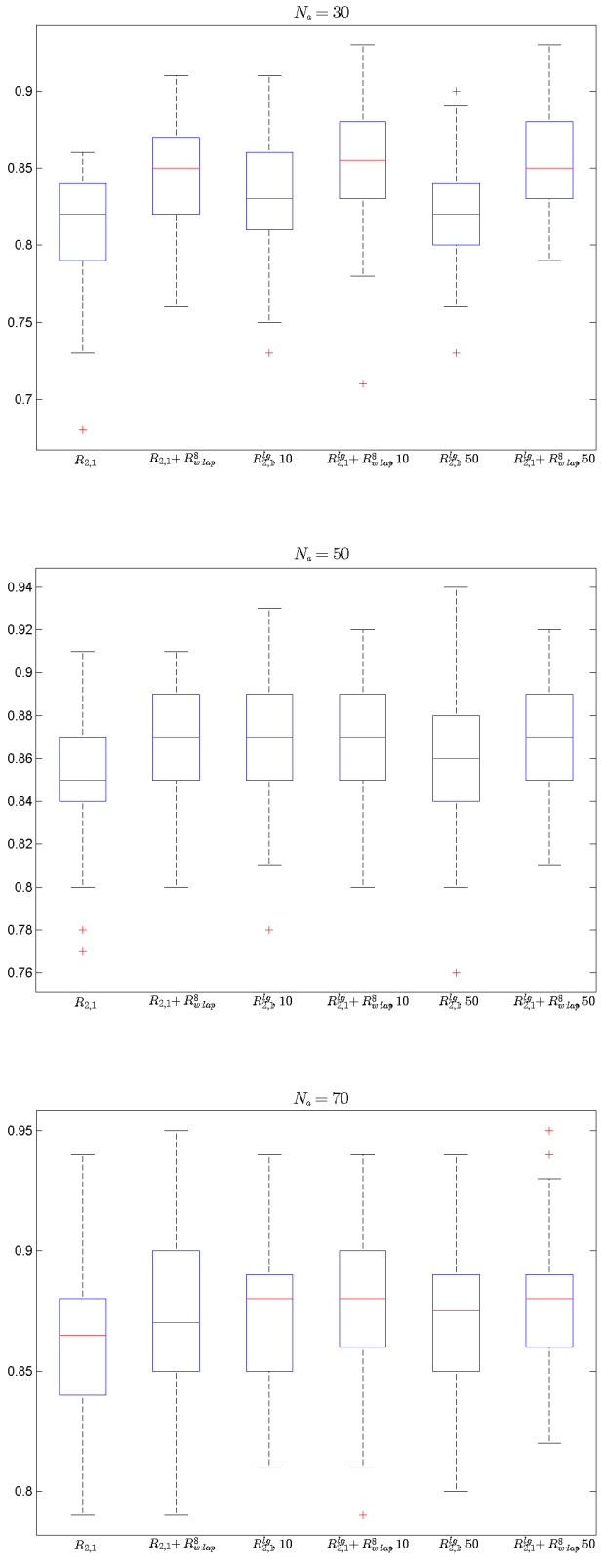


FIGURE 5.8 – Taux de classification correcte pour la détection de P300 après décomposition régularisée des enregistrements EEG sur un dictionnaire temps-fréquence. Les régularisations étudiées sont comparées pour différentes valeurs du nombre d'atomes sélectionnées N_a.

significativement (même test) meilleurs par rapport à ceux obtenus pour les signaux bruts. Les parties des signaux non reconstruites par les approches peuvent alors être considérées comme des bruits perturbant la classification et notre approche permet alors un débruitage des signaux. Cette observation est présentée dans la figure 5.9 qui compare les scores de classification des signaux brutes avec ceux des scores mis en avant plus haut pour $N_a = 70$ lorsque les régularisations de lissage sont utilisées. Rappelons que pour $N_a = 70$, l'approche par groupe est significativement plus efficace que celle utlisant tous les canaux (en particulier la régularisation $R^{lg}_{2,1}50$). Ce type de décomposition peut ainsi être utilisé pour le débruitage

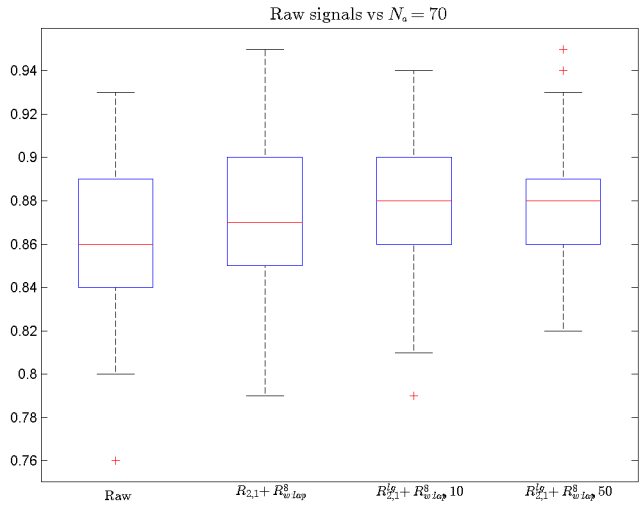


FIGURE 5.9 – Taux de classification correcte pour la détection de P300. Comparaison entre la classification des signaux bruts et de celles obtenues après décompositions régularisées spatialement des enregistrements EEG pour différents termes de régularisation. Le nombre d'atomes sélectionnés par ces décompositions est fixé à $N_a = 70$.

de tels signaux dans un contexte d'interface cerveau-machine.

5.4 Conclusion

Dans ce chapitre nous nous sommes attachés à la mise en place d'algorithmes permettant de réaliser la décomposition temps-fréquence de signaux EEG régularisée spatialement dont nous avions discuté l'interêt dans le chapitre précédent. Des algorithmes gloutons permettant d'approcher la solution de la formulation stricte du problème ont ainsi été proposés de même qu'une adaptation de l'algorithme FISTA pour la forme convexe du problème.

Dans un premier temps, nous avons vu que ces algorithmes sont efficaces pour le recouvrement des différentes composantes sous-jacentes de signaux synthétiques réalistes conçus pour respecter les *a priori* physiologiques supposés lors de la création des termes de régularisation. Ces expériences ont donc permis d'assurer le bon comportement des algorithmes

proposés et l'efficacité des régularisations proposées lorsque les signaux considérés respectent effectivement les caractéristiques supposées des signaux EEG. Par ailleurs, ces tests ont permis l'obtention de trois résultats importants :

— les approches convexes développées présentent de meilleures performances que les approches gloutonnes lorsque le bruit est fort,
— la régularisation de lissage proposée permet d'améliorer significativement l'extraction des composantes,
— la régularisation du « Latent Group LASSO » est plus efficace que celle imposant une structure parcimonieuse commune pour l'extraction de ces mêmes composantes.

L'application de ces méthodes dans le contexte de la détection du PE P300 a ensuite mis en avant l'intérêt de ces régularisations (et des algorithmes associés) pour la décomposition de signaux réels. Malgré le fondement génératif des approches proposées, elles se révèlent efficaces pour le « débruitage » de signaux EEG dans un contexte discriminatif grâce à un choix cohérent des atomes durant les décompositions.

De futurs travaux pourraient envisager l'utilisation d'une régularisation de lissage apprise comme nous l'avons fait dans le chapitre 4 pour améliorer les résultats de l'approche proposée dans ce chapitre.

Chapitre 6

Décomposition régularisée en analyse

Sommaire

Nous considérons maintenant le problème d'optimisation associé à la régularisation temporelle introduite dans le chapitre 4. La résolution de ce problème est nécessaire à la mise en place de l'extension du modèle des micro-états décrite dans le chapitre 7. Ce problème rentre dans le cadre plus général des décompositions régularisées en analyse et sa résolution peut donc avoir un intérêt pour d'autres applications. Nous montrerons notamment à la fin de ce chapitre comment il est possible de réaliser une régularisation spatiale similaire au chapitre précédent pour une décomposition temps-fréquence de signaux EEG. Cette régularisation sera appliquée à la détection de P300.

6.1 Modèle et problème d'optimisation

Soit $Y = [\mathbf{y}(1), \ldots, \mathbf{y}(T)] \in \mathbb{R}^{C \times T}$ un ensemble de T signaux ordonnés (par exemple enregistrés à des instants consécutifs) de dimension C et $\Phi \in \mathbb{R}^{C \times N_\Phi}$ un dictionnaire re-

dondant composé de N_Φ ($N_\Phi \gg C$) atomes normés. Nous considérons le modèle linéaire classique suivant :

$$\begin{aligned} t \in \{1, \ldots, T\}, \;\; \mathbf{y}(t) &= \Phi \mathbf{x}(t) + \mathbf{e}(t), \\ Y &= \Phi X + E \;\; , \end{aligned} \tag{6.1}$$

dans lequel $X = [\mathbf{x}(1), \cdots, \mathbf{x}(T)] \in \mathbb{R}^{N_\Phi \times T}$ est la matrice de décomposition et $E = [\mathbf{e}(1), \cdots, \mathbf{e}(T)] \in \mathbb{R}^{C \times T}$ une matrice de bruit gaussien.

Nous étudions ici l'approximation parcimonieuse structurée d'un tel signal. Plus précisément, nous supposons que le signal multidimensionnel Y est composé d'un ensemble de signaux $\mathbf{y}(t)$, $t \in \{1, \cdots, T\}$ C-dimensionnel dont la concaténation présente une structure particulière connue. Notre but est d'utiliser la connaissance de cette structure afin de guider la décomposition parcimonieuse vers une solution plausible dépendant moins du bruit en imposant à la matrice de décomposition X le respect de cette structure.
Cette contrainte est exprimée ici à travers une régularisation combinant un terme de régularisation en analyse et un terme de régularisation parcimonieuse classique. Notre problème de décomposition structurée est alors formalisé à travers la minimisation suivante :

$$\hat{X} = \underset{X \in \mathbb{R}^{N_\Phi \times T}}{\arg\min} \; \|Y - \Phi X\|_F^2 + \lambda_1 \|X\|_1 + \lambda_2 \|XP\|_1 \;\; , \tag{6.2}$$

avec λ_1, λ_2 les coefficients de régularisation et $P \in \mathbb{R}^{T \times N_P}$ la matrice encodant l'*a priori* de structure. Le terme de régularisation $\|XP\|_1$ peut être interprété dans le cadre des problèmes parcimonieux en analyse [69] (voir chapitre 3). Dans ce contexte de décomposition avec contrainte en analyse, les atomes du dictionnaire Φ^* peuvent être vus comme des filtres, sur lesquels la projection des signaux décomposés est supposée parcimonieuse. De ce point de vue, la matrice P est un ensemble de filtres conçus de telle façon que la projection des lignes de la matrice de décomposition X sur ceux-ci soit parcimonieuse, conservant ainsi les régularités supposées du signal : régularités répondant aux connaissances *a priori* des signaux du domaine d'application.

Dans ce cadre, nous nous intéressons spécifiquement à un *a priori* de constance par morceaux d'une série temporelle. La contrainte associée peut alors être obtenue via la régularisation précédente en choisissant pour P un opérateur de différence finie :

$$P^1 = \begin{pmatrix} -1 & & & \\ 1 & -1 & & \\ & 1 & \ddots & \\ & & \ddots & -1 \\ & & & 1 \end{pmatrix}, \quad P^2 = \begin{pmatrix} 1 & & & \\ -2 & 1 & & \\ 1 & -2 & \ddots & \\ & 1 & \ddots & 1 \\ & & \ddots & -2 \\ & & & 1 \end{pmatrix}$$

$$\|XP^1\|_1 = \sum_{t=2}^{T} \|\mathbf{x}_t - \mathbf{x}_{t-1}\|_1,$$
$$\|XP^2\|_1 = \sum_{t=2}^{T-1} \|\mathbf{x}_{t+1} - 2\mathbf{x}_t + \mathbf{x}_{t-1}\|_1.$$

La régularisation obtenue à l'aide d'une combinaison de la régularisation ℓ_1 et d'une régularisation en analyse de ce type permet l'obtention d'une décomposition en blocs parcimonieux dont une visualisation est proposée dans la figure 6.1(avec en gris, l'atome dont les coefficients sont représentés).

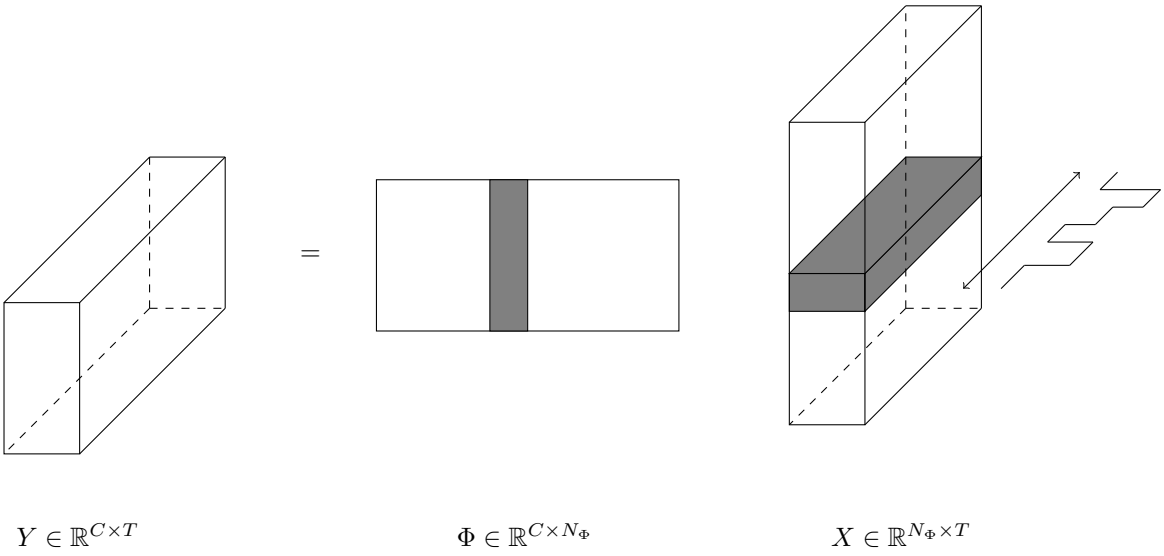


FIGURE 6.1 – Décomposition par bloc parcimonieux d'une série temporelle multidimensionnelle, la suite des coefficients de décomposition des colonnes de Y présente une structure constante par morceaux.

6.2 Fused-LASSO

Ce type de régularisation, composée d'une régularisation parcimonieuse et d'une régularisation de type VT (Variation Totale) a été introduit dans un cas monodimensionnel pour un problème de régression sous le nom de « Fused-LASSO » dans [211]. Le terme de régularisation VT vise à préserver les singularités du signal et à permettre la détection des changements abrupts ; il a notamment été étudié de manière intensive dans le modèle ROF [198, 51], très utilisé pour le débruitage d'images.
Malgré la convexité du problème d'optimisation associé, les deux termes ℓ_1 non-differentiables le rendent difficile à résoudre. Dans leur papier séminal, Tibshirani et al. [211] transforme ce problème en un problème quadratique et utilise ensuite des outils standards d'optimisation quadratique (SQOPT) pour sa résolution. Cette approche bien que résolvant effectivement

le problème, n'est néanmoins applicable que pour des problèmes de petites dimensions étant donné le fait qu'elle repose sur une reformulation augmentant grandement l'espace de recherche de la solution. Une approche calculant itérativement les solutions du problème pour différentes valeurs des paramètres de régularisation a ensuite été proposée dans [110] dans le cas où le dictionnaire est la matrice identité ($\Phi = I_C$) et Tibshirani et al [212] ont ensuite conçu une approche similaire pour le cas du LASSO généralisé (avec pour cas particulier le Fused-LASSO). Ces approches offrent de meilleures performances que la reformulation quadratique, mais sont tout de même peu efficaces lorsque les dimensions du problème augmentent. Ainsi, plus récemment, des approches permettant un meilleur passage à l'échelle ont été développées. On trouve notamment parmi celles-ci des méthodes basées sur le sous-gradient [140], l'algorithme des directions alternées (ADMM en anglais : « Alternating Direction Method of Multipliers ») [226] ou la méthode du « split Bregman » [236].
À notre connaissance, dans un cadre multidimensionnel avec un dictionnaire redondant, ce problème n'a pas été étudié, néanmoins une méthode permettant la résolution du problème décrit plus haut pour un problème de régression multi-tâches non-redondante a été proposée dans [44]. Cette dernière étude considère l'application d'une méthode proximale [170] pour une version approchée différentiable de la fonction à minimiser.
Nous proposons ici une approche alternative fondée sur l'extension multidimensionnelle de l'approche utilisant la technique d'optimisation du « split Bregman » décrite dans [236]. Le schéma d'optimisation du « split Bregman » s'est montré particulièrement efficace pour la résolution de problèmes impliquant des régularisations ℓ_1 grâce notamment à sa capacité à déterminer de manière rapide le support de décomposition [90]. Une comparaison en matière de vitesse de convergence de cette approche avec celle (proximale) présentée dans [44] est décrite dans la section 6.4.

6.3 Stratégie d'optimisation

Afin de résoudre le problème de minimisation introduit plus haut, nous proposons d'étendre le schéma d'optimisation fondé sur les itérations du « split Bregman » décrit dans [236] au cas multidimensionnel. La méthode proposée est nommée MultiSSSA (pour « Multidimensionnal Sparse Structured Signal Approximation »).

6.3.1 Schéma d'optimisation

Rappelons dans un premier temps le problème :

$$\hat{X} = \underset{X \in \mathbb{R}^{N_\Phi \times T}}{\arg\min} \|Y - \Phi X\|_F^2 + \lambda_1 \|X\|_1 + \lambda_2 \|XP\|_1 \ .$$

Afin de mettre en place le schéma d'optimisation du « split Bregman » le problème est premièrement reformulé comme suit :

$$(\hat{X}, \hat{A}, \hat{B}) = \underset{\substack{X \in \mathbb{R}^{N_\Phi \times T} \\ A \in \mathbb{R}^{N_\Phi \times T}, B \in \mathbb{R}^{N_\Phi \times N_P}}}{\arg\min} \|Y - \Phi X\|_F^2 + \lambda_1 \|A\|_1 + \lambda_2 \|B\|_1$$
$$\text{t.q.} \quad A = X \quad \text{et} \quad B = XP \ . \tag{6.3}$$

Cette reformulation est une étape clé permettant de découpler les 3 termes de la fonction objectif afin de les optimiser par la suite séparément. La mise en place du schéma d'optimisation nécessite ensuite le passage du problème sous une forme non contrainte :

$$\underset{\substack{X\in\mathbb{R}^{N_\Phi\times T}\\ A\in\mathbb{R}^{N_\Phi\times T},B\in\mathbb{R}^{N_\Phi\times N_P}}}{\arg\min} \|Y-\Phi X\|_F^2+\lambda_1\|A\|_1+\lambda_2\|B\|_1 \quad +\frac{\mu_1}{2}\|X-A\|_F^2+\frac{\mu_2}{2}\|XP-B\|_F^2 \ .$$

Les itérations du « split Bregman » [90] peuvent enfin être mises en place comme suit :

$$\begin{aligned}(X^{i+1},A^{i+1},B^{i+1}) = &\underset{\substack{X\in\mathbb{R}^{N_\Phi\times T}\\ A\in\mathbb{R}^{N_\Phi\times T},B\in\mathbb{R}^{N_\Phi\times N_P}}}{\arg\min} \|Y-\Phi X\|_F^2+\lambda_1\|A\|_1+\lambda_2\|B\|_1\\ &+\frac{\mu_1}{2}\|X-A+D_A^i\|_F^2+\frac{\mu_2}{2}\|XP-B+D_B^i\|_F^2 \qquad (6.4)\\ D_A^{i+1} &= D_A^i+(X^{i+1}-A^{i+1})\\ D_B^{i+1} &= D_B^i+(X^{i+1}P-B^{i+1}) \ .\end{aligned}$$

Ce schéma est équivalent à celui obtenu avec le lagrangien augmenté (ou l'ADMM) lorsque les contraintes sont linéaires [234]. Le lien avec ces méthodes peut être fait en remarquant que les variables duales U et V des contraintes $A=X$ et $B=XP$ peuvent être exprimées en fonction de D_A et D_B comme suit : $U=\mu_1 D_A$ et $V=\mu_2 D_B$.
Grâce à la séparation des trois termes, la minimisation du problème primal Eq. (6.4) peut être réalisée itérativement en alternant la mise à jour des 3 variables :

$$\begin{aligned}X^{i+1} = &\underset{X\in\mathbb{R}^{N_\Phi\times T}}{\arg\min} \|Y-\Phi X\|_F^2+\frac{\mu_1}{2}\|X-A^i+D_A^i\|_F^2\\ &+\frac{\mu_2}{2}\|XP-B^i+D_B^i\|_F^2 \qquad (6.5)\\ A^{i+1} = &\underset{A\in\mathbb{R}^{N_\Phi\times T}}{\arg\min} \lambda_1\|A\|_1+\frac{\mu_1}{2}\|X^{i+1}-A+D_A^i\|_F^2 \qquad (6.6)\\ B^{i+1} = &\underset{B\in\mathbb{R}^{N_\Phi\times N_P}}{\arg\min} \lambda_2\|B\|_1+\frac{\mu_2}{2}\|X^{i+1}P-B+D_B^i\|_F^2 \qquad (6.7)\end{aligned}$$

Empiriquement, il a été montré que la convergence de ce système nécessite seulement quelques itérations [90] avant d'être atteinte. Par conséquent, notre schéma d'optimisation ne réalise cette étape qu'une seule fois à chaque itération de l'algorithme global.
Les equations . (6.6) et (6.7) peuvent être résolues à l'aide de l'opérateur de seuillage doux [1]

$$\begin{aligned}A^{i+1} = &\underset{\frac{\lambda_1}{\mu_1}}{\text{SoftThreshold}}(X^{i+1}+D_A^i), \qquad (6.8)\\ B^{i+1} = &\underset{\frac{\lambda_2}{\mu_2}}{\text{SoftThreshold}}(X^{i+1}P+D_B^i) \qquad (6.9)\end{aligned}$$

avec

$$(\underset{\lambda}{\text{SoftThreshold}}(X))(i,j)=\max(0,\ 1-\frac{\lambda}{\|X(i,j)\|_1})X(i,j).$$

1. Opérateur proximal de la régularisation ℓ_1.

Résoudre l'eq. (6.5) nécessite la minimisation d'une fonction convexe différentiable qui peut être réalisée par des méthodes classiques d'optimisation. Nous proposons ici de le résoudre de façon déterministe. Cette étape est la principale difficulté dans l'extension de [236] au cas multidimensionnel.
Définissons dans un premier temps H à partir de l'eq. (6.5) tel que :

$$X^{i+1} = \underset{X \in \mathbb{R}^{N_\Phi \times T}}{\arg\min}\ H(X)\ . \tag{6.10}$$

La différenciation de cette expression par rapport à X donne :

$$\begin{aligned}\frac{d}{dX}H = \ & (2\Phi^T\Phi + \mu_1 I_{N_\Phi})X + X(\mu_2 PP^T) \\ & - 2\Phi^T Y + \mu_1(D_A^i - A^i) + \mu_2(D_B^i - B^i)P^T\ .\end{aligned} \tag{6.11}$$

Le minimum $\hat{X} = X^{i+1}$ de l'eq. (6.5) peut alors être obtenu en résolvant $\frac{d}{dX}H(\hat{X}) = 0$, qui est une équation de Sylvester,

$$W\hat{X} + \hat{X}Z = M^i\ , \tag{6.12}$$

avec $W = 2\Phi^T\Phi + \mu_1 I_{N_\Phi}$, $Z = \mu_2 PP^T$ et $M^i = 2\Phi^T Y + \mu_1(A^i - D_A^i) + \mu_2(B^i - D_B^i)P^T$.

Pour rappel, la résolution de cette équation dans notre cadre est décrite dans l'annexe A. Dans la suite de ce chapitre, les matrices de passage des diagonalisations de W et de Z sont notées F et G. La matrice permettant la résolution du système diagonal est notée O.
La gestion des paramètres de régularisation est importante pour la stabilité de cette résolution et sera décrite dans la section 6.3.3 ainsi que d'autres détails d'implémentation tel que le calcul préliminaire de différents termes permettant d'accélérer les calculs de chaque itération.
Grâce à la résolution exacte du sous-problème de l'eq. (6.5), la convergence du schéma ci-dessus est assurée (voir l'annexe C). Néanmoins, il a été montré que le schéma du « split Bregman » peut converger même si le problème primal n'est pas résolu exactement. Une approche résolvant approximativement le problème primal pourrait donc être considérée pour accélérer l'algorithme. Lorsque N_Φ ou N_P prennent des valeurs élevées, cette possibilité pourrait permettre de réduire de façon significative le temps de calcul. En effet, lorsque ces dimensions augmentent, les dimensions des matrices F et G augmentent également et les transformations $\hat{X}' = F^T\hat{X}G$ et $M^{i\prime} = F^T M^i G$ deviennent coûteuses en temps de calcul.

6.3.2 Convergence

Suivant l'analyse de convergence de Osher et al [35], la convergence du schéma présenté précédemment peut être démontrée :

Théorème 5. *Sous l'hypothèse de positivité des paramètres du schéma* $\lambda_1 \geq 0$, $\lambda_2 \geq 0$, $\mu_1 > 0$ *et* $\mu_2 > 0$, *nous avons :*

$$\begin{aligned}&\lim_{i\to\infty} \|Y - \Phi X^i\|_F^2 + \lambda_1\|X^i\|_1 + \lambda_2\|X^i P\|_1 \\ &= \|Y - \Phi\hat{X}\|_F^2 + \lambda_1\|\hat{X}\|_1 + \lambda_2\|\hat{X}P\|_1\end{aligned} \tag{6.13}$$

dans lequel $\hat{X}$ est une solution du problème initial 6.2.
De plus, lorsque ce problème possède une solution unique, nous pouvons déduire de la convexité de la fonction $E(X) = \|Y - \Phi X\|_F^2 + \lambda_1\|X\|_1 + \lambda_2\|XP\|_1$ et de l'eq. (6.13) le résultat suivant :

$$\lim_{i\to\infty} X^i = \hat{X} \ . \tag{6.14}$$

La preuve de ce théorème est présentée dans l'annexe C.

6.3.3 Détails d'implémentation et gestion des paramètres de l'optimisation

Les termes W et Z (dans l'eq. (6.12)) sont indépendants de l'itération considérée (i). Leurs diagonalisations (pour la résolution de l'équation de Sylvester) peuvent donc être réalisées une fois seulement de même que le calcul de O :

$$\begin{aligned} &\forall n \in \{1, \cdots, N_\Phi\}, \quad \forall t \in \{1, \cdots, T\}, \\ &\qquad O(n,t) = D_w(n,n) + D_z(t,t). \end{aligned} \tag{6.15}$$

Ces diagonalisations peuvent être facilement obtenues depuis celles de $2\Phi^T\Phi$ et PP^T qui sont réalisées hors-ligne (pré-calculées).

$$2\Phi^T\Phi = F\Delta_w F^T, \qquad PP^T = G\Delta_z G^T \tag{6.16}$$

$$D_w = \Delta_w + \mu_1 I_{N_\Phi} \text{ and} \qquad D_z = \mu_2\Delta_z. \tag{6.17}$$

Ainsi, la mise à jour de Eq. (6.10) ne nécessite pas de calculs lourds, même lorsque les paramètres de pénalisation μ_1 et μ_2 sont modifiés au cours des itérations (voir le paragraphe suivant).
$Y_\Phi = 2F^T\Phi^T YG$ et $P_G = P^T G$ peuvent aussi être précalculés afin d'éviter d'inutiles calculs à chaque itération. De plus, le calcul de la fonction de coût réalisé à chaque itération dépend de multiplications en chaîne de trois matrices (par exemple $FX^{temp}G^T$). La durée du calcul de ces produits peut dépendre de l'ordre dans lequel les multiplications sont réalisées (les matrices étant rectangulaires). Les coûts de calcul des chaînes de produit apparaissant dans le schéma proposé dans la section précédente sont présentés dans la figure 6.2.

Chaine	Coût LR	Coût RL
$F^T(D_A^i - A^i)G$	$N_\Phi T(N_\Phi + T)$	$N_\Phi T(N_\Phi + T)$
FX^iG^T	$N_\Phi T(N_\Phi + T)$	$N_\Phi T(N_\Phi + T)$
$F^T(D_B^i - B^i)P_G$	$N_\Phi N_P(T + N_\Phi)$	$N_\Phi T(N_\Phi + N_P)$

FIGURE 6.2 – Coûts de calcul des chaînes de produits présentent dans l'algorithme. LR : ABC calculé avec $(AB)C$; RL : ABC calculé avec $A(BC)$.

Les coûts de calcul des deux premières chaînes ne dépendent pas de l'ordre des produits effectués, alors que la dernière en dépend. Ainsi, si $T \geq N_P$, alors $F^T((D_B^i - B^{temp})P_G)$ est calculé alors que dans le cas contraire $(F^T(D_B^i - B^{temp}))P_G$ est calculé.

Les choix de μ_1 et μ_2 ont un impact important sur le taux de convergence. Ils doivent d'une part être choisis de manière à bien conditionner le problème primal. D'autre part, ces paramètres peuvent être mis à jour au cours des itérations afin d'accélérer la convergence de la même manière que dans les techniques d'optimisation fondées sur le lagrangien augmenté [22].

Un mauvais conditionnement du problème primal apparaît lorsque ces paramètres sont trop petits du fait d'instabilités numériques dans la résolution du sous-problème de mise à jour de X (eq. (6.5)) lorsque les éléments de la matrice $O(n,t) = D_w(n,n) + D_z(t,t)$ sont proches de 0. Les valeurs propres de $W = 2\Phi^T\Phi + \mu_1 I_{N_\Phi}$ et de $Z = \mu_2 PP^T$ étant non négatives, ce conditionnement dépend directement des paramètres de pénalité μ_1 et μ_2.
Soit $\{\lambda_{\mathbf{w}}(n), \forall n \in \{1, \cdots, N_\Phi\}\}$ et $\{\lambda_{\mathbf{z}}(t), \forall t \in \{1, \cdots, T\}\}$ les valeurs propres respectives de $2\Phi^T\Phi$ et de PP^T. Nous avons :

$$\min_{n,t} O(n,t) = \min_n \lambda_{\mathbf{w}}(n) + \mu_1 + \mu_2 \min_t \lambda_{\mathbf{z}}(t) \ . \tag{6.18}$$

Lorsque Φ est un dictionnaire redondant, $\min_n \lambda_{\mathbf{w}}(n) = 0$ car cette valeur minimale correspond à la plus petite valeur singulière de Φ. Ainsi, en fonction des valeurs propres de PP^T, μ_1 et μ_2 doivent être choisis avec précaution pour éviter les instabilités numériques.

Une modification de ces paramètres au cours des itérations peut de plus accélérer la convergence de l'algorithme. Nous avons mentionnés précédemment que le schéma du « split Bregman » été équivalent à celui du lagrangien augmenté lorsque les contraintes du problème été linéaires. Ainsi, une stratégie communément utilisée avec ces approches a été choisie afin de mettre à jour ces paramètres après chaque itération : dès que le coût associé à une contrainte ne diminue plus suffisamment entre deux itérations, le paramètre correspondant est augmenté. Formellement, avec $h_1(X,A) = \|X - A\|_F$ et $h_2(X,B) = \|XP - B\|_F$ les coûts associés aux contraintes, les paramètres sont mis à jour comme suit :

$$\mu_1^i = \begin{cases} \mu_1^{i-1} & \text{if } \ h_1(X^i, A^i) < r_1 h_1(X^{i-1}, A^{i-1}) \\ \rho_1 \mu_1^{i-1} & \text{sinon} \end{cases} \tag{6.19}$$

et

$$\mu_2^i = \begin{cases} \mu_2^{i-1} & \text{if } \ h_2(X^i, B^i) < r_2 h_2(X^{i-1}, B^{i-1}) \\ \rho_2 \mu_2^{i-1} & \text{sinon} \end{cases} \tag{6.20}$$

avec r_1, r_2 les seuils associés aux coûts des contraintes et ρ_1, ρ_2 les coefficients des progressions géométriques de μ_1 et μ_2.
Afin d'obtenir une convergence rapide, l'initialisation de ces paramètres ne doit pas imposer les contraintes de façon trop stricte au cours des premières itérations. Ces paramètres ne

doivent pas non plus être initialisés à des valeurs trop faibles pour ne pas affecter la résolution du problème primal.
Une heuristique est présentée ici afin de faciliter l'initialisation de μ_1 et μ_2. Cette procédure s'est révélée efficace empiriquement :

1. création d'une grille g de paramètres pour μ_1 et μ_2,
2. calcul de la première itération du schéma d'optimisation pour chaque couple de paramètres $[g(j), g(l)]$,
3. évaluation de $t_1(j,l) = \frac{\mu_1}{2} h_1(X^1_{j,l}, A^1_{j,l})^2$ et $t_2(j,l) = \frac{\mu_2}{2} h_2(X^1_{j,l}, B^1_{j,l})^2$ pour chaque couple,
4. initialisation de μ_1^0 et μ_2^0 avec

$$\begin{aligned} \mu_1^0 &= g(\arg\max_j \sum_l t_1(j,l)), \\ \mu_2^0 &= g(\arg\max_l \sum_j t_2(j,l)). \end{aligned}$$

6.3.4 Algorithme complet détaillé

Paramètres : λ_1, λ_2, μ_1^0, μ_2^0, ε, $iterMax$, $kMax$, r_1, r_2, ρ_1, ρ_2

procedure MULTI-SSSA(Y, Φ, P)

Initialiser D_A^0, D_B^0, X^0 et définir $B^0 = X^0 P$, $A^0 = X^0$,
Diagonaliser $2\Phi^T\Phi$ et PP^T pour obtenir Δ_w, Δ_z, F et G (Eq. (6.16)).
Calculer D_w, D_z, depuis Δ_w, Δ_z, μ_1^0 et μ_2^0 (Eq. (6.17)).
Calculer O depuis D_w et D_z (voir annexe A).
Pré-calculer $Y_\Phi = 2F^T\Phi^T YG$ and $P_G = P^T G$.
$i = 0$
while $i \leq iterMax$ et $\frac{\|X^i - X^{i-1}\|_F}{\|X^i\|_F} \geq \varepsilon$ **do**
  $k = 0$
  $X^{temp} = X^i$; $A^{temp} = A^i$; $B^{temp} = B^i$
  for $k \to kMax$ **do**
    $M' = Y_\Phi - F^T(\mu_1^i(D_A^i - A^{temp})G$
      $+ \ \mu_2^i(D_B^i - B^{temp})P_G)$
    $X^{temp} = M' \oslash O$
    $X^{temp} = FX^{temp}G^T$
    $A^{temp} = \text{SoftThreshold}_{\frac{\lambda_1}{\mu_1^i}\|.\|_1}(X^{temp} + D_A^i)$
    $B^{temp} = \text{SoftThreshold}_{\frac{\lambda_2}{\mu_2^i}\|.\|_1}(X^{temp}P + D_B^i)$
  end for
  $X^{i+1} = X^{temp}$; $A^{i+1} = A^{temp}$; $B^{i+1} = B^{temp}$
  $D_A^{i+1} = D_A^i + (X^{i+1} - A^{i+1})$
  $D_B^{i+1} = D_B^i + (X^{i+1}P - B^{i+1})$
  Calculer μ_1^{i+1} et μ_2^{i+1} (eq. (6.19, 6.20)).

Mettre à jour D_w, D_z et O (eq. (6.17, 6.15))
$i = i + 1$
end while
return X^i
end procedure

6.4 Évaluation expérimentale de la rapidité du schéma proposé

Cette section est dédiée à l'évaluation de l'approche proposée en terme de rapidité. D'après notre étude de la littérature, la seule méthode proposée permettant de résoudre efficacement le problème du Fused-LASSO multidimensionnel (cf. l'eq. (6.2)) est une méthode de type gradient proximal [44]. Dans cette approche, la régularisation ℓ_1 non-différentiable est approchée par un terme différentiable et une méthode classique de descente de gradient accéléré est utilisée pour minimiser la fonction de coût approchée. Les formulations quadratiques proposées auparavant étant beaucoup plus lente que l'approche proximale [44] (en particulier lorsque les dimensions du problème grandissent), nous ne comparerons pas notre algorithme à celles-ci mais uniquement à cette approche proximale.

6.4.1 Protocole expérimental

Détails d'implémentation. Les deux approches ont été implémentées sous MATLAB (64 bits). Les expériences ont été réalisées sur un ordinateur possédant 16 GB de mémoire vive et un processeur à 8 cœurs.
Concernant l'approche proximale, la minimisation du coût est réalisée par l'algorithme FISTA comme présenté dans [44]. Plus précisément, la version de FISTA avec « backtracking » présentée dans [18] est utilisée car durant nos tests elle s'est avérée converger légèrement plus vite que la version sans backtracking. Le coefficient de Lipschitz L du gradient de la partie différentiable du coût est approché par une variable suivant une progression géométrique : $L^i = \rho^k L^{i-1}$ avec $\rho = 1.05$ et $L^0 = 1$. De plus, le paramètre μ modifiant la proximité entre le terme de régularisation et sa version différentiable est choisi de façon à ce que la précision souhaitée sur le résultat puisse être atteinte (voir ci-dessous).
En ce qui concerne l'approche proposée, les valeurs initiales des paramètres de pénalités (obtenus avec la méthode décrite dans la section 6.3.3) ont été observées stables pour les signaux considérés. Ces valeurs ont donc été fixées pour chacun de nos tests hors-ligne en réalisant une recherche sur une grille logarithmique de taille 20×20. Les mises à jour de ces paramètres de pénalité sont effectuées via l'approche présentée dans la section 6.3.3 avec $\rho_1 = \rho_2 = 1.05$ et $r_1 = r_2 = 0.95$. Les diagonalisations de $2\Phi^T\Phi$ et de PP^T sont réalisées hors-ligne.
Afin de comparer équitablement ces approches, la solution de notre problème convexe est dans un premier temps calculée de façon précise via l'approche proposée qui est arrêtée lorsque le changement relatif du coût entre deux itérations passe en dessous de 10^{-10} (le nombre maximum d'itération est fixé à 10000). Ensuite, les deux approches sont utilisées

	C	N_Φ	T
T1	100	200	$50 \ldots_{50} 500$ $600 \ldots_{100} 1000$ $2000 \ldots_{1000} 6000$
T2	100	$50 \ldots_{50} 500$ $600 \ldots_{100} 1000$ $2000 \ldots_{1000} 5000$	300
T3	$50 \ldots_{50} 500$ $600 \ldots_{100} 1000$ $2000 \ldots_{1000} 8000$	200	300

FIGURE 6.3 – Paramètres de l'expérience permettant d'évaluer la vitesse de l'approche proposée $m \ldots_n p = \{m + kn, \forall k \in \mathbb{N} \;\; s.t \;\; m + kn \leq p\}$.

pour retrouver cette solution et arrêtées lorsque l'erreur relative entre la valeur du coût à l'itération courante et la valeur du coût obtenu pour la solution calculée précédemment devient inférieure à une valeur de précision fixée. Comme expliqué plus haut, le paramètre μ du gradient proximal est calculé en fonction de cette précision. En effet, une borne fine de la distance entre le coût strict et le coût approché (avec le terme différentiable) est théoriquement connue [44] et est proportionnelle à μ. Par conséquent, afin d'obtenir $\mathit{coûtFinal} \times \mathit{précision} \geq \mathit{majorantÉcart} = K_g \mu$ où $\mathit{coûtFinal}$ est la valeur minimale de la fonction de coût et $K_g = \frac{1}{2} N_\Phi (T + N_P)$, μ est défini comme suit :

$$\mu = 0.95 \times \frac{\mathit{coûtFinal} \times \mathit{précision}}{K_g} \tag{6.21}$$

Tests. Le but ici est d'évaluer la vitesse de convergence des méthodes lorsque différentes dimensions du problème varient. Trois expériences ont été réalisées, leurs paramètres sont décrits ci-dessous dans la figure. 6.3. De plus, chaque test est réalisé pour différentes valeurs de la précision à atteindre définie plus haut : 10^{-4}, 10^{-5} et 10^{-6}.

Signaux et paramètres de régularisation. Les expériences ont été effectuées sur les signaux construits durant l'expérience décrite dans la section 6.5. La durée moyenne des activités est de 0.45 et leur nombre est fixé à 65 pour le *test 1*, le *test 2* et est défini en fonction du nombre de canaux dans le *test 3* tel que le ratio entre ces paramètres est égal à 0.65.
Le paramètre de régularisation λ_1 a été fixé tel que $\frac{\|Y - \Phi \hat{X}\|_F}{\|Y\|_F} \approx 0.1$ où $\hat{X}$ est le résultat

de l'exécution de la méthode et λ_2 choisi par validation croisée sur la distance entre les matrices de décomposition utilisées pour construire les signaux et celles obtenues en sortie des méthodes.

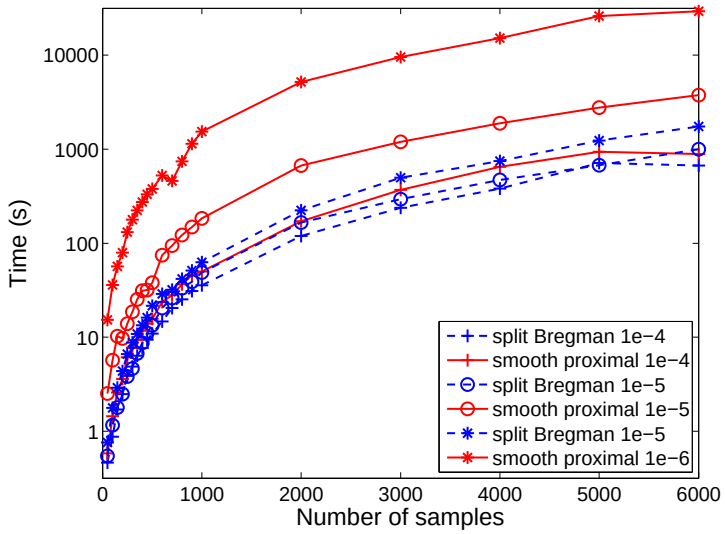


FIGURE 6.4 – Test 1 – « Split Bregman *vs* Smooth Proximal Gradient ». Comparaison des temps necessaires pour atteindre la solution avec une précision fixée lorsque le nombre d'échantillons du signal varie.

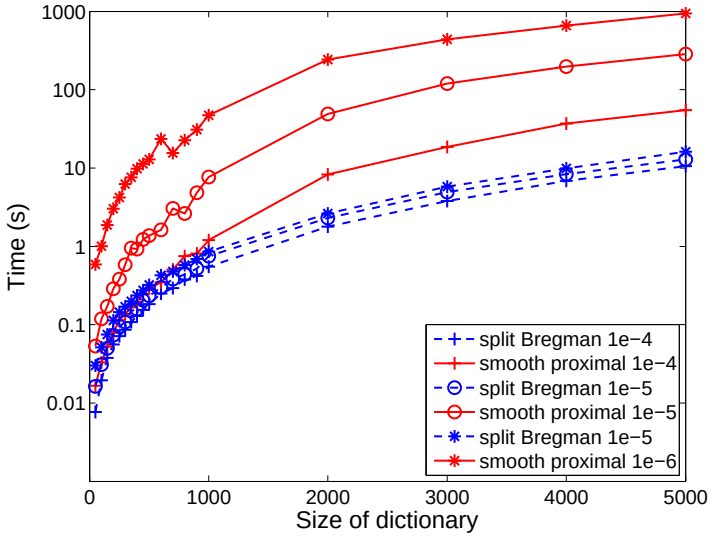


FIGURE 6.5 – Test 2 – « Split Bregman *vs* Smooth Proximal Gradient ». Comparaison des temps necessaires pour atteindre la solution avec une précision fixée lorsque la taille du dictionnaire varie.

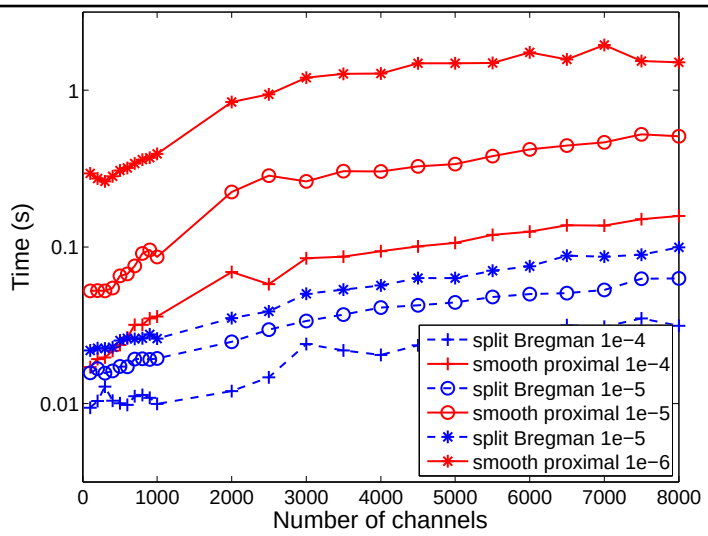


FIGURE 6.6 – Test 3 – « Split Bregman *vs* Smooth Proximal Gradient ». Comparaison des temps necessaires pour atteindre la solution avec une précision fixée lorsque le nombre de canaux varie.

6.4.2 Résultats et discussion

Les résultats sont résumés dans les figures 6.4, 6.5 et 6.6. Les temps d'exécution sont exprimés dans une échelle logarithmique.
Trois observations principales peuvent être faites à propos de ces courbes :

- Les temps de calcul des deux approches ne sont pas affectés par le nombre de canaux. Leurs augmentations sont dues au calcul des critères d'arrêt à chaque itération.
- L'approche proposée est plus rapide que l'approche proximale dans tous les cas testés ici (pour toutes les précisions : 10^{-4}, 10^{-5} et 10^{-6}) et les courbes ont les mêmes formes/tendances.
- Les différences de vitesse observées entre les approches deviennent plus importantes lorsque la précision souhaitée devient plus petite. Cette dernière observation peut être comprise en notant que le coefficient de Lipschitz du gradient de l'approximation du terme d'analyse présentée dans [44] est inversement proportionnel au paramètre μ. Ce coefficient correspondant à l'inverse du pas de gradient de FISTA (réalisé sur la partie différentiable), lorsque μ devient petit (assurant un petit gap et l'obtention de la précision souhaitée) le pas du gradient devient petit et la méthode est donc moins rapide.

6.5 Évaluation du modèle pour le recouvrement des structures sous-jacentes de signaux artificiels

La capacité du modèle présenté plus haut à retrouver la structure parcimonieuse en bloc sous-jacente des signaux artificiels est maintenant évaluée et comparée à celle d'autres régularisations classiques. Le Multi-SSSA est comparé à la fois à des méthodes décom-

posant les différents canaux séparément avec des termes de régularisation ℓ_0 ou ℓ_1 et les méthodes réalisant la décomposition simultanément sur les différents canaux avec des termes de régularisation $\ell_{2,0}$, $\ell_{2,1}$ et $\ell_{2,1} + \ell_1$. En ce qui concerne les régularisations ℓ_0 et $\ell_{2,0}$, les solutions sont respectivement approchées par l'OMP [181] et le SOMP [219]. Les solutions du problème avec contrainte ℓ_1 sont obtenues par l'algorithme LARS [66] et une approche proximale (FISTA [18]) a été choisie pour les régularisations $\ell_{2,1}$ et $\ell_{2,1} + \ell_1$ (utilisée dans [96]). Les problèmes de décomposition associés aux régularisations ℓ_1 et $\ell_{2,1}$ sont respectivement appelés dans les sections suivantes les problèmes du LASSO [210] et du groupe-LASSO [239].

6.5.1 Création des données

Pour cette expérience, un dictionnaire redondant aléatoire Φ fixé a été premièrement généré puis un ensemble de K signaux structurés par bloc créés à partir de ce dictionnaire. Chaque signal $Y^k, \quad k \in \{1, \cdots, K\}$ de cet ensemble est synthétisé à partir d'une matrice de décomposition X^k comme suit : $Y^k = \Phi X^k$.
Le terme de régularisation VT du fused-LASSO est particulièrement efficace pour retrouver la structure de données ayant des changements abrupts. Ainsi, des matrices de décomposition parcimonieuses par bloc ont été créées afin de mesurer l'efficacité et les limites du modèle du fused-LASSO multidimensionnel comparé à d'autres régularisations classiques sur de tels signaux. Les matrices de décomposition ont donc été construites comme combinaison linéaire d'activités spécifiques générées comme suit :

$$\Theta_{ind,m,d}(i,j) = \begin{cases} 0 & \text{si } i \neq ind \\ \mathcal{H}(j - (m - \frac{d \times T}{2})) - \mathcal{H}(j - (m + \frac{d \times T}{2})) & \text{si } i = k \end{cases}$$

où $\Theta \in \mathbb{R}^{N_\Phi \times T}$, $\mathcal{H}$ est la fonction de Heaviside, $ind \in \{1, \cdots, N_\Phi\}$ l'index d'un atome, m le centre de l'activité et d sa durée. Une matrice de décomposition X peut alors être écrit :

$$X = \sum_{i=1}^{n_a} a_i \Theta_{ind_i, m_i, d_i} \ ,$$

où n_a correspond au nombre d'activités apparaissant dans ce signal et les a_i sont les poids d'activation. Un exemple d'un tel signal est présenté dans la figure. 6.7.

6.5.2 Paramètres de l'expérience

Pour chaque signal Y, la distance entre la matrice de décomposition $\hat{X}$ obtenue à l'aide des différents algorithmes et la matrice utilisée pour construire le signal X est calculée par $\varepsilon(X, \hat{X}) = \frac{\|X - \hat{X}\|_F}{\|X\|_F}$. Cette distance est suffisante dans ce contexte étant donné que le dictionnaire créé de façon aléatoire possède une cohérence relativement faible. Le but est de comprendre l'influence du nombre d'activités (N_a) et de leurs durées (d) sur l'efficacité des différents modèles à retrouver les structures sous-jacentes des signaux.
Ces distances ont été calculées pour toutes les méthodes comparées sur la grille de paramètres suivante :

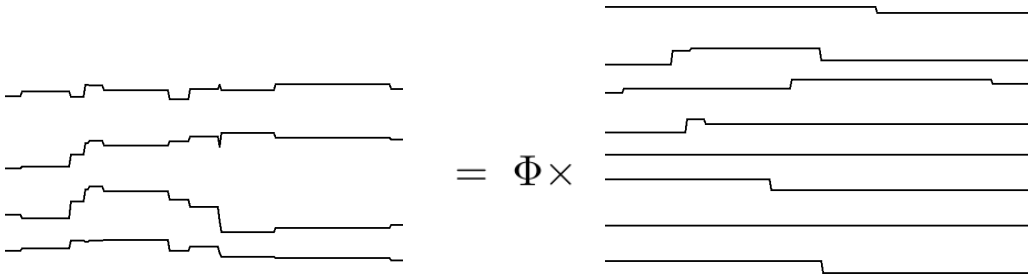


FIGURE 6.7 – Exemple de signal par blocs construit avec $C = 4$ canaux et $N_\Phi = 8$ atomes.

— $N_a \in \{20, 30, \cdots, 110\}$,
— $d \sim \mathcal{U}(d_{min}, d_{max})$
$(d_{min} d_{max}) \in \{(0.1, 0.15), (0.2, 0.25), \cdots, (1, 1)\}$.

Pour chaque point dans cette grille de paramètres, l'ensemble des signaux a été divisé en deux : un ensemble d'entraînement permettant de déterminer pour chaque méthode les meilleurs paramètres de régularisation et un ensemble de tests permettant l'évaluation proprement dite avec les paramètres calculés.
Le choix des autres paramètres est présenté dans la figure 6.8

Modèle		Activités	
$C = 20$	$T = 300$	$m \sim \mathcal{U}(0, T)$	$a \sim \mathcal{N}(0, 2)$
$N_\Phi = 30$	$K = 100$	$ind \sim \mathcal{U}(1, N_\Phi)$	

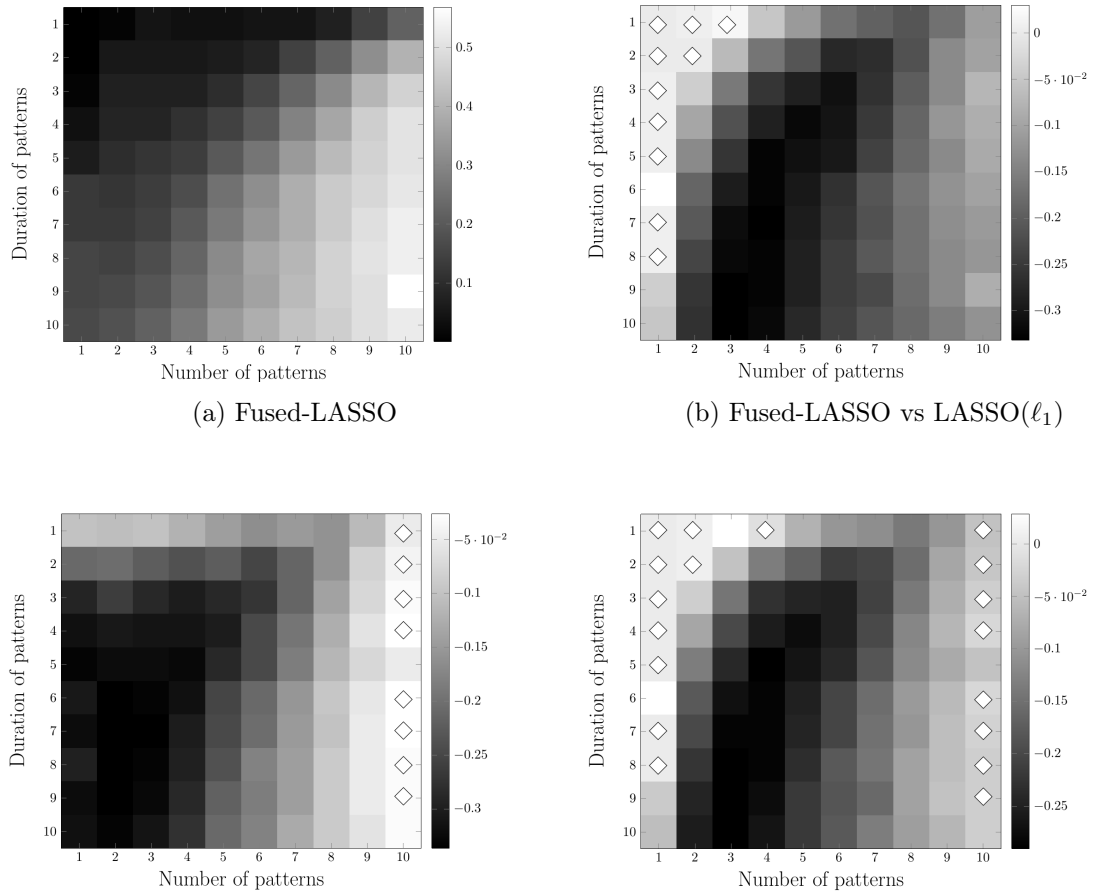
FIGURE 6.8 – Paramètres de l'expérience de récupération de la structure sous-jacente des signaux.

Le dictionnaire a été généré à partir d'atomes tirés indépendamment dans une distribution gaussienne et sa cohérence est donc faible.
Pour tester les différents algorithmes auxquels est comparé notre méthode, nous avons implémenté l'OMP et le SOMP tandis que la toolbox SPAMS [2] [118] a été utilisée pour les autres méthodes.

6.5.3 Résultats

Pour chaque point dans la grille de paramètres précédente, la moyenne (parmi les signaux de l'ensemble de tests) du critère ε décrit plus haut a été calculée pour chaque méthode et comparée à la moyenne obtenue par le Multi-SSSA. De plus, pour chaque couple de paramètres une analyse de variance suivie d'un test de Student apparié avec correction de

2. http ://spams-devel.gforge.inria.fr/



(a) Fused-LASSO

(b) Fused-LASSO vs LASSO(ℓ_1)

(c) Fused-LASSO vs Group-LASSO($\ell_{2,1}$

(d) Fused-LASSO vs $\ell_1 + \ell_{2,1}$

FIGURE 6.9 – (a) Distances moyennes obtenues avec le Multi-SSSA. (b), (c), (d) Différences entre les distances moyennes obtenues avec le Multi-SSSA et celles obtenues avec les autres méthodes. Les diamants blancs correspondent aux différences non-significatives.

Bonferroni ($p < 0.05$) ont été réalisés afin de vérifier si les différences de moyennes étaient significatives.
Les résultats sont présentés dans la figure 6.9. En ordonné, la durée des activités augmente du haut vers le bas et en abscisse le nombre de ces activités augmente depuis la gauche vers a droite. L'image (a) présente la distance moyenne obtenue par l'algorithme du Multi-SSSA. Sans surprise, la difficulté à approcher la vraie décomposition augmente avec le nombre d'activités et leurs durées. Les autres figures permettent la visualisation des comparaisons de l'algorithme du Multi-SSSA avec les autres approches en présentant les différences (point à point) des distances moyennes en échelle de gris. Ces différences sont calculées telles que les valeurs négatives (blocs sombres) correspondent aux couples de paramètres où la méthode proposée obtient de meilleurs résultats que les autres régularisations. Les diamants blancs correspondent aux différences de moyenne non-significatives. Les résultats de l'OMP et de l'algorithme résolvant le LASSO (ℓ_1) sont très similaires de même que ceux du SOMP et de l'algorithme résolvant le groupe-LASSO ($\ell_{2,1}$), les approches gloutonnes approchant les solutions ℓ_0 et $\ell_{2,0}$ étant tout de même moins efficaces. Par conséquent, nous ne présentons ici que les comparaisons de performances entre les solutions obtenues par le modèle du fused-LASSO et celles obtenues avec les modèles du LASSO, du groupe-LASSO et de la régularisation $\ell_1 + \ell_{2,1}$.

6.5.4 Discussion

Premièrement, il peut être noté que lorsque le nombre d'atomes actifs en même temps est petit, la régularisation ℓ_1 (et ℓ_0) et la régularisation $\ell_1 + VT$ obtiennent des résultats très similaires. Cela se produit dans nos données artificielles lorsque les signaux sont composés de peu d'activités ou lorsque ces activités ont des durées faibles. Au contraire, lorsque de nombreux atomes sont actifs en même temps, le fused-LASSO présente de bien meilleures performances que le LASSO retrouvant plus facilement la matrice de décomposition utilisée pour construire les signaux, le terme VT permettant de prendre en compte des informations intersignaux.
Concernant la régularisation $\ell_{2,1}$ (et $\ell_{2,0}$) les résultats dépendent principalement de la durée des activités. Lorsque la durée des activités est importante, les performances de cette régularisation sont similaires à celles du modèle présenté et au contraire lorsque la durée des activités est faible ou moyenne le fused-LASSO permet un meilleur recouvrement de la matrice de décomposition sous-jacente.
Ces résultats peuvent être facilement expliqués puisque cette régularisation choisie les atomes en fonction de tous les signaux C-dimensionnels et est donc beaucoup plus efficace lorsque les activités sont longues et que les atomes sont les mêmes pour ces signaux.
Enfin, concernant la comparaison avec la dernière régularisation $\ell_1 + \ell_{2,1}$, il peut être noté qu'elle combine les avantages des régularisations ℓ_1 et $\ell_{2,1}$, ayant des performances similaires au fused-LASSO lorsque le nombre d'activités simultanées est faible ou que ces activités sont longues. Son efficacité est meilleure que ces deux régularisations pour les couples de paramètres du milieu de notre grille mais reste en dessous de celle du fused-LASSO qui se comporte mieux pour les signaux ayant des changements brusques.

6.6 Application à la détection de potentiels évoqués P300

Le modèle décrit plus haut peut être adapté à différents contextes. Nous nous y sommes principalement intéressés ici dans l'optique d'une extension du modèle des micro-états ($\ell_1 + VT$) qui a été introduit dans le chapitre 4 et qui sera étudié dans le chapitre suivant. Néanmoins, d'autres types de modèle pour les EEG peuvent être considérés. Dans cette section, un modèle de décomposition temps-fréquence régularisé spatialement, proche de celui étudié dans le chapitre précédent, est appliqué pour le débruitage de P300 afin d'illustrer l'interêt de l'approche dans un cadre plus général.

6.6.1 Objectif et modèle

Nous souhaitons dans cette expérience décomposer des signaux EEG sur un dictionnaire temps-fréquence de Gabor de manière à obtenir des composantes plausibles physiologiquement grâce au modèle présenté plus haut. Pour cela, la matrice de régularisation en analyse P est construite comme le dual d'un dictionnaire spatial composé de topographies EEG réalistes.

Nous avons choisi ici de concevoir ce dictionnaire spatial en suivant les étapes suivantes :

- un modèle de tête réaliste a premièrement été créé à partir de données d'imagerie par résonance magnétique (d'un sujet quelconque), puis divisé en voxels,
- ensuite, pour chaque voxel, les topographies associées à différentes orientations d'activités électriques du voxel ont été calculées par résolution du problème direct EEG, à l'aide de l'algorithme OpenMEEG [95] et regroupées dans un dictionnaire de grande taille Φ_s ($\approx$ 5000 atomes),
- enfin, un sous-ensemble de ce dictionnaire a été sélectionné de manière gloutonne afin que la cohérence de celui-ci ne dépasse pas 0.9 ($\approx$ 350 elements).

Cette dernière étape permet d'améliorer le conditionnement du dictionnaire et de conserver des temps de décomposition raisonnables (temps de décomposition d'un signal $\approx$ 2 secondes).

Le dual de ce dictionnaire utilisé en tant que matrice de régularisation en analyse P est approché par le pseudo-inverse de Moore-Penrose [69].

6.6.2 Paramètres de l'expérience et protocole

Signaux. Comme dans le chapitre précédent, le jeu de données IIb de la seconde compétition BCI a été choisi pour l'évaluation de ce modèle. Ces signaux ont été enregistrés sur 64 électrodes à 240 Hz durant plusieurs sessions. Les essais considérés ici ont été extraits entre 150 et 450 ms après chaque stimulus et filtrés entre 0.3 et 10 Hz à l'aide d'un filtre de Butterworth d'ordre quatre.

Protocole. Deux régularisations sont comparées pour le débruitage de signaux EEG : une régularisation simple ℓ_1 ($L1$) et la régularisation étudiée ($MSSSA$). L'évaluation de ces régularisations est réalisée en deux étapes : chaque essai Y^k est premièrement décomposé à l'aide du MultiSSSA sur le dictionnaire Φ afin d'obtenir la matrice de décomposition X^k, puis les signaux reconstruits $\{\tilde{Y}^k = \Phi X^k | k \in \{1, \ldots, K\}\}$ sont classifiés.

Les paramètres de régularisation de ces décompositions sont appris sur la seconde session du jeu de données tandis que la première est utilisée pour la validation. Cette validation consiste à entraîner un classifieur à partir d'un ensemble de signaux choisi de manière aléatoire dans cette session puis à le tester sur les signaux restants.
L'expérience a été réalisée pour différentes tailles de l'ensemble d'entraînement : *taille ensemble d'entrainement* = *ratio train/test* × *nombre d'essais de la session*. Le *ratio train/test* prenant les valeurs suivantes : 0.5, 0.7, 0.9. Pour chaque ratio la validation est effectuée 50 fois avec des ensembles d'entraînement différents.

Classification. Le BLDA (Bayesian Linear Discriminant Analysis) [111] a été choisi pour l'étape de classification du fait de son efficacité pour la détection de potentiels évoqués P300.

6.6.3 Résultats et discussion

Les résultats de cette expérience sont présentés dans la figure 6.10.

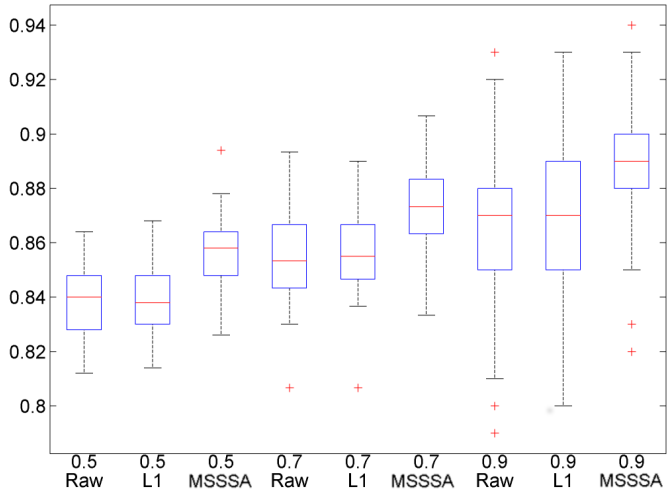


FIGURE 6.10 – Comparaison des taux de classification obtenus sur les signaux bruts et sur les signaux reconstruits après décomposition avec les régularisations $\ell_1(L1)$ et en analyse $(MSSSA)$ pour différents *ratio train/test*.

Il apparaît que la régularisation ℓ_1 ne permet pas d'améliorer les taux de classification obtenus sur les signaux bruts, au contraire de la régularisation étudiée. Un test de type « Wilcoxon signed-rank » montre en effet que pour l'ensemble des *ratio train/test* considérés, les résultats de l'approche ℓ_1 ne diffère pas significativement de ceux obtenus sur les signaux bruts, tandis que la régularisation proposée permet une amélioration significative ($(p < 0.01)$. De plus, lorsque le *ratio train/test* augmente, la variance des résultats obtenus sur les signaux bruts s'accroît de même que celle des résultats de la régularisation ℓ_1 mais pas celle des résultats de la régularisation proposée.

La régularisation proposée semble permettre une extraction de composantes plus intéressante pour la classification que celle obtenue à l'aide d'une régularisation parcimonieuse simple. Elle peut être envisagée comme étape de débruitage avant la classification d'essais

de P300 et plus généralement de signaux EEG.
De plus, les résultats obtenus ici pour la détection de P300 sont meilleurs que ceux obtenus dans la partie précédente avec les autres régularisations spatiales étudiées. Cette régularisation en analyse construite à partir de topographies réalistes semble ainsi plus efficace que les régularisations analytiques conçues précédemment, car plus « spécialisée » pour l'extraction de ce type de composantes.
Afin d'améliorer encore cette approche, il serait intéressant de construire le dictionnaire spatial en se basant non plus sur des données IRM d'un sujet quelconque pour la construction du modèle de tête, mais des données du sujet utilisant l'ICM. Par manque de temps, cette possibilité n'a pas été étudiée dans cette thèse.

6.7 Discussion globale

L'intérêt de l'algorithme du Multi-SSSA proposé est double. D'une part, il permet de réaliser une décomposition structurée de signaux sur un dictionnaire redondant : structure encodée dans la matrice P du terme de régularisation $\|XP\|_1$. D'autre part, la minimisation de la fonction de coût est réalisée efficacement, permettant d'utiliser cette approche en grandes dimensions grâce au schéma d'optimisation primale-duale du « split Bregman ».

Concernant la décomposition structurée, l'utilité de l'approche a été démontrée dans un premier temps à travers le cas du Fused-LASSO multidimensionnel et la capacité de ce modèle à retrouver la structure parcimonieuse par bloc de signaux artificiels. Le recouvrement de ce type de structures est significativement plus efficace avec cette régularisation qu'avec d'autres régularisations classiques. Dans un second temps, nous avons montré l'intérêt de cette régularisation dans le cadre d'une décomposition de signaux EEG sur un dictionnaire temps-fréquence pour le débruitage de signaux P300. Nous avons choisi pour cela de construire la matrice de régularisation en analyse de manière à obtenir des composantes possédant des topographies plausibles.
Le terme $\|XP\|_1$ permet ainsi d'imposer des propriétés structurelles sur la décomposition d'une façon générique en fonction des régularités supposées des signaux et il est même possible d'envisager l'apprentissage de cette matrice pour une classe particulière de signaux [183, 195] afin d'améliorer son efficacité.

Le second intérêt de la méthode réside dans l'utilisation du schéma d'optimisation du « split Bregman » pour la résolution du problème d'optimisation sous-jacent. Comme présenté dans la section 6.4, l'algorithme du Multi-SSSA est plus rapide que l'approche proximale, tout particulièrement lorsque les dimensions du problème grandissent. Sa vitesse est aussi moins sensible à la précision souhaitée. De plus, une heuristique s'étant montrée efficace empiriquement est proposée afin de régler les paramètres de pénalité de la méthode au mieux, préservant ainsi les bonnes propriétés de la méthode.
L'efficacité de ce type de schéma d'optimisation pour les problèmes ℓ_1 peut s'expliquer par une capacité à détecter rapidement les coefficients nuls et les coefficients non-nuls (voir l'annexe de [90]).
La principale limite du schéma proposé vient de la diagonalisation de certaines matrices

(Eq. (6.16), réalisée avec une complexité temporelle cubique par rapport à leurs tailles. Si les dimensions du dictionnaire Φ (N_Φ) et/ou de la matrice de régularisation venaient à augmenter fortement la diagonalisation pourrait devenir longue à réaliser. Une seconde limitation concerne la consommation de mémoire de la méthode lorsque les dimensions du problème grandissent.

Chapitre 7

Extension du modèle des micro-états

Sommaire

À l'aide de la décomposition régularisée temporellement introduite dans le chapitre précédent, nous proposons maintenant une extension du modèle des micro-états [137]. Nous allons nous intéresser dans un premier temps à ce modèle de manière formelle avant de voir comment il peut être étendu via une approche redondante et un algorithme d'apprentissage de dictionnaires. L'intérêt de l'extension proposée est ensuite évalué sur des signaux synthétiques, puis celle-ci est illustrée sur des signaux EEG réels pour l'extraction des micro-états du potentiel évoqué P300.

7.1 Modèle des micro-états

Comme nous l'avons déjà évoqué, le modèle des micro-états repose sur l'observation d'intervalles de stabilité spatiale des signaux EEG [134, 138, 136]. Ces signaux peuvent ainsi être modélisés par des suites de topographies restant stables quelques dizaines de

millisecondes chacune et présentant des transitions brusques entre elles. Les hypothèses sous-jacentes ont été discutées plus en détail dans la section 4.2.

7.1.1 Modèle

La modélisation classique de ces suites de topographies par des micro-états est présentée dans [180]. Celle-ci considère un modèle linéaire s'écrivant pour un signal $Y \in \mathbb{R}^{C \times T}$ comme suit :

$$Y = \Phi X + N \tag{7.1}$$

avec $\Phi \in \mathbb{R}^{C \times N_\Phi}$ une matrice non-redondante ($C \geq N_\Phi$) regroupant un ensemble de topographies normalisées, $i \in \{1, \ldots, N_\Phi\}$, $\|\Phi(i)\|_2 = 1$, N une matrice de bruit et X une matrice de coefficients dont chaque colonne ne comporte qu'un seul élément non nul : $\forall t \in \{1, \ldots, T\},\ \|X(t)\|_0 = 1$.
Dans ce modèle, les éléments de Φ sont appelés les micro-états bien qu'ils soient plus précisément des topographies représentant les états sous-jacents. Chaque topographie du signal Y est associée à l'un de ces états via un coefficient de X. Ce modèle ne suppose pas directement la stabilité temporelle des topographies, celle-ci est néanmoins assurée durant l'association de ces états avec les topographies des signaux.

7.1.2 Application de ce modèle à l'étude des potentiels évoqués

Nous avions évoqué dans le chapitre 2 que le modèle des micro-états était utilisé dans le cadre de la détection et de la caractérisation de maladies mentales comme la schizophrénie [135] ou la dépression [205]. Ce modèle peut également être considéré dans l'étude des PE (potentiels évoqués) [182, 228]. Les PE sont généralement étudiés en moyennant de nombreux enregistrements de réponses à un stimulus et en observant sur la moyenne obtenue les pics et les creux des différents canaux. Cette approche présente plusieurs inconvénients dont les principaux sont :

- la dépendance à la référence choisie pour les signaux, *i.e.* la manière dont sont exprimés les signaux par rapport à un potentiel particulier,
- l'impossibilité de différencier un changement de topographie d'un changement dans l'amplitude d'une topographie spécifique, même si certains travaux ont proposé des normalisations permettant de limiter cet effet [156].

L'étude présentée dans [166] propose une méthodologie pour l'étude des PE à l'aide du modèle des micro-états mettant notamment en avant l'avantage d'un tel modèle par rapport aux approches classiques. Les indices du GFP et du GMD que nous avons introduit dans la section 4.2 sont particulièrement intéressants dans ce contexte, le premier est indépendant de la référence et mesure la « force/amplitude » d'une topographie tandis que le second mesure la différence entre deux topographies sans tenir compte de leurs « forces/amplitudes ».

7.1.3 Algorithmes d'extraction des micro-états

L'étude présentée dans [180] propose de réaliser l'extraction des micro-états présents dans Φ grâce à un algorithme de partitionnement de données basé sur l'algorithme des K-

moyennes [60] permettant de diviser l'espace des signaux en N_Φ parties.
Pour un ensemble de K signaux $\{Y^k, \forall k \in \{1,\dots,K\}\}$ concaténés dans $Z \in \mathbb{R}^{C\times KT}$ le problème d'optimisation associé peut être écrit comme suit :

$$\begin{aligned}(\hat{X},\hat{\Phi}) = &\underset{X\in\mathbb{R}^{N_\Phi\times KT},\Phi\in\mathbb{R}^{C\times N_\Phi}}{\arg\min} \|Z-\Phi X\|_F^2 \\ t.q.\ \ &\forall t\in\{1,\dots,KT\},\quad \|X(t)\|_0 = 1 \\ &\forall n\in\{1,\dots,N_\Phi\},\quad \|\Phi(n)\|_2^2 \le 1\end{aligned} \tag{7.2}$$

Cette minimisation est effectuée de manière itérative. À chaque itération l'algorithme comprend deux étapes principales :

— l'association de chaque topographie du signal Y à un état de Φ,
— la mise à jour de ces états.

Durant la première étape de l'itération j, le choix de l'état associé à la t-ième topographie du signal $Z(t)$ est réalisé comme suit :

$$\mathbf{index}^j(t) = \underset{n\in\{1,\dots,N_\Phi\}}{\arg\max} \langle Z(t), \Phi^{j-1}(n)\rangle^2 \tag{7.3}$$

$$X^j(\mathbf{index}^j(t),t) = \langle Z(t), \Phi^{j-1}(\mathbf{index}^j(t))\rangle \tag{7.4}$$

Une fois ce choix réalisé pour toutes les topographies de Z, l'étape de mise à jour des atomes est effectuée en choisissant pour chaque état le vecteur propre associé à la plus grande valeur propre de la matrice de covariance des topographies associées à cet atome :

$$S_n^j = \sum_{n=1 \mid \mathbf{index}^j(t)=n}^{N_\Phi} Z(t)Z(t)^T \tag{7.5}$$

$$\Phi(n)^j = \underset{\mathbf{x}\in\mathbb{R}^{N_\Phi}}{\arg\max} \frac{\mathbf{x}^T S_n^j \mathbf{x}}{\mathbf{x}^T\mathbf{x}} \tag{7.6}$$

La méthode est arrêtée lorsque les éléments de Φ restent stables entre deux itérations.

Après extraction de ces états, un algorithme de lissage est appliqué afin d'obtenir des suites de coefficients respectant l'hypothèse de stabilité présentée plus haut pour la représentation des signaux. L'association des états aux topographies de Z est alors effectuée à l'aide du critère précédent (eq. (7.3)) auquel la régularisation suivante est ajoutée :

$$\Theta_{n,t}^b = \#\{\mathbf{index}(i) = n, i \in \{t-b,\dots,t+b\}\}\ . \tag{7.7}$$

Pour la topographie indicée par t et un atome n, ce terme est égal au nombre de topographies ayant été associées à n dont les indices se situent à une distance temporelle inférieure à b de t, avec b un paramètre à déterminer. Un paramètre de régularisation λ gérant l'équilibre entre le critère de l'eq. (7.3) et la régularisation doit également être déterminé.

Une étude plus récente a également proposée des approches de partitionnement pour réaliser l'extraction [166] de ces états. Une variante de l'algorithme des K-moyennes est

proposée ainsi qu'une méthode de partitionnement hiérarchique. Dans les deux cas, les algorithmes sont conçus de manière à maximiser un indice mesurant la variance expliquée par les micro-états extraits. Cet indice s'exprime à l'itération j comme suit :

$$GEV(Z,\Phi) = \frac{\sum_{i=1}^{KT}(GFP(Z(i))\langle Z(i), \Phi^j(\mathbf{index}(i)^j)\rangle)^2}{\sum_{i=1}^{KT} GFP(Z(i))^2}.$$

Il mesure la qualité de la description des données par les micro-états en prenant en compte la valeur de l'indice GFP, les topographies ayant des valeurs élevées étant considérées comme celles étant les moins bruitées et donc les plus informatives sur l'état sous-jacent.

Le nombre de micro-états à extraire est un paramètre important de ce modèle. Dans le papier fondateur [180], un critère de validation croisée généralisée sur la variance du résidu est utilisé pour la sélection de ce paramètre. Ce dernier étant très sensible au nombre d'électrodes, les travaux présentés dans [166] proposent de sélectionner ce paramètre en fonction du GEV ou d'un critère de Krzanowski-Lai [213].

7.2 Extension du modèle des micro-états à l'aide des dictionnaires

Le modèle des micro-états peut être étendu de manière naturelle dans un cadre redondant, l'algorithme de partitionnement étant alors remplacé par une méthode d'apprentissage de dictionnaire. Nous présentons ici une telle extension.

7.2.1 Modèle proposé

Dans le modèle des micro-états, chaque topographie d'un signal EEG est associée à un état et l'évolution de ces états est une fonction constante par morceaux. L'extraction des micro-états est donc un problème de quantification vectorielle (QV) contrainte temporellement. C'est-à-dire, un problème où l'on cherche à coder un ensemble de mesures successives (topographies) en les associant à un nombre limité de mots de code (micro-états) et en contraignant la succession des mots de code choisis.
L'apprentissage de dictionnaires peut être vu comme une généralisation de la quantification vectorielle. Au lieu de représenter un signal avec le mot de code le plus proche, il est représenté par une combinaison linéaire de quelques éléments du dictionnaire. L'algorithme K-SVD introduit dans la section 3.4 est notamment inspiré de l'algorithme des K-moyennes [5]. Cette généralisation semble encore plus naturelle dans le cas des micro-états car les mots de code sont normalisés et les signaux sont associés à ceux-ci par des coefficients.
Nous proposons donc d'étendre le modèle des micro-états à l'aide d'une représentation redondante dont le dictionnaire est obtenu via un algorithme d'apprentissage. Le modèle considéré présente ainsi deux différences majeures avec le modèle des micro-états : une contrainte de parcimonie moins stricte et la possibilité d'un dictionnaire contenant plus d'éléments que la dimension des topographies ($N_\Phi \geq C$). Ce modèle peut être décrit qualitativement par les propriétés suivantes :

— chaque topographie est associée à une combinaison de quelques éléments du dictionnaire,
— le nombre de micro-états actifs en même temps est faible,
— le dictionnaire des micro-états n'est pas limité en taille.

Formellement, ce modèle s'écrit comme suit :

$$\begin{aligned} & Y = \Phi X + N \\ & \text{t.q. } \forall t \in \{t, \dots, T\} \quad \|X(t)\|_0 \ll C \\ & \qquad \forall i \in \{1, \dots, N_\Phi\} \quad \|\Phi(i)\|_2^2 \leq 1 \end{aligned} \tag{7.8}$$

Dans ce modèle que nous proposons, la stabilité des micro-états est un *a priori* permettant l'obtention de représentations plausibles physiologiquement. Cette caractéristique est ainsi imposée dans les matrices de décomposition grâce à une régularisation. Dans ce cadre, le problème d'apprentissage de dictionnaires peut alors s'exprimer comme suit :

$$\begin{aligned} (\hat{X}^1, \dots, \hat{X}^K, \hat{\Phi}) = \underset{X^1,\dots,X^K \in \mathbb{R}^{N_\Phi \times T}, \Phi \in \mathbb{R}^{C \times N_\Phi}}{\arg\min} & \sum_{k=1}^{K} \|Y^k - \Phi X^k\|_F^2 + R(X^k) \\ \text{t.q } \forall k \in \{1, \dots, K\}, \forall t \in \{1, \dots, T\} & \quad \|X(t)^k\|_0 \leq J \\ \forall i \in \{1, \dots, N_\Phi\} & \quad \|\Phi(i)\|_2^2 \leq 1 \end{aligned} \tag{7.9}$$

avec $R(X)$ une régularisation encodant l'*a priori* temporel.
La conception d'une régularisation temporelle respectant les hypothèses du modèle des micro-états a été discutée dans le chapitre 4. Pour cela nous proposons, une régularisation combinant un terme de parcimonie et une pénalisation de type Variation Totale (VT). Nous considérons donc finalement le problème d'apprentissage de dictionnaires suivant :

$$\begin{aligned} (\hat{X}^1, \dots, \hat{X}^K, \hat{\Phi}) = \underset{X^1,\dots,X^K \in \mathbb{R}^{N_\Phi \times T}, \Phi \in \mathbb{R}^{C \times N_\Phi}}{\arg\min} & \sum_{k=1}^{K} \|Y^k - \Phi X^k\|_F^2 + \lambda \|X^k\|_1 + \mu \|X^k P\|_1 \\ \text{t.q } \forall i \in \{1, \dots, N_\Phi\} & \quad \|\Phi(i)\|_2^2 \leq 1 \end{aligned} \tag{7.10}$$

avec P une matrice de différence finie encodant la régularisation VT.

Intérêt de la régularisation temporelle pour l'apprentissage. L'extraction des micro-états à partir de cette minimisation prend en compte l'*a priori* de stabilité des suites de topographies (régularisation temporelle) contrairement aux approches de partitionnement que nous avons vu précédemment qui ne considèrent cet aspect qu'une fois les micro-états extraient, lors de la phase de décomposition des signaux sur ceux-ci. Grâce à cela, l'apprentissage est moins sensible aux bruits.
Comme nous l'avons vu précédemment, ces méthodes (incluant celle que nous proposons) ont en commun une approche itérative dont chaque itération est composée de deux étapes principales : une étape d'association des micro-états avec les topographies et une étape de mise à jour des états. Lorsque la stabilité des micro-états est prise en compte, chaque

topographie est associée de façon plus cohérente avec un micro-état car ce choix prend en compte les topographies voisines. La mise à jour des états se fait alors par rapport aux « bonnes » topographies et l'apprentissage donne de meilleurs résultats. Ce phénomène sera mis en lumière à travers une expérience sur des signaux artificiels dans la section 7.3.
Cette caractéristique de notre méthode constitue par conséquent une autre différence importante avec l'approche traditionnellement utilisée pour l'extraction des micro-états. Ne dépendant pas du nouveau modèle proposé, cette idée peut être appliquée directement au modèle classique des micro-états pour améliorer leur extraction.
À noter que l'utilisation des topographies possédant les plus grandes valeurs de l'indice GFP par les méthodes de partitionnement proposées récemment [166] vise également à réduire l'effet du bruit en ne considérant que les topographies les moins bruitées pour l'apprentissage.

Variance expliquée. Afin d'évaluer si la variance du signal est correctement décrite avec ce modèle, l'indice GEV peut être étendu pour le modèle proposé. Nous redéfinissons ainsi cette mesure pour un signal multidimensionnel $Y \in \mathbb{R}^{C\times T}$ décomposé sur le dictionnaire $\Phi \in \mathbb{R}^{C\times N_\Phi}$ avec la matrice de coefficients $X \in \mathbb{R}^{N_\Phi \times T}$ comme suit :

$$GEV(X,\Phi) = \frac{\sum_{t=1}^{T}(GFP(Y(t))\langle Y(t), \Phi(t)X(t)\rangle)^2}{\sum_{t=1}^{T} GFP(Y(t))^2}$$

Le modèle proposé vise à obtenir une meilleure description des signaux EEG tout en conservant une représentation respectant les régions de stabilité spatiale de ceux-ci. Chaque topographie EEG est approchée via plusieurs atomes du dictionnaire permettant ainsi de décrire plus efficacement les variances du signal. Nous pouvons donc espérer une augmentation de cet indice avec ce modèle, nous vérifierons cela dans la section 7.3.

7.2.2 Stratégie d'optimisation

Le problème d'apprentissage de dictionnaires qui nous occupe est un problème non-convexe que nous avons choisi de résoudre via un algorithme d'apprentissage en ligne. Ces approches sont moins sensibles aux *minima* locaux et le traitement successif des signaux permet un « redémarrage à chaud » de l'algorithme (l'algorithme peut être relancé dans le même état où il avait été arrêté).
Plus précisément, nous considérons l'algorithme proposé par Mairal et al. [149] présenté dans le chapitre 3 auquel nous nous référerons par l'abréviation ODL. Comme la plupart des algorithmes d'apprentissage de dictionnaires en ligne, pour chaque nouveau signal, deux tâches sont effectuées : la décomposition du signal sur le dictionnaire et la mise à jour de celui-ci. Les deux sous-problèmes associés étant convexes, ils possèdent des solutions uniques et peuvent être résolus efficacement.
Cet algorithme est défini pour un apprentissage monodimensionnel, néanmoins l'extension « mini-batch » proposé dans [149] permet un apprentissage multidimensionnel. À l'itération j, le sous-problème de décomposition peut être écrit pour le k-ième signal comme suit :

$$\hat{X}^k = \underset{X\in\mathbb{R}^{N_\Phi\times T}}{\arg\min} \quad \|Y^k - \Phi^j X\|_F^2 + \lambda\|X\|_1 + \mu\|XP\|_1$$

Ce problème a été étudié dans le chapitre précédent et nous utiliserons l'algorithme MultiSSSA pour le résoudre.
Concernant l'étape de mise à jour du dictionnaire, elle prend en compte les décompositions réalisées précédemment. A l'itération j, le problème d'optimisation associé s'écrit :

$$\Phi^{j+1} = \underset{\Phi \in \mathbb{R}^{C \times N_\Phi}}{\arg\min} \sum_{k=1}^{j} \|Y^k - \Phi X^k\|_F^2 \quad \text{t.q.} \quad \forall i \in \{1, \dots, N_\Phi\}, \quad \|\Phi(i)\|_2^2 \leq 1 \ . \tag{7.11}$$

L'algorithme ODL considère pour cette étape une reformulation de ce problème à l'aide de deux matrices B, C mise à jour à chaque itération comme suit :

$$B^{j+1} = B^j + X^k X^{k^T} \ ,$$
$$C^{j+1} = C^j + Y^k X^{k^T} \ ,$$

le problème s'écrivant alors :

$$\Phi^{j+1} = \underset{\Phi}{\arg\min} \ \ Tr(\Phi^T \Phi B^{j+1}) - 2Tr(\Phi^T C^{j+1})$$
$$\text{t.q.} \quad \forall i \in \{1, \dots, N_\Phi\}, \quad \|\Phi(i)\|_2^2 \leq 1 \ .$$

Cette minimisation est effectuée itérativement via une méthode de descente par bloc [149]. À chaque itération, les atomes du dictionnaire sont mis à jour grâce à une descente de gradient. Les valeurs des pas de ces descentes sont les éléments diagonaux de B qui approximent la valeur du hessien du coût à minimiser (pour les blocs de variables que constituent les atomes). Formellement, à chaque itération, le j-ième atome est mis à jour comme suit :

$$\text{Descente :} \qquad \mathbf{u} = \frac{1}{B(j,j)}(C(j) - \Phi B(j)) + \Phi(j)$$
$$\text{Normalisation :} \qquad \Phi(j) = \frac{1}{max(1, \|\mathbf{u}\|_2)}\mathbf{u}$$

Les matrices B et C représentent des éléments de mémoire (en plus du dictionnaire lui-même) de l'algorithme permettant de mettre à jour le dictionnaire en fonction de l'ensemble des décompositions effectuées jusque-là et pas uniquement en fonction de celle obtenue pour le signal courant.

Remarque. Il est également possible d'envisager pour cette dernière minimisation une résolution exacte. La solution du système linéaire suivant doit alors être calculée :

$$\Phi B^{j+1} = C^{j+1}.$$

7.2.3 Détails d'implémentation

Dans nos expériences, les signaux sont créés avant l'apprentissage des dictionnaires et n'arrivent donc pas progressivement. Par conséquent, nous avons choisi dans notre implémentation d'effectuer cet apprentissage en plusieurs cycles utilisant à chaque fois l'ensemble

des signaux. Dans chaque cycle, tous les signaux sont traités dans un ordre aléatoire. D'un cycle à l'autre, les éléments constituant la mémoire de l'algorithme sont conservés (Φ, B et C), la décomposition d'un même signal sera donc différente d'un cycle à l'autre du fait des changements réalisés sur le dictionnaire.
Une fois le premier cycle terminé, nous avons également choisi de supprimer après chaque décomposition les composantes des matrices de mémoire B et C correspondantes aux plus anciennes décompositions, ces anciennes décompositions ayant été effectuées avec un dictionnaire très différent du dictionnaire courant.

Le problème n'étant pas convexe, l'initialisation du dictionnaire influe directement sur le résultat de l'apprentissage. Par conséquent, nous avons choisi d'initialiser les atomes de notre dictionnaire avec les topographies du jeu de données présentant les plus grandes valeurs pour le GFP, celles supposées les moins bruitées.

7.3 Evaluations expérimentales

Nous allons maintenant considérer différentes expériences permettant d'évaluer l'efficacité de l'approche proposée. Cette évaluation est réalisée dans un premier temps sur des signaux synthétiques réalistes construits de manière à respecter l'*a priori* de stabilité supposé dans les signaux EEG. L'application de notre apprentissage à des signaux réels est ensuite effectuée pour l'analyse du potentiel évoqué P300.

7.3.1 Signaux synthétiques

Deux expériences sur signaux synthétiques sont premièrement effectuées pour l'évaluation de notre approche. Celles-ci évaluent l'intérêt des modifications apportées au modèle des micro-états dans l'approche proposée. La première évalue l'intérêt de la prise en compte de la stabilité des topographies au cours de l'apprentissage des micro-états dont nous avons discuté plus haut. La deuxième s'attache ensuite à examiner l'intérêt du relâchement de la contrainte de parcimonie lorsque plusieurs atomes sont actifs simultanément dans les signaux.

Critère. Dans ces deux expériences, la mesure de l'efficacité des méthodes est effectuée en comparant le dictionnaire appris avec le dictionnaire utilisé pour la synthèse des signaux. Dans la littérature relative à l'apprentissage de dictionnaires, le dictionnaire appris est souvent comparé au dictionnaire optimal en comptant le nombre d'atomes du dictionnaire optimal retrouvé dans le dictionnaire appris. Un atome est alors considéré comme retrouvé lorsque sa corrélation avec un atome du dictionnaire optimal est supérieure à un seuil. Ce critère présentant une dépendance très forte au seuil et étant non continu nous avons choisi d'utiliser une approche différente avec une métrique présentée récemment dans [45]. Nous considérons plus précisément la distance de Wasserstein construite à partir d'une distance euclidienne entre atomes pour notre critère (voir [45] pour la description de cette métrique).

7.3.1.1 Expérience 1

Protocole expérimental Afin d'évaluer l'avantage procuré par la prise en compte de la structure temporelle des signaux dans l'extraction des micro-états, nous nous plaçons ici dans un contexte simple correspondant au modèle initial des micro-états avec des signaux pour lesquels un seul état est actif à un instant donné.

Les signaux créés pour cette expérience ont été synthétisés à partir d'un dictionnaire fixe Φ. Chaque signal est construit à partir de ce dictionnaire par la création d'une matrice de décomposition possédant des coefficients constants par morceaux de telle façon qu'un seul atome soit actif à un instant donné. Le nombre de changements d'état des signaux est fixé (arbitrairement) à 4. Le dictionnaire choisi possède une faible cohérence $\mu_1(\Phi) \leq 0.2$. Nous supposons en effet que les micro-états EEG sont faiblement corrélés au vu des états extraits dans les différentes études de la littérature.
Ces signaux sont bruités pour différentes valeurs de RSB : $5db$, $0db$, $-5db$, $-10db$. Pour chaque niveau de bruit, deux apprentissages de dictionnaire sont effectués à l'aide de l'algorithme ODL dans lequel la phase de décomposition est effectuée via l'algorithme OMP avec une parcimonie de 1 (l'atome ayant la plus grande corrélation avec la topographie est choisi). Dans l'un de ces deux apprentissages, après chaque décomposition, les coefficients de décompositions sont lissés à l'aide de l'algorithme décrit dans [180] dont le terme de régularisation a été introduit dans la section 7.1.3. Les dictionnaires ainsi appris sont évalués en les comparants avec le dictionnaire utilisé pour la création des signaux via le critère introduit plus haut.

Dans chacun des cas étudiés, les paramètres λ et b associés à la régularisation de lissage sont déterminés dans un premier temps sur un ensemble de K signaux d'entraînement. Une fois ces paramètres fixés, l'évaluation des approches est réalisée sur K autres signaux.
Pour chaque apprentissage, la taille du dictionnaire appris est fixée à la même valeur que celle du dictionnaire utilisé pour la création des signaux. Le problème considéré étant non convexe, l'obtention d'un minimum global n'est pas assuré et donc les performances des méthodes sont évaluées en effectuant la moyenne des scores obtenus sur $N_l = 20$ apprentissages.
Les choix effectués pour les autres paramètres du modèle sont présentés dans la figure 7.1.

Modèle			
$C = 20$	$T = 300$	$N_\Phi = 30$	$K = 100$

FIGURE 7.1 – Paramètres du modèle pour l'évaluation de l'intêret de la régularisation temporelle dans la récupération des structures par blocs sous-jacentes.

Résultats et discussion Le résultat de cette expérience est présenté dans la figure 7.2. Des tests de type « wilcoxon signed rank » révèlent que des différences significatives de

	5 db	0 db	-5 db	-10 db	-15 db	-20 db	Moy
ODL, OMP	**0.2622**	**0.3135**	0.5498	0.7961	0.9852	1.0911	0.6663
ODL, OMP + lissage	0.2689	0.3137	**0.4085**	**0.6315**	**0.9450**	**1.0330**	**0.5887**

FIGURE 7.2 – Comparaison d'apprentissage de dictionnaires avec et sans régularisation temporelle pour différents niveaux de bruit.

performances n'aparaissent qu'à partir de $-5db$. Il apparaît ainsi que pour des valeurs élevées du RSB la régularisation temporelle est inutile et qu'une décomposition à l'aide de l'OMP seule donne des résultats similaires. Au contraire, lorsque le RSB diminue, la régularisation de lissage devient intéressante et permet un meilleur apprentissage du dictionnaire. Cela s'explique par un choix plus cohérent des atomes malgré le bruit lorsque la régularisation de lissage est utilisée.

7.3.1.2 Expérience 2

Nous souhaitons maintenant évaluer le modèle proposé dans son entier et en particulier l'intérêt du relâchement de la contrainte de parcimonie permettant à plusieurs atomes d'être actifs simultanément.

Protocole expérimental Cette évaluation est réalisée sur les mêmes signaux que ceux utilisés pour la validation du MultiSSSA dans le chapitre 6, *i.e.* des signaux présentant des structures sous-jacentes constantes par morceaux. De la même façon qu'auparavant, l'expérience est effectuée pour différents nombres d'activités présentes dans les signaux et différentes valeurs de la durée de ces activités avec un dictionnaire de cohérence faible. Le processus de création de ces signaux a été décrit dans la section 6.5.

Pour chaque cas étudié (nombre d'activités, durées des activités), deux dictionnaires sont appris via l'algorithme ODL avec des algorithmes de décomposition différents. Dans l'un de ces apprentissages, la décomposition est effectuée de manière similaire à l'expérience 1 avec l'algorithme OMP arrêté après la première itération, suivi d'un algorithme de lissage alors que pour l'autre, l'algorithme MultiSSSA est appliqué. Les dictionnaires appris sont comparés avec le dictionnaire utilisé pour la création des signaux via la métrique que nous avons mentionnée auparavant (Wasserstein).

Les paramètres associés aux deux régularisations des algorithmes de décomposition utilisés sont déterminés pour chaque configuration sur un ensemble de $K = 100$ signaux d'entraînement. L'évaluation proprement dite est réalisée ensuite sur K signaux tests.
Comme précédemment, les dictionnaires appris possèdent le même nombre d'atomes que le dictionnaire utilisé pour la création des signaux. De plus, du fait de la non-convexité du problème, l'efficacité des méthodes a été évaluée en moyennant les résultats obtenus sur $N_l = 20$ apprentissages.

Résultats et discussion. La figure 7.3 décrit les résultats de cette expérience.

ODL	Durée des activités				
FusedLasso/OMP+lissage	**0.1**	**0.3**	**0.5**	**0.7**	**0.9**
Nb Act : 20	0.0948/0.2348	0.2184/0.4077	0.4188/0.5548	0.5300/0.7156	0.5949/0.7730
Nb Act : 30	0.0847/0.1888	0.3134/0.4634	0.4982/0.6449	0.5917/0.7298	0.5973/0.7864
Nb Act : 40	0.0812/0.1616	0.3979/0.5867	0.5754/0.6713	0.5863/0.7903	0.6434/0.8247
Nb Act : 50	0.0994/0.2048	0.4619/0.5838	0.5746/0.7401	0.6189/0.8063	0.6364/0.7805

FIGURE 7.3 – Comparaison d'apprentissage de dictionnaires : modèle des micro-états (en rouge) *versus* extension proposée (en bleu).

Quelque soit le cas observé, l'approche proposée présente de meilleures performances que l'algorithme d'extraction des micro-états. Le relâchement de la contrainte de parcimonie permet donc d'améliorer l'apprentissage du dictionnaire lorsque les structures de décomposition sous-jacentes présentent effectivement plusieurs atomes actifs simultanément.
Dans le cadre des signaux EEG, il semble raisonnable de prendre comme hypothèse que les différentes activités cérébrales ne sont pas disjointes temporellement. Notre approche semble donc adaptée à l'apprentissage des composantes spatiales sous-jacentes à ces signaux et nous allons donc l'appliquer maintenant à des signaux EEG réels.

7.3.2 Signaux réels

Nous allons maintenant appliquer la méthode proposée sur des enregistrements du potentiel évoqué P300. N'ayant pas de « vérité terrain » concernant les composantes sous-jacentes du PE P300, il est difficile d'évaluer les résultats obtenus. Par conséquent, nous comparerons dans cette section les décompositions obtenues avec l'extension proposée à celles obtenues avec le modèle classique des micro-états du point de vue de la capacité de description de ces deux approches.

7.3.3 Données

Pour cette expérience nous considérons le même jeu de données (IIb de la deuxième compétition ICM [26]) que celui utilisé pour l'évaluation des régularisations spatiales étudiées dans le chapitre 5. Pour rappel, ces signaux ont été mesurés sur 64 électrodes à une fréquence d'échantillonnage de 240Hz. Ce jeu de données comporte des enregistrements obtenus sur un seul sujet au cours de 3 sessions de mesure.
Pour cette expérience, les essais ont été extraits dans l'intervalle 0-600 ms après chaque stimulus lumineux et filtrés pour ne conserver que les fréquences entre 0.3 et 20 hz avec un filtre de Butterworth d'ordre 4.
Une fois cette extraction effectuée, la moyenne des essais des deux classes (P300 et non P300) est calculée pour chaque session de mesure. Pour l'analyse d'ERP, c'est sur ces signaux moyens que l'analyse des micro-états est effectuée et donc c'est sur eux que nous allons évaluer l'approche proposée.

7.3.4 Protocole expérimental

Nous comparons dans cette expérience les deux mêmes approches que dans l'expérience précédente, *i.e.* des apprentissages de dictionnaire réalisés à l'aide de l'algorithme ODL pour deux méthodes de décomposition différentes : OMP avec lissage temporel et fused-LASSO multidimensionnel.
L'objectif de l'expérience est d'observer les décompositions obtenues, les atomes appris, ainsi que la quantité de variance expliquée avec ces deux approches lorsque les ruptures des signaux sont respectées dans les décompositions. Pour cela, les paramètres de régularisation de ces deux approches ont été déterminés sur les essais de notre jeu de données en calculant dans un premier temps les ruptures de ces signaux à l'aide de l'indice GMD, puis en apprenant des dictionnaires avec différents paramètres de régularisation (grille de paramètres) et en observant les décalages temporels entre les ruptures obtenus dans les décompositions et celles déterminés précédemment avec l'indice GMD.
Une fois ces paramètres fixés, la capacité des approches à expliquer les variations du PE P300 est évaluée en apprenant les dictionnaires sur les signaux moyens de deux sessions et en décomposant ensuite le signal moyen de la troisième sur ceux-ci. Cette évaluation est réalisée en prenant comme session de test chacune des trois sessions et en moyennant les résultats pour 20 apprentissages de dictionnaires.
La description des signaux moyens des deux classes d'essais (P300 et non P300) est ensuite obtenue pour les deux approches en apprenant les dictionnaires sur les signaux moyens des deux classes pour l'ensemble des sessions, puis en décomposant les grandes moyennes (moyenne sur tous les essais) de chaque classe à l'aide des dictionnaires appris.

7.3.5 Résultats et discussion

La figure 7.4 présente la mesure de l'indice GEV obtenue pour les deux approches lorsque le taille du dictionnaire appris augmente. Il apparait que l'extension proposée permet d'expliquer la variance du signal plus efficacement que le modèle classique des micro-états. Ce résultat confirme l'hypothèse que nous avions faite auparavant et provient directement du relâchement de la contrainte de parcimonie du modèle redondant étudié.
À noter pour la lecture de cette figure le fait que l'indice GEV est calculé sur des décompositions régularisées temporellement et donc prend des valeurs plus faibles que les valeurs de la littérature présentées généralement pour des décompositions non contraintes. Le calcul de l'indice GEV pour des décompositions non contraintes n'est pas possible pour le modèle considéré étant donné que ni le nombre d'atomes actifs simultanément ni la taille du dictionnaire redondant ne sont contraints directement dans le modèle.

Il est également interessant d'observer la représentation obtenue via cette approche pour le potentiel évoqué étudié. Cette représentation est présentée dans la figure 7.5 pour les deux approches comparées et les signaux moyens des deux classes.

Il est premièrement intéressant de noter que les deux approches font apparaître les mêmes ruptures dans les signaux et que les zones de stabilité correspondent effectivement aux instants où l'indice GFP est maximum.

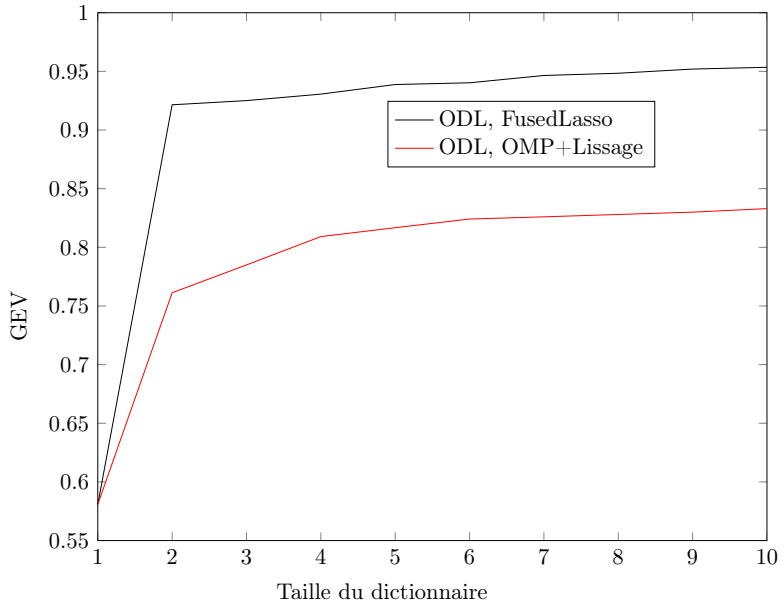


FIGURE 7.4 – Valeur de l'indice GEV pour la description du PE P300 à l'aide de décompositions régularisées temporellement sur des dictionnaires de micro-états.

Concernant le modèle classique, les micro-états extraits ressemblent fortement à ceux de la littérature [137]. De plus, la séparation en deux parties du P300 est une caractéristique connue de ce dernier qui a notamment été décrite dans [180]. Ainsi, l'approche considérée pour les extraire (ODL+OMP+lissage) semble efficace et permet d'obtenir des résultats conforment à ceux obtenus avec les algorithmes classiques d'extraction de micro-états.
En ce qui concerne le modèle étendu, la décomposition obtenue comporte plus d'activités de faible durée et de ruptures que celle du modèle classique. Le modèle étendu prenant en compte la possibilité de superposition des activités, il est possible de détecter les ruptures entre celles-ci avec plus de finesse même lorsque les topographies de celles-ci sont similaires. Cette caractéristique permet notamment de détecter des changements d'activités n'apparaîssant pas dans les évolutions de l'indice GFP. Pour le P300 par exemple, nous pouvons observer qu'il est composé dans cette décomposition de 4 activités. Les deux atomes les plus marqués (coefficients plus forts pour le 4 et le 6) ressemblent fortement aux atomes extraits dans le modèle classique mais deux autres atomes sont activés également avec des coefficients plus faibles. Les intervalles temporels d'activités de ces atomes n'étant pas disjoints, le modèle classique ne peut les faire apparaître.

7.4 Conclusion

Dans ce chapitre nous nous sommes intéressés au modèle des micro-états et à la manière dont il peut être étendu naturellement à l'aide d'une représentation redondante.
Un modèle des micro-états étendu a été premièrement proposé pour cela. Ce modèle présente deux différences majeures avec le modèle classique des micro-états :

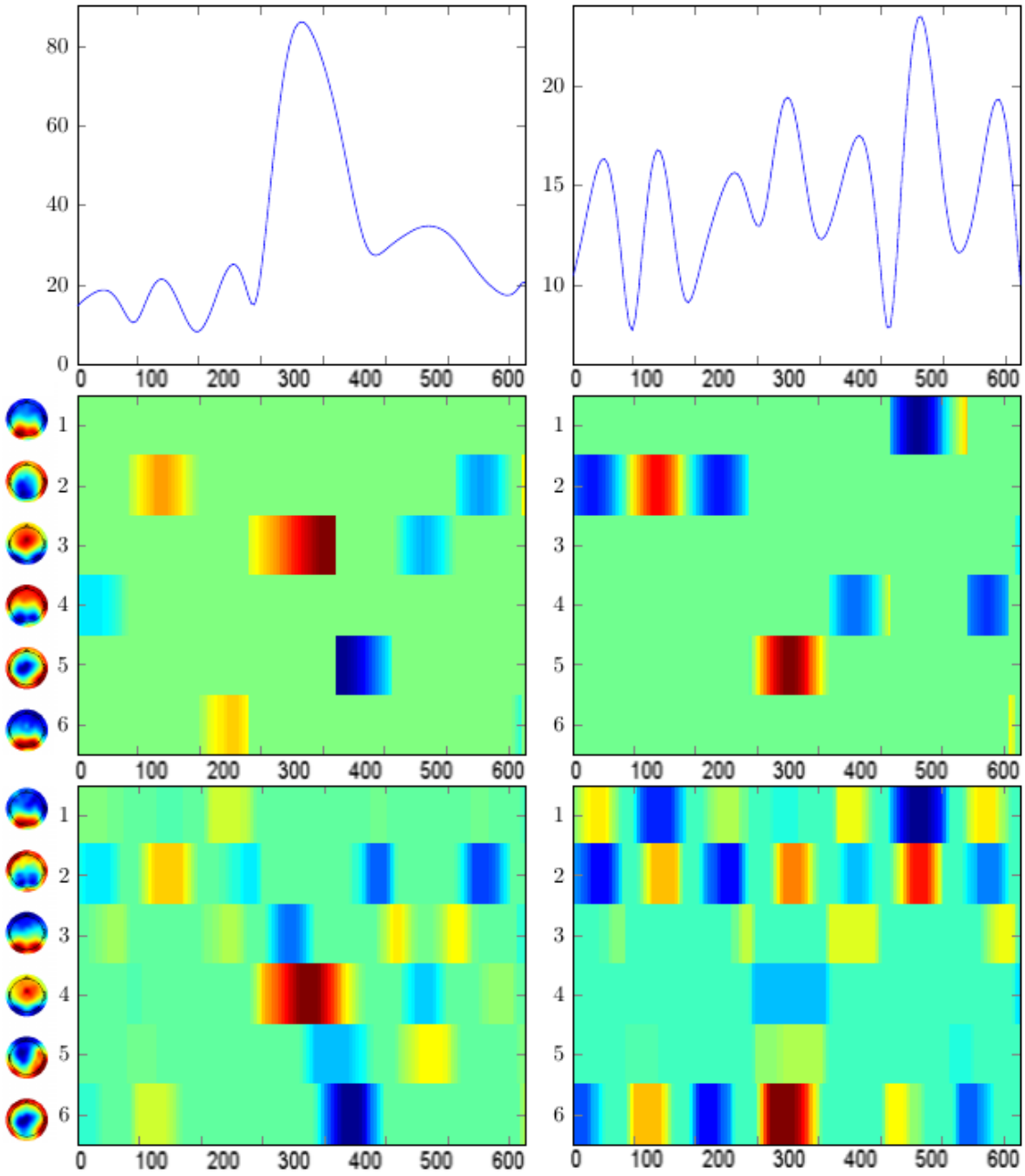


FIGURE 7.5 – Description du PE P300 à l'aide du modèle des micro-états. La colonne de gauche correspond au PE P300 et celle de droite à la réponse standard aux flashs lumineux. Dans chaque colonne : en haut, l'évolution de l'indice GFP est présentée, au centre, la matrice de décomposition pour le modèle des micro-états et en bas, cette même matrice pour une décomposition avec le modèle étendu. Les 144 points de mesure correspondent aux 600 milisecondes après stimulus.

— il autorise les topographies des signaux EEG à être représentées par des combinaisons linéaires d'atomes (relâchement de la contrainte de parcimonie),
— le nombre d'atomes possibles n'est pas limité à la taille des topographies (nombre d'électrodes).

Nous avons ensuite proposé pour le problème d'apprentissage de dictionnaires associé à ce modèle une stratégie d'optimisation fondée sur l'algorithme d'apprentissage online proposé dans [149]. L'étape de décomposition parcimonieuse de cette approche est remplacée par l'algorithme MultiSSSA développé dans le chapitre précédent afin d'effectuer une régularisation temporelle permettant d'obtenir des matrices de coefficients constantes par morceaux.

L'approche développée a finalement été évaluée sur des signaux synthétiques puis illustrée sur données réelles à travers l'analyse du potentiel évoqué P300. L'évaluation sur données artificielles a permis de mettre en avant l'intérêt de la régularisation temporelle pour l'apprentissage du dictionnaire ainsi que l'intérêt du relâchement de la contrainte de parcimonie. L'application de cette approche pour l'analyse du PE P300 a permis ensuite de mettre en avant la souplesse du modèle proposé ainsi que sa capacité à expliquer la variance des signaux et à faire apparaître leurs ruptures même lorsque les activités des atomes ne sont pas disjointes temporellement.

L'approche proposée semble permettre une description plus fine des phénomènes apparaissant dans les EEG. Néanmoins, afin de s'assurer que cette décomposition reflète réellement les phénomènes neurophysiologiques sous-jacents une validation sur plusieurs sujets devra être envisagée dans de futurs travaux.

7.5 Pistes d'amélioration

Nous proposons maintenant deux pistes d'amélioration du modèle étudié ainsi que des idées d'heuristiques pouvant améliorer l'algorithme d'apprentissage du dictionnaire. La première piste considère l'intégration de la mesure de l'indice GFP dans l'apprentissage et la seconde l'incorporation d'une régularisation spatiale permettant l'obtention d'atomes plausibles physiologiquement. Ces extensions du modèle ne sont pas incompatibles et peuvent être considérées simultanément. Leur principal inconvénient réside dans l'ajout de paramètres supplémentaires dont les valeurs peuvent être difficile à choisir. En ce qui concerne l'algorithme d'apprentissage, nous discuterons dans la suite d'heuristiques permettant de régler d'une part le degré de parcimonie des décompositions et d'autre part de contrôler les éléments de mémoire de l'algorithme.

7.5.1 Prise en compte de l'indice GFP

L'indice GFP d'un signal présente des valeurs élevées dans les intervalles où celui-ci possède des topographies stables, ce qui en fait une mesure importante pour le modèle des micro-états. Les instants correspondant à des maximums de cette mesure sont ceux associés avec les topographies les moins bruitées et sont donc ceux utilisés en priorité pour apprendre une représentation. C'est la raison pour laquelle les algorithmes décrits dans [166] tentent de

maximiser l'indice GEV qui est une mesure de la variance expliquée pondéré par les valeurs du GFP, donnant ainsi une importance plus grande à la représentation des topographies associées à ces maximums.
Il semble par conséquent naturel de vouloir faire de même dans notre extension en pondérant l'importance de la bonne approximation des topographies en fonction de l'indice GFP. Pour cela, le problème d'optimisation que nous avons introduit plus haut peut être modifié comme suit :

$$(\hat{X}^1, \dots, \hat{X}^K, \hat{\Phi}) = \underset{X^1,\dots,X^K,\Phi}{\arg\min} \sum_{k=1}^{K} \|(Y^k - \Phi X^k)D\|_F^2 + \lambda\|X^k\|_1 + \mu\|X^k P\|_1$$
$$\text{t.q.} \quad \forall i \in \{1, \dots, N_\Phi\} \quad \|\Phi(i)\|_2^2 \leq 1$$

avec D une matrice diagonale de pondération dont les éléments dépendent de la valeur du GFP à chaque instant :

$$\forall i \in \{1, \dots, T\}, \quad D(i,i) = f(GFP(Y(i))),$$

f réglant l'influence de l'indice GFP sur l'apprentissage. Dans un premier temps, la fonction identité peut être envisagée pour f.
Dans un tel problème d'optimisation, les erreurs d'approximation obtenues sur les topographies ayant une valeur élevée du GFP pénalise le coût plus fortement que les autres et sont donc diminuées en priorité, aboutissant ainsi à de meilleures approximations de celles-ci. La minimisation de ce coût peut être effectuée via la même approche que celle proposée plus haut, néanmoins certaines modifications des algorithmes considérés doivent être effectuées.

7.5.1.1 Décomposition parcimonieuse

Concernant la décomposition parcimonieuse, c'est la mise à jour primale du schéma d'optimisation (voir section 6.3.1) qui doit être modifiée. Plus précisément, c'est la mise à jour de X (équation 6.5) qui doit être adaptée, la différenciation de l'expression étant différente. Ainsi, en utilisant les mêmes notations que celle de la section 6.3.1, soit H la fonction matricielle définie par :

$$H = \|(Y - \Phi X)D\|_F^2 + \frac{\mu_1}{2}\|X - A^i + D_A^i\|_F^2 + \frac{\mu_2}{2}\|XP - B^i + D_B^i\|_F^2,$$

nous avons :

$$\begin{aligned}\frac{d}{dX}H = \; & 2\Phi^T\Phi X D D^T + X(\mu_1 I_{N_\Phi} + \mu_2 P P^T) \\ & - 2\Phi^T Y D D^T + \mu_1(D_A^i - A^i) + \mu_2(D_B^i - B^i)P^T \quad .\end{aligned}$$

Le minimum $\hat{X}$ de H peut alors être obtenu en résolvant $\frac{d}{dX}H(\hat{X}) = 0$. L'équation obtenue rentre dans le cadre des équations de Sylvester généralisées :

$$AXB + XC = M,$$

avec $A = 2\Phi^T\Phi$, $B = DD^T$, $C = \mu_1 I_{N_\Phi} + \mu_2 PP^T$ et $M = 2\Phi^T YDD^T + \mu_1(A^i - D_A^i) + \mu_2(B^i - D_B^i)P^T$.
B est une matrice diagonale donc les valeurs sont les valeurs aux carrés de D :

$$\forall i \in \{1, \ldots, T\}, B(i,i) = D(i,i)^2.$$

Lorsque les éléments diagonaux de D ne sont pas nuls (caractéristique qui dépend de la fonction f), B est inversible avec pour inverse une matrice diagonale :

$$\forall i \in \{1, \ldots, T\}, B^{-1}(i,i) = \frac{1}{D(i,i)^2}.$$

Le problème ci-dessus peut alors être transformé en un problème de Sylvester classique :

$$AX + XCB^{-1} = MB^{-1}$$

Les matrices A et CB^{-1} étant toujours symétriques réelles, la résolution de ce dernier problème peut être effectuée via l'approche décrite dans l'annexe A.
Dans le cas où certains des éléments de D sont nuls, le problème d'optimisation défini plus haut peut être modifié en supprimant les colonnes de Y (et donc de X) correspondantes aux éléments nuls de D et résolu avec la même approche.

7.5.1.2 Mise à jour du dictionnaire

L'étape de mise à jour du dictionnaire choisie plus haut doit également être adaptée. Le problème d'optimisation à résoudre pour cette mise à jour s'écrit maintenant :

$$\Phi^{j+1} = \arg\min_{\Phi} \sum_{k=1}^{j} \|(Y^k - \Phi X^k)D\|_F^2 \text{ t.q. } \forall i \in \{1, \ldots, N_\Phi\} \ \|\Phi(i)\|_2^2 \leq 1 \ .$$

De la même façon que précédemment, il est possible de reformuler ce problème en considérant les matrices $\bar{B}$ et $\bar{C}$ suivantes :

$$\bar{B}^{j+1} = \bar{B}^j + X^k DD^T X^{kT} \ ,$$
$$\bar{C}^{j+1} = \bar{C}^j + Y^k DD^T X^{kT} \ .$$

Le problème à résoudre est alors le même que celui présenté plus haut :

$$\Phi^{j+1} = \arg\min_{\Phi} Tr(\Phi^T\Phi\bar{B}^{j+1}) - 2Tr(\Phi^T\bar{C}^{j+1})$$
$$\text{t.q. } \forall i \in \{1, \ldots, N_\Phi\} \ \|\Phi(i)\|_2^2 \leq 1 \ ,$$

est peut donc être résolu de la même façon.
Ainsi, seul le calcul des matrices mémorisant les anciennes décompositions diffère pour la résolution de ce nouveau problème.

7.5.2 Régularisation spatiale des atomes

Afin de guider l'apprentissage proposé vers une solution physiologiquement plausible, la mise en place d'une régularisation spatiale appliquée aux atomes peut être envisagée.
De la même façon que pour la régularisation spatiale considérée pour une décomposition temporelle et décrite dans le chapitre 4, il est possible ici d'imposer une régularité aux atomes appris à l'aide d'une régularisation de lissage. Le problème d'apprentissage de dictionnaire s'écrit alors :

$$(\hat{X}^1,\dots,\hat{X}^K,\hat{\Phi}) = \underset{X^1,\dots,X^K,\Phi}{\arg\min} \quad \sum_{k=1}^{K} \|Y^k - \Phi X^k\|_F^2 + \lambda\|X^k\|_1 + \mu\|X^k P\|_1 + \eta\|\Phi^T L\|_F^2$$
$$\text{t.q.} \quad \forall i \in \{1,\dots,N_\Phi\} \quad \|\Phi(i)\|_2^2 \leq 1$$

avec L une matrice de régularisation encodant l'*a priori* de régularité spatiale (opérateur laplacien par exemple) et η le paramètre de régularisation associé.
Cette contrainte n'affecte pas l'étape de décomposition parcimonieuse des signaux mais la méthode effectuant la mise à jour du dictionnaire doit être adaptée pour la prendre en compte. En conservant les structures de mémoire utilisées dans l'approche précédente, le problème d'optimisation à résoudre pour cette mise à jour est maintenant le suivant :

$$\Phi^{j+1} = \underset{\Phi}{\arg\min} \quad Tr(\Phi^T\Phi B^{j+1}) - 2Tr(\Phi^T C^{j+1}) + \eta\|\Phi^T L\|_F^2$$
$$\text{t.q} \quad \forall i \in \{1,\dots,N_\Phi\}, \quad \|\Phi(i)\|_2^2 \leq 1 \ .$$

Cette minimisation est réalisable de manière exacte via la résolution de l'équation de Sylvester (obtenue après différentiation du coût) suivante :

$$\Phi B^{j+1} + \eta LL^T\Phi = C^{j+1},$$

dont une méthode de résolution est présentée dans A. Où en adaptant l'algorithme de descente par bloc proposé dans [149] :

$$\text{Descente :} \quad \mathbf{u} = \frac{1}{B(j,j) + \eta(LL^T)(j,j)}(C(j) - \Phi B(j) - \eta LL^T\Phi(j)) + \Phi(j)$$
$$\text{Normalisation :} \quad \Phi(j) = \frac{1}{\max(1, \|\mathbf{u}\|_2)}\mathbf{u}$$

7.5.3 Amélioration de l'ODL

Nous proposons maintenant des améliorations de l'algorithme d'apprentissage pouvant être intéressantes pour notre application, mais également dans un cadre plus général. Ces améliorations concernent la gestion de la parcimonie et la gestion de la mémoire aux cours des itérations.

Gestion de la parcimonie. Au cours de nos expériences sur données artificielles, nous avons observé les évolutions de la distance entre le dictionnaire utilisé pour créer les signaux et le dictionnaire en cours d'apprentissage.

Nous avons remarqué qu'une parcimonie forte dans les décompositions permettait d'améliorer l'apprentissage du dictionnaire dans les premiers cycles (boucle sur l'ensemble des signaux) mais qu'ensuite l'apprentissage « stagnait ». Au contraire, avec une parcimonie moins forte, l'apprentissage est moins rapide au démarrage, mais arrive parfois (quand il ne tombe pas dans un minimum local avant) à s'approcher plus de la solution que celle obtenue avec une parcimonie forte. Cela peut s'expliquer par un meilleur apprentissage des composantes principales des signaux avec une parcimonie forte et un meilleur apprentissage des détails (composantes plus faibles) de ces signaux lorsque la parcimonie est plus faible.
Ainsi, une heuristique faisant évoluer la parcimonie des décompositions au cours des itérations en modifiant le paramètre de régularisation associée peut être envisagée de manière à assurer une parcimonie forte dans les premiers cycles, puis une parcimonie de plus en plus faible.
Considérons par exemple que nous souhaitons obtenir un dictionnaire sur lequel les signaux ont une parcimonie correcte (celle supposée pour la classe de signaux considérée) lorsque la décomposition est effectuée avec $\lambda = \hat{\lambda}$ et que $\bar{\lambda}$ est la valeur de λ limite au-dessus de laquelle la décomposition d'un signal (de l'ensemble étudié) donne une matrice de décomposition nulle. Une heuristique linéaire simple pourrait alors s'écrire :

$$\lambda(t) = \hat{\lambda}(t - t_i) + \bar{\lambda}(t - t_f)$$

ou t représente une variable d'avancement de l'apprentissage progressant entre les valeurs t_i (initiale) et t_f (finale), comme le nombre d'itérations ou de cycle déjà effectués ou alors une variable déterminée par exemple en fonction de la stabilité du dictionnaire courant entre deux itérations (cette stabilité devenant plus forte lorsque l'apprentissage progresse).

Gestion de la mémoire. Concernant la mémoire de l'algorithme, elle est présente dans le dictionnaire lui-même ainsi que dans les matrices B et C. Or, les contributions des anciennes décompositions aux matrices B et C ne sont plus pertinentes pour la mise à jour du dictionnaire courant lorsqu'elles ont été effectuées avec un dictionnaire très éloigné de celui-ci. Par conséquent, une stratégie peut être mise en place pour supprimer les contributions trop anciennes. Dans notre cas, comme nous l'avons expliqué plus tôt, après le premier cycle, à chaque décomposition effectuée, la contribution de la décomposition la plus ancienne est supprimée. Les matrices B et C sont donc constituées des K dernières contributions.
Afin d'améliorer cela, une heuristique de gestion de la mémoire basée sur la distance entre le dictionnaire courant et les précédents dictionnaires pourrait permettre de ne conserver que les composantes de B et C pertinentes pour la mise à jour du dictionnaire courant.

Chapitre 8

Conclusion et perspectives

Nous concluons maintenant ce document en effectuant un rappel du déroulement de l'étude et des contributions effectuées pendant cette thèse, avant de considérer les perspectives ouvertes ainsi que les travaux envisageables afin de la prolonger.

8.1 Déroulement de l'étude et contributions

Cette thèse s'est attachée à l'étude de représentations redondantes pour les signaux EEG.
Une étude bibliographique détaillée des applications ainsi que des méthodes d'analyse de ces signaux a été effectuée dans un premier temps. Celle-ci a permis de mettre en évidence l'intérêt de ce type de représentations pour leur étude, ces approches étant trés flexibles et permettant une description précise des signaux (en plus de leur interêt pour la résolution du problème inverse).
Nous nous sommes ensuite intéressés de manière approfondie à ces représentations à travers une seconde étude bibliographique. La flexibilité de ces approches fait apparaître un problème mal posé pour l'obtention de telles représentations. Des contraintes de parcimonie sont généralement choisies pour y remédier et nous avons donc étudié les caractéristiques théoriques de telles décompositions et passé en revue les algorithmes liés à celles-ci. La dernière partie de cette étude bibliographique a enfin été consacrée à l'application de ces méthodes aux signaux EEG, et les différentes approches de la littérature ont été passées en revue.

Ces études nous ont conduit à centrer nos travaux sur la conception de régularisations pour la décomposition de signaux EEG sur des dictionnaires redondants ; ces contraintes étant très importantes pour l'obtention de représentations stables de ces signaux. Plus spécifiquement, nous avons choisi de créer ces régularisations pour la réalisation de décompositions dans l'espace des électrodes n'utilisant qu'un seul enregistrement du phénomène étudié (« single-trial »), ces choix ayant été effectués notamment dans l'optique d'une application de ces décompositions pour la conception d'interfaces cerveau-machine. Ces régularisations ont été conçues dans le but d'obtenir des décompositions plausibles physiologiquement et sont fondées sur les caractéristiques spatiales et temporelles supposées de ces signaux.
Pour une décomposition temporelle régularisée spatialement, nous avons ainsi proposé une régularisation permettant l'obtention de décompositions plausibles en prenant en compte la régularité des topographies des activités cérébrales de même que la localisation des maxima de ces topographies. Des algorithmes ont ensuite été développés pour la résolution du problème de décomposition associé sous ses formes ℓ_0 et ℓ_1. Ces méthodes ont été validées sur

des données artificielles réalistes puis appliquées dans un contexte discriminatif à la détection de potentiels évoqués PE P300. Cette approche s'est révélée efficace pour un choix plus consistant des atomes du dictionnaire temporel et intéressantes pour une meilleure compréhension des phénomènes étudiés.
La régularisation temporelle que nous proposons ensuite se fonde sur le modèle des micro-états et plus particulièrement sur l'hypothèse consistant à supposer que les suites de topographie des signaux EEG présentent des changements brusques et donc une évolution par blocs. Le problème d'optimisation ℓ_1 associé à cette régularisation rentre dans le cadre des décompositions régularisées en analyse dans un cadre multidimensionnel. Afin de pouvoir effectuer rapidement ce type de décompositions, nous avons proposé alors dans un contexte plus général non lié aux signaux EEG un algorithme primal-dual fondé sur le schéma d'optimisation du « split Bregman » pour sa résolution. L'algorithme développé s'est montré plus rapide que les algorithmes de la littérature. Cet algorithme a été validé dans un premier temps sur des signaux synthétiques, pour lesquels, la régularisation considérée s'est révélée plus efficace que d'autres régularisations classiques dans le recouvrement des composantes d'un signal respectant l'hypothèse d'évolution par blocs. Puis, dans un second temps, l'intérêt du modèle a été illustré sur des signaux EEG réels pour la détection de P300, en considérant pour la régularisation en analyse le dual d'un dictionnaire composé de topographies réalistes afin de guider les décompositions vers des composantes plausibles.
Enfin, nous avons proposé une extension du modèle des micro-états grâce à cette dernière régularisation et à l'algorithme conçu pour effectuer la décomposition associée. Cette extension se fonde sur un apprentissage de dictionnaire spatial et permet une description plus fine des signaux en autorisant notamment l'activité simultanée de plusieurs atomes. Notre approche a été illustrée sur des signaux EEG réels pour l'extraction des états composant le potentiel évoqué P300.

8.2 Perspectives et travaux futurs

Il est possible d'envisager à partir de ces travaux de futures études permettant d'étendre les approches développées ici ou bien de les utiliser dans d'autres contextes. Nous en décrivons ici quelques-unes.

8.2.1 Apprentissage de dictionnaire multicanal temporel invariant par translation

Nous avons déja évoqué dans la section 3.5.2.4 une étude ayant appliqué un algorithme d'apprentissage de dictionnaires invariant par translation aux signaux EEG mais pour un modèle multivarié dans lequel les atomes sont bi-dimensionnelles (section 3.5.2.1). Après l'étude effectuée dans le chapitre 5, il semble naturel d'envisager cet apprentissage en multicanal afin de comparer les dictionnaires appris avec ces deux approches. Les régularisations spatiales proposées étant utilisées pour guider les décompositions des signaux vers des solutions plausibles physiologiquement et permettant de guider l'apprentissage.
L'algorithme proposé pour l'apprentissage multivarié dans [16] est un algorithme « online » alternant de façon classique entre la décomposition des signaux sur le dictionnaire

et la mise à jour de ce dernier. Sa particularité réside dans l'utilisation de noyaux pouvant être translatés sur toute la durée du signal pour la représentation de celui-ci, le dictionnaire utilisé pouvant alors être vu comme la concaténation des différents noyaux translatés pour chaque pas temporel. Afin de gérer l'ensemble de ces translations efficacement, les décompositions parcimonieuses sont réalisées à l'aide d'un algorithme de type OMP dans lequel le critère de choix des atomes repose sur la corrélation croisée entre chaque noyau pour tous les décalages temporels et le signal à décomposer, permettant à chaque itération le choix d'un noyau et de la translation optimale de celui-ci pour la représentation du signal. La mise à jour des noyaux est ensuite réalisée grâce à une descente de gradient.
L'adaptation de cet algorithme à un cas multicanal peut être réalisée directement en modifiant l'algorithme de décomposition pour réaliser la corrélation croisée de chaque noyau monodimensionnel pour tous les décalages temporels avec les canaux du signal.
Ainsi pour une telle décomposition, avec régularisation $\ell_{2,0}$, soit un ensemble de N_ψ noyaux $\{\Psi_n, \forall n \in \{1, \dots, N_\psi\}\}$ (Pas necessairement de même taille) dont toutes les translations sont réunies dans $\Phi \in \mathbb{R}^{T \times TN_\Psi}$, à l'itération i le choix des atomes peut être réalisé pour un résidu U^i en résolvant le problème suivant :

$$\begin{aligned} & \underset{\mathbf{x},j,v}{\arg\min} \quad \|U^i - \Psi_j^v \mathbf{x}^T\|_F^2 \\ \Leftrightarrow \quad & \underset{j,v}{\arg\max} \quad \sum_{c=1}^{C} (\langle U^i(c), \Phi^v(j) \rangle)^2 \\ \Leftrightarrow \quad & \underset{j,v}{\arg\max} \quad \|(\Psi_j^v)^T U^i\|_F^2 \end{aligned}$$

où l'on note $\Psi_j^v = \Phi(v + T(j-1))$ l'atome du dictionnaire Φ correspondant au noyau j translaté de v pas temporels. La résolution de ce problème peut être facilitée par le calcul de la corrélation croisée entre les noyaux et les canaux du résidu qui peut être effectué rapidement par un produit de convolution grâce à l'algorithme classique TFR (« transformée de fourier rapide »). Une fois les atomes choisis, la projection sur ceux-ci est réalisée de la même façon que dans l'algorithme SOMP.

Lorsqu'une régularisation spatiale est considérée, nous avons vu dans le chapitre 5 que le choix des atomes pour la décomposition associée est effectué en modifiant la norme utilisée grâce à la matrice $Q = (I_C + LL^T)^{-1}$ (pour une matrice de régularisation spatiale L). En appliquant cela ici, le problème devient :

$$\begin{aligned} & \underset{j,v}{\arg\max} \quad \langle\ (\Psi^v(j))^T U^i Q^T,\ \ (\Psi^v(j))^T U^i\ \rangle, \\ \Leftrightarrow \quad & \underset{j,v}{\arg\max} \quad \|\Psi^v(j))^T U^i\|_{Q^T}^2. \end{aligned}$$

Comme précédemment, le calcul de la corrélation croisée entre le résidu et les noyaux permet d'obtenir la solution de ce problème rapidement même si cette fois-ci la norme quadratique $\|.\|_{Q^T}$ doit être appliquée ensuite sur le résultat de cette corrélation. La projection sur les atomes peut ensuite être réalisée comme décrit dans le chapitre 5.
L'adaptation de la régularisation du « Latent Group Lasso » dans ce cadre est plus complexe mais est réalisable avec le même type de méthode.

La mise à jour des noyaux n'est pas impactée par le passage à un modèle multicanal où les régularisations et peut être réalisée par une descente de gradient comme dans l'algorithme initial.

Il pourrait s'avérer par exemple intéressant d'appliquer cette approche à l'apprentissage des composantes du PE P300 et de les comparer avec celles apprises à l'aide de l'algorithme multivarié.

8.2.2 Comparaison de modèles de régularisation spatiale

Plusieurs types de régularisation peuvent être envisagés pour l'obtention de décompositions sur des dictionnaires temporels plausibles physiologiquement. Nous avons proposé dans le chapitre 5 différentes régularisations spatiales fondées sur des hypothèses de localisation spatiale des activités cérébrales et de régularité des topographies associées. À partir des mêmes hypothèses, d'autres modèles de régularisation sont possibles.
Parmi ceux-ci, il nous semble intéressant d'en étudier tout particulièrement deux utilisant un dictionnaire spatial d'atomes gaussien 2D noté ici Φ_s. Une visualisation de quelques atomes d'un tel dictionnaire est présentée dans la figure 8.1.
Le respect des hypothèses rappelées plus haut par les composantes obtenues durant une dé-

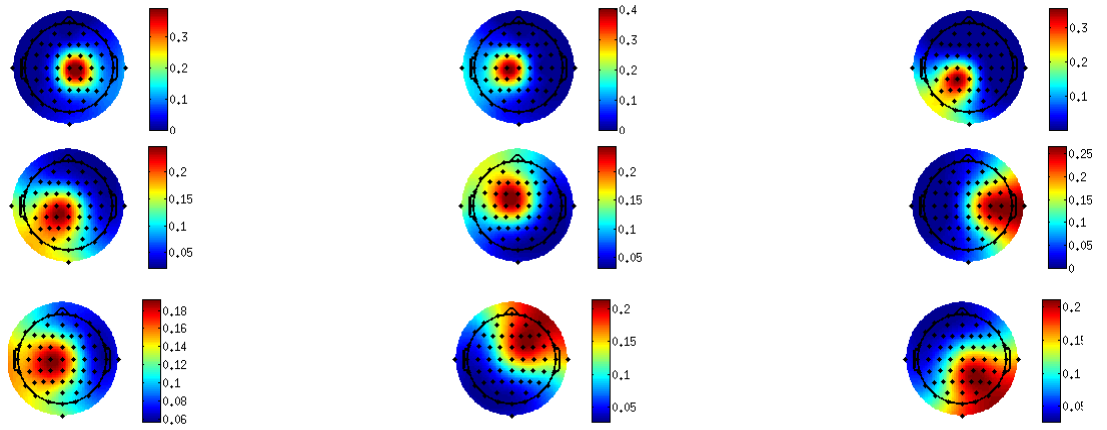


FIGURE 8.1 – Exemple d'atomes gaussiens 2D de Φ_s (pour différents rayons de variance).

composition peut être assuré directement à l'aide d'un modèle de décomposition impliquant deux dictionnaires comme nous l'avons déjà évoqué dans la section 3.5.2.1. Les composantes de décomposition sont alors des atomes spatio-temporels de rang 1 obtenus par multiplication d'atomes temporels et d'atomes spatiaux. Le dictionnaire spatial peut alors être vu comme une régularisation spatiale imposant des topographies strictes aux composantes. Le problème d'optimisation associé pour un signal Y et un dictionnaire temporel Φ s'écrit alors (sous forme ℓ_1) :

$$\hat{X} = \arg\min_X \|Y - \Phi X \Phi_s\|_F^2 + \lambda \|X\|_1,$$

et peut être résolu sans études supplémentaires en adaptant par exemple l'algorithme FISTA ou bien avec une approche gloutonne de manière similaire à l'algorithme OMP en choisissant à chaque itération les atomes temporels et spatiaux les plus adaptés (à l'aide d'un produit scalaire matriciel).

Cette structure spatiale des composantes peut être également obtenue lorsque ce dictionnaire gaussien est utilisé dans un terme de régularisation en analyse comme celui utilisé pour le Fused-Lasso multidimensionnel considéré dans le chapitre 6. Le problème de décomposition s'écrit alors :

$$\hat{X} = \arg\min_{X} \|Y - \Phi X\|_F^2 + \lambda \|X\Phi_s^*\|_1.$$

avec Φ_s^* une estimation du dual du dictionnaire spatial Φ_s : par exemple $\Phi_s^* = \Phi_s^\dagger$. La résolution d'un tel problème est réalisable via le schéma d'optimisation proposé dans le chapitre 6.

Par conséquent, il serait intéressant de comparer ces régularisations afin notamment de déterminer leur efficacité pour le recouvrement des composantes sous-jacentes des signaux EEG.

8.2.3 Décompositions parcimonieuses discriminantes

Dans le cadre des interfaces cerveau-machine, les représentations parcimonieuses utilisant un dictionnaire redondant pourrait permettre de séparer efficacement les activités cérébrales d'un individu.
Nous avons déjà vu dans le chapître 5 que le taux de détection des potentiels évoqués P300 pouvait être augmenté grâce à une telle représentation, particulièrement lorsque le choix des atomes et des coefficients est amélioré grâce à des régularisations tenant compte des propriétés des signaux EEG.
De plus, ces dernières années, de nombreuses études se sont intéressées à l'obtention de représentations parcimonieuses permettant de classifier efficacement des signaux. Nous ne ferons pas une revue exhaustive ici mais le lecteur intéressé pourra trouver l'état de l'art de ces représentations dans le chapitre 6 de [15]. Les méthodes développées pour ces représentations proposent à la fois des approches permettant de guider les décompositions des signaux vers des ensembles de coefficients discriminants sur des dictionnaires fixes mais aussi des approches d'apprentissage de dictionnaires discriminants.

L'incorporation de régularisations adaptées aux signaux EEG dans ces méthodes de représentations parcimonieuses discriminantes pourrait être très efficace pour la classification de ces signaux. De plus, ce type d'approches pourrait aboutir à des représentations interprétables pouvant permettre une meilleure compréhension des mécanismes cérébraux et de leurs contributions à l'activité électrique capturée dans les signaux EEG.

Bibliographie

[1] H. Abdi and L. J. Williams. Principal component analysis. *Wiley Interdisciplinary Reviews : Computational Statistics*, 2(4) :433–459, 2010. (Cité en page 25.)

[2] A. M. Abdulghani, A. J. Casson, and E. Rodriguez-Villegas. Compressive sensing scalp eeg signals : implementations and practical performance. *Medical & biological engineering & computing*, 50(11) :1137–1145, 2012. (Cité en page 59.)

[3] P. Achermann and A. A. Borbély. Mathematical models of sleep regulation. *Front Biosci*, 8(Cited April 25, 2003) :683–693, 2003. (Cité en page 21.)

[4] M. Aharon and M. Elad. Sparse and redundant modeling of image content using an image-signature-dictionary. *SIAM Journal on Imaging Sciences*, 1(3) :228–247, 2008. (Cité en page 49.)

[5] M. Aharon, M. Elad, and A. Bruckstein. K-svd : An algorithm for designing overcomplete dictionaries for sparse representation. *Signal Processing, IEEE Transactions on*, 54(11) :4311–4322, 2006. (Cité en pages 49 et 134.)

[6] K. K. Ang, Z. Y. Chin, H. Zhang, and C. Guan. Filter bank common spatial pattern (fbcsp) in brain-computer interface. In *Neural Networks, 2008. IJCNN 2008.(IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on*, pages 2390–2397. IEEE, 2008. (Cité en page 27.)

[7] D. Angelosante, G. Giannakis, and N. Sidiropoulos. Multiple frequency-hopping signal estimation via sparse regression. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pages 3502–3505. IEEE, 2010. (Cité en page 76.)

[8] A. Argyriou, T. Evgeniou, and M. Pontil. Multi-task feature learning. *Advances in neural information processing systems*, 19 :41, 2007. (Cité en page 46.)

[9] M. Arnold, X. Milner, H. Witte, R. Bauer, and C. Braun. Adaptive ar modeling of nonstationary time series by means of kalman filtering. *Biomedical Engineering, IEEE Transactions on*, 45(5) :553–562, 1998. (Cité en page 20.)

[10] S. Aviyente. Compressed sensing framework for eeg compression. In *Statistical Signal Processing, 2007. SSP'07. IEEE/SP 14th Workshop on*, pages 181–184. IEEE, 2007. (Cité en page 59.)

[11] F. Bach, R. Jenatton, J. Mairal, G. Obozinski, et al. Convex optimization with sparsity-inducing norms. *Optimization for Machine Learning*, pages 19–53, 2011. (Cité en pages 46, 85 et 86.)

[12] R. Baraniuk. Compressive sensing. *IEEE signal processing magazine*, 24(4), 2007. (Cité en pages 2 et 50.)

[13] H. B. Barlow. Possible principles underlying the transformation of sensory messages. *Sensory communication*, pages 217–234, 1961. (Cité en page 34.)

[14] R. Bartels and G. Stewart. Solution of the matrix equation ax+ xb= c [f4]. *Communications of the ACM*, 15(9) :820–826, 1972. (Cité en page 177.)

[15] Q. Barthélemy. *Représentations parcimonieuses pour les signaux multivariés*. PhD thesis, Université de Grenoble, 2013. (Cité en page 155.)

[16] Q. Barthélemy, C. Gouy-Pailler, Y. Isaac, A. Souloumiac, A. Larue, and J. I. Mars. Multivariate temporal dictionary learning for eeg. *Journal of neuroscience methods*, 215(1) :19–28, 2013. (Cité en pages 12, 26, 58, 59 et 152.)

[17] Q. Barthélemy, A. Larue, A. Mayoue, D. Mercier, and J. I. Mars. Shift & 2d rotation invariant sparse coding for multivariate signals. *Signal Processing, IEEE Transactions on*, 60(4) :1597–1611, 2012. (Cité en page 49.)

[18] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1) :183–202, 2009. (Cité en pages 46, 86, 87, 88, 118 et 122.)

[19] C. G. Bénar, T. Papadopoulo, B. Torrésani, and M. Clerc. Consensus matching pursuit for multi-trial eeg signals. *Journal of neuroscience methods*, 180(1) :161–170, 2009. (Cité en page 54.)

[20] H. Berger. Über das elektrenkephalogramm des menschen. *European archives of psychiatry and clinical neuroscience*, 98(1) :231–254, 1933. (Cité en pages 11, 1, 8 et 14.)

[21] O. Bertrand, F. Perrin, and J. Pernier. A theoretical justification of the average reference in topographic evoked potential studies. *Electroencephalography and Clinical Neurophysiology/Evoked Potentials Section*, 62(6) :462–464, 1985. (Cité en page 23.)

[22] D. Bertsekas. Constrained optimization and lagrange multiplier methods. *Computer Science and Applied Mathematics, Boston : Academic Press, 1982*, 1, 1982. (Cité en page 116.)

[23] M. Besserve. *Analyse de la dynamique neuronale pour les Interfaces Cerveau-Machines : un retour aux sources.* PhD thesis, Université Paris Sud-Paris XI, 2007. (Cité en pages 63 et 179.)

[24] A. Bidet-Caulet, C. Fischer, J. Besle, P.-E. Aguera, M.-H. Giard, and O. Bertrand. Effects of selective attention on the electrophysiological representation of concurrent sounds in the human auditory cortex. *The Journal of neuroscience*, 27(35) :9252–9261, 2007. (Cité en page 12.)

[25] J. M. Bioucas-Dias and M. A. Figueiredo. A new twist : two-step iterative shrinkage/thresholding algorithms for image restoration. *Image Processing, IEEE Transactions on*, 16(12) :2992–3004, 2007. (Cité en page 46.)

[26] B. Blankertz, K. Muller, G. Curio, T. M. Vaughan, G. Schalk, J. R. Wolpaw, A. Schlogl, C. Neuper, G. Pfurtscheller, T. Hinterberger, et al. The bci competition 2003 : progress and perspectives in detection and discrimination of eeg single trials. *Biomedical Engineering, IEEE Transactions on*, 51(6) :1044–1051, 2004. (Cité en pages 104 et 141.)

[27] B. Blankertz, R. Tomioka, S. Lemm, M. Kawanabe, and K.-R. Muller. Optimizing spatial filters for robust eeg single-trial analysis. *Signal Processing Magazine, IEEE*, 25(1) :41–56, 2008. (Cité en page 26.)

[28] K. Bleakley and J.-P. Vert. The group fused lasso for multiple change-point detection. *arXiv preprint arXiv :1106.4199*, 2011. (Cité en page 76.)

[29] G. Bodenstein and H. M. Praetorius. Feature extraction from the electroencephalogram by adaptive segmentation. *Proceedings of the IEEE*, 65(5) :642–652, 1977. (Cité en page 19.)

[30] P. Bofill and M. Zibulevsky. Underdetermined blind source separation using sparse representations. *Signal processing*, 81(11) :2353–2362, 2001. (Cité en pages 2 et 50.)

[31] J. Bolduc-Teasdale, P. Jolicoeur, and M. McKerral. Multiple electrophysiological markers of visual-attentional processing in a novel task directed toward clinical use. *Journal of ophthalmology*, 2012, 2012. (Cité en pages 11 et 14.)

[32] A. Bolstad, B. V. Veen, and R. Nowak. Space–time event sparse penalization for magneto-/electroencephalography. *NeuroImage*, 46(4) :1066–1081, 2009. (Cité en page 58.)

[33] B. Bromm and E. Scharein. Principal component analysis of pain-related cerebral potentials to mechanical and electrical stimulation in man. *Electroencephalography and clinical neurophysiology*, 53(1) :94–103, 1982. (Cité en page 25.)

[34] C. Brunner, R. Leeb, G. Müller-Putz, A. Schlögl, and G. Pfurtscheller. Bci competition 2008–graz data set a. *Institute for Knowledge Discovery (Laboratory of Brain-Computer Interfaces), Graz University of Technology*, 2008. (Cité en page 16.)

[35] J. Cai, S. Osher, and Z. Shen. Split Bregman methods and frame based image restoration. *Multiscale modeling and simulation*, 8(2) :337, 2009. (Cité en pages 114, 181, 184 et 185.)

[36] E. Callaway III, R. T. Jones, and E. Donchin. Auditory evoked potential variability in schizophrenia. *Electroencephalography and Clinical Neurophysiology*, 29(5) :421–428, 1970. (Cité en page 17.)

[37] E. Candes and J. Romberg. l1-magic : Recovery of sparse signals via convex programming. *URL : www. acm. caltech. edu/l1magic/downloads/l1magic. pdf*, 4, 2005. (Cité en page 46.)

[38] E. J. Candes. The restricted isometry property and its implications for compressed sensing. *Comptes Rendus Mathematique*, 346(9) :589–592, 2008. (Cité en pages 40 et 41.)

[39] E. J. Candes, Y. C. Eldar, D. Needell, and P. Randall. Compressed sensing with coherent and redundant dictionaries. *Applied and Computational Harmonic Analysis*, 31(1) :59–73, 2011. (Cité en page 43.)

[40] E. J. Candes and T. Tao. Decoding by linear programming. *Information Theory, IEEE Transactions on*, 51(12) :4203–4215, 2005. (Cité en pages 40 et 50.)

[41] J. Chen and X. Huo. Theoretical results on sparse representations of multiple-measurement vectors. *Signal Processing, IEEE Transactions on*, 54(12) :4634–4643, 2006. (Cité en page 42.)

[42] S. Chen and J. Wigger. Fast orthogonal least squares algorithm for efficient subset model selection. *IEEE Transactions on Signal Procesing*, 43(7) :1713–1715, 1995. (Cité en page 45.)

[43] S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM journal on scientific computing*, 20(1) :33–61, 1998. (Cité en pages 37 et 45.)

[44] X. Chen, Q. Lin, S. Kim, J. Carbonell, and E. Xing. Smoothing proximal gradient method for general structured sparse regression. *The Annals of Applied Statistics*, 6(2) :719–752, 2012. (Cité en pages 44, 112, 118, 119 et 121.)

[45] S. Chevallier, Q. Barthélemy, and J. Atif. On the need for metrics in dictionary learning assessment. 2014. (Cité en page 138.)

[46] O. Christensen. *An introduction to frames and Riesz bases*. Springer, 2003. (Cité en pages 34, 35 et 39.)

[47] P. L. Combettes and J.-C. Pesquet. Proximal splitting methods in signal processing. In *Fixed-point algorithms for inverse problems in science and engineering*, pages 185–212. Springer, 2011. (Cité en pages 46 et 85.)

[48] P. L. Combettes and V. R. Wajs. Signal recovery by proximal forward-backward splitting. *Multiscale Modeling & Simulation*, 4(4) :1168–1200, 2005. (Cité en pages 46 et 85.)

[49] N. E. Crone, D. L. Miglioretti, B. Gordon, J. M. Sieracki, M. T. Wilson, S. Uematsu, and R. P. Lesser. Functional mapping of human sensorimotor cortex with electrocorticographic spectral analysis. i. alpha and beta event-related desynchronization. *Brain*, 121(12) :2271–2299, 1998. (Cité en page 1.)

[50] P. Danaher, P. Wang, and D. Witten. The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 2013. (Cité en page 76.)

[51] J. Darbon and M. Sigelle. A fast and exact algorithm for total variation minimization. In *Pattern recognition and image analysis*, volume 3522 of *Lecture Notes in Computer Science*, pages 351–359, 2005. (Cité en pages 75 et 111.)

[52] I. Daubechies, M. Defrise, and C. De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on pure and applied mathematics*, 57(11) :1413–1457, 2004. (Cité en pages 46 et 50.)

[53] O. David, S. J. Kiebel, L. M. Harrison, J. Mattout, J. M. Kilner, and K. J. Friston. Dynamic causal modeling of evoked responses in eeg and meg. *NeuroImage*, 30(4) :1255–1272, 2006. (Cité en page 20.)

[54] L. Ding and B. He. Sparse source imaging in electroencephalography with accurate field modeling. *Human brain mapping*, 29(9) :1053–1067, 2008. (Cité en page 56.)

[55] D. L. Donoho and M. Elad. Optimally sparse representation in general (nonorthogonal) dictionaries via 1 minimization. *Proceedings of the National Academy of Sciences*, 100(5) :2197–2202, 2003. (Cité en page 40.)

[56] D. L. Donoho, M. Elad, and V. N. Temlyakov. Stable recovery of sparse overcomplete representations in the presence of noise. *Information Theory, IEEE Transactions on*, 52(1) :6–18, 2006. (Cité en page 84.)

[57] D. L. Donoho and J. M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3) :425–455, 1994. (Cité en page 36.)

[58] D. L. Donoho, Y. Tsaig, I. Drori, and J.-L. Starck. Sparse solution of underdetermined systems of linear equations by stagewise orthogonal matching pursuit. *Information Theory, IEEE Transactions on*, 58(2) :1094–1121, 2012. (Cité en page 45.)

[59] G. Dornhege. *Toward brain-computer interfacing*. MIT press, 2007. (Cité en page 13.)

[60] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern classification*. John Wiley & Sons, 2012. (Cité en pages 29 et 133.)

[61] G. Dumermuth and L. Molinari. Spectral analysis of the eeg. *Neuropsychobiology*, 17(1-2) :85–99, 1987. (Cité en page 21.)

[62] P. Durka and K. Blinowska. Analysis of eeg transients by means of matching pursuit. *Annals of biomedical engineering*, 23(5) :608–611, 1995. (Cité en pages 22 et 52.)

[63] P. Durka and A. House. *Matching pursuit and unification in EEG analysis*. Artech House Norwood, 2007. (Cité en pages 11, 21, 22 et 52.)

[64] P. J. Durka, D. Ircha, C. Neuper, and G. Pfurtscheller. Time-frequency microstructure of event-related electro-encephalogram desynchronisation and synchronisation. *Medical and biological engineering and computing*, 39(3) :315–321, 2001. (Cité en page 22.)

[65] P. J. Durka, A. Matysiak, E. M. Montes, P. V. Sosa, and K. J. Blinowska. Multichannel matching pursuit and eeg inverse solutions. *Journal of neuroscience methods*, 148(1) :49–59, 2005. (Cité en page 53.)

[66] B. Efron, T. Hastie, I. Johnstone, R. Tibshirani, et al. Least angle regression. *The Annals of statistics*, 32(2) :407–499, 2004. (Cité en pages 46 et 122.)

[67] M. Elad. *Sparse and redundant representations : from theory to applications in signal and image processing*. Springer, 2010. (Cité en page 34.)

[68] M. Elad and M. Aharon. Image denoising via sparse and redundant representations over learned dictionaries. *Image Processing, IEEE Transactions on*, 15(12) :3736–3745, 2006. (Cité en pages 2 et 50.)

[69] M. Elad, P. Milanfar, and R. Rubinstein. Analysis versus synthesis in signal priors. *Inverse problems*, 23(3) :947, 2007. (Cité en pages 38, 110 et 126.)

[70] Y. C. Eldar, P. Kuppinger, and H. Bolcskei. Block-sparse signals : Uncertainty relations and efficient recovery. *Signal Processing, IEEE Transactions on*, 58(6) :3042–3054, 2010. (Cité en pages 43 et 45.)

[71] Y. C. Eldar and H. Rauhut. Average case analysis of multichannel sparse recovery using convex relaxation. *Information Theory, IEEE Transactions on*, 56(1) :505–519, 2010. (Cité en page 42.)

[72] K. Engan, S. O. Aase, and J. H. Husøy. Multi-frame compression : Theory and design. *Signal Processing*, 80(10) :2121–2140, 2000. (Cité en page 49.)

[73] K. Engan, K. Skretting, and J. H. Husøy. Family of iterative ls-based dictionary learning algorithms, ils-dla, for sparse signal representation. *Digital Signal Processing*, 17(1) :32–49, 2007. (Cité en pages 2 et 50.)

[74] L. A. Farwell and E. Donchin. Talking off the top of your head : toward a mental prosthesis utilizing event-related brain potentials. *Electroencephalography and clinical Neurophysiology*, 70(6) :510–523, 1988. (Cité en page 13.)

[75] R. P. Feldman and J. T. Goodrich. The edwin smith surgical papyrus. *Child's Nervous System*, 15(6-7) :281–284, 1999. (Cité en page 1.)

[76] D. J. Field. What is the goal of sensory coding ? *Neural computation*, 6(4) :559–601, 1994. (Cité en page 36.)

[77] R. Flamary, N. Jrad, R. Phlypo, M. Congedo, and A. Rakotomamonjy. Mixed-norm regularization for brain decoding. *Computational and mathematical methods in medicine*, 2014, 2014. (Cité en page 59.)

[78] M. R. Ford, J. W. Goethe, and D. K. Dekker. Eeg coherence and power in the discrimination of psychiatric disorders and medication effects. *Biological psychiatry*, 21(12) :1175–1188, 1986. (Cité en pages 8 et 17.)

[79] J. H. Friedman and W. Stuetzle. Projection pursuit regression. *Journal of the American statistical Association*, 76(376) :817–823, 1981. (Cité en page 44.)

[80] O. Friman, I. Volosyak, and A. Graser. Multiple channel detection of steady-state visual evoked potentials for brain-computer interfaces. *Biomedical Engineering, IEEE Transactions on*, 54(4) :742–750, 2007. (Cité en pages 11 et 13.)

[81] W. J. Fu. Penalized regressions : the bridge versus the lasso. *Journal of computational and graphical statistics*, 7(3) :397–416, 1998. (Cité en page 46.)

[82] K. Fukunaga. *Introduction to statistical pattern recognition*. Academic press, 1990. (Cité en page 25.)

[83] H. Gamboa. Rythmes cérébraux, `http://fr.wikipedia.org/wiki/Rythme_cerebral`., 2013. (Cité en pages 11 et 16.)

[84] X. Gao, D. Xu, M. Cheng, and S. Gao. A bci-based environmental controller for the motion-disabled. *Neural Systems and Rehabilitation Engineering, IEEE Transactions on*, 11(2) :137–140, 2003. (Cité en page 13.)

[85] S. D. Georgiadis, P. O. Ranta-aho, M. P. Tarvainen, and P. A. Karjalainen. Single-trial dynamical estimation of event-related potentials : a kalman filter-based approach. *Biomedical Engineering, IEEE Transactions on*, 52(8) :1397–1406, 2005. (Cité en page 20.)

[86] W. Gersch. Spectral analysis of eeg's by autoregressive decomposition of time series. *Mathematical Biosciences*, 7(1) :205–222, 1970. (Cité en page 19.)

[87] A. B. Geva and D. H. Kerem. Forecasting generalized epileptic seizures from the eeg signal by wavelet analysis and dynamic unsupervised fuzzy clustering. *Biomedical Engineering, IEEE Transactions on*, 45(10) :1205–1216, 1998. (Cité en page 21.)

[88] A. Gholami and S. M. Hosseini. A balanced combination of tikhonov and total variation regularizations for reconstruction of piecewise-smooth signals. *Signal Processing*, 93(7) :1945–1960, 2013. (Cité en page 44.)

[89] A. Gholami and H. Siahkoohi. Regularization of linear and non-linear geophysical ill-posed problems with joint sparsity constraints. *Geophysical Journal International*, 180(2) :871–882, 2010. (Cité en page 76.)

[90] T. Goldstein and S. Osher. The split bregman method for l1-regularized problems. *SIAM Journal on Imaging Sciences*, 2(2) :323–343, 2009. (Cité en pages 46, 112, 113 et 128.)

[91] G. H. Golub, P. C. Hansen, and D. P. O'Leary. Tikhonov regularization and total least squares. *SIAM Journal on Matrix Analysis and Applications*, 21(1) :185–194, 1999. (Cité en page 67.)

[92] I. F. Gorodnitsky, J. S. George, and B. D. Rao. Neuromagnetic source imaging with focuss : a recursive weighted minimum norm algorithm. *Electroencephalography and clinical Neurophysiology*, 95(4) :231–251, 1995. (Cité en page 56.)

[93] I. F. Gorodnitsky and B. D. Rao. Sparse signal reconstruction from limited data using focuss : A re-weighted minimum norm algorithm. *Signal Processing, IEEE Transactions on*, 45(3) :600–616, 1997. (Cité en page 37.)

[94] A. Gramfort, M. Kowalski, and M. Hämäläinen. Mixed-norm estimates for the m/eeg inverse problem using accelerated gradient methods. *Physics in medicine and biology*, 57(7) :1937, 2012. (Cité en page 57.)

[95] A. Gramfort, T. Papadopoulo, E. Olivi, M. Clerc, et al. Openmeeg : opensource software for quasistatic bioelectromagnetics. *Biomed. Eng. Online*, 9(1) :45, 2010. (Cité en pages 24, 64 et 126.)

[96] A. Gramfort, D. Strohmeier, J. Haueisen, M. Hamalainen, and M. Kowalski. Functional brain imaging with m/eeg using structured sparsity in time-frequency dictionaries. In *Information Processing in Medical Imaging*, pages 600–611. Springer, 2011. (Cité en pages 58 et 122.)

[97] M. Gratkowski, J. Haueisen, L. Arendt-Nielsen, A. Cn Chen, and F. Zanow. Decomposition of biomedical signals in spatial and time-frequency modes. *Methods of Information in Medicine*, 47(1) :26–37, 2008. PMID : 18213425. (Cité en page 55.)

[98] M. Gratkowski, J. Haueisen, L. Arendt-Nielsen, and F. Zanow. Topographic matching pursuit of spatio-temporal bioelectromagnetic data. *Przeglad Elektrotechniczny*, 83(11) :138–141, 2007. (Cité en page 54.)

[99] R. Grech, T. Cassar, J. Muscat, K. P. Camilleri, S. G. Fabri, M. Zervakis, P. Xanthopoulos, V. Sakkalis, and B. Vanrumste. Review on solving the inverse problem in eeg source analysis. *Journal of neuroengineering and rehabilitation*, 5(1) :25, 2008. (Cité en page 24.)

[100] R. Gribonval. Piecewise linear source separation. In *Optical Science and Technology, SPIE's 48th Annual Meeting*, pages 297–310. International Society for Optics and Photonics, 2003. (Cité en page 53.)

[101] R. Gribonval, H. Rauhut, K. Schnass, and P. Vandergheynst. Atoms of all channels, unite ! average case analysis of multi-channel sparse recovery using greedy algorithms. *Journal of Fourier analysis and Applications*, 14(5-6) :655–687, 2008. (Cité en page 42.)

[102] R. Gribonval and P. Vandergheynst. On the exponential convergence of matching pursuits in quasi-incoherent dictionaries. *Information Theory, IEEE Transactions on*, 52(1) :255–261, 2006. (Cité en pages 39 et 40.)

[103] D. Hagemann, E. Naumann, and J. F. Thayer. The quest for the eeg reference revisited : A glance from brain asymmetry research. *Psychophysiology*, 38(5) :847–857, 2001. (Cité en page 23.)

[104] T. Harmony, T. Fernández, J. Gersenowies, L. Galán, A. Fernández-Bouzas, E. Aubert, and L. Dıaz-Comas. Specific eeg frequencies signal general common cognitive processes as well

as specific task processes in man. *International journal of psychophysiology*, 53(3) :207–216, 2004. (Cité en page 21.)

[105] M. Hart. Placement des électrodes dans le système 10-20, http://www.mariusthart.net/downloads/eeg_electrodes_10-20.pdf., 2013. (Cité en page 11.)

[106] T. Hastie, R. Tibshirani, J. Friedman, T. Hastie, J. Friedman, and R. Tibshirani. *The elements of statistical learning*, volume 2. Springer, 2009. (Cité en page 27.)

[107] S. Haufe, V. V. Nikulin, A. Ziehe, K.-R. Müller, and G. Nolte. Combining sparsity and rotational invariance in eeg/meg source reconstruction. *NeuroImage*, 42(2) :726–738, 2008. (Cité en page 56.)

[108] S. Haufe, R. Tomioka, T. Dickhaus, C. Sannelli, B. Blankertz, G. Nolte, and K.-R. Müller. Large-scale eeg/meg source localization with spatial flexibility. *NeuroImage*, 54(2) :851–859, 2011. (Cité en page 57.)

[109] S. Hitziger, M. Clerc, A. Gramfort, S. Saillet, C. Bénar, and T. Papadopoulo. Jitter-adaptive dictionary learning-application to multi-trial neuroelectric signals. *arXiv preprint arXiv :1301.3611*, 2013. (Cité en page 58.)

[110] H. Hoefling. A path algorithm for the fused Lasso signal approximator. *Journal of Computational and Graphical Statistics*, 19(4) :984–1006, 2010. (Cité en page 112.)

[111] U. Hoffmann, J. Vesin, T. Ebrahimi, and K. Diserens. An efficient p300-based brain–computer interface for disabled subjects. *Journal of Neuroscience methods*, 167(1) :115–125, 2008. (Cité en pages 104 et 127.)

[112] R. W. Homan, J. Herman, and P. Purdy. Cerebral location of international 10–20 system electrode placement. *Electroencephalography and clinical neurophysiology*, 66(4) :376–382, 1987. (Cité en page 11.)

[113] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent component analysis*, volume 46. John Wiley & Sons, 2004. (Cité en page 26.)

[114] L. D. Iasemidis and J. C. Sackellares. Review : Chaos theory and epilepsy. *The Neuroscientist*, 2(2) :118–126, 1996. (Cité en page 20.)

[115] L. Jacob, G. Obozinski, and J.-P. Vert. Group lasso with overlap and graph lasso. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 433–440. ACM, 2009. (Cité en page 67.)

[116] R. Jenatton. *Structured sparsity-inducing norms : Statistical and algorithmic properties with applications to neuroimaging*. PhD thesis, École normale supérieure de Cachan-ENS Cachan, 2011. (Cité en pages 43 et 67.)

[117] R. Jenatton, R. Gribonval, and F. Bach. Local stability and robustness of sparse dictionary learning in the presence of noise. *arXiv preprint arXiv :1210.0685*, 2012. (Cité en page 49.)

[118] R. Jenatton, J. Mairal, F. Bach, and G. Obozinski. Proximal methods for sparse hierarchical dictionary learning. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 487–494, 2010. (Cité en page 123.)

[119] P. Jost, P. Vandergheynst, S. Lesage, and R. Gribonval. Motif : an efficient algorithm for learning translation invariant dictionaries. In *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, volume 5, pages V–V. IEEE, 2006. (Cité en page 58.)

[120] P. Jost, P. Vandergheynst, S. Lesage, R. Gribonval, et al. Learning redundant dictionaries with translation invariance property : the motif algorithm. In *SPARS'05-Workshop on Signal Processing with Adaptive Sparse Structured Representations*, pages 1–3, 2005. (Cité en page 58.)

[121] M. A. Kennard, S. Rabinovitch, and W. P. Fister. The use of frequency analysis in the interpretation of the eegs of patients with psychological disorders. *Electroencephalography and clinical neurophysiology*, 7(1) :29–38, 1955. (Cité en page 21.)

[122] S. Kim and E. Xing. Statistical estimation of correlated genome associations to a quantitative trait network. *PLoS genetics*, 5(8) :e1000587, 2009. (Cité en page 76.)

[123] S.-J. Kim, K. Koh, S. Boyd, and D. Gorinevsky. ℓ_1 trend filtering. *Siam Review*, 51(2) :339–360, 2009. (Cité en page 76.)

[124] T. Koenig, F. Marti-Lopez, and P. Valdes-Sosa. Topographic time-frequency decomposition of the eeg. *NeuroImage*, 14(2) :383–390, 2001. (Cité en page 55.)

[125] R. Kohavi and G. H. John. Wrappers for feature subset selection. *Artificial intelligence*, 97(1) :273–324, 1997. (Cité en page 27.)

[126] K. Kreutz-Delgado, J. F. Murray, B. D. Rao, K. Engan, T.-W. Lee, and T. J. Sejnowski. Dictionary learning algorithms for sparse representation. *Neural computation*, 15(2) :349–396, 2003. (Cité en pages 47 et 49.)

[127] A. Kübler, A. Furdea, S. Halder, E. M. Hammer, F. Nijboer, and B. Kotchoubey. A brain–computer interface controlled auditory event-related potential (p300) spelling system for locked-in patients. *Annals of the New York Academy of Sciences*, 1157(1) :90–100, 2009. (Cité en pages 1 et 14.)

[128] A. Kübler, B. Kotchoubey, J. Kaiser, J. R. Wolpaw, and N. Birbaumer. Brain–computer communication : Unlocking the locked in. *Psychological bulletin*, 127(3) :358, 2001. (Cité en page 17.)

[129] T. D. Lagerlund, F. W. Sharbrough, and N. E. Busacker. Spatial filtering of multichannel electroencephalographic recordings through principal component analysis by singular value decomposition. *Journal of Clinical Neurophysiology*, 14(1) :73–82, 1997. (Cité en page 25.)

[130] J. A. Lasserre, C. M. Bishop, and T. P. Minka. Principled hybrids of generative and discriminative models. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 1, pages 87–94. IEEE, 2006. (Cité en page 31.)

[131] A. Lécuyer, F. Lotte, R. B. Reilly, R. Leeb, M. Hirose, M. Slater, et al. Brain-computer interfaces, virtual reality, and videogames. *IEEE Computer*, 41(10) :66–72, 2008. (Cité en page 17.)

[132] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755) :788–791, 1999. (Cité en page 44.)

[133] H. Lee, A. Battle, R. Raina, and A. Y. Ng. Efficient sparse coding algorithms. *Advances in neural information processing systems*, 19 :801, 2007. (Cité en page 46.)

[134] D. Lehmann. Multichannel topography of human alpha eeg fields. *Electroencephalography and clinical neurophysiology*, 31(5) :439–449, 1971. (Cité en pages 73 et 131.)

[135] D. Lehmann, P. L. Faber, S. Galderisi, W. M. Herrmann, T. Kinoshita, M. Koukkou, A. Mucci, R. D. Pascual-Marqui, N. Saito, J. Wackermann, et al. Eeg microstate duration and syntax in acute, medication-naive, first-episode schizophrenia : a multi-center study. *Psychiatry Research : Neuroimaging*, 138(2) :141–156, 2005. (Cité en pages 17 et 132.)

[136] D. Lehmann, H. Ozaki, and I. Pal. Eeg alpha map series : brain micro-states by space-oriented adaptive segmentation. *Electroencephalography and clinical neurophysiology*, 67(3) :271–288, 1987. (Cité en pages 12, 16, 73, 74 et 131.)

[137] D. Lehmann, R. Pascual-Marqui, and C. Michel. Eeg microstates,`http://www.scholarpedia.org/article/EEG_microstates`., 2013. (Cité en pages 131 et 143.)

[138] D. Lehmann and W. Skrandies. Spatial analysis of evoked potentials in man—a review. *Progress in neurobiology*, 23(3) :227–250, 1984. (Cité en pages 73 et 131.)

[139] M. S. Lewicki and T. J. Sejnowski. Learning overcomplete representations. *Neural computation*, 12(2) :337–365, 2000. (Cité en pages 2, 29, 47 et 48.)

[140] J. Liu, L. Yuan, and J. Ye. An efficient algorithm for a class of fused Lasso problems. In *Proc. 16th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, pages 323–332. ACM, 2010. (Cité en page 112.)

[141] Y. Liu, S. Li, T. Mi, H. Lei, and W. Yu. Performance analysis of l1-synthesis with coherent frames. In *Information Theory Proceedings (ISIT), 2012 IEEE International Symposium on*, pages 2042–2046. IEEE, 2012. (Cité en page 38.)

[142] F. Lopes da Silva, K. Van Hulten, J. Lommen, W. Storm van Leeuwen, C. Van Veelen, and W. Vliegenthart. Automatic detection and localization of epileptic foci. *Electroencephalography and Clinical Neurophysiology*, 43(1) :1–13, 1977. (Cité en page 19.)

[143] F. Lotte. *Study of Electroencephalographic Signal Processing and Classification Techniques towards the use of Brain-Computer Interfaces in Virtual Reality Applications*. PhD thesis, PhD thesis, 2007. (Cité en page 1.)

[144] F. Lotte, M. Congedo, A. Lécuyer, F. Lamarche, B. Arnaldi, et al. A review of classification algorithms for eeg-based brain–computer interfaces. *Journal of neural engineering*, 4, 2007. (Cité en pages 27 et 29.)

[145] F. Lotte and C. Guan. Spatially regularized common spatial patterns for eeg classification. In *Pattern Recognition (ICPR), 2010 20th International Conference on*, pages 3712–3715. IEEE, 2010. (Cité en page 26.)

[146] L. Lustick, B. Saltzberg, J. Buckley, and R. Heath. Autoregressive model for simplified computer generation of eeg correlation functions. In *Proceedings of the IEEE annual conference on engineering in medicine and biology*, volume 10, pages 78–94, 1968. (Cité en page 19.)

[147] B. Mailhé, D. Barchiesi, and M. D. Plumbley. Ink-svd : Learning incoherent dictionaries for sparse representations. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pages 3573–3576. IEEE, 2012. (Cité en page 48.)

[148] J. Mairal. *Représentations parcimonieuses en apprentissage statistique, traitement d'image et vision par ordinateur*. PhD thesis, PhD thesis, 2010. (Cité en pages 11, 37, 38 et 50.)

[149] J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online dictionary learning for sparse coding. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 689–696. ACM, 2009. (Cité en pages 49, 50, 136, 137, 145 et 148.)

[150] S. Makeig, A. J. Bell, T.-P. Jung, T. J. Sejnowski, et al. Independent component analysis of electroencephalographic data. *Advances in neural information processing systems*, pages 145–151, 1996. (Cité en page 26.)

[151] A. Maleki and D. L. Donoho. Optimally tuned iterative reconstruction algorithms for compressed sensing. *Selected Topics in Signal Processing, IEEE Journal of*, 4(2) :330–341, 2010. (Cité en page 45.)

[152] S. Mallat. *A wavelet tour of signal processing : the sparse way*. Academic press, 2008. (Cité en pages 13, 22, 80 et 81.)

[153] S. G. Mallat and Z. Zhang. Matching pursuits with time-frequency dictionaries. *Signal Processing, IEEE Transactions on*, 41(12) :3397–3415, 1993. (Cité en page 44.)

[154] K. Matsuura and Y. Okabe. Selective minimum-norm solution of the biomagnetic inverse problem. *Biomedical Engineering, IEEE Transactions on*, 42(6) :608–615, 1995. (Cité en page 56.)

[155] A. Matysiak, P. J. Durka, E. Martinez Montes, M. Barwiński, P. Zwoliński, M. Roszkowski, and K. J. Blinowska. Time-frequency-space localization of epileptic eeg oscillations. *Acta neurobiologiae experimentalis*, 65(4) :435, 2005. (Cité en page 54.)

[156] G. McCarthy and C. C. Wood. Scalp distributions of event-related potentials : an ambiguity associated with analysis of variance models. *Electroencephalography and Clinical Neurophysiology/Evoked Potentials Section*, 62(3) :203–208, 1985. (Cité en page 132.)

[157] D. J. McFarland, L. M. McCane, S. V. David, and J. R. Wolpaw. Spatial filter selection for eeg-based communication. *Electroencephalography and clinical Neurophysiology*, 103(3) :386–394, 1997. (Cité en pages 11 et 24.)

[158] J. d. R. Millán, B. Hamner, and R. Chavarriaga. Learning dictionaries of spatial and temporal eeg primitives for brain-computer interfaces. In *Workshop on Structured Sparsity : Learning and Inference, ICML 2011*, number EPFL-CONF-166740, 2011. (Cité en page 58.)

[159] F. Miwakeichi, E. Martınez-Montes, P. A. Valdés-Sosa, N. Nishiyama, H. Mizuhara, and Y. Yamaguchi. Decomposing eeg data into space–time–frequency components using parallel factor analysis. *NeuroImage*, 22(3) :1035–1045, 2004. (Cité en page 26.)

[160] J. Möcks. Decomposing event-related potentials : A new topographic components model. *Biological Psychology*, 26(1) :199–215, 1988. (Cité en page 26.)

[161] H. Mohimani, M. Babaie-Zadeh, and C. Jutten. A fast approach for overcomplete sparse decomposition based on smoothed norm. *Signal Processing, IEEE Transactions on*, 57(1) :289–301, 2009. (Cité en page 37.)

[162] J. Montoya-Martinez, A. Artes-Rodriguez, L. K. Hansen, and M. Pontil. Structured sparsity regularization approach to the eeg inverse problem. In *Cognitive Information Processing (CIP), 2012 3rd International Workshop on*, pages 1–6. IEEE, 2012. (Cité en page 57.)

[163] J. C. Mosher, R. M. Leahy, and P. S. Lewis. Eeg and meg : forward solutions for inverse methods. *Biomedical Engineering, IEEE Transactions on*, 46(3) :245–259, 1999. (Cité en pages 24, 63 et 180.)

[164] G. Muller-Putz, R. Scherer, C. Neuper, and G. Pfurtscheller. Steady-state somatosensory evoked potentials : suitable brain signals for brain-computer interfaces ? *Neural Systems and Rehabilitation Engineering, IEEE Transactions on*, 14(1) :30–37, 2006. (Cité en page 12.)

[165] G. R. Müller-Putz, R. Scherer, C. Brauneis, and G. Pfurtscheller. Steady-state visual evoked potential (ssvep)-based communication : impact of harmonic frequency components. *Journal of neural engineering*, 2(4) :123, 2005. (Cité en page 12.)

[166] M. M. Murray, D. Brunet, and C. M. Michel. Topographic erp analyses : a step-by-step tutorial review. *Brain topography*, 20(4) :249–264, 2008. (Cité en pages 132, 133, 134, 136 et 145.)

[167] S. Nam, M. E. Davies, M. Elad, and R. Gribonval. The cosparse analysis model and algorithms. *Applied and Computational Harmonic Analysis*, 34(1) :30–56, 2013. (Cité en pages 38 et 45.)

[168] B. K. Natarajan. Sparse approximate solutions to linear systems. *SIAM journal on computing*, 24(2) :227–234, 1995. (Cité en page 37.)

[169] D. Needell and R. Vershynin. Uniform uncertainty principle and signal recovery via regularized orthogonal matching pursuit. *Foundations of computational mathematics*, 9(3) :317–334, 2009. (Cité en page 45.)

[170] Y. Nesterov. Smooth minimization of non-smooth functions. *Mathematical Programming*, 103(1) :127–152, 2005. (Cité en page 112.)

[171] Y. Nesterov et al. Gradient methods for minimizing composite objective function, 2007. (Cité en pages 46 et 86.)

[172] E. Niedermeyer and F. L. da Silva. *Electroencephalography : basic principles, clinical applications, and related fields*. Lippincott Williams & Wilkins, 2005. (Cité en page 17.)

[173] P. Nunez and R. Srinivasan. *Electric fields of the brain : the neurophysics of EEG*. Oxford University Press, USA, 2006. (Cité en pages 63 et 64.)

[174] B. Obermaier, C. Guger, C. Neuper, and G. Pfurtscheller. Hidden markov models for online classification of single trial eeg data. *Pattern recognition letters*, 22(12) :1299–1309, 2001. (Cité en pages 11 et 20.)

[175] G. Obozinski, L. Jacob, and J.-P. Vert. Group lasso with overlaps : the latent group lasso approach. *arXiv preprint arXiv :1110.0413*, 2011. (Cité en pages 12, 67, 68, 87 et 88.)

[176] B. A. Olshausen and D. J. Field. Sparse coding with an overcomplete basis set : A strategy employed by v1 ? *Vision research*, 37(23) :3311–3325, 1997. (Cité en pages 36, 47 et 48.)

[177] R. Oostenveld, P. Fries, E. Maris, and J.-M. Schoffelen. Fieldtrip : open source software for advanced analysis of meg, eeg, and invasive electrophysiological data. *Computational intelligence and neuroscience*, 2011, 2010. (Cité en page 64.)

[178] W. Ou, M. S. Hämäläinen, and P. Golland. A distributed spatio-temporal eeg/meg inverse solver. *NeuroImage*, 44(3) :932–946, 2009. (Cité en pages 56 et 57.)

[179] R. D. Pascual-Marqui, C. M. Michel, and D. Lehmann. Low resolution electromagnetic tomography : a new method for localizing electrical activity in the brain. *International Journal of psychophysiology*, 18(1) :49–65, 1994. (Cité en page 56.)

[180] R. D. Pascual-Marqui, C. M. Michel, and D. Lehmann. Segmentation of brain electrical activity into microstates : model estimation and validation. *Biomedical Engineering, IEEE Transactions on*, 42(7) :658–665, 1995. (Cité en pages 16, 26, 132, 134, 139 et 143.)

[181] Y. C. Pati, R. Rezaiifar, and P. Krishnaprasad. Orthogonal matching pursuit : Recursive function approximation with applications to wavelet decomposition. In *Signals, Systems and Computers, 1993. 1993 Conference Record of The Twenty-Seventh Asilomar Conference on*, pages 40–44. IEEE, 1993. (Cité en pages 45 et 122.)

[182] A. J. Pegna, A. Khateb, L. Spinelli, M. Seeck, T. Landis, and C. M. Michel. Unraveling the cerebral dynamics of mental imagery. *Human brain mapping*, 5(6) :410–421, 1997. (Cité en page 132.)

[183] G. Peyré, J. Fadili, et al. Learning analysis sparsity priors. *Sampta'11*, 2011. (Cité en pages 38 et 128.)

[184] G. Pfurtscheller and F. H. Lopes da Silva. Event-related eeg/meg synchronization and desynchronization : basic principles. *Clinical neurophysiology*, 110(11) :1842–1857, 1999. (Cité en page 14.)

[185] G. Pfurtscheller, C. Neuper, A. Schlogl, and K. Lugger. Separability of eeg signals recorded during right and left motor imagery using adaptive autoregressive parameters. *Rehabilitation Engineering, IEEE Transactions on*, 6(3) :316–325, 1998. (Cité en pages 15 et 20.)

[186] M. Pham, T. Hinterberger, N. Neumann, A. Kübler, N. Hofmayer, A. Grether, B. Wilhelm, J.-J. Vatine, and N. Birbaumer. An auditory brain-computer interface based on the self-regulation of slow cortical potentials. *Neurorehabilitation and Neural Repair*, 19(3) :206–218, 2005. (Cité en page 15.)

[187] M. D. Plumbley, T. Blumensath, L. Daudet, R. Gribonval, and M. E. Davies. Sparse representations in audio and music : from coding to source separation. *Proceedings of the IEEE*, 98(6) :995–1005, 2010. (Cité en pages 2 et 50.)

[188] J. Polich. Updating p300 : an integrative theory of p3a and p3b. *Clinical neurophysiology*, 118(10) :2128–2148, 2007. (Cité en pages 13 et 101.)

[189] H. Ramoser, J. Muller-Gerking, and G. Pfurtscheller. Optimal spatial filtering of single trial eeg during imagined hand movement. *Rehabilitation Engineering, IEEE Transactions on*, 8(4) :441–446, 2000. (Cité en pages 15 et 26.)

[190] J. Rapin, J. Bobin, A. Larue, and J.-L. Starck. Sparse redundant formulations and non-negativity in blind source separation. In *Signal Processing Conference (EUSIPCO), 2013 Proceedings of the 21st European*, pages 1–5. IEEE, 2013. (Cité en page 44.)

[191] D. Regan. Chromatic adaptation and steady-state evoked potentials. *Vision research*, 8(2) :149–158, 1968. (Cité en page 12.)

[192] B. Rivet, A. Souloumiac, V. Attina, and G. Gibert. xdawn algorithm to enhance evoked potentials : application to brain–computer interface. *Biomedical Engineering, IEEE Transactions on*, 56(8) :2035–2043, 2009. (Cité en pages 26 et 51.)

[193] N. Rougier. Schéma d'un neuronne, `http://fr.wikipedia.org/wiki/Neurone#mediaviewer/File:Neuron-figure-fr.svg`., 2013. (Cité en pages 11 et 10.)

[194] R. Rubinstein, A. M. Bruckstein, and M. Elad. Dictionaries for sparse representation modeling. *Proceedings of the IEEE*, 98(6) :1045–1057, 2010. (Cité en pages 34 et 46.)

[195] R. Rubinstein, T. Faktor, and M. Elad. K-svd dictionary-learning for the analysis sparse model. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pages 5405–5408. IEEE, 2012. (Cité en page 128.)

[196] R. Rubinstein, T. Peleg, and M. Elad. Analysis k-svd : a dictionary-learning algorithm for the analysis sparse model. *Signal Processing, IEEE Transactions on*, 61(3) :661–677, 2013. (Cité en page 50.)

[197] R. Rubinstein, M. Zibulevsky, and M. Elad. Double sparsity : Learning sparse dictionaries for sparse signal approximation. *Signal Processing, IEEE Transactions on*, 58(3) :1553–1564, 2010. (Cité en page 50.)

[198] L. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D : Nonlinear Phenomena*, 60(1-4) :259–268, 1992. (Cité en pages 44, 75 et 111.)

[199] M. Salinsky, R. Kanter, and R. M. Dasheiff. Effectiveness of multiple eegs in supporting the diagnosis of epilepsy : an operational curve. *Epilepsia*, 28(4) :331–334, 1987. (Cité en page 17.)

[200] A. Schlögl. *The electroencephalogram and the adaptive autoregressive model : theory and applications*. Citeseer, 2000. (Cité en page 19.)

[201] K. Schnass and P. Vandergheynst. Dictionary preconditioning for greedy algorithms. *Signal Processing, IEEE Transactions on*, 56(5) :1994–2002, 2008. (Cité en page 40.)

[202] C. Sieluzycki, R. Konig, A. Matysiak, R. Kus, D.Ircha, and P. J. Durka. Single-trial evoked brain responses modeled by multivariate matching pursuit. *Biomedical Engineering, IEEE Transactions on*, 56(1) :74–82, 2009. (Cité en page 54.)

[203] K. Skretting and K. Engan. Recursive least squares dictionary learning algorithm. *Signal Processing, IEEE Transactions on*, 58(4) :2121–2130, 2010. (Cité en page 49.)

[204] R. Srinivasan, P. L. Nunez, D. M. Tucker, R. B. Silberstein, and P. J. Cadusch. Spatial sampling and filtering of eeg with spline laplacians to estimate cortical potentials. *Brain topography*, 8(4) :355–366, 1996. (Cité en page 24.)

[205] W. Strik, T. Dierks, T. Becker, and D. Lehmann. Larger topographical variance and decreased duration of brain electric microstates in depression. *Journal of Neural Transmission/General Section JNT*, 99(1-3) :213–222, 1995. (Cité en pages 17 et 132.)

[206] M. W. Tangermann. Feature selection for brain-computer interfaces. 2007. (Cité en page 27.)

[207] M. Tappaz. Dans l'intimité du cerveau : Organisation, développement et plasticité, http://www.upsavoie-mb.fr/wp/wp-content/plugins/calendrierUP/getDetails.php?eventID=4094., 2013. (Cité en pages 11 et 9.)

[208] C. E. Tenke and J. Kayser. Reference-free quantification of eeg spectra : combining current source density (csd) and frequency principal components analysis (fpca). *Clinical Neurophysiology*, 116(12) :2826–2846, 2005. (Cité en page 25.)

[209] T. S. Tian and Z. Li. A spatio-temporal solution for the eeg/meg inverse problem using group penalization methods. *Stat. Interface*, 4 :521–533, 2011. (Cité en page 58.)

[210] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996. (Cité en pages 37 et 122.)

[211] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight. Sparsity and smoothness via the fused Lasso. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 67(1) :91–108, 2005. (Cité en pages 76 et 111.)

[212] R. Tibshirani and J. Taylor. The solution path of the generalized lasso. *The Annals of Statistics*, 39(3) :1335–1371, 2011. (Cité en page 112.)

[213] R. Tibshirani and G. Walther. Cluster validation by prediction strength. *Journal of Computational and Graphical Statistics*, 14(3) :511–528, 2005. (Cité en page 134.)

[214] R. Tomioka and K.-R. Müller. A regularized discriminative framework for eeg analysis with application to brain–computer interface. *Neuroimage*, 49(1) :415–432, 2010. (Cité en page 27.)

[215] I. Tosic and P. Frossard. Dictionary learning. *Signal Processing Magazine, IEEE*, 28(2) :27–38, 2011. (Cité en page 34.)

[216] G. Townsend, B. LaPallo, C. Boulay, D. Krusienski, G. Frye, C. Hauser, N. Schwartz, T. Vaughan, J. Wolpaw, and E. Sellers. A novel p300-based brain–computer interface stimulus presentation paradigm : moving beyond rows and columns. *Clinical Neurophysiology*, 121(7) :1109–1120, 2010. (Cité en page 14.)

[217] J. Tropp. *Topic in sparse approximation*. PhD thesis, PhD thesis, 2004. (Cité en pages 37, 41 et 47.)

[218] J. Tropp. Algorithms for simultaneous sparse approximation. Part II : Convex relaxation. *Signal Processing*, 86(3) :589–602, 2006. (Cité en page 84.)

[219] J. Tropp, A. Gilbert, and M. Strauss. Algorithms for simultaneous sparse approximation. Part I : Greedy pursuit. *Signal Processing*, 86(3) :572–588, 2006. (Cité en pages 84 et 122.)

[220] J. A. Tropp. Greed is good : Algorithmic results for sparse approximation. *Information Theory, IEEE Transactions on*, 50(10) :2231–2242, 2004. (Cité en pages 39, 41 et 89.)

[221] J. A. Tropp. Algorithms for simultaneous sparse approximation. part ii : Convex relaxation. *Signal Processing*, 86(3) :589–602, 2006. (Cité en page 42.)

[222] J. A. Tropp, A. C. Gilbert, and M. J. Strauss. Algorithms for simultaneous sparse approximation. part i : Greedy pursuit. *Signal Processing*, 86(3) :572–588, 2006. (Cité en pages 42 et 45.)

[223] S. Vaiter, G. Peyré, C. Dossal, and J. Fadili. Robust sparse analysis regularization. *Information Theory, IEEE Transactions on*, 59(4) :2001–2016, 2013. (Cité en page 43.)

[224] J.-J. Vidal. Toward direct brain-computer communication. *Annual review of Biophysics and Bioengineering*, 2(1) :157–180, 1973. (Cité en pages 1 et 8.)

[225] R. N. Vigário. Extraction of ocular artefacts from eeg using independent component analysis. *Electroencephalography and clinical neurophysiology*, 103(3) :395–404, 1997. (Cité en page 26.)

[226] B. Wahlberg, S. Boyd, M. Annergren, and Y. Wang. An ADMM algorithm for a class of total variation regularized estimation problems. *arXiv preprint arXiv :1203.1828*, 2012. (Cité en page 112.)

[227] W. WG. Discussion on recent advances in the eeg diagnosis of epilepsy. *Proceedings of the Royal Society of Medicine*, 44(4) :315–332, 1951. (Cité en pages 1 et 8.)

[228] M. Wirth, H. Horn, T. Koenig, A. Razafimandimby, M. Stein, T. Mueller, A. Federspiel, B. Meier, T. Dierks, and W. Strik. The early context effect reflects activity in the temporo-prefrontal semantic system : evidence from electrical neuroimaging of abstract and concrete word reading. *NeuroImage*, 42(1) :423–436, 2008. (Cité en page 132.)

[229] P. J. Wolfe, S. J. Godsill, and M. Dorfler. Multi-gabor dictionaries for audio time-frequency analysis. In *Applications of Signal Processing to Audio and Acoustics, 2001 IEEE Workshop on the*, pages 43–46. IEEE, 2001. (Cité en page 46.)

[230] J. R. Wolpaw, N. Birbaumer, D. J. McFarland, G. Pfurtscheller, and T. M. Vaughan. Brain–computer interfaces for communication and control. *Clinical neurophysiology*, 113(6) :767–791, 2002. (Cité en pages 1 et 17.)

[231] K. G. Woodgate. Least-squares solution of< i> f</i>=< i> pg</i> over positive semidefinite symmetric< i> p</i>. *Linear algebra and its applications*, 245 :171–190, 1996. (Cité en page 72.)

[232] J. Wright, R. Kydd, and A. Sergejew. Autoregression models of eeg. *Biological cybernetics*, 62(3) :201–210, 1990. (Cité en page 19.)

[233] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(2) :210–227, 2009. (Cité en page 50.)

[234] C. Wu and X. Tai. Augmented Lagrangian method, dual methods, and split Bregman iteration for ROF, vectorial TV, and high order models. *SIAM Journal on Imaging Sciences*, 3(3) :300–339, 2010. (Cité en page 113.)

[235] S. R. y Cajal. *Comparative study of the sensory areas of the human cortex*. 1899. (Cité en pages 11 et 10.)

[236] G. Ye and X. Xie. Split Bregman method for large scale fused Lasso. *Computational Statistics & Data Analysis*, 55(4) :1552–1569, 2011. (Cité en pages 112 et 114.)

[237] X. Yong, R. K. Ward, and G. E. Birch. Sparse spatial filter optimization for eeg channel reduction in brain-computer interface. In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pages 417–420. IEEE, 2008. (Cité en page 59.)

[238] H. Yuan, V. Zotev, R. Phillips, W. C. Drevets, and J. Bodurka. Spatiotemporal dynamics of the brain at rest—exploring eeg microstates as electrophysiological signatures of bold resting state networks. *Neuroimage*, 60(4) :2062–2072, 2012. (Cité en page 17.)

[239] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 68(1) :49–67, 2006. (Cité en pages 42 et 122.)

[240] J. Zhou, V. Liu, J.and Narayan, and J. Ye. Modeling disease progression via fused sparse group lasso. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1095–1103. ACM, 2012. (Cité en page 76.)

[241] D. Zhu, J. Bieger, G. G. Molina, and R. M. Aarts. A survey of stimulation methods used in ssvep-based bcis. *Computational intelligence and neuroscience*, 2010 :1, 2010. (Cité en page 12.)

[242] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 67(2) :301–320, 2005. (Cité en pages 44 et 58.)

Publications

Les travaux réalisés durant cette thèse ont fait/feront l'objet de publications dans des conférences et des revues à comité de lecture.
L'étude décrite dans le chapitre 5 a ainsi été présentée dans une conférence française de traitement du signal (GRETSI 2013) et un article complet sera soumis dans les semaines à venir à une revue internationale (*IEEE Transactions on Biomedical Engineering*).
Le schéma d'optimisation proposé dans le chapitre 6 a été présenté dans une conférence internationale (ICASSP 2013), un article journal plus détaillé sera également soumis dans quelques jours (resoumission à *IEEE Transactions on Signal Processing* après correction).
Par ailleurs, cette thèse a été également l'occasion de travailler avec Quentin Barthélemy sur un apprentissage de dictionnaire multivarié invariant par translation pour les signaux EEG, entraînant la publication d'un article dans la revue *Journal of neuroscience methods*.

Revues

Q. Barthélemy, C. Gouy-Pailler, **Y. Isaac**, A. Souloumiac, A. Larue, and J.I. Mars. Multivariate temporal dictionary learning for EEG. *Journal of neuroscience methods*, 215(1) :19–28, 2013.

Conférences

Y. Isaac, Q. Barthélemy, J. Atif, C. Gouy-Pailler, and M. Sebag. Multi-dimensional sparse structured signal approximation using split bregman iterations. In *Acoustics Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 3826–3830. IEEE, 2013.

Y. Isaac, Q. Barthélemy, J. Atif, C. Gouy-Pailler, M. Sebag. Régularisations spatiales pour la décomposition de signaux EEG sur un dictionnaire temps-fréquence. In *Colloque Gretsi XXIV*, 2013.

En cours

Y. Isaac, Q. Barthélemy, J. Atif, C. Gouy-Pailler, M. Sebag. Multi-dimensional signal approximation with sparse structured priors using split Bregman iterations. *Signal Processing (EURASIP)*, 2015.

Y. Isaac, Q. Barthélemy, J. Atif, C. Gouy-Pailler, M. Sebag. Spatially constrained sparse decomposition of electroencephalographic signals : application to P300 detection. *Biomedical Engineering, IEEE Transactions on*, 2015.

Y. Isaac, Q. Barthélemy, C. Gouy-Pailler, J. Atif, M. Sebag. Généralisation des micro-états EEG par apprentissage

régularisé temporellement de dictionnaires topographiques. Régularisations spatiales pour la décomposition de signaux EEG sur un dictionnaire temps-fréquence. In *Colloque Gretsi XXV*, 2015.

Représentations redondantes pour les signaux d'électroencéphalographie

Yoann Isaac

Résumé : L'électroencéphalographie permet de mesurer l'activité du cerveau à partir des variations du champ électrique à la surface du crâne. Cette mesure est utilisée pour le diagnostic médical, la compréhension du fonctionnement du cerveau ou dans les systèmes d'interface cerveau-machine. De nombreux travaux se sont attachés au développement de méthodes d'analyse de ces signaux en vue d'en extraire différentes composantes d'intérêt, néanmoins leur traitement pose encore de nombreux problèmes.
Cette thèse s'intéresse à la mise en place de méthodes permettant l'obtention de représentations redondantes pour ces signaux. Ces représentations se sont avérées particulièrement efficaces ces dernières années pour la description de nombreuses classes de signaux grâce à leur grande flexibilité. L'obtention de telles représentations pour les mesures EEG présente certaines difficultés du fait d'un faible rapport signal à bruit des composantes recherchées. Nous proposons dans cette thèse de les surmonter en guidant les méthodes considérées vers des représentations physiologiquement plausibles des signaux EEG à l'aide de régularisations. Ces dernières sont construites à partir de connaissances *a priori* sur les propriétés spatiales et temporelles de ces signaux. Pour chacune d'entres elles, des algorithmes sont proposés afin de résoudre les problèmes d'optimisation associés à l'obtention de ces représentations. L'évaluation des approches proposées sur des signaux EEG souligne l'efficacité des régularisations proposées et l'intérêt des représentations obtenues.

Mots clés : électroencéphalographie, représentation redondante, régularisation, décomposition parcimonieuse, apprentissage de dictionnaires

Redundant representations
for electroencephalography signals

Yoann Isaac

Abstract :

The electroencephalography measures the brain activity by recording variations of the electric field on the surface of the skull. This measurement is usefull in various applications like medical diagnosis, analysis of brain functionning or whithin brain-computer interfaces. Numerous studies have tried to develop methods for analyzing these signals in order to extract various components of interest, however, none of them allows to extract them with sufficient reliabilty.

This thesis focuses on the development of approaches considering redundant (overcomoplete) representations for these signals. During the last years, these representations have been shown particularly efficient to describe various classes of signals due to their flexibility. Obtaining such representations for EEG presents some difficuties due to the low signal-to-noise ratio of these signals.

We propose in this study to overcome them by guiding the methods considered to physiologically plausible representations thanks to well-suited regularizations. These regularizations are built from prior knowledge about the spatial and temporal properties of these signals. For each regularization, an algorithm is proposed to solve the optimization problem allowing to obtain the targeted representations. The evaluation of the proposed EEG signals approaches highlights their effectiveness in representing them.

Keywords : electroencephalography, overcomplete representation, regularization, sparse decomposition, dictionnary learning

Annexe A

Equation de Sylvester

Les régularisations mises en place pour les différents modèles de décompositions étudiés agissent sur les colonnes de la matrice de décomposition. Les développements mathématiques utilisés pour la résolution des problèmes d'optimisation associés aux décompositions correspondantes aboutissent alors à des équations matricielles (en X) ayant la forme suivante :

$$AX + XB = C$$

avec $X \in \mathbb{R}^{n,m}$, $A \in \mathbb{R}^{n,n}$, $B \in \mathbb{R}^{m,m}$ et $C \in \mathbb{R}^{n,m}$.
Celle-ci est nommée équation de Sylvester et apparaît notamment dans l'étude des systèmes dynamiques. Sans informations supplémentaires sur les matrices A, B et C, résoudre une équation de Sylvester peut nécessiter de lourds calculs, particulièrement pour des matrices de grandes dimensions. Une solution sous forme fermée peut être dérivée de la formulation vectorielle du problème :

$$(I_m \otimes A + B^T \otimes I_n)vect(X) = vect(C) \ ,$$

où $\otimes$ est le produit de Kronecker, I_n la matrice identité de dimension n et $vect(.)$ l'opérateur de vectorisation. Toutefois, cette formulation sous la forme d'un système linéaire de taille $nm \times nm$ peut poser des problèmes (en temps et en mémoire nécessaire) de résolution lorsque la taille des matrices est grande. Ainsi, l'algorithme de Bartels–Stewart [14] est souvent privilégié à la résolution directe du système linéaire précédent.
Lors de la résolution des problèmes de décomposition qui nous occupe, la structure des matrices impliquées facilite la résolution de ce problème. Les matrices A et B apparaissant pour nos problèmes d'optimisation sont en effet symétriques réelles et rentrent dans le cadre de la proposition qui suit.

Proposition 1. *Soit A et B deux matrices symétriques réelles. Résoudre en X l'équation de Sylvester $AX + XB = C$ revient à résoudre en X' le système diagonal suivant :*

$$D_A X' + X' D_B = C' \ , \tag{A.1}$$

où

$$\begin{aligned} A &= F D_A F^T, & B &= G D_B G^T \ , \\ X' &= F^T X G, & C' &= F^T C G \ . \end{aligned} \tag{A.2}$$

Démonstration. A et B étant des matrices symétriques réelles, elles sont diagonalisables dans des bases orthogonales (F et G) Eq. (A.2). L'équation de Sylvester peut alors être réécrite comme suit :

$$FD_AF^TX + XGD_BG^T = M \ ,$$

et l'application des propriétés d'orthogonalités de F et G permet d'obtenir la proposition précédente. □

La résolution du système diagonal obtenu peut alors être réalisée comme suit :

$$\forall i \in \{1, \ldots, m\} \quad \hat{X}'(i) = (D_A + D_B(i,i)I_n)^{-1}C'(i)$$

avec $\hat{X}'$ la solution du problème diagonal.
De façon équivalente, nous avons :

$$\begin{aligned} \forall j \in \{1, \cdots, n\}, \quad \forall i \in \{1, \cdots, m\}, \\ (D_A(i,i) + D_B(j,j))\ \hat{X}'(i,j) = \ C'(i,j), \end{aligned} \tag{A.3}$$

et la solution peut être calculée via l'expression suivante :

$$\hat{X}' = C' \oslash O, \tag{A.4}$$

où $\oslash$ correspond à une division terme à terme et

$$O(j,i) = D_A(j,j) + D_B(i,i) \ . \tag{A.5}$$

La solution $\hat{X}$ de l'équation de Sylvester présentée plus haut est alors obtenue par :

$$\hat{X} = F\hat{X}'G^T.$$

Cette inversion n'est possible que lorsque O ne contient pas d'éléments nuls et donc lorsque les ensembles de valeurs propres des matrices A et B ne contiennent pas d'éléments opposés. En fait, dans un cadre plus général (matrices A et B quelconques), il est possible de montrer que l'équation de Sylvester possède une unique solution seulement lorsque cette condition est respectée, ce qui est toujours le cas dans les problèmes d'optimisations associés aux décompositions que nous étudions ici.

Annexe B

Problème direct

Nous nous interessons ici à la relation présente entre l'activité électrique des sources cérébrales et les mesures de cette activité obtenues sur les électrodes. Il est possible de montrer que sous certaines hypothèses cette relation peut être approchée par un modèle linéaire. Pour cela nous suivons ici le raisonnement décrit dans [23] [1].

Localement, soit $\mathbf{B}$ le champ magnétique et $\mathbf{j}$ la densité volumique du courant, l'équation de Maxwell-Ampère (sous hypothèse quasi-statique) s'écrit :

$$\mathbf{rot}(\mathbf{B}) = \mu_0 \mathbf{j}$$

où **rot** est l'opérateur rotationnel et μ_0 la perméabilité magnétique du vide. De plus, pour un champ électrique $\mathbf{E}$, la loi d'Ohm s'écrit localement : $\mathbf{j} = \sigma \mathbf{E}$ avec σ la conductivité du milieu considéré et le potentiel électrique est défini par $\nabla \mathbf{V} = -\mathbf{E}$. Il est alors possible d'écrire l'équation de Poisson :

$$\begin{aligned} & div(\mathbf{rot}(\mathbf{B})) = \mu_0 div(\mathbf{j}) = \mu_0 div(\sigma \mathbf{E}) \\ \Leftrightarrow \quad & div(\mathbf{j}) = -div(\sigma \nabla \mathbf{V}) \\ \Leftrightarrow \quad & \frac{div(\mathbf{j})}{\sigma} = \Delta V \end{aligned}$$

avec div l'opérateur de divergence.
Soit $\mathbf{y} \in \mathbb{R}^C$ le vecteur des potentiels électriques mesurés sur un ensemble de C électrodes. Le potentiel électrique $\mathbf{y}(j) = V(\mathbf{x})$ mesuré sur la j-ième électrode placée en $\mathbf{x}$ peut s'exprimer en utilisant le principe de superposition par intégration volumique comme suit :

$$\mathbf{y}(j) = \int \mathbf{g}\ \mathbf{j}(\mathbf{r})\ d\mathbf{r},$$

où $\mathbf{g}$ correspond au champ de sensibilité.
En choisissant un ensemble de S dipôles représentant des sources ponctuels placés aux positions $\{\mathbf{r}_i, i \in \{1, \ldots, S\}\}$ réparties uniformément dans le cerveau (correspondant à des diracs dans la densité volumique du courant), ce potentiel électrique s'écrit :

$$\mathbf{y}(j) = \sum_{i=1}^{S} \mathbf{g}(\mathbf{r}_i)^T \mathbf{q}(\mathbf{r}_i),$$

1. Exceptionnelement ici, afin de conserver les noms classiques des variables physiques, les lettres majuscules en gras correspondent à des vecteurs.

avec $\mathbf{q}(\mathbf{r}_i) \in \mathbb{R}^3$ le moment dipôlaire associé au dipôle placé en $\mathbf{r}_i$.
Pour une matrice de gain $G = [\mathbf{g}(\mathbf{r}_1)(1), \mathbf{g}(\mathbf{r}_1)(2), \mathbf{g}(\mathbf{r}_1)(3), \ldots, \mathbf{g}(\mathbf{r}_S)(1), \mathbf{g}(\mathbf{r}_S)(2), \mathbf{g}(\mathbf{r}_S)(3)] \in \mathbb{R}^{C\times 3S}$ et $\bar{\mathbf{q}} = [\mathbf{q}(\mathbf{r}_1), \ldots, \mathbf{q}(\mathbf{r}_S)]^T \in \mathbb{R}^{3S}$ un vecteur où sont concaténés les moments dipôlaires, cette expression s'écrit de manière matricielle :

$$y = G\ \bar{\mathbf{q}}.$$

Dans le cas d'un modèle de tête sphérique, une solution analytique existe pour le calcul de la matrice de gain G. Ce modèle étant peu précis, un modèle réaliste lui est souvent préféré. Le calcul de la matrice de gain peut alors être effectué de manière numérique [163] par la méthode des éléments finis de frontière (« Boundary Element Method » (BEM)) ou la méthode des éléments finis (« Finite Element Method » (FEM)).

Annexe C

Convergence du Multi-SSSA

La preuve de la convergence de l'approche proposée dans le chapitre 6 est présentée ici. Rappelons dans un premier temps le problème d'optimisation considéré :

$$\min_X \|Y - \Phi X\|_F^2 + \lambda_1 \|X\|_1 + \lambda_2 \|XP\|_1 \ .$$

Ainsi que le schéma d'optimisation proposé :

$$\begin{aligned}
X^{i+1} &= \underset{X \in \mathbb{R}^{N_\Phi \times T}}{\arg\min} \|Y - \Phi X\|_F^2 + \frac{\mu_1}{2}\|X - A^i + D_A^i\|_F^2 + \frac{\mu_2}{2}\|XP - B^i + D_B^i\|_F^2 \\
A^{i+1} &= \underset{\frac{\lambda_1}{\mu_1}}{\text{SoftThreshold}}(X^{i+1} + D_A^i) \\
B^{i+1} &= \underset{\frac{\lambda_2}{\mu_2}}{\text{SoftThreshold}}(X^{i+1}P + D_B^i) \\
D_A^{i+1} &= D_A^i + (X^{i+1} - A^{i+1}) \\
D_B^{i+1} &= D_B^i + (X^{i+1}P - B^{i+1}) \quad .
\end{aligned}$$

Nous suivons ici l'analyse de convergence de Osher et al [35] en l'adaptant à notre cas. Plus précisément nous démontrons le théorème suivant :

Théorème 6. *Sous l'hypothèse de positivité des paramètres du schéma* $\lambda_1 \geq 0$, $\lambda_2 \geq 0$, $\mu_1 > 0$ *et* $\mu_2 > 0$, *nous avons :*

$$\begin{aligned}
&\lim_{i \to \infty} \|Y - \Phi X^i\|_F^2 + \lambda_1 \|X^i\|_1 + \lambda_2 \|X^i P\|_1 \\
&= \|Y - \Phi \hat{X}\|_F^2 + \lambda_1 \|\hat{X}\|_1 + \lambda_2 \|\hat{X} P\|_1
\end{aligned} \tag{C.1}$$

dans lequel $\hat{X}$ *est une solution du problème.*
Lorsque ce problème possède une unique solution, nous pouvons déduire de la convexité de la fonction $E(X) = \|Y - \Phi X\|_F^2 + \lambda_1 \|X\|_1 + \lambda_2 \|XP\|_1$ *et de Eq. (6.13) le résultat suivant :*

$$\lim_{i \to \infty} X^i = \hat{X} \ . \tag{C.2}$$

Prenons pour hypothèse que nos paramètres sont positifs : $\lambda_1 \geq 0$, $\lambda_2 \geq 0$, $\mu_1 > 0$ et $\mu_2 > 0$. La condition d'optimalité du premier ordre des sous-problèmes convexes du schéma permet

d'écrire :

$$\begin{aligned}
0 &= (2\Phi^T\Phi + \mu_1 I)X^{i+1} + \mu_2 X^{i+1}PP^T \\
&\quad - 2\Phi^T Y + \mu_1(D_A^i - A^i) + \mu_2(D_B^i - B^i)P^T, \\
0 &= \lambda_1 Q_A^{i+1} - \mu_1(D_A^i - A^{i+1} + X^{i+1}), \\
0 &= \lambda_2 Q_B^{i+1} - \mu_2(D_B^i - B^{i+1} + X^{i+1}P), \\
D_A^{i+1} &= D_A^i + (X^{i+1} - A^{i+1}), \\
D_B^{i+1} &= D_B^i + (X^{i+1}P - B^{i+1}),
\end{aligned} \tag{C.3}$$

où $Q_A^{i+1} \in \partial\|A^{i+1}\|_1$ et $Q_B^{i+1} \in \partial\|B^{i+1}\|_1$.

De plus, la convexité du problème principal assure l'existence d'une solution $\hat{X}$ respectant les conditions KKT. Le lagrangien L de notre problème peut donc être écrit comme suit :

$$L = \|Y - \Phi X\|_F^2 + \lambda_1\|X\|_1 + \lambda_2\|XP\|_1,$$

et $\exists\hat{X}$ tel que

$$\begin{aligned}
0 &= -2\Phi^T(Y - \Phi\hat{X}) + \lambda_1\hat{Q}_A + \lambda_2\hat{Q}_B P^T, \\
\hat{A} &= \hat{X}, \quad \text{and} \quad \hat{B} = \hat{X}P,
\end{aligned} \tag{C.4}$$

où $\hat{Q}_A \in \partial\|\hat{A}\|_1$ et $\hat{Q}_B \in \partial\|\hat{B}\|_1$.

La solution est un point fixe du schéma d'optimisation et vérifie :

$$\begin{aligned}
0 &= (2\Phi^T\Phi + \mu_1 I)\hat{X} + \mu_2\hat{X}PP^T - 2\Phi^T Y \\
&\quad + \mu_1(\hat{D}_A - \hat{A}) + \mu_2(\hat{D}_B - \hat{B})P^T, \\
0 &= \lambda_1\hat{Q}_A - \mu_1(\hat{D}_A - \hat{A} + \hat{X}), \\
0 &= \lambda_2\hat{Q}_B - \mu_2(\hat{D}_B - \hat{B} + \hat{X}P), \\
\hat{D}_A &= \hat{D}_A + (\hat{X} - \hat{A}), \\
\hat{D}_B &= \hat{D}_B + (\hat{X}P - \hat{B}).
\end{aligned} \tag{C.5}$$

En soustrayant Eq. (C.5) à Eq. (C.3), le système d'équations précédent est conservé pour les variables d'erreurs suivantes :

$$\begin{aligned}
\tilde{X}^i = X^i - \hat{X},\ \tilde{A}^i = A^i - \hat{A},\ \tilde{B}^i = B^i - \hat{B}, \\
\tilde{D}_A^i = D_A^i - \hat{D}_B,\ \tilde{D}_B^i = D_B^i - \hat{D}_B \\
\tilde{Q}_A^i = Q_A^i - \hat{Q}_A,\ \tilde{Q}_B^i = Q_B^i - \hat{Q}_B.
\end{aligned}$$

En réalisant maintenant le produit scalaire de la première ligne par $\tilde{X}^{i+1}$, le produit scalaire de la seconde ligne par $\tilde{A}^{i+1}$, le produit scalaire de la troisième ligne par $\tilde{B}^{i+1}$ et en

prenant la norme de Frobenius au carré des deux dernières lignes le système suivant peut être dérivé :

$$\begin{aligned}
0 &= 2\|\Phi\tilde{X}^{i+1}\|_F^2 + \mu_1\|\tilde{X}^{i+1}\|_F^2 + \mu_2\langle\tilde{X}^{i+1}, \tilde{X}^{i+1}PP^T\rangle \\
&\quad + \mu_1(\langle\tilde{X}^{i+1}, \tilde{D}_A^i\rangle - \langle\tilde{X}^{i+1}, \tilde{A}^i\rangle) + \mu_2(\langle\tilde{X}^{i+1}, \tilde{D}_B^i P^T\rangle - \langle\tilde{X}^{i+1}, \tilde{B}^i P^T\rangle), \\
0 &= \lambda_1\langle\tilde{A}^{i+1}\tilde{Q}_A^{i+1}\rangle - \mu_1(\langle\tilde{A}^{i+1}\tilde{D}_A^i\rangle - \|\tilde{A}^{i+1}\|_F^2 + \langle\tilde{A}^{i+1}, \tilde{X}^{i+1}\rangle), \\
0 &= \lambda_2\langle\tilde{B}^{i+1}, \tilde{Q}_B^{i+1}\rangle - \mu_2(\langle\tilde{B}^{i+1}, \tilde{D}_B^i\rangle - \|\tilde{B}^{i+1}\|_F^2 + \langle\tilde{B}^{i+1}, \tilde{X}^{i+1}P\rangle), \\
\|\tilde{D}_A^{i+1}\|_F^2 &= \|\tilde{D}_A^i\|_F^2 + (\|\tilde{X}^{i+1}\|_F^2 + \|\tilde{A}^{i+1}\|_F^2 \\
&\quad - 2\langle\tilde{X}^{i+1}, \tilde{A}^{i+1}\rangle) - 2\langle\tilde{D}_A^i, \tilde{X}^{i+1} - \tilde{A}^{i+1}\rangle, \\
\|\tilde{D}_B^{i+1}\|_F^2 &= \|\tilde{D}_B^i\|_F^2 + (\|\tilde{X}^{i+1}P\|_F^2 + \|\tilde{B}^{i+1}\|_F^2 \\
&\quad - 2\langle\tilde{X}^{i+1}P, \tilde{B}^{i+1}\rangle) - 2\langle\tilde{D}_B^i, \tilde{X}^{i+1}P - \tilde{B}^{i+1}\rangle.
\end{aligned}$$

La sommation des trois premières équations et la réécriture des deux autres permet d'écrire le système comme suit :

$$\begin{aligned}
0 = \quad & 2\|\Phi\tilde{X}^{i+1}\|_F^2 + \mu_1\|\tilde{X}^{i+1}\|_F^2 + \mu_2\langle\tilde{X}^{i+1}, \tilde{X}^{i+1}PP^T\rangle \\
& + \lambda_1\langle\tilde{A}^{i+1}, \tilde{Q}_A^{i+1}\rangle + \lambda_2\langle\tilde{B}^{i+1}, \tilde{Q}_B^{i+1}\rangle \\
& + \mu_1(\langle\tilde{X}^{i+1}, \tilde{D}_A^i\rangle - \langle\tilde{X}^{i+1}, \tilde{A}^i\rangle - \langle\tilde{A}^{i+1}, \tilde{D}_A^i\rangle + \|\tilde{A}^{i+1}\|_F^2 - \langle\tilde{A}^{i+1}, \tilde{X}^{i+1}\rangle) \\
& + \mu_2(\langle\tilde{X}^{i+1}, \tilde{D}_B^i P^T\rangle - \langle\tilde{X}^{i+1}, \tilde{B}^i P^T\rangle - \langle\tilde{B}^{i+1}, \tilde{D}_B^i\rangle + \|\tilde{B}^{i+1}\|_F^2 - \langle\tilde{B}^{i+1}, \tilde{X}^{i+1}P\rangle), \\
\langle\tilde{D}_A^i, \tilde{X}^{i+1} - \tilde{A}^{i+1}\rangle = \quad & \frac{1}{2}(\|\tilde{D}_A^{i+1}\|_F^2 - \|\tilde{D}_A^i\|_F^2 - \|\tilde{X}^{i+1} - \tilde{A}^{i+1}\|_F^2), \\
\langle\tilde{D}_B^i, \tilde{X}^{i+1}P - \tilde{B}^{i+1}\rangle = \quad & \frac{1}{2}(\|\tilde{D}_B^{i+1}\|_F^2 - \|\tilde{D}_B^i\|_F^2 - \|\tilde{X}^{i+1}P - \tilde{B}^{i+1}\|_F^2).
\end{aligned}$$

La combinaison des équations de ce système permet alors en sommant entre $i = 1$ et $i = S$ d'obtenir :

$$\begin{aligned}
& \frac{\mu_1}{2}(\|\tilde{D}_A^1\|_F^2 - \|\tilde{D}_A^S\|_F^2) + \frac{\mu_2}{2}(\|\tilde{D}_B^1\|_F^2 - \|\tilde{D}_B^S\|_F^2) \\
& \quad = 2\sum_{i=1}^{S}\|\Phi\tilde{X}^i\|_F^2 + \sum_{i=1}^{S}\lambda_1\langle\tilde{A}^{i+1}, \tilde{Q}_A^{i+1}\rangle + \lambda_2\langle\tilde{B}^{i+1}, \tilde{Q}_B^{i+1}\rangle \\
& \quad + \frac{\mu_1}{2}(-\|\tilde{A}^1\|_F^2 + \sum_{i=1}^{S}\|\tilde{X}^{i+1} - \tilde{A}^{i+1}\|_F^2 + \|\tilde{X}^{i+1} - \tilde{A}^i\|_F^2 + \|A^S\|_F^2) \\
& \quad + \frac{\mu_2}{2}(-\|\tilde{B}^1\|_F^2 + \sum_{i=1}^{S}\|\tilde{X}^{i+1}P - \tilde{B}^{i+1}\|_F^2 + \|\tilde{X}^{i+1}P - \tilde{B}^i\|_F^2 + \|B^S\|_F^2).
\end{aligned}$$

La norme $\|.\|_1$ étant convexe, les termes $\langle\tilde{A}^i, \tilde{Q}_A^i\rangle$ et $\langle\tilde{B}^i, \tilde{Q}_B^i\rangle$ sont positifs ($\forall i$). Ainsi, μ_1, μ_2, λ_1 et λ_2 étant non-négatifs, tous les termes de l'équation ci-dessus sont non-négatifs

et nous avons :

$$\begin{aligned}\frac{\mu_1}{2}(\|\tilde{D}_A^1\|_F^2+\|\tilde{A}^1\|_F^2)+\frac{\mu_2}{2}(\|\tilde{D}_B^1\|_F^2+\|\tilde{B}^1\|_F^2)\\ \geq\ 2\sum_{i=1}^{S}\|\Phi\tilde{X}^i\|_F^2+\sum_{i=1}^{S}\lambda_1\langle\tilde{A}^{i+1},\tilde{Q}_A^{i+1}\rangle+\lambda_2\langle\tilde{B}^{i+1},\tilde{Q}_B^{i+1}\rangle\\ +\frac{\mu_1}{2}\|\tilde{X}^{i+1}-\tilde{A}^i\|_F^2+\frac{\mu_2}{2}\|\tilde{X}^{i+1}P-\tilde{B}^i\|_F^2.\end{aligned}$$

De cette dernière expression, suit :

$$\sum_{i=1}^{\infty}\|\Phi\tilde{X}^i\|_F^2<\infty, \tag{C.6}$$

$$\sum_{i=1}^{\infty}\langle\tilde{A}^{i+1},\tilde{Q}_A^{i+1}\rangle\ <\infty,\ \sum_{i=1}^{\infty}\langle\tilde{B}^{i+1},\tilde{Q}_B^{i+1}\rangle\ <\infty, \tag{C.7}$$

$$\sum_{i=1}^{\infty}\|\tilde{X}^{i+1}-\tilde{A}^i\|_F^2\ <\infty,\ \sum_{i=1}^{\infty}\|\tilde{X}^{i+1}P-\tilde{B}^i\|_F^2\ <\infty, \tag{C.8}$$

permettant d'obtenir le théorème 6. En effet, en considérant l'eq. (C.7) ainsi que les propriétés de la distance de Bregman (Osher et al [35] eq. (3.16)) nous obtenons :

$$\lim_{i\to\infty}\|A^i\|_1-\|\hat{A}\|_1-\langle A^i-\hat{A},\ \hat{Q}_A\rangle=0, \tag{C.9}$$

$$\lim_{i\to\infty}\|B^i\|_1-\|\hat{B}\|_1-\langle B^i-\hat{B},\ \hat{Q}_B\rangle=0, \tag{C.10}$$

qui, combinées avec l'eq. (C.8) donne :

$$\lim_{i\to\infty}\|X^i\|_1-\|\hat{X}\|_1-\langle X^i-\hat{X},\ \hat{Q}_A\rangle=0, \tag{C.11}$$

$$\lim_{i\to\infty}\|X^iP\|_1-\|\hat{X}P\|_1-\langle X^i-\hat{X},\ \hat{Q}_BP^T\rangle=0, \tag{C.12}$$

et qui finalement en prenant λ_1 eq. (C.11) + λ_2 eq. (C.12) et en utilisant la relation obtenue dans l'eq. (C.4) permet d'écrire :

$$\begin{aligned}\lim_{i\to\infty}\|X^i\|_1-\|\hat{X}\|_1+\|X^iP\|_1-\|\hat{X}P\|_1\\ -\langle X^i-\hat{X},\ 2\Phi^T(Y-\Phi\hat{X})\rangle=0.\end{aligned} \tag{C.13}$$

De plus, $\|\Phi\tilde{X}^i\|_F^2=\langle\nabla f(X^i)-\nabla f(\hat{X}),X^i-\hat{X}\rangle$ pour $f(X)=\|Y-\Phi X\|_F^2$. f étant une fonction convexe, Eq. (C.6) associée à la propriété de la distance de Bregman utilisée précédemment, permet d'obtenir :

$$\begin{aligned}\lim_{i\to\infty}\|Y-\Phi X^i\|_F^2-\|Y-\Phi\hat{X}\|_F^2\\ -\langle X^i-\hat{X},-2\Phi^T(Y-\Phi\hat{X})\rangle,\end{aligned} \tag{C.14}$$

qui, avec l'eq. (C.13) permet l'obtention du premier résultat du théorème.

$$\begin{aligned}\lim_{i\to\infty}\|Y-\Phi X^i\|_F^2+\lambda_1\|X^i\|_1+\lambda_2\|X^iP\|_1\\ =\|Y-\Phi\hat{X}\|_F^2+\lambda_1\|\hat{X}\|_1+\lambda_2\|\hat{X}P\|_1=0.\end{aligned} \tag{C.15}$$

La seconde partie prend pour hypothèse l'existence d'un unique minimum du problème précédent.
La fonction $g(X) = \|Y - \Phi X\|_F^2 + \lambda_1 \|X\|_1 + \lambda_2 \|XP\|_1$ est convexe et continue. Par conséquent, lorsque g possède un unique minimum, nous avons :

$$\lim_{i \to \infty} g(X^i) = g(\hat{X}) \Rightarrow \lim_{i \to \infty} X^i = \hat{X}. \tag{C.16}$$

Pour la preuve de ce dernier point voir [35].