Comparaison des Méthodes de Classification

Lucas Maisonnave - Dorian Joubaud - Richard Szeberin

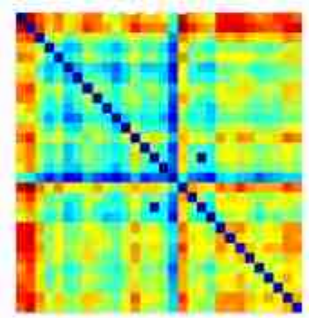
lucas.maisonnave@etu.emse.fr
djoubaud@gmail.com
rszeberin@gmail.com

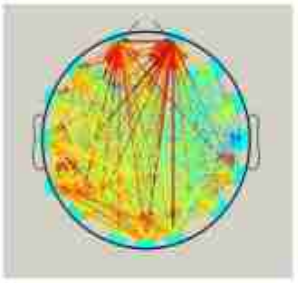
Table des matières

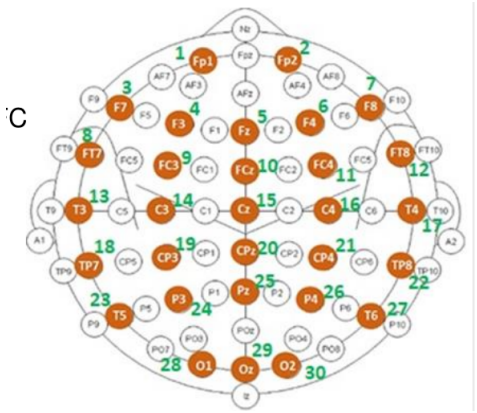
Introduction

L'objectif de ce projet est d'étudier des signaux provenant d'électroencéphalogrammes de patients de l'hôpital de Charles-Foix. Les signaux ont été prélevé grâce à 30 électrodes sur 102 personnes pour 4 bandes de fréquence : **Alpha**, **Beta**, **Delta** et **Theta**, on dispose donc de 408 observations. Les données sont des matrices de connectivité extraites depuis 30 capteurs posés sur chaque patients.
On classe nos données en 3 groupes correspondant au groupe à laquelle appartient nos patients : **AD**(Possible Alzheimer's Disease patients), **SCI** (Subjective Cognitive Impairment patients) et **MCI** (Mild Cognitive Impairments patients) contenant respectivement 28, 22 et 52 patients.
Notre objectif est d'étudier les matrices de connectivités des patients afin de prédire si un patients peut développer la maladie d'Alhzeimer et d'étudier les comportements entre les classes.

On ne s'intéressera ici, qu'à un niveau de résolution de 100% pour nos matrices de connectivités.







1 Analyse des données

La base de données utilisées comporte 3 types d'individus différents : AD, MCI et SCI correspondant respectivement à des individus atteints d'alzheimer, atteints d'une autre maladie et des patient sains. Pour discriminer ces 3 types de patients on s'intéresse aux valeurs de connectivités entre les capteurs déposés sur le crâne du patient pour différentes bandes de fréquence : Alpha, Bêta, Delta, Theta. Chaque patient est donc décris par 4 matrices de connectivités 30x30 (la figure 6 qui correspond à la bande bêta d'un

individu de classe AD). Les valeurs de connectivités sont comprises entre -1 et 1 et correspondent à la corrélation entre les capteurs, plus la valeur absolue est élevée plus les signaux mesurés sont similaires.

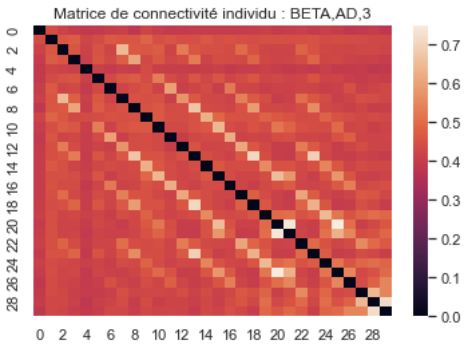
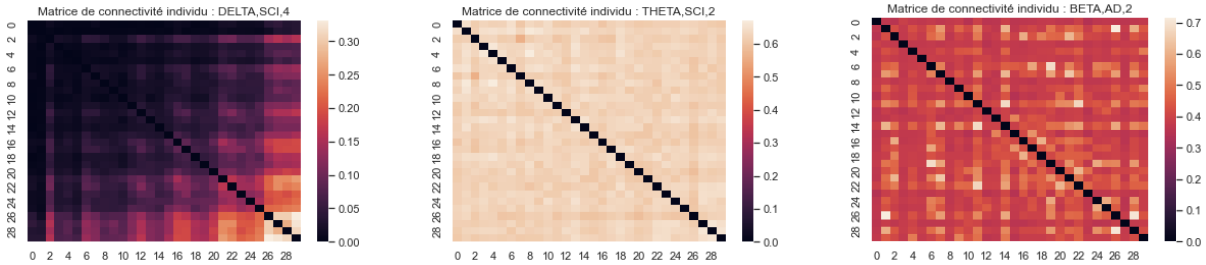


FIGURE 1 – Matrice de connectivité

Dans un premier temps il est important visualiser les différences entre les 3 type de patients. Pour cela on identifie 4 types de matrice de connectivités (figure 6 et 2), et on remarque que certaines matrices comme figure 2b ne contient quasiement pas d'information utile pour discriminer nos classes. La matrice 6 contient des diagonales plus claires qui correspondent à une forte relation entre un capteur et le 5ème plus loin, qui s'explique par la disposition des électrodes. En effet l'électrode 5 et 10 (figure **??**) par exemple sont juste à coté ce qui explique les observations de connectivité. Ensuite, le pattern de la matrice 2a contient peu de relations entre les capteurs et l'information semble être plutôt situé dans la partie occipitale du cerveau. L'information de la matrice 2c est plus éparse et semble contenir plus de relations entre les différents capteurs.



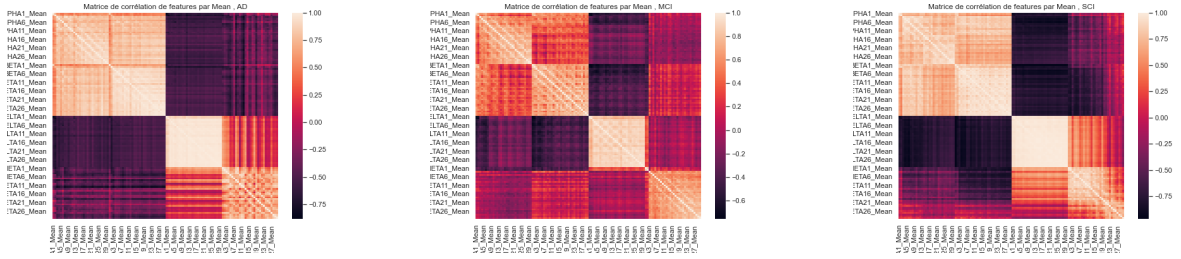
(a) DELTA, SCI, 4 (b) THETA, SCI, 2 (c) BETA, AD, 2

FIGURE 2 – 3 types de matrices de connectivité

Finalement on observe que les matrices de connectivités sont très différentes celon le type de signal, la classe du patient et même celon l'individu étudié. Il faut donc raffiner l'analyse pour mieux comprendre les différences entre les 3 types de patients. Pour cela, on transforme les matrices de connectivité en vecteur de features en prennant dans un premier temps la moyenne des connectivités de chaque électrode. En rassemblant les vecteurs ainsi obtenu pour chaque bande de fréquence, on obtient un vecteur de taille $4 * 30 = 120$ dimensions pour chaque patient (102 individus).

La base de données est donc constitué de 102 individus décris par un vecteur de 120 dimensions, dont 28 AD, 22 SCI et 52 MCI. Pour mieux comprendre ce qui peut

discriminer ces 3 classes on calcul les matrices de corrélations sur les individus de chaque type indépendamment (figure 3). On constate dans un premier temps que certaines variables sont très corrélées pour toutes les classes comme les variables delta qui ne semble pas pouvoir discriminer à elles seules. De même, on remarque de fortes similaritées entre les matrices 3a et 3c, donc l'information est identiquement distribué pour ces 2 classes ce qui semble signifier qu'elles seront plus difficiles à discriminer. Cependant la matrice 3b est quant à elle très différente des 2 autres ce qui pourrait vouloir dire que les individus classés MCI sont plus faciles à discriminer des 2 autres.



(a) AD (b) MCI (c) SCI

FIGURE 3 – Matrices de corrélations

Finalement l'objectif de discriminer les 3 types de patients est envisageable mais pourrait être déséquilibré à première vue. En effet, la classe MCI semble être plus facilement détectable, à cela s'ajoute un déséquilibre du nombre d'individus entre les types de patients. Il est donc nécessaire, avant de procéder à la classification de correctement rééquilibre les données pour ne pas favoriser les MCI et obtenir un modèle biaisé. Pour cela, on se contentera de supprimer les classes en trop afin d'obtenir le même nombre d'individus pour chaque classe.

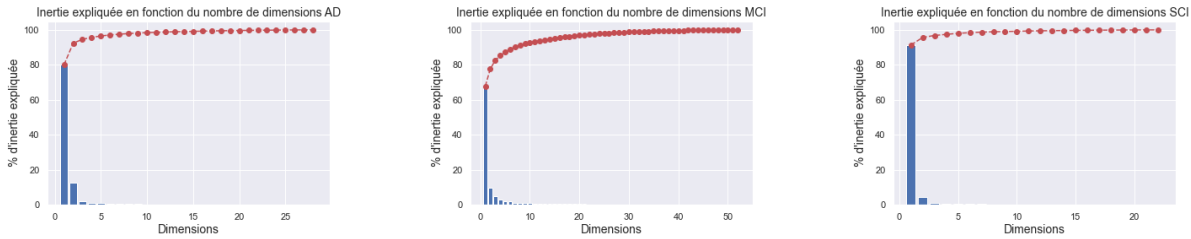
2 Méthodes de sélection de variables et de réduction de dimension

Avant de s'intéresser à la classification il est important de mettre en forme nos données de façon à ce qu'elles représentent au mieux la répartition des patients. Pour cela on s'intéressera majoritairement à l'Analyse en Composantes Principales (ACP) afin de réduire la dimension de nos données ainsi que pour de la sélection de variables. Ces 2 méthodes seront ensuite comparé à la méthode LDA qui est sensé être plus performante.

2.1 ACP pour la sélection de variables

Cette méthode à pour objectif de sélectionner les variables les plus importantes pour les 3 types de patients. Pour cela, on isole les individus selon leurs types (AD, SCI, MCI) et on réalise une ACP sur chacun des groupes. La figure 5 permet de visualiser l'inertie cumulée des 3 groupes et on constate que la première composante contient plus

de 70% de l'information pour les 3 classes. La première composante semble donc être essentiel et contient la majorité de l'information, on peut donc se demander quelles variables contribuent le plus à ce nouvel axe.



(a) AD (b) MCI (c) SCI

FIGURE 4 – Inertie cumulée de l'ACP

Afin de connaitre les variables qui ont le plus d'influence sur le premier axe on calcul les contributions de chaque variable et on les classes par ordre décroissant pour un contribution cumulée égale à 50%. Le tableau 1 permet de visualiser ces variables les plus importantes pour chaque classe. On constate que le signal Delta monopolise l'entièreté des variables les plus importantes ce qui fait écho à l'analyse des corrélations en partie 1 où l'on constatait de fortes corrélations entre les variables delta. Maintenant on isole ces variables pour chaque classe et on les combine pour la classification binaire (AD vs MCI, AD vs SCI, MCI vs SCI)

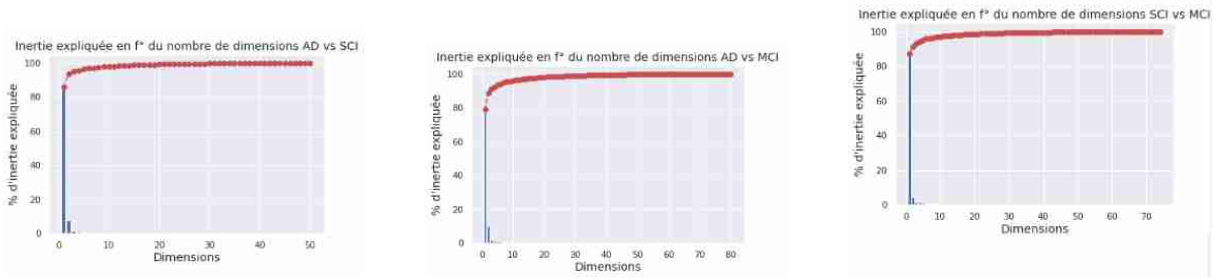
AD	**MCI**	**SCI**
DELTA1	DELTA2	DELTA1
DELTA12	DELTA1	DELTA2
DELTA6	DELTA6	DELTA6
DELTA2	DELTA5	DELTA4
DELTA5	DELTA4	DELTA5
DELTA4	DELTA11	DELTA11
DELTA9	DELTA16	DELTA9
DELTA11	DELTA10	DELTA10
DELTA8	DELTA12	DELTA8
DELTA10	DELTA3	DELTA12
DELTA3	DELTA26	DELTA14
DELTA16	DELTA21	DELTA13
DELTA13	DELTA7	DELTA3
DELTA18		DELTA19
DELTA14		DELTA15

TABLE 1 – Variables qui contribuent le plus au premier axe

2.2 ACP pour la réduction de dimensions

Utilisons maintenant l'ACP, pour réduire la dimensions de notre espace de travail. En effet, dans cette partie, on regroupe les individus par un couple de type, dans le

but d'étudier le comportement d'un type contre un autre type seulement. Pour ce faire on réunit nos données en 3 bases : AD vs SCI, AD vs MCI et SCI vs MCI. Les nouvelles bases contiennent respectivement 50, 80 et 74 observations.
On détermine le nombre de dimension à retenir tel que notre ACP conserve 99% de l'inertie de nos données.



(a) AD vs SCI (b) AD vs MCI (c) SCI vs MCI

FIGURE 5 – Inertie cumulée de l'ACP

	AD vs SCI	AD vs MCI	MCI vs SCI
Dimensions retenues	17	30	25

TABLE 2 – Dimensions retenues par bases

2.3 LDA

Afin d'augmenter la performance de notre modèle de classificatiion, on compare les resultats d'ACP avec un autre mèthode de réduction de dimension, notamment l'Analyse Discriminant Linéaire (LDA). Ce technique permet de maximiser la séparabilité des catégories connues dans notre variable cible ("MCI", "SCI", "AD") tout en réduisant les dimensions. Le choix des nouvels axes, dans ce scénario, est très différent de celui de l'ACP. Dans ce cas, nous finissons par perdre une quantité importante de variance en projettant les données sur les axes retenus. Cependant, nous obtenons une meilleure séparation des deux catégories. Êtant donné qu'on a 3 classes cibles, on peut garder un maximum de 2 axes (nombre de classes - 1) axes principals. Afin de trouver la meilleure projection, il est nécessaire de comparer les performances des projections unidimensionelle et bidimensionelle.

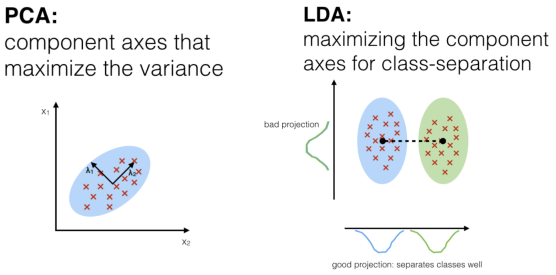


FIGURE 6 – PCA v. LDA

3 Classification

3.1 MLP Classifier

Pour réaliser notre classification, on utilise un perceptron multicouche (MLP). Notre modèle est composé de 2 couches de 10 neurones suivit de fonction d'activation Sigmoïde, puis d'un softmax pour la dernière couche. On prendra un learning rate α de 0.001 et *Adam* comme optimizer. Afin d'empêcher un sur-apprentissage de nos données, on utilisera des couches de DropOut.

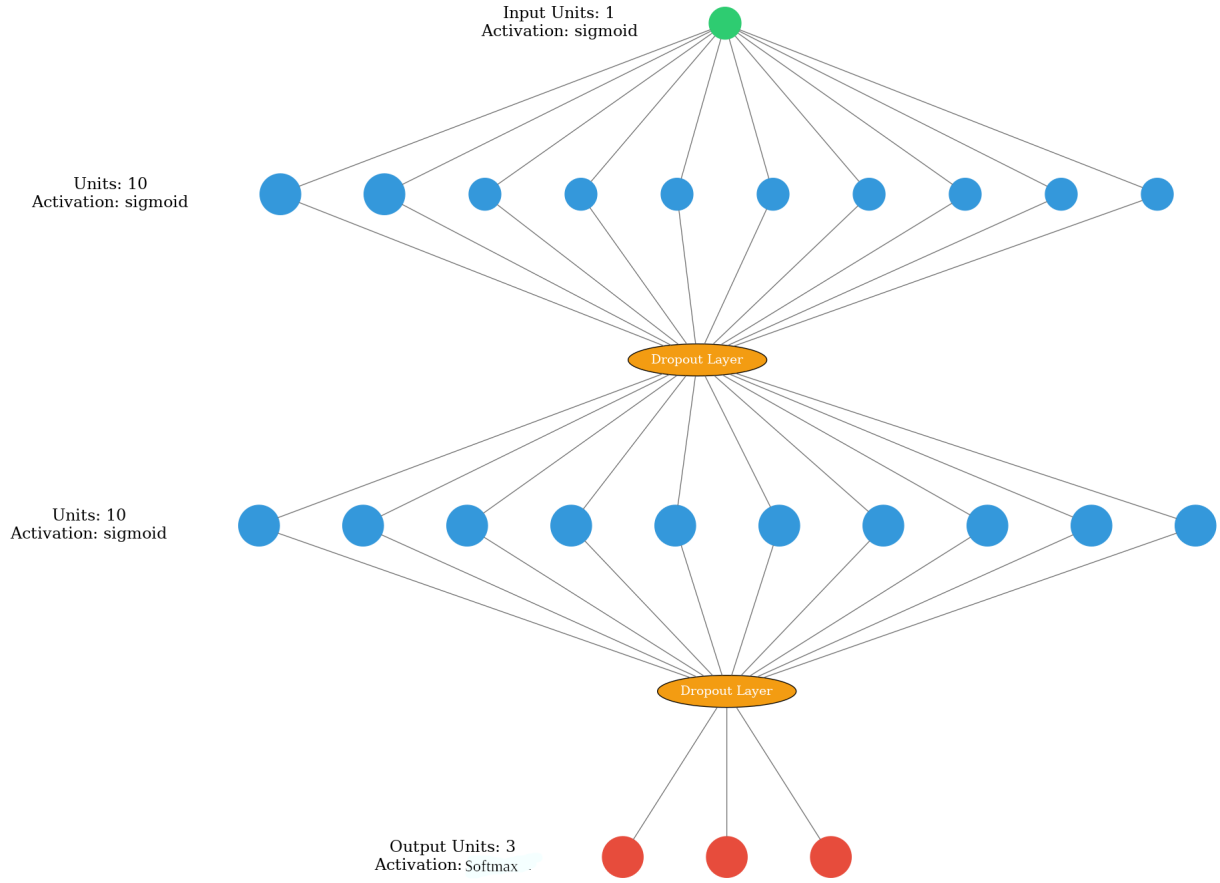


FIGURE 7 – Représentation du MLP utilisé

3.2 Validation croisée

Afin de comparer des différentes configurations du modéle MLP, on utilise des méthodes de validation croisée K-Fold. En plus d'une validation croisée 5-fold standard, nous utilisons LOOCV qui est la manière la plus robuste de tester des modèles de classification. La méthode LOOCV est recommandée pour valider les modèles construits sur des ensembles de données plus petits où une division standard test/train peut introduire un biais significatif dans le modèle. Cependant, LOOCV est également le plus coûteux en

termes de calcul. En utilisant les hyperparamètres dans le tableau 1, on a trouvé que la meilleure performance et obtenu par des configurations différentes pour chaque problème dee classification (tableau 2).

K	N	5			
Activation	Tanh	ReLU			
Couches cachées	(5, 5, 2)	(10, 5, 5, 2)	(5, 2)	(20)	(10, 30, 10)
Optimiser	LBFGS	SGD	Adam		
α	0.00001	0.0001	0.0005	0.01	0.05
Learning rate	constant	adaptive			

TABLE 3 – Validation croisée : configurations considerées

K	N			5		
Type	MCI/SCI	MCI/AD	SCI/AD	MCI/SCI	MCI/AD	SCI/AD
Activation	Tanh	Tanh	ReLU	Tanh	Tanh	ReLU
Couches cachées	(5, 5, 2)	(10, 5, 5, 2)	(5, 2)	(10, 5, 5, 2)	(10, 5, 5, 2)	(5, 10, 5)
Optimiser	LBFGS	LBFGS	LBFGS	LBFGS	LBFGS	LBFGS
α	0.0001	0.05	0.05	0.0001	0.0001	0.0005
Learning rate	constant	constant	adaptative	adaptative	constant	constant

TABLE 4 – Meilleures configuration pour chaque problème de classification

4 Résultats

Le modèle de classification ainsi défini obtient les résultats table 5 par Leave-One-Out et pour les 3 méthodes de mise en forme des données. On constate dans un premier temps que seule la réduction de dimension (ACP) obtient des performances acceptables (accuracy supérieure à 60%). La sélection de variable ne semble pas du tout fonctionner, et en particulier pour AD vs SCI. De plus, avec l'ACP et la sélection, la classe MCI semble être mieux classifié par ces modèles (VN $\geq$ 85%) et semble avoir plus de difficultés pour les autres classes. On a donc un modèle très confiant pour la classe MCI et beaucoup pour les autres. Avec l'ACP notre modèle permet donc de discriminer précisément un patient malade mais par forcément d'alzheimer. La méthode LDA ne semble pas très performante avec les paramètres choisis, il pourrait être intéressant de mieux raffiner le modèle afin d'obtenir de meilleurs paramètres pour discriminer.

	LDA			ACP			Sélection		
Type	Accuracy	VP	VN	Accuracy	VP	VN	Accuracy	VP	VN
AD vs MCI	48.21	42	54	**71.42**	57	86	69.64	50	89
AD vs SCI	59	54.5	63.6	**63.63**	60	72	13.63	4.5	22.7
SCI vs MCI	61	50	72.7	**70.5**	55.2	85.7	74.8	50	100

TABLE 5 – Tableau des résultats en %

Le modèle de classification discrimine correctement les MCI comme nous l'avions remarqué en partie 1 sur les matrices de corrélation. En effet, l'ACP étant basé sur les corrélations en maximisant la variance des axes, il semble cohérent de penser que notre modèle discrimine mieux les MCI grâce à cette différence de corrélations entre les variables. Les autres classes nécessitent cependant une meilleure mise en forme pour obtenir de meilleurs résultats en classification.

5 Conclusion

Pour conclure, nous avons réalisé 3 modèles différents afin de d'étudier les matrices de connectivités et de prédire si un patient peut développer la maladie d'Alhzeimer. Nous avons pour cela étudier les comportement entre les classes AD, SCI et MCI.
Nos modèles ne donne pas d'excellent résultats mais on peut en déduire un comportement de nos données.Le modèle utilisant l'ACP comme réduction de dimension nous donne les meilleurs résultats. On arrive tout de même à mieux déceler et caractériser le groupe des MCI que les autres.