Implémentation d'une procédure visant à retirer l'influence des artéfacts oculaires (EMG) parasitant les enregistrements de l'activité électrique du cerveau (électroencéphalographie – EEG)

Résumé

La présente procédure, devant permettre le retrait de l'influence des artéfacts oculaires d'un signal EEG, a été implémentée sous le logiciel R et se base sur une Séparation Aveugle de Sources (SAS). À l'issue de cette SAS mise en œuvre grâce à l'Analyse en Composantes Indépendantes (ACI), différentes statistiques (distances, corrélations) ont permis d'identifier le groupe auquel appartient chacune des sources estimées. Une fois les sources associées aux artéfacts identifiées, leurs valeurs ont été annulées et un nouveau mélange a permis d'obtenir les observations exemptées des artéfacts oculaires. Cet outil se présente sous forme de deux fonctions R acceptant différents arguments et ayant comme sortie finale, des données prétraitées.

Mots clés: Électroencéphalographie, Séparation Aveugle de Sources, corrélations, distances, Artéfacts oculaires.

Abstract

This tool, that must allow the removal of eye artefacts, was developed under the R software and is based on a Blind Sources Separation (BSS). At the end of this BSS by Independent Components Analysis (ICA), various statistics (distances, correlations) allowed to identify the group of each estimated component. Once the independent components associated with the artefacts were identified, their values were replaced by zero and a new mixture is done to obtain the observations without eye artefacts. This tool contains two R functions accepting arguments and its last output is a cleaned dataset.

Keywords: electroencephalography, Blind Sources Separation, correlations, distances, eye artefacts.

Introduction

La psychologie, discipline appartenant à la catégorie des sciences humaines peut être définie selon le dictionnaire de français Larousse comme l'étude des activités mentales et des comportements en fonction des conditions de l'environnement. Suivant les objets d'investigation, cette discipline se divise en plusieurs sous-disciplines parmi lesquelles la psychologie cognitive. Cette branche d'étude de la psychologie s'intéresse au langage, à la mémoire, au raisonnement ou encore à l'intelligence et utilise l'expérimentation ou les mesures comportementales comme principale méthode d'étude. Parmi les techniques d'expérimentation, l'enregistrement de l'activité électroencéphalographique (EEG) suite à un événement donné c'est à dire les potentiels évoqués (en anglais Event-Related Potentials - ERP) occupe une place de choix en psychologie cognitive.

L'électroencéphalographie est un examen qui permet l'enregistrement de l'activité électrique spontanée des neurones du cortex cérébral. Selon Perret et Laganaro (2013), les données issues d'un tel enregistrement sont constituées de deux types d'information : le signal correspondant à l'activité électrique du cerveau en rapport avec un stimulus ERP et un ensemble d'autres éléments nommés de façon générique « Bruit ». Ce bruit peut provenir de signaux aberrants parmi lesquels les artéfacts associés aux clignements et mouvements d'yeux qui en représentent une grande partie. Avant toute analyse de ces données EEG/ERP, ce bruit doit être identifié puis éliminé. Tel est l'objectif visé par la procédure présentée ci-dessous.

Ce prétraitement est parfois fait manuellement et consiste à identifier puis à supprimer grâce à une évaluation visuelle les parties jugées associées à une activité électromyographique, ces dernières étant souvent matérialisées par de brusques fluctuations remarquables sur un électroencéphalogramme. Cette activité électromyographique provoquée par la sollicitation de muscles lors de clignements et mouvements d'yeux génère un potentiel électrique qui est aussi enregistré lors de l'EEG.

La mise en œuvre de la procédure proposée permet d'obtenir en fin de processus, un jeu de données débarrassé de l'influence électromyographique des artéfacts oculaires. Pour y parvenir, l'idée est de séparer aveuglément au moyen d'algorithmes existants, les différentes sources composant l'enregistrement obtenu grâce à l'EEG. Une fois cette séparation faite, on identifie les sources génératrices du bruit, on annule leur influence et on ne retient que le signal provenant de la source cérébrale. Comment se fait cette séparation aveugle de sources ? Comment identifie-t-on les différentes sources et que fait-on pour annuler l'influence des sources indésirables ?

Avant de répondre à ces différentes questions, il est primordial de remarquer que la séparation aveugle de sources avait déjà été utilisée par différents auteurs pour le retrait des artéfacts électroencéphalographiques. En 2000, Jung et *al*. proposaient en lieu et place de la suppression des parties des signaux associées aux artéfacts qui génère une grande perte de données, l'utilisation de la séparation aveugle de sources pour éliminer tous les artéfacts enregistrés lors d'une EEG comme les clignements et mouvements d'yeux ou aussi les mouvements du cœur ou d'un autre muscle.

Aussi Flexer, Bauer et Pripfl ont montré en 2005 que l'Analyse en Composantes Indépendantes (ACI) qui est une technique de séparation aveugle de sources permettait d'éliminer avec succès les artéfacts oculaires présents dans un enregistrement EEG. Pour ce faire, ils avaient comparé les enregistrements EEG/ERP nettoyés au moyen de l'ICA de sujets aveugles et un enregistrement EEG/ERP d'un sujet sans globes oculaires qui par conséquent ne présentait pas d'artéfacts oculaires.

Afin d'ôter des données EEG l'influence de leurs artéfacts, le logiciel Matlab de moins en moins utilisé par les psychologues a été très utilisé par la communauté scientifique jusqu'en 2010. De plus en plus, de nouveaux packages sont disponibles sous le logiciel R, prisé des nouveaux chercheurs pour l'analyse de données EEG ou particulièrement de données EEG/ERP.

Il sera présenté dans un premier temps la structure des données EEG puis les techniques utilisées pour la séparation, l'identification et l'annulation des différentes sources.

1. Electroencéphalographie et structure des données provenant de l'EEG

L'électroencéphalographie peut être définie comme une méthode qui a pour but principal de mesurer l'activité électrique du cerveau grâce à des électrodes disposées sur le cuir chevelu d'un individu. L'enregistrement de l'EEG procure aux chercheurs et praticiens des renseignements instantanés sur l'activité neurophysiologique du cerveau.

Le tracé issu de l'EEG est appelé électroencéphalogramme. Pour tracer cet électroencéphalogramme, un flux de données est enregistré suivant un taux d'échantillonnage bien déterminé. Ce taux peut être de l'ordre de 128 Hz, 512 Hz, etc. Un hertz (Hz) étant équivalent à un événement par seconde, un taux d'échantillonnage de 512 Hz impliquent donc 512 enregistrements/seconde soit 1 enregistrement toutes les 1,95 millisecondes (environ 2 ms) pour une électrode. En utilisant donc un casque à 128 électrodes pour une électroencéphalographie à 512 Hz qui dure 10 secondes, le flux de données serait identique à une matrice comportant 128 lignes et 512 x 10 = 5120 colonnes ou inversement en fonction de la disposition des données. On peut en déduire que la taille des données obtenues lors

d'une EEG est fonction du nombre d'électrodes utilisées, du taux d'échantillonnage et de la durée de l'enregistrement.

Parmi ces trois paramètres, la durée de l'enregistrement varie fortement d'une expérience à l'autre. Particulièrement quand il s'agit d'observer l'activité électrique du cerveau en temps réel suite à un stimulus, la durée d'enregistrement peut être longue. Par exemple, l'utilisation d'un enregistrement EEG pour l'étude d'une production de texte conceptuellement dirigée dure environ une demi-heure. Demi-heure au cours de laquelle, outre l'activité électrique associée au stimulus qui est la production de texte dans le cadre de l'exemple choisi, il sera aussi enregistré l'activité électrique associée à différents artéfacts comme les clignements ou mouvements d'yeux, les mouvements parasites des membres, etc. Il en résulte donc que le signal enregistré provient d'un mélange de différentes sources. De ces artéfacts, les clignements d'yeux sont fréquents, irréguliers et remarquables lorsqu'on représente les données issues de l'enregistrement. La figure 1 présente un jeu de données EEG issu de 130 électrodes représenté à l'aide du logiciel CARTOOL. Au cours de cet enregistrement, en plus des 128 électrodes placées sur le cuir chevelu, deux électrodes ont été placées autour des yeux pour capter l'activité électrique de ces derniers. On peut remarquer sur la figure 1 que le tracé de ces deux électrodes e129, e130 (les 2 derniers tracés de couleur noire en allant du haut vers le bas) diffère de celui des électrodes e1 à e128. Aussi, remarquons sur les 130 électrodes la figure 1, des variations brusques du tracé qui correspondent à l'enregistrement de l'activité électrique associée aux clignements d'yeux.

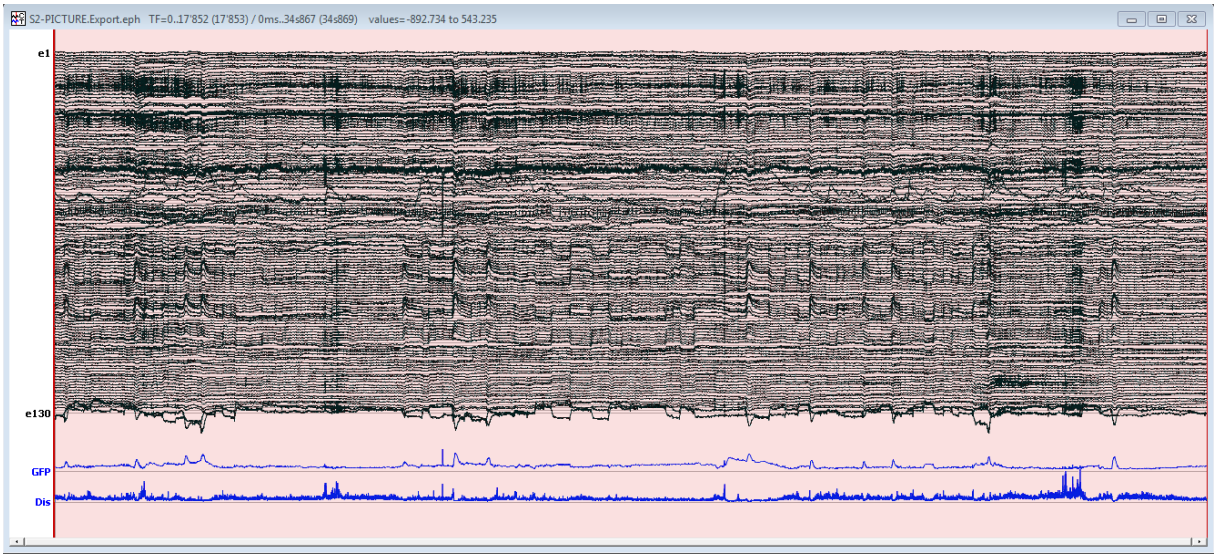

Figure 1 : Représentation d'un jeu de données EEG dans le logiciel CARTOOL

La structure des données représentées dans la figure 1 est celle prise en charge par la procédure de prétraitement proposée. Elles se présentent sous la forme d'une matrice de n lignes et de m+2 colonnes. Le nombre de lignes n résulte du produit du temps d'enregistrement en secondes et du taux d'échantillonnage en hertz tandis que m désigne le nombre d'électrodes présentes sur le casque EEG. Les 2 dernières colonnes situées en fin de matrice renseignent sur le signal provenant des deux électrodes placées autour des yeux. La figure 2 est une illustration de la structure des données décrite ci-dessus.

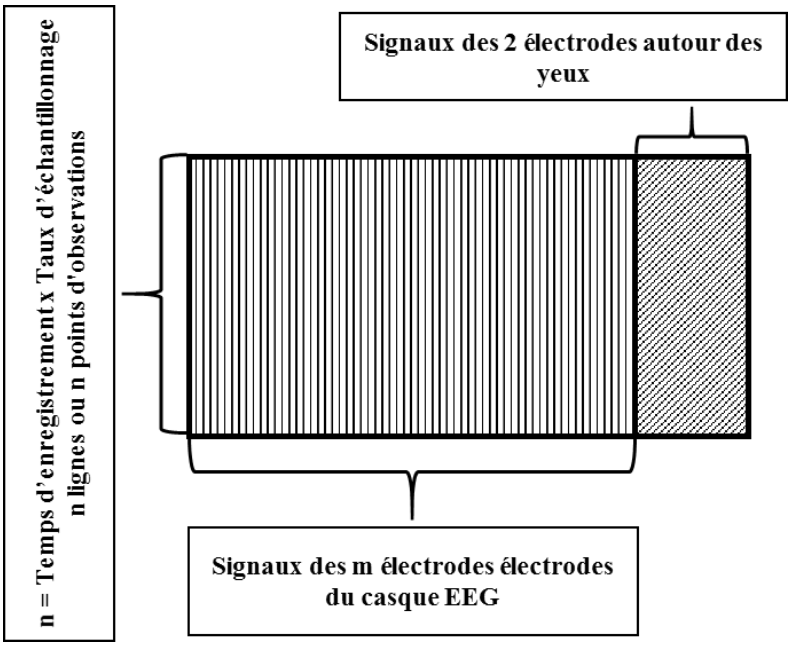

Figure 2 : Illustration de la structure des données acceptées par l'outil de prétraitement

2. Méthodologie utilisée pour le développement de l'outil

Quatre (04) étapes sont nécessaires pour la mise en œuvre de la procédure proposée. Partant des observations obtenues suite à l'expérimentation, il faut :

- Effectuer une séparation Aveugle de Sources (SAS) en considérant le nombre d'électrodes utilisées pour l'expérimentation comme nombre de sources primitives
- Identifier le groupe auquel appartient chacune des sources primitives estimées (Electromyographie (EMG) ou Electroencéphalographie (EEG))
- Annuler l'influence des sources appartenant au groupe EMG
- Mélanger à nouveau toutes les sources primitives

2.1. Séparation Aveugle de Sources (SAS) et Analyse en composantes Indépendantes (ACI)

2.1.1. Définitions et théories

Avant de définir chacun de ces deux termes, on peut dire que l'Analyse en Composantes Indépendantes est un outil permettant originellement la séparation aveugle de sources. Pour mieux comprendre cette dernière, utilisons comme illustration le problème de la « soirée cocktail ». Lors d'une soirée, même en présence du bruit issu d'un mélange de voix et de musique, l'ouïe humaine est capable de discerner naturellement la voix qui l'intéresse, celle de son interlocuteur par exemple. Cette séparation qui consiste à retrouver un certain nombre de sources à partir de l'observation d'un mélange bruité de celles-ci posent le problème de la séparation de sources. Jutten & Gribonval (2003) ajoutent que deux idées fondamentales définissent la SAS à savoir l'utilisation de plusieurs capteurs (principe de diversité) qui fourniront chacun un mélange différent des sources et l'hypothèse que les sources à extraire soient statistiquement indépendantes.

La représentation du modèle des données est illustrée par la figure 3. Si l'on observe N signaux EEG $s_j(t)$ à l'aide de E électrodes (capteurs), on obtient E mélanges EEG $x_i(t)$ tels que :

$$\mathbf{x = F(s)}$$

Avec x les vecteurs des observations et s celui des sources. La fonction F est appelée la fonction de mélange. Si le nombre d'électrodes est supérieur ou égal au nombre de sources ($E \geq N$) et si la fonction F est inversible, estimer son inverse $F^{-1} = H$ permet de séparer les sources. La fonction H est appelée fonction de séparation et on peut écrire :

$$\mathbf{y = H(x) = H\,(F(s)) = s}$$

La seule hypothèse sur les sources étant leur indépendance statistique, la fonction H est estimée de sorte que les sources estimées y(t) soient indépendantes entre elles (Jutten & Gribonval, 2003). Selon ces mêmes auteurs, ce serait cette dernière hypothèse d'indépendance qui met en évidence le lien entre la SAS et l'ACI.

Aussi selon Le Borgne (2004), la restriction des fonctions F et H à des transformations linéaires permettrait de simplifier le problème d'un point de vue calculatoire mais aussi de pouvoir visualiser les sources comme les coordonnées des observations dans une base particulière. Sous cette restriction, les fonctions de mélange F et de séparation H pourront être assimilées à des applications linéaires et se présenteront respectivement sous la forme de matrice de mélange A et de séparation W. Les sources exprimées pourront donc s'écrire :

$$\mathbf{y = Wx = WAs.}$$

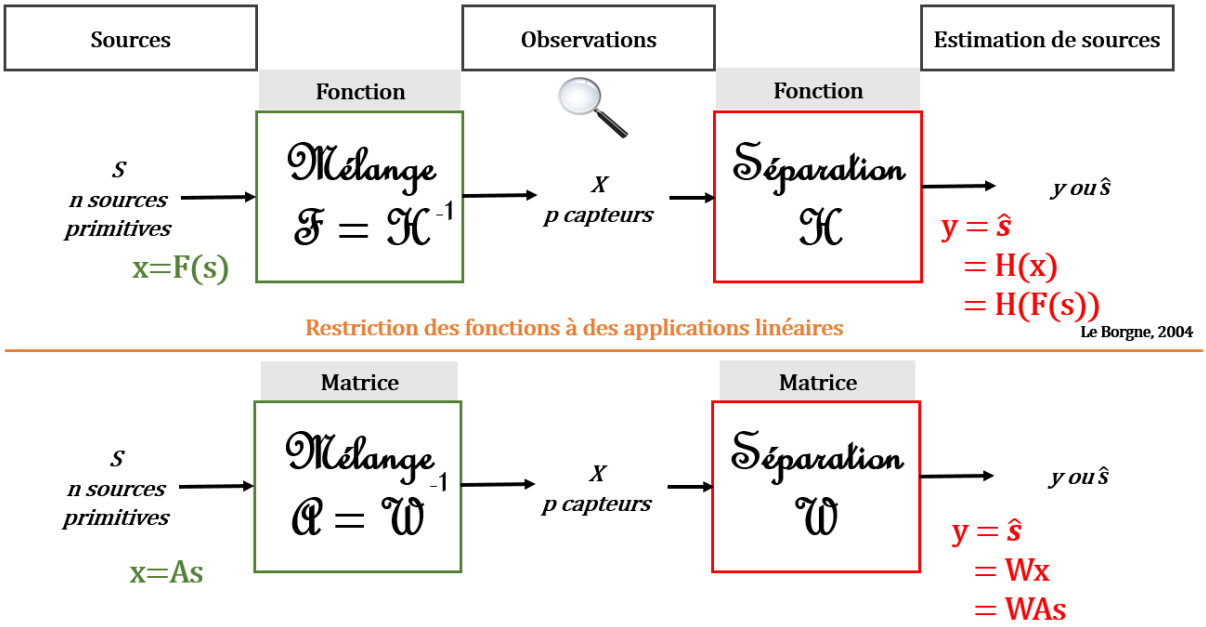

Figure 3 : Représentation du modèle des données pour l'ACI

Faire une ACI revient donc à trouver les différentes matrices W et A qui permettraient la séparation en composantes indépendantes des mélanges observés. Pour y arriver, il est primordial de trouver une fonction de coût ainsi qu'un algorithme d'optimisation. Même si ces deux paramètres sont souvent liés, plusieurs algorithmes d'optimisation et fonctions de coût existent et permettent la mise en œuvre de l'ACI. Dans le cadre du développement du présent outil, deux algorithmes au choix ont été étudiés à savoir l'algorithme FastICA avec la néguentropie comme fonction de coût et l'algorithme « Information-Maximisation » (Infomax).

2.1.2. L'algorithme rapide du point fixe ou l'algorithme FastICA

L'algorithme FastICA, comme on peut le deviner à son nom est l'un des algorithmes les plus rapides utilisés dans la mise en œuvre de l'ACI. La version originale de cet algorithme basée sur les cumulants a été proposée en 1997 par deux auteurs Hyvärinen et Oja tandis qu'une version statistiquement robuste et efficiente d'un point de vue calculatoire a été proposée par Hyvärinen en 1999. Cette dernière version est celle qui est implémentée dans les différents logiciels en général et utilisée dans la mise en œuvre de la présente procédure.

Que toutes les composantes soient estimées en parallèle ou de manière itérative comme le proposent les différentes versions de l'algorithme, une mesure de la non-gaussianité peut être utilisée comme une fonction de coût. L'utilisation de la non-gaussianité dans la résolution de l'ACI repose sur l'idée fondamentale selon laquelle les observations qui sont issues d'une combinaison linéaire des sources sont plus gaussiennes que n'importe laquelle des sources (théorème centrale limite) et deviennent de moins en moins gaussiennes lorsqu'elles se rapprochent de ces sources primitives recherchées. Plusieurs mesures de non-gaussianité existent parmi lesquelles la néguentropie utilisée comme fonction de coût de l'algorithme rapide du point fixe dans la présente procédure.

Considérons un vecteur aléatoire $y= (y_1, \dots , y_n)^T$ ayant pour densité la fonction f et y_{gauss} un vecteur aléatoire gaussien ayant la même matrice de variance-covariance que y. La néguentropie J s'écrit :

$$\boldsymbol{J(y) = H(y_{gauss}) - H(y)}\text{ , avec }\boldsymbol{H\,(y) = -\int f(y)\log(f(y))\,dy}$$

H désigne l'entropie d'un vecteur aléatoire.

L'algorithme FastICA permettant de retrouver une direction maximisant la non-gaussianité est donnée par (Hyvärinen et *al*., 2001):

1. Centrer les données
2. Blanchir les données pour obtenir z
3. Choisir (aléatoirement par exemple) un vecteur unitaire w de norme unitaire
4. Calculer $w \leftarrow E\{zg(w^Tz)\} - E\{g'(w^Tz)\}w$, où la fonction g peut être choisie entre $g(y) = \tanh(y)$, $g(y) = y^3$ ou $g(y) = ye^{-y^2/2}$
5. Calculer $w \leftarrow w/\|w\|$
6. Si une convergence n'est pas observée, repartir à l'étape 4

Outre la néguentropie, le maximum de vraisemblance aussi peut être utilisé comme fonction de coût avec l'algorithme FastICA. Seulement, l'utilisation du maximum de vraisemblance présente des connections explicites avec un autre algorithme très utilisé, l'algorithme infomax (Hyvärinen & Oja, 2000).

2.1.3. L'algorithme INFOMAX

Cet algorithme est proposé en 1995 par Bell & Sejnowski. Selon ces auteurs, il est issu de la convergence des résultats de deux domaines de recherche à savoir le développement de règles d'apprentissage non-supervisé d'information théorique pour les réseaux de neurones dont le pionnier est Linsker en 1992 puis l'utilisation des statistiques d'ordre supérieur dans le domaine du traitement de signal pour la séparation aveugle de sources ou la déconvolution aveugle. Dans cet algorithme, la mesure de la dépendance utilisée est l'information mutuelle. L'information mutuelle I(Y,X) est obtenue grâce à l'expression :

$$\boldsymbol{I(Y,X) = H(Y) - H(Y|X)}$$

Avec H(Y) l'entropie de l'output et H(Y|X) l'entropie conditionnelle ou encore toute l'entropie contenue dans Y et qui ne provient pas de X.

L'utilisation de la minimisation de l'information mutuelle pour la résolution de l'ACI est inspirée de la théorie de l'information. Cette approche est motivée par le fait que dans certains cas, il ne sera pas toujours réaliste de prétendre que les données suivent le modèle non-bruité de l'ACI. De ce fait, il a été développé cette approche qui ne fait aucune hypothèse sur les données et permet d'obtenir une mesure de la dépendance des composantes d'un vecteur aléatoire. En utilisant cette mesure, nous pouvons définir l'ACI comme la combinaison linéaire qui minimise cette mesure de dépendance (Hyvärinen et *al.*, 2001).

L'information mutuelle peut être optimisée dans le cadre de l'ACI par un algorithme du gradient ou par celui de Newton-Raphson qui est utilisé dans le développement de la présente procédure. L'algorithme Infomax se présente comme suit (Amari et *al.*, 1996):

1. Choisir (aléatoirement par exemple) un vecteur unitaire W(0)
2. Calculer $w(t+1) = w(t) + \eta(t)(I - f(Y)Y^T)w(t)$, avec t une étape d'approximation donnée, $\eta(t)$ une fonction générale qui spécifie la taille des étapes pour les actualisations de la matrice de séparation (habituellement une fonction exponentielle ou une constante), I une matrice identité de dimension n x n, et f(Y) une fonction non linéaire habituellement choisie entre $f(Y) = \tanh(Y)$ ou $f(Y) = Y - \tanh(Y)$
3. Si une convergence n'est pas observée, repartir à l'étape 2

Ces différentes analyses ont été mises en œuvre dans la procédure grâce au package « ica » (Helwig, 2015).

2.2. Identification des différentes sources

Une fois l'ACI effectuée, on obtient autant de composantes indépendantes qu'il y avait d'électrodes lors de l'enregistrement. On peut donc considérer qu'en lieu et place des mélanges obtenus via les capteurs qui sont ici les électrodes à la suite des observations, on est plutôt en présence d'autant de composantes indépendantes à la suite de l'ACI. La première étape de la procédure est donc franchie et il faut maintenant identifier le groupe auquel appartient chaque composante ou source primitive.

Pour ce faire, les observations fournies par les deux électrodes posées autour des yeux sont exploitées. Qu'il vous souvienne, lors de l'enregistrement du signal EEG, outre le casque à électrodes, deux électrodes supplémentaires enregistraient l'activité électrique des yeux. Pour identifier le groupe auquel appartient chaque composante indépendante, les corrélations (corrélation de Bravais-Pearson, corrélation de Spearman) et distances (distance de Minkowski avec une incrémentation de la norme p) sont utilisées pour établir des niveaux d' association entre chacun des signaux des électrodes disposées autour des yeux et chacune des composantes indépendantes obtenues.

Considérant deux variables aléatoires X et Y puis E(X) et $\sigma(X)$ respectivement comme l'espérance et l'écart type de X, la corrélation de Bravais-Pearson r a pour expression :

$$\boldsymbol{r(X,Y) = \frac{E[(X - E(X))(Y - E(Y))]}{\sigma(X)\sigma(Y)}}$$

Souvent noté ρ et basé sur un test de rang, la corrélation de Spearman est calculée grâce à la formule :

$$\boldsymbol{\rho(X,Y) = 1 - \frac{6\sum_{i=1}^{N}[r(X_i) - r(Y_i)]^2}{N^3 - N}}$$

Avec

$r(X_i)$ le rang de X_i dans la distribution X_i … X_N et

$r(Y_i)$ le rang de Y_i dans la distribution Y_i … Y_N , N étant le nombre d'observations des variables X et Y.

Outre les corrélations comme moyen d'identification du groupe auquel appartient chaque composante, un deuxième paramètre de décision a été proposé pour l'identification du groupe. Ce deuxième paramètre est la distance de Minkowski d_p. Considérée comme une généralisation de la distance de Manhattan et de la distance euclidienne, elle équivaut respectivement à ces distances lorsque sa norme p prend les valeurs 1 ou 2. Elle est obtenue grâce à l'expression :

$$\boldsymbol{d_p(X,Y) = (\sum_{i=1}^{n} |x_i - y_i|^p)^{1/p}}$$

Pour le développement de la précédente procédure, la norme p prend 81 valeurs différentes, soit de 1 à 9 avec un pas de 0,1 afin de choisir la norme de la distance permettant d'identifier un maximum de composantes indépendantes appartenant au groupe EMG.

Au terme du calcul de ces corrélations et de ces distances, on obtient pour chaque type de corrélation et pour chaque norme de distance deux vecteurs v1 et v2 de dimension m+2, m étant le nombre d'électrodes présentes sur le casque EEG. Deux vecteurs sont obtenus à raison d'un vecteur par électrode posée autour des yeux. Juger de l'association entre l'information fournie par chaque capteur oculaire et celle contenue dans chaque composante indépendante revient donc à une interprétation de chacune des valeurs de corrélations et de distance. Les composantes indépendantes les moins distantes et les plus distantes (cas de deux signaux négativement corrélées) des signaux enregistrés autour des yeux sont considérées comme appartenant au groupe des artéfacts EMG. D'un autre côté, plus la valeur absolue de la corrélation est proche de 1, mieux le signal de l'électrode oculaire est corrélé à la composante indépendante obtenue en fin d'ICA et cette dernière considérée comme appartenant au groupe EMG. Cette technique a souvent été utilisée par diverses études utilisant l'ACI pour le pré-traitement des données EEG. Les études de Jung et *al.* (2000) et Flexer et *al.* (2005) basées sur la corrélation de Pearson, avaient retenu que lorsque la valeur absolue de la corrélation était supérieure à 0,4, la composante était identifiée comme « bruit ». En fonction de la spécificité des données EEG/ERP, fixer une valeur de référence amène parfois à considérer peu ou trop de composantes comme artéfacts.

Pour remédier à cette difficulté, la présente procédure s'appuie sur le fait que l'amplitude des signaux des artéfacts oculaires et musculaires est très différente de l'amplitude des signaux provenant du cortex cérébral. Et donc, dans chacun des vecteurs des corrélations ou des distances, les corrélations et distances entre signaux des électrodes oculaires et composantes indépendantes comportant du signal provenant du cortex cérébral sont très différentes des corrélations et distances entre signaux des électrodes oculaires et composantes indépendantes comportant du signal provenant des muscles. Les limites de Tuckey encore appelées frontières basses ou hautes des boîtes à moustaches (boxplot) sont utilisées pour identifier les valeurs extrêmes des distributions des corrélations et des distances. Les composantes dont les corrélations et distances sont supérieures à la frontière haute et inférieures à la frontière basse sont donc identifiées comme appartenant aux groupes EMG.

On peut donc en déduire que les longueurs des moustaches jouent un rôle primordial dans l'identification du groupe des composantes. Ces longueurs dépendent des valeurs des frontières basses et hautes obtenues respectivement grâce aux formules max (min, Q_3 Q_1-1,5(Q_3-Q_1)) et min (max,

Q_3+1,5(Q_3-Q_1)). Q_1 et Q_3 représentent les premiers et troisièmes quartiles. Ces limites sont programmées dans l'outil proposé avec en option la possibilité de modifier la valeur par défaut 1,5. Cette valeur qui affecte les limites de Tuckey, les longueurs des moustaches et par conséquent les composantes identifiées peut être modifiée dans l'outil proposé.

Une fois des composantes EMG identifiées grâce à chaque type de corrélation et de distance, il faut sortir une liste des composantes retenues en fin d'identification. Alors pour chaque électrode placée autour des yeux, l'intersection des composantes EMG identifiées par l'ensemble des corrélations d'un côté et celles identifiées par la "meilleure distance" (c'est-à-dire celle dont la norme p permet d'identifier le plus de composantes EMG) d'un autre permet d'obtenir la liste des composantes EMG associées au signal de cet électrode. Une fois cela fait pour chacune des deux électrodes oculaires, l'union des composantes retenues pour chacune des électrode permet d'identifier l'ensemble des composantes EMG présentes dans les sources primitives estimées.

2.3. Annulation de l'influence des composantes EMG et obtention d'observations dépourvues d'artéfacts oculaires

Pour obtenir les observations dépourvus d'artéfacts oculaires, il suffit de remplacer les composantes indépendantes du groupe EMG associées aux artéfacts par des signaux nuls (c'est à dire remplacer toutes les valeurs de ces colonnes de la matrice par 0) avant de procéder à un ré-mélange de ces sources primitives estimées à l'aide de la matrice de mélange (inverse de la matrice de séparation) obtenue lors de la mise en œuvre de l'Analyse en Composantes Indépendantes. Après ce nouveau mélange, nous obtenons des données dépourvues de toute influence des artéfacts oculaires.

3. Présentation de la procédure implémentée

Deux fonctions implémentées sous le langage R composent la procédure de prétraitement des données EEG proposée.

La première fonction prend comme arguments la matrice x des observations provenant de l'enregistrement EEG, l'algorithme à utiliser pour la séparation aveugle de sources (algorithme fastICA avec la néguentropie comme fonction de coût ou l'algorithme Infomax) et le paramètre q égal par défaut à 1,5 intervenant dans le calcul des limites de Tuckey d'une boîte à moustache. Parmi les sorties de la première fonction, nous avons pour chacune des deux électrodes oculaires, un tableau récapitulatif comportant le paramètre utilisé pour l'identification de sources (corrélation ou distance), le numéro de l'électrode oculaire concerné, le nombre total de composantes indépendantes identifiées comme appartenant au groupe EMG et les numéros de ces composantes. Outre ce tableau, on a aussi une liste des composantes retenues pour chaque électrode, une liste des composantes retenues pour toutes les deux électrodes oculaires, la matrice des composantes indépendantes ainsi que les matrices de mélange et de séparation issues de l'ACI. Cette première fonction est donc capable de séparer et d'identifier le groupe auquel appartiennent les sources estimées.

La deuxième fonction accepte comme arguments la matrice des composantes indépendantes, la matrice de mélange des sources primitives, et un vecteur des composantes indépendantes appartenant au groupe EMG. Cette fonction retourne comme résultat un jeu de données (observations ou mélanges) exempt de toutes influences des artéfacts oculaires mais aussi une matrice des composantes indépendantes dans laquelle les valeurs des composantes identifiées sont égales à 0. Cette dernière fonction permet d'annuler l'influence des composantes EMG et de mélanger à nouveau les composantes indépendantes telles qu'elles nous intéressent, c'est-à-dire sans artéfact oculaire.

Conclusion

L’enregistrement de données EEG/ERP dans le cadre de l’étude de la réponse à un stimulus est souvent effectuée dans la recherche en science cognitive. L’analyse de ces données souvent influencées par différents artéfacts parmi lesquels les artéfacts oculaires, oblige les chercheurs à prétraiter ces données avant leur analyse proprement dite. Il est proposé dans le présent document de retirer l’influence de ces artéfacts oculaires à l’aide de deux fonctions implémentées sous le logiciel R. Il faut certes que les données collectées respectent une certaine structure mais cette procédure a l’avantage d’être rapide d’utilisation, performant, accessible, et d’éviter des pertes de données considérables limitant ainsi le biais de prétraitement. Il permettra aussi aux chercheurs d’envisager de plus en plus l’utilisation de l’EEG dans leur protocole de recherche, le retrait de l’influence des artéfacts étant plus aisé et pertinent.

Le présent outil impose une structure des données nécessitant l’enregistrement d’un signal électro-oculographique. À l’avenir, il serait intéressant de pouvoir se passer de l’enregistrement de ce signal électro-oculographique et de pouvoir atteindre le même ou un meilleur résultat.

Bibliographie

Amari S., Cichocki A., & Yang H.H., A new learning algorithm for blind signal separation. *Advances in Neural Information Processing Systems*, 8, 757-763, 1996.

Bell A.J., Sejnowski T.J., An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7, 1129-1159, 1995

Brunet D., Murray M.M., Michel C.M., Spatiotemporal analysis of multichannel EEG: CARTOOL. Computational Intelligence and Neuroscience, doi.10.1155/2011/813870, 2011.

Flexer A., Bauer H., Pripfl J., Dorffner G., Using ICA for removal of ocular artifacts in EEG recorded from blind subjects. *Neural Networks*, 18(7), 998-1005, 2005

Helwig N.E., ica: Independent Component Analysis. R package version 1.0-1, URL https://CRAN.R project.org/package=ica, 2015

Hyvärinen A., Fast and Robust Fixed-Point Algorithms for Independent Component Analysis. *IEEE Transactions on Neural Networks*, 10(3), 626-634, 1999

Hyvärinen A., Karhunen J., and Oja E., Independent Component Analysis. Wiley Interscience, 2001. 481 pages.

Hyvärinen A., Oja E., A Fast Fixed-Point Algorithm for Independent Component Analysis. *Neural Computation*, 9(7), 1483-1492, 1997

Hyvärinen A., Oja E., Independent Component Analysis: Algorithms and Applications. *Neural Networks* 13(4-5), 411-430, 2000

Jung T-P., Humphries C., Lee T-W., McKeown M. J., Iragui V., Makeig S. and Sejnowski T. J., Removing electroencephalographic artifacts by blind source separation. *Psychophysiology*, 37, 163-178, 2000

Jutten C., Gribonval R., L'analyse en composantes indépendantes: un outil puissant pour le traitement de l'information. Proc. XIXe colloque GRETSI (traitement du signal et des images), 8-11 septembre 2003, Sep 2003, Paris, France. GRETSI, Groupe d'Etudes du Traitement du Signal et des Images, I, pp.11-16, 2003

Le Borgne H., Analyse de scènes par composantes indépendantes. Thèse de doctorat, INP Grenoble, Chapitre 3, 2004

Linsker R., Self-organisation in a perceptual network. *IEEE Computer*, 21, 105-117, 1988

Perret C., Laganaro M., Dynamique de préparation de la réponse verbale et Electroencéphalographie: une revue. *L'Année Psychologique*/Topics in Cognitive Psychology, 113, 667-698, 2013.

Appendix: les fonctions

Fonction 1 : Séparation et Identification des différentes sources.

```
f1<-function(x,algorithm=c("infomax", "fastICA"),q=1.5){

  ######### SEPARATION AVEUGLE DES SOURCES EEG et EMG #################

  library(ica)
  if(algorithm=="infomax"){r1=icaimax(x,dim(x)[2],center = FALSE)} else
  { if(algorithm=="fastICA"){r1=icafast(x,dim(x)[2],center = FALSE)}  }

  ######### IDENTIFICATION DES DIFFERENTES SOURCES EEG et EMG #########

  # Corrélations (2 différentes méthodes)
  m=2
  cpea<-matrix(data = NA, nrow = dim(x)[2], ncol = m, byrow = FALSE)
  cspe<-matrix(data = NA, nrow = dim(x)[2], ncol = m, byrow = FALSE)
  abj2<-matrix(data = NA, nrow = m, ncol = 4)
  abj3<-matrix(data = NA, nrow = m, ncol = 4)
  for(j in 1:m){
    cpea[,j]<-cor(r1$S,x[,(dim(x)[2])+1-j],method ="pearson")
    cspe[,j]<-cor(r1$S,x[,(dim(x)[2])+1-j],method ="spearman")
    al2<-c(which(cpea[,j]<boxplot(cpea[,j],plot=F,range=q)$stats[1,]),
           which(cpea[,j]>boxplot(cpea[,j],plot=F,range=q)$stats[5,]))
    al3<-c(which(cspe[,j]<boxplot(cspe[,j],plot=F,range=q)$stats[1,]),
           which(cspe[,j]>boxplot(cspe[,j],plot=F,range=q)$stats[5,]))
    abj2[j,]<-cbind("Pearson",j,length(al2),paste(as.character(al2),collapse=", "))
    abj3[j,]<-cbind("Spearman",j,length(al3),paste(as.character(al3),collapse=", "))
  }

  # P_Distance de Minkowski (p à pas 0.1 de 1 à 9)

  al1<-as.vector(NULL)
  abjt<-matrix(data = NA, nrow =1, ncol = 4)
  for(alpha in seq(1,9,0.1)){
    mpd<-matrix(data = NA, nrow = dim(x)[2], ncol = m, byrow = FALSE)
    abj<-matrix(data = NA, nrow = m, ncol = 4)
    for(j in 1:m){
      for(i in 1:dim(x)[2]){
        mpd[i,j]<- ((sum(abs(r1$S[,i]-x[,(dim(x)[2])+1-j])^alpha)))^(1/alpha)
      }
      al1<-c(which(mpd[,j]<boxplot(mpd[,j],plot=F,range=q)$stats[1,]),
             which(mpd[,j]>boxplot(mpd[,j],plot=F,range=q)$stats[5,]))
      abj[j,]<-cbind(alpha,j,length(al1),paste(as.character(al1),collapse=", "))
    }
    abjt<-rbind2(abjt,abj)
  }
  abjt<-rbind(abj2,abj3,abjt[-1,]) ; abjt<-as.data.frame(abjt)
  colnames(abjt)<- c("statistique","EOC","n.electrodes","electrodes")
  abjt1<-abjt[which(abjt[,2]==1),] ; abjt2<-abjt[which(abjt[,2]==2),]

  elint1<-abjt1[which.max(as.numeric(as.character(abjt1$n.electrodes[-c(1:2)])))+2,4]
  elv1<-as.numeric(unlist(strsplit(as.character(elint1), split = ", " , fixed = TRUE)))

  elint2<-abjt1[1,4]
  elv2<-as.numeric(unlist(strsplit(as.character(elint2), split = ", " , fixed = TRUE)))
```

```
  elint3<-abjt1[2,4]
  elv3<-as.numeric(unlist(strsplit(as.character(elint3), split = ", " , fixed = TRUE)))

  elv2.3<-c(elv2,elv3)
  elv1.23<-intersect(elv1,elv2.3)

  elint2_1<-abjt2[which.max(as.numeric(as.character(abjt2$n.electrodes[-c(1:2)])))+2,4]
  elv2_1<-as.numeric(unlist(strsplit(as.character(elint2_1), split = ", " , fixed = TRUE)))

  elint2_2<-abjt2[1,4]
  elv2_2<-as.numeric(unlist(strsplit(as.character(elint2_2), split = ", " , fixed = TRUE)))

  elint2_3<-abjt2[2,4]
  elv2_3<-as.numeric(unlist(strsplit(as.character(elint2_3), split = ", " , fixed = TRUE)))

  elv2_2.3<-c(elv2_2,elv2_3)
  elv2_1.23<-intersect(elv2_1,elv2_2.3)

  resds<-list(e1=abjt2,e1.common=elv2_1.23,e2=abjt1,e2.common=elv1.23,
              e.common=union(elv2_1.23,elv1.23),comp.ind=r1$S,mixmat=r1$M)
  resds
}
```

Fonction 2 : Retrait de l'influence des artéfacts et Mélange des composantes

```
############# deuxième fonctions: Remettre à zéro les composantes identifiées

# CI       Composantes independantes: matrice
# CI.EM    Composantes independantes Comportant de l'activité EMG retenue: vecteur
# MATMEL   Matrice de mélange

f2<-function(CI,CI.EM,MATMEL){

  ### RETRAIT DE L'INFLUENCE DE L'ACTIVITE ELECTROMYOGRAPHIQUE DES DONNEES ###

  CI[,CI.EM]<-0
  NCI=CI
  NX<-tcrossprod(NCI,MATMEL) ## NCI %*% t(MATMEL)
  RES=list(SET0=NCI,CLEANED.OBS=NX)

  return(RES)

}
```