Débruitage de signaux électroencéphalographiques par analyse de corrélation canonique

ESIR 2 spécialité Ingénierie Biomédicale - Séance 1/3

Laboratoire LTSI - UMR INSERM U1099 - Université de Rennes 1

1 Contexte applicatif

L'épilepsie est une pathologie neurologique complexe, caractérisée par la répétition de crises qui vont fortement altérer la qualité de vie des patients, surtout lorsqu'elles résistent aux traitements médicamenteux (épilepsie pharmaco-résistante). L'épilepsie est l'une des premières causes d'hospitalisation en neurologie (après l'accident vasculaire cérébral) et le trouble neurologique du cerveau le plus courant, indépendamment de considérations d'âge, d'origine ethnique, ou de région géographique. En présence d'épilepsie pharmaco-résistante (environ 30% des épilepsies), une approche chirurgicale peut être envisagée afin de réaliser une ablation de la zone épileptogène du cerveau. D'où la nécessité de localiser avec précision cette zone. Une telle localisation peut être effectuée lors du bilan préchirurgical à l'aide d'algorithmes de traitement de signaux ElectroEncéphaloGraphiques (EEG) enregistrés à la surface de la tête. Néanmoins, les signaux EEG peuvent être considérablement affectés par la présence d'artefacts variés, tels

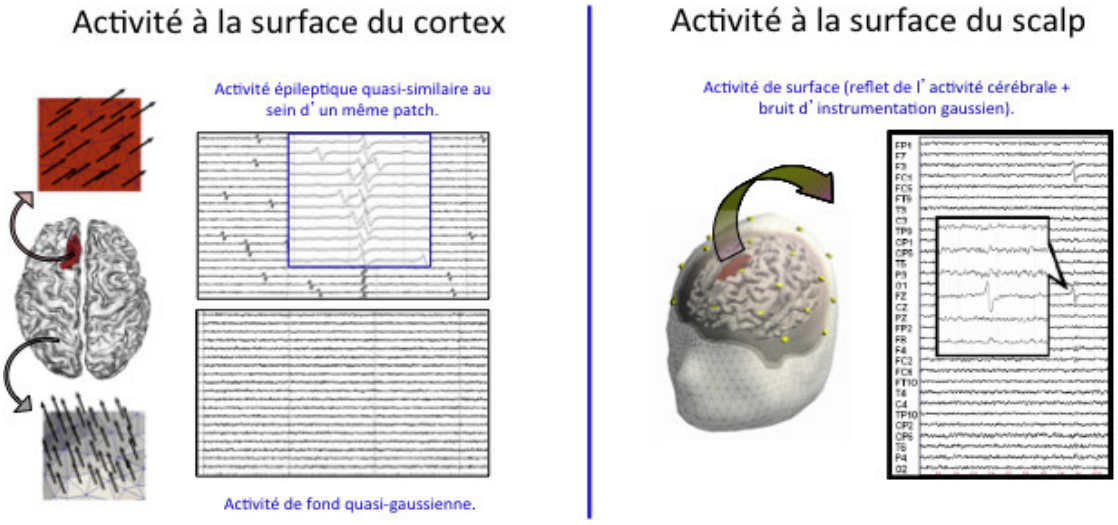
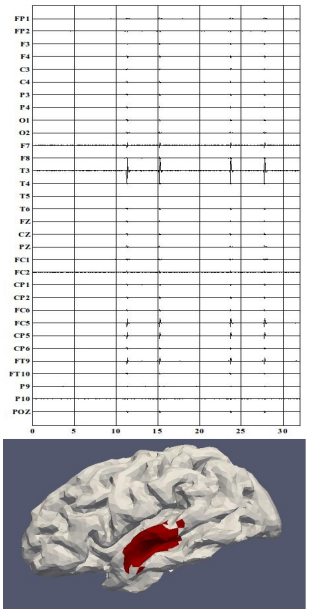


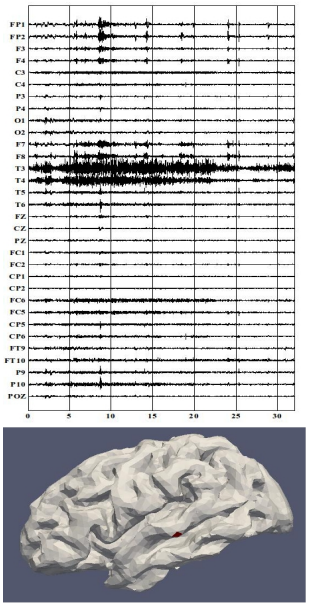
Figure 1: Mesure de l'activité ElectroEncéphaloGraphique (EEG)

que les mouvements oculaires, l'électrocardiogramme, l'activité musculaire. Ces artefacts peuvent considérablement changer les résultats de localisation comme le montre la figure 2. Parmi tous ces artefacts, l'élimination de l'activité musculaire est particulièrement ardue. Ceci peut être essentiellement attribué au fait que i) l'activité musculaire est largement distribuée à la fois aux niveaux spatial et spectral et ii) l'activité musculaire est moins stéréotypée que les autres artefacts. Considérons dans la suite de ce TP que la zone épileptogène à localiser est une surface connexe que nous nommerons *patch*. On peut considérer que les signaux EEG enregistrés à la surface de la tête à l'instant m suivent le modèle suivant :

$$\boldsymbol{x}[m] = \boldsymbol{A}_e\, \boldsymbol{s}_e[m] + \boldsymbol{A}_f\, \boldsymbol{s}_f[m] + \boldsymbol{A}_b\, \boldsymbol{s}_b[m] + \boldsymbol{\nu}[m] \tag{1}$$



(a) Localisation de la région épileptogène à partir d'un segment d'EEG intercritique sans activité musculaire



(b) Localisation de la région épileptogène à partir d'un segment d'EEG intercritique avec activité musculaire

Figure 2: Influence de l'activité musculaire sur la localisation de région épileptogène

où $\boldsymbol{x}[m]$ est le vecteur aléatoire d'observations représentant les mesures électriques enregistrées à l'instant m au niveau des N électrodes de surface, où $\boldsymbol{s}_e[m]$, $\boldsymbol{s}_f[m]$ et $\boldsymbol{s}_b[m]$ sont des vecteurs aléatoires de longueurs respectives P_e, P_f et P_b représentant, à l'instant m, l'activité épileptique intercritique des dipôles électriques constituant le patch épileptogène, l'activité électrique de fond des dipôles électriques affectés au reste du cerveau et l'activité électrique des dipôles musculaires situés par exemple au niveau de la machoire. $\boldsymbol{A}_e$, $\boldsymbol{A}_f$ et $\boldsymbol{A}_b$ sont des matrices de mélange de tailles $(N \times P_e)$, $(N \times P_f)$ et $(N \times P_b)$, modélisant les transferts d'activité électrique des dipôles vers les électrodes de surface. Enfin, $\boldsymbol{\nu}[m]$ est un vecteur de bruit d'instrumentation Gaussien de longueur N. Notons qu'en pratique le nombre de dipôles électriques utilisé pour modéliser l'activité du cerveau est beaucoup plus grand que le nombre d'électrodes EEG. Néanmoins, sous l'hypothèse admise d'une très forte synchronicité des dipôles électriques du patch épileptogène, on peut faire l'approximation suivante :

$$\boldsymbol{x}[m] = \bar{\boldsymbol{a}}_e\, \bar{\boldsymbol{s}}_e[m] + \boldsymbol{A}_f\, \boldsymbol{s}_f[m] + \boldsymbol{A}_b\, \boldsymbol{s}_b[m] + \boldsymbol{\nu}[m] \tag{2}$$

où $\bar{\boldsymbol{a}}_e = \sum_p^{P_e} \boldsymbol{A}_e(:,p)$ et où $\bar{\boldsymbol{s}}_e[m]$ représente l'activité électrique commune à tous les dipôles du patch épileptogène. Le problème de débruitage de l'activité intercritique de surface peut donc être reformulé comme un problème d'identification du vecteur colonne $\bar{\boldsymbol{a}}_e$ et du processsus $\{\bar{\boldsymbol{s}}_e[m]\}$ associé. En effet, le processus aléatoire vectoriel $\{\boldsymbol{x}_e[m]\}$ défini par $\boldsymbol{x}_e[m] = \bar{\boldsymbol{a}}_e\, \bar{\boldsymbol{s}}_e[m]$ pour tout m ne contiendra plus ni activité de fond ni artefact musculaire, et constituera ainsi notre activité EEG intercritique débruitée. La solution étudiée au travers de ces trois séances de TP, solution qui permettra d'identifier le couple $(\bar{\boldsymbol{a}}_e, \bar{\boldsymbol{s}}_e[m]\})$ de paramètres d'intérêt et de construire $\{\boldsymbol{x}_e[m]\}$, est la méthode d'Analyse de Corrélation Canonique (CCA).

2 Rappels statistiques

Soient x et y deux variables aléatoires réelles telles que $\mathsf{E}[x^2] < \infty$ et $\mathsf{E}[y^2] < \infty$. La covariance entre x et y est définie par :

$$\mathrm{Cov}(x,y) = \mathsf{E}[(xy - \mathsf{E}[xy])^2] = E(xy) - \mathsf{E}[x]\mathsf{E}[y] \tag{3}$$

La covariance est une mesure de similarité ou d'interdépendance linéaire entre deux variables aléatoires. Néanmoins, cette mesure se voit affectée par les unités de chaque variable. Pour résoudre ce problème on utilise une mesure normalisée nommée coefficient de corrélation défini par :

$$\mathrm{Corr}(x,y) = \rho(x,y) = \frac{\mathsf{E}[(xy - \mathsf{E}[xy])^2]}{\sqrt{\mathrm{Var}(x)\mathrm{Var}(y)}} = \frac{\mathsf{E}[xy] - \mathsf{E}[x]\mathsf{E}[y]}{\sqrt{(E(x^2) - E[x]^2)(E(y^2) - E[y]^2)}}. \tag{4}$$

où $\mathrm{Var}(x) = \mathsf{E}[x^2] - \mathsf{E}[x]^2$ désigne la variance de la variable aléatoire x.

Remarques

- Si $\mathsf{E}[x] = \mathsf{E}[y] = 0$ (i.e. si x et y sont deux variables aléatoires centrées), alors $\mathrm{Cov(x, y)} = \mathsf{E}[xy]$ et $\rho(x,y) = \frac{\mathsf{E}[xy]}{\sqrt{\mathsf{E}[x^2]\mathsf{E}[y^2]}}$.
- Si $\mathsf{E}[x] = \mathsf{E}[y] = 0$ et $\mathsf{E}[x^2] = \mathsf{E}[y^2] = 1$ (i.e. si x et y sont deux variables aléatoires normalisées), alors $\mathrm{Cov(x, y)} = \mathrm{Corr(x, y)} = \mathsf{E}[xy]$.
- Il peut être démontré en utilisant l'inégalité de Chauchy-Schwarz que $-1 \leq \rho(x,y) \leq 1$.

Les définitions de covariance et corrélation peuvent être généralisés au cas d'un processus stochastique $\{x[m]\}$ engendrant les fonctions d'autocovariance et autocorrélation. Notons C la fonction d'autocovariance du processus $\{x[m]\}$, elle est définie par :

$$C_x[m,k] = \mathrm{Cov}(x[m], x[m-k]) = \mathsf{E}[x[m]x[m-k]] - \mathsf{E}[x[m]]\mathsf{E}[x[m-k]] \tag{5}$$

En particulier, si $k = 0$ alors $C_x[m,0] = \mathrm{Cov}(x[m], x[m]) = \mathrm{Var}(x[m])$. Rappelons également que si i) $\mathsf{E}[x[m]]$ est constante quel que soit m et ii) la fonction C_x ne dépend pas de m (i.e. $\forall (m,k)$, $C_x[m,k] = C_x[k]$), alors le processus aléatoire $\{x[m]\}$ peut être qualifié de *stationnaire au sens large à l'ordre deux*. Supposons à présent qu'il en soit ainsi du processus $\{x[m]\}$ considéré. En normalisant la fonction d'autocovariance, la fonction d'autocorrélation est définie par :

$$\rho[k] = \frac{C_x[k]}{C_x[0]} = \frac{\mathrm{Cov}(x[m], x[m-k]))}{\mathrm{Var}(x[m])} \tag{6}$$

Remarques

- Les fonctions d'autocovariance et d'autocorrélation sont symétriques, c'est à dire que $C_x[k] = C_x[-k]$ et que $\rho[k] = \rho[-k]$. C'est pourquoi on calcule généralement les valeurs de ces fonctions uniquement pour $k \geq 0$.
- Par définition on a $\rho[0] = 1$.

Les définitions de covariance et corrélation peuvent être généralisés cette fois au cas d'un processus vectoriel stochastique $\{\boldsymbol{x}[m]\}$ de dimension N. Considérons à nouveau par soucis de simplicité que le processus $\{\boldsymbol{x}[m]\}$ soit stationnaire au sens large à l'ordre deux. La fonction $\boldsymbol{C}_x$ d'autocovariance du processus $\{\boldsymbol{x}[m]\}$ est alors à valeurs dans $\mathbb{R}^{N\times N}$ et définie par :

$$\begin{aligned}
\boldsymbol{C}_x[k] &= (\,\mathrm{Cov}(x_{n_1}[m], x_{n_2}[m-k])\,) \\
&= (\,\mathsf{E}[x_{n_1}[m]x_{n_2}[m-k]] - \mathsf{E}[x_{n_1}[m]]\mathsf{E}[x_{n_2}[m-k]]\,) \\
\boldsymbol{C}_x[k] &= \mathsf{E}[\boldsymbol{x}[m]\boldsymbol{x}[m-k]^\intercal] - \mathsf{E}[\boldsymbol{x}[m]]\mathsf{E}[\boldsymbol{x}[m-k]]^\intercal
\end{aligned}$$

La fonction $\boldsymbol{\rho}_x$ d'autocorrélation de $\{\boldsymbol{x}[m]\}$ est également à valeurs dans $\mathbb{R}^{N\times N}$ et est obtenue à partir de la matrice $\boldsymbol{C}_x[k]$ en normalisant chacune de ses composantes :

$$\boldsymbol{\rho}_x[k] = \left(\frac{C_x[k]_{n_1,n_2}}{\sqrt{C_x[0]_{n_1,n_1} C_x[0]_{n_2,n_2}}} \right)$$

3 Chargement, représentation graphique et étude statistique des données

- Charger les jeux de données "data.dat" à l'adresse "perso.univ-rennes1.fr/laurent.albera/" dans l'onglet "Teaching" à la rubrique "ESIR 2 module de TS3" (indication : utiliser la fonction MATLAB "load").
- Afficher sur une première figure les vecteurs lignes de la matrice "Xe" définie par "Xe = Ae*Se" (indication : utiliser les fonctions MATLAB "figure", "hold on/off" et "plot").
- Mettre un titre à cette figure comme par exemple "Activités intercritiques des dipôles du patch épileptique" (indication : utiliser la fonction MATLAB "title").
- Reformater l'axe des abscisses (Ox) et des ordonnées (Oy), tels que : i) l'axe (Oy) aient des valeurs comprises entre la valeur minimale et la valeur maximale de la matrice Xe, et ii) l'axe (Ox) soit limité entre 1 et le nombre de colonnes de la matrice Se (indication : utiliser les fonctions "axis", "min" et "max").
- Réaliser le même travail de représentation graphique des activités EEG de fond et des activités musculaires, contenues respectivement dans les matrices "Xb" et "Xb".
- Montrer que l'estimée de la fonction d'autocorrélation $\rho[1]$ des activités intercritiques est supérieure à celle des activités EEG de fond et musculaires (indication : utiliser la fonction MATLAB "corrcoef").