

Performances des approches déterministes et stochastiques pour le débruitage des signaux épileptiques intercritiques

DOHA SAFIEDDINE^{1,2}, AMAR KACHENOURA^{1,2}, LAURENT ALBERA^{1,2}, GWENAËL BIROT^{1,2},

FABRICE WENDELING^{1,2}, LOTFI SENHADJI^{1,2}, ISABELLE MERLET^{1,2}

¹INSERM, U 642, Rennes, F-35000 France

²Université de Rennes 1, LTSI, Rennes, F-35000 France

amar.kachenoura@univ-rennes1.fr

Résumé - On propose dans cette étude de comparer les performances d'une approche déterministe (Décomposition Modale Empirique, EMD), et deux approches stochastiques (l'Analyse en Composantes Indépendantes, ICA et l'Analyse de Corrélation Canonique, CCA) pour l'élimination des artéfacts musculaires (EMG) affectant les signaux ElectroEncéphaloGraphiques (EEG, électrodes de scalp) des patients épileptiques. L'objectif final est de choisir la méthode qui rehausse au mieux les signaux d'intérêt (transitoires de type pointes épileptiques et activité de fond) afin d'améliorer leur analyse qualitative et quantitative. La démarche entreprise pour la comparaison de ces trois méthodes est originale puisqu'elle exploite des signaux EEG de surface simulés, qui reproduisent de façon très réaliste les pointes épileptiques ainsi que l'activité EEG de fond; on ajoute à ces simulations des activités musculaires provenant de tracés EEG réels. Ces simulations nous apportent la "vérité terrain" concernant les sources d'intérêt et permettent donc de quantifier les performances de chacune des trois méthodes étudiées. Les résultats quantitatifs montrent que l'EMD offre des performances plus intéressantes que celles de l'ICA et de la CCA dans le contexte étudié lorsque le rapport signal sur bruit est très faible.

Abstract - ElectroEncephaloGram (EEG) recordings are often contaminated with muscle artifacts (EMG). These artifacts obscure the EEG and make its interpretation difficult or even unfeasible. In this paper, realistic EEG epileptic spikes are simulated from the activation of an epileptic patch. Realistic background activities are also generated. Real muscle artifacts are then added to the simulated data. Such data allow us to quantify the performance of denoising methods since we have the "ground truth". We compare the efficiency of Empirical Mode Decomposition (EMD), Independent Component Analysis (ICA) and Canonical Correlation Analysis (CCA) to remove muscle artifacts from the EEG signals. The quantitative comparison indicates that the EMD approach exhibits a better performance than ICA and CCA, especially in the case of very low signal to noise ratio.

1 Introduction

Au cours du bilan préchirurgical des épilepsies partielles pharmacorésistantes, l'analyse visuelle des signaux ElectroEncéphaloGraphiques (EEG) pendant les crises (décharges critiques) et entre les crises (pointes intercritiques) est primordiale pour guider des explorations plus invasives telles que l'implantation d'électrodes intracérébrales (SEEG) et pour aider à la chirurgie. Néanmoins, les signaux EEG de surface peuvent être considérablement affectés par des artéfacts variés, tels que les mouvements oculaires, l'électrocardiogramme, l'activité musculaire (EMG). Parmi tous ces artéfacts, l'élimination de l'activité musculaire est particulièrement ardue [1]. 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) par rapport aux autres artéfacts, l'activité musculaire est moins stéréotypée. Dans cette étude, on propose de comparer trois méthodes : i) deux approches stochastiques que sont l'Analyse en Composantes Indépendantes (ICA) [2,11],

et l'Analyse de Corrélation Canonique (CCA) [3,4], et ii) la Décomposition Modale Empirique (EMD) [5,6], une approche dite déterministe. Les performances de ces trois méthodes ont été évaluées quantitativement sur des simulations EEG réalistes [7] reproduisant à la fois l'activité de fond et des activités de type pointes épileptiques intercritiques, bruitées par des activités musculaires issues de tracés EEG de scalp réels.

2 Formulation du problème

Le problème traité dans cette étude peut être décrit par l'équation suivante :

$$\begin{aligned} x[m] &= A_e s_e[m] + A_f s_f[m] + A_b s_b[m] + v[m] \\ x[m] &= A s[m] + v[m] \end{aligned} \quad (1)$$

où $x[m]$ est le vecteur d'observations représentant N mesures électriques enregistrées au niveau des électrodes de scalp, et $s_e[m]$, $s_f[m]$ et $s_b[m]$ sont des vecteurs de longueurs respectives P_e , P_f et P_b représentant, à l'instant m , l'activité épileptique intercritique, l'EEG de fond et les artéfacts musculaires.

A_e , A_f et B sont des matrices de mélange de tailles $(N \times P_e)$, $(N \times P_f)$ et $(N \times P_b)$, modélisant les transferts entre les différentes sources d'activité électrique et les électrodes de surface. Enfin, $v[m]$ est un vecteur bruit d'instrumentation Gaussien de longueur N .

Le modèle linéaire instantané découle directement de la formulation mathématique du problème direct en EEG. En effet, la mise en équation du problème direct fait appel aux équations de Maxwell sous leur forme quasi-statique [8, 9, 10]. Cette considération consiste à négliger toutes les dérivées temporelles intervenant dans les équations de Maxwell [8]. Autrement dit, connaissant la position, l'orientation, l'étendue et l'intensité des sources neuronales actives, le calcul de l'activité électrique recueillie au niveau des électrodes de surface est indépendante du temps. L'hypothèse quasi-statique est validée en considérant les ordres de grandeur de certains paramètres intervenant dans les équations de Maxwell [8] tels que : i) la conductivité et la permittivité des tissus de la tête, et ii) la gamme des fréquences des signaux observés en EEG (inférieur à 1000 Hz). Dès lors, l'approximation quasi-statique nous permet de montrer (au travers des équations de Poisson [8,9,10]) que le potentiel électrique engendré sur une électrode de surface dépend linéairement des amplitudes des courants émis au niveau des muscles et/ou du cerveau.

Le problème étudié dans ce papier peut être formulé comme un problème de Séparation Aveugle de Sources (SAS). Il consiste à estimer le vecteur source $s[m]$, à partir du vecteur observation $x[m]$ (1), avec de faibles *a priori* sur les dites sources, le mélange et le bruit. Pour l'ICA et la CCA, qui supposent que $\{s[m]\}$ et $\{v[m]\}$ sont deux réalisations respectives de deux processus aléatoires que nous nommerons ci-après $\{s'[m]\}$ et $\{v'[m]\}$, certaines des hypothèses suivantes sont faites :

- H1.** l'indépendance statistique spatiale des sources, i.e. l'indépendance mutuelle des composantes de $s'[m]$ à tout instant m ;
- H2.** le rang colonne plein de la matrice de mélange globale A ;
- H3.** l'indépendance statistique entre les sources et le bruit, i.e. entre les composantes de $s'[m]$ et de $v'[m]$ à tout instant m ;
- H4.** la couleur temporelle des sources, i.e. une dépendance statistique entre toute composante p de $s'[m]$ et la p -ième composante de $s'[m+1]$ quel que soit l'instant m ;

Dans le cas de l'EMD, on suppose :

- H5.** la décomposition de chaque source comme somme de composantes Modulées en Amplitude et en Fréquence (AM-FM), qui plus est différente d'une source à l'autre.

A noter qu'outre l'interprétation aléatoire des signaux observés, les deux méthodes stochastiques étudiées ici

se distinguent de notre utilisation de l'EMD de par leur exploitation conjointe des différentes mesures électriques de surface. En d'autres termes, l'ICA et la CCA cherchent à estimer le vecteur source $s[m]$ en exploitant conjointement les différentes composantes de $x[m]$ (1), contrairement à l'utilisation de l'EMD faite dans ce papier.

3 Les méthodes

3.1 L'analyse en composantes indépendantes

L'ICA est un outil d'analyse statistique de données bidimensionnelles dont le concept a été formalisé dans les années 90 [2]. L'ICA permet de résoudre le problème de SAS sous les hypothèses **H1** à **H3**. L'ICA va donc séparer les P sources d'intérêt du mélange observé, $x[m]$, en maximisant leur indépendance. Plusieurs algorithmes d'ICA ont été proposés dans la littérature [11,12,13]. Nos différentes études ont montré le bon compromis performances/complexité de la méthode CoM2 [2] dans le contexte biomédical [11,12,13]. Dans cette nouvelle étude, on a donc décidé d'utiliser l'algorithme CoM2 qui cherche à reconstruire des sources indépendantes deux à deux en maximisant, après blanchiment des observations, une fonction de contraste basée sur les cumulants d'ordre 4.

Justifier que les activités électrophysiologiques considérées dans ce papier sont statistiquement indépendantes (**H1**) n'est pas trivial [13]. Lorsque les activités qu'on cherche à séparer sont induites par des organes différents, on pourrait penser que ces dernières étant physiologiquement indépendantes le sont aussi statistiquement. Mais qu'en est-il d'activités émanant d'un même organe? Néanmoins, nous avons vérifié sur les données exploitées dans la section 4 que l'estimateur statistique, non biaisé et consistant dans le cadre de signaux stationnaires et ergodiques vérifie la propriété d'indépendance des cumulants. L'hypothèse **H2** implique que l'on ait $N \geq P_e + P_f + P_b$. En modélisant les sources de courant à l'aide de dipôles de courant, le bien fondé de cette inégalité dépend du niveau de synchronisation des dipôles caractérisant chaque type d'activité. Quant à **H3**, elle est vraie car on considère ici uniquement le bruit d'instrumentation qui est un bruit externe.

3.2 L'analyse de corrélation canonique

Initialement proposée dans [14], la CCA est une méthode statistique permettant d'exploiter les relations qui peuvent exister entre deux vecteurs aléatoires, telles que : i) une décorrélation des composantes d'un même vecteur et ii) une corrélation deux à deux des composantes des deux vecteurs. Friman et al. [3] ont montré que la CCA peut être exploitée pour résoudre le problème de SAS en prenant par exemple comme premier vecteur aléatoire le vecteur source et comme second vecteur aléatoire une version retardée de ce dernier. Sous les hypothèses **H1** à **H4**, la CCA vise donc

à séparer les sources d'intérêt du mélange présenté dans l'équation (1), en minimisant leur corrélation spatiale et en maximisant leur autocorrélation temporelle. Plus précisément, les hypothèses **H1** et **H4** doivent être vérifiées au sens large à l'ordre 2. Les discussions concernant **H1**, **H2** et **H3** sont les mêmes que celles présentées dans la section précédente. Concernant l'hypothèse **H4**, le calcul des autocorrélations des sources d'intérêt (activités de type pointes épileptiques) de l'EEG de fond et des activités musculaires montre une réelle différence entre ces trois familles d'activités. En effet, pour un retard $\tau=1$, les activités de type pointes épileptiques sont fortement colorées (coefficient d'autocorrélation moyen sur toute la base de données de 0.9975), contrairement aux activités musculaires (coefficient d'autocorrélation moyen de 0.3496).

3.3 La décomposition modale empirique

L'EMD a été introduite par Huang et al. [5] à la fin des années 90. Elle est actuellement utilisée dans des domaines aussi variés que la sismographie, l'océanographie et le génie biomédical. L'objectif de l'EMD est de décomposer tout signal oscillant comme la somme de composantes hautes fréquences et basses fréquences extraites directement du signal [5,6]. L'idée de base de l'EMD est de considérer un signal localement comme la somme d'une tendance et d'une somme de composantes AM-FM, appelées IMF, pour Intrinsic Mode Functions. La tendance est le résidu du signal après avoir extrait toutes les composantes oscillantes. Ces dernières sont obtenues en utilisant un opérateur, dit de tamisage (ou sifting process), qui calcule itérativement une enveloppe moyenne du signal en interpolant des extrema par splines cubiques. Notons que les sources d'intérêt (pointes intercritiques et EEG de fond) sont obtenues en sommant certaines IMFs, les IMFs restantes et la tendance représentant l'EMG.

4 Etude comparative des performances

4.1 Génération de données

Les données EEG simulées, sont générées en utilisant un modèle réaliste de genèse de l'EEG décrit dans [7]. Ce modèle combine une description réaliste des sources (maillage du cortex cérébral et champ dipolaire associé) et une description réaliste de l'activité des dipôles (modèle de populations de neurones). Les données EEG ont été simulées au niveau de 31 électrodes de scalp à partir d'une source corticale distribuée ou « patch » de 5 cm^2 situé au niveau du gyrus temporal supérieur gauche. Des activités de type de pointes hyper-synchrones, générées par un modèle de populations de neurones, sont attribuées aux dipôles du patch, alors que des activités de fond non-corrélées sont associées au reste des dipôles du cortex. L'EEG est ensuite généré en résolvant le problème direct (Boundary Element Method, BEM) dans un modèle réaliste de tête constitué d'un maillage du cerveau, du crâne et du scalp. 50 réalisations de ces signaux EEG simulés sont générées

pour l'étude. Pour chaque réalisation on ajoute une activité musculaire différente issue de tracés EEG de scalp réels. Au préalable, ces activités sont filtrées passe haut (15Hz) afin d'éliminer les rythmes intrinsèques de l'EEG (delta, theta, alpha, et sigma) potentiellement sous-jacents et de ne garder majoritairement que de l'activité musculaire.

4.2 Critère de performance

La performance des trois méthodes CoM2 [2], CCA [3] et 2T-EMD [6], est évaluée en calculant l'Erreur Quadratique Moyenne Normalisée (EQMN) suivante :

$$EQMN_{x_i} = \frac{\sum_{j=1}^J \sum_{m=1}^M \left[(x_i[m] - \hat{x}_i^{(j)}[m])^2 \right]}{J \sum_{m=1}^M (x_i[m])^2} \quad (2)$$

où $x_i[m]$ est l'EEG initial (EEG sans activité musculaire) observé sur la $i^{\text{ème}}$ électrode, $\hat{x}_i^{(j)}[m]$ et le signal EEG de surface reconstruit après débruitage de la $j^{\text{ème}}$ réalisation, J est le nombre total des réalisations et M le nombre d'échantillons.

4.3 Résultats

Avant de discuter des résultats obtenus, il est important de noter que : i) la sélection des composantes de l'ICA représentant les sources d'intérêt est basée sur une inspection visuelle et le signal est reconstruit en éliminant les composantes relatives à l'artéfact musculaire, ii) pour la méthode CCA, la sélection des sources d'intérêt est fondée sur des critères de seuillage pour la courbe d'auto-corrélation des composantes canoniques [3], et iii) pour l'approche EMD, les IMFs d'intérêt sont sélectionnées en utilisant la procédure présentée dans [15] qui a l'avantage de ne pas nécessiter un signal de référence.

A titre d'exemple, la figure 1 montre, pour un RSB de -10dB, les résultats obtenus au niveau de l'électrode de surface T3 qui se trouve en face du patch. Clairement, les performances de 2T-EMD sont supérieures à celles de CoM2 et de CCA. Ce résultat est confirmé dans le tableau 1 où on a reporté les valeurs de l'EQMN calculé pour les méthodes 2T-EMD, CoM2 et CCA en fonction du RSB. En effet, l'erreur de reconstruction de 2T-EMD est inférieure à celle de l'ICA et de CCA. Ce résultat est particulièrement visible pour des RSB faibles. Les résultats obtenus peuvent, en partie, être expliqués par le fait que certaines hypothèses (voir **H1** et **H2** pour l'ICA et **H1**, **H2** et **H4** pour la CCA) ne semblent pas être complètement satisfaites dans le cadre de notre application.

5 Conclusions et perspectives

Les résultats préliminaires montrent que la méthode EMD est un outil efficace dans le contexte du débruitage des pointes intercritiques épileptiques. Ces résultats prometteurs, laissent entrevoir la possibilité de

proposer une méthode de prétraitement des données EEG de scalp permettant par exemple d'améliorer les performances des méthodes de localisation des zones épileptogènes dans le cerveau. En ce sens des simulations futures envisageront plusieurs patches et des tests sur des données cliniques réelles permettront d'évaluer la robustesse de la méthode. De plus, des critères pour la sélection automatique des sources d'intérêt basés sur les statistiques d'ordre supérieur, et plus spécifiquement sur l'analyse de la dépendance temporelle des composantes extraites pour des ordres $q > 2$ sont à envisager.

Électrode T3 localisée en face du patch : Résultat pour un RSB de -10 dB

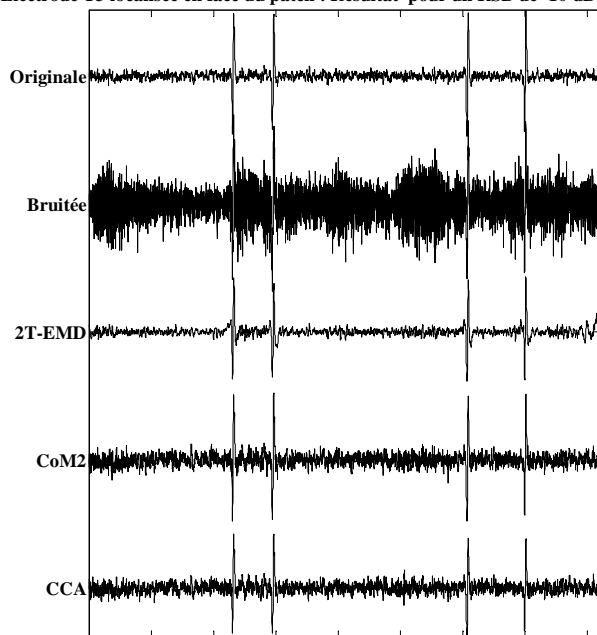


Figure 1 : l'EEG de surface de l'électrode T3 : Signal original, signal bruité par l'activité musculaire et signal débruité par 2T-EMD, CoM2 et CCA.

Tab 1 : *EQMN* obtenue pour l'électrode T3, calculée sur 50 réalisations, avec 2T-EMD, CCA et CoM2 en fonction du RSB.

	-25dB	-20dB	-15dB	-10dB	-5 dB	0 dB
2T-EMD	0.393	0.308	0.207	0.121	0.073	0.052
CoM2	0.618	0.423	0.261	0.154	0.088	0.050
CCA	0.866	0.592	0.366	0.216	0.124	0.071

Références

[1] B. W. McMenamin, A. J. Shackman, J. S. Maxwell, D. R. W. Bachhuber, A. M. Koppenhaver, L. L. Greischar, and R. J. Davidson, "Validation of ICA-based myogenic artifact correction for scalp and source-localized EEG," *Neuroimage*, vol. 49, pp. 2416-2432, 2011.

[2] P. Comon, "Independent component analysis, A new concept?," *Signal Processing*, vol. 36, pp. 287-314, 1994.

[3] O. Friman, M. Borga, P. Lundberg, H. Knutsson, Exploratory fMRI analysis by autocorrelation maximization, *NeuroImage*, vol. 16, no. 2, pp. 454-464, 2002.

[4] W. De Clercq, A. Vergult, B. Vanrumste, W. Van Paesschen, S. Van Huffel, "Canonical correlation analysis applied to remove muscle artifacts from the electroencephalogram," *IEEE Trans Biomed Eng*, vol. 53, pp. 2583-7, Dec 2006.

[5] N. E. Huang, Z. Shen, S. R. Long, M. C. Wu, H. H. Shin, Q. Zheng, N. C. Yen, C. C. Tung, and H. H. Liu, "The Empirical Mode Decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis," *Proc. Royal Soc*, vol. 454, pp. 903-995, 1998.

[6] J. Fleureau, J. C. Nunes, A. Kachenoura, L. Albera, L. Senhadji, Turning Tangent Empirical Mode Decomposition: a framework for mono and multivariate signals, *IEEE Transactions on Signal Processing*, vol. 59, no. 3, pp. 1309-1316, 2011.

[7] D. Cosandier-Rimele, J. M. Badier, P. Chauvel, F. Wendling, "A physiologically plausible spatio-temporal model for EEG signals recorded with intracerebral electrodes in human partial epilepsy," *IEEE Trans Biomed Eng*, vol. 54, pp. 380-8, Mar 2007.

[8] G. Adde, "Méthodes de Traitement d'Image Appliquées au Problème Inverse en Magnéto-Electro-Encéphalographie", Thèse, Ecole Nationale des Ponts et Chaussées, 2005.

[9] M. Besserve "Analyse de la dynamique neuronale pour les Interfaces Cerveau-Machines : un retour aux sources", Thèse, Université Paris Sud, 2007.

[10] E. Kenneth, II Hild, and S. Srikantan Nagarajan, "Source localization of EEG/MEG data by correlating columns of ICA and lead field matrices", in *IEEE Trans. on Biomedical Engineering*, vol. 56, no. 11, pp. 2619-2626, 2009.

[11] L. Albera, P. Comon, L. Parra, A. Karfoul, A. Kachenoura, L. Senhadji, Biomedical applications, in *Handbook of blind source separation*, P. COMON and C. JUTTEN, Eds. Academic Press, 2010.

[12] A. Kachenoura, L. Albera and L. Senhadji, P. Comon, "ICA: a potential tool for bci systems", in *IEEE Signal Processing Magazine*, , vol. 25, no. 1, pp. 57-68, 2008.

[13] A. Kachenoura, "Traitement Aveugle de Signaux Biomédicaux", Thèse, Université de Rennes1, 2006.

[14] H. Hotelling, Relations between two sets of variates, *Biometrika*, vol. 28, pp. 321-377, 1936.

[15] Y. Kopsinis, E. Aboutanios, D. A. Waters, S. McLaughlin, "Development of EMD-based Denoising Methods Inspired by Wavelet Thresholding," *IEEE Trans. on Signal Processing*, pp. 1351-1362, 2009.