Débruitage de signaux physiologiques par agrégation de méthodes gloutonnes

Statistique Appliquée - Note de Synthèse

0.1 Introduction

La magnétoencéphalographie (MEG) et l'électroencéphalographie (EEG) mesurent les champs électromagnétiques induits par l'activité neuronale ; pour faciliter la quantification et la caractérisation de ces signaux cérébraux, nous cherchons à les débruiter tout en préservant au mieux leur énergie. Nous avons donc recours à un algorithme bien connu en traitement du signal : celui du *matching pursuit*. Il consiste à rechercher les projections les plus corrélées avec le signal étudié au sein d'un dictionnaire redondant fixé ; elles se superposent jusqu'à former un signal décrivant convenablement les données initiales. Se posent alors les questions du critère d'arrêt à retenir pour l'inclusion de ces variables explicatives dans notre modèle et du dictionnaire le plus adapté à notre étude.

0.2 Statistique descriptive

Nous disposons de signaux électromagnétiques issus de 203 capteurs, échantillonnés à 540 instants distincts sur une durée totale de 300 ms. Ces mesures comprennent trois phases de même longueur (180 échantillons chacune) :
– une phase pré-stimulation, où le sujet est à l'état de repos initial
– une phase de stimulation, au cours de laquelle on mesure l'activité cérébrale provoquée par une stimulation électrique sur le sujet
– une phase post-stimulation.
Ces données sont comprises sous la forme de deux types de matrices, `data_cond` et `times`, contenant respectivement les données et les temps d'échantillonnage, pour l'ensemble de l'expérimentation. Pour éviter toute erreur d'approximation induite par les méthodes de calcul numérique employées nous travaillerons sur les données augmentées d'un facteur 10^{12}.

0.3 Modèle et hypothèses

Nous considérons le modèle suivant :

$$Y = D\lambda + \epsilon, \tag{1}$$

$Y \in \mathscr{M}_{n,1}(\mathbb{R})$ est notre observation pour un capteur, échantillonnée à divers instants $t_1, ...t_n$. On note $D \in \mathscr{M}_{n,q}(\mathbb{R})$ le dictionnaire initial ; ses atomes - colonnes de D - engendrent un sous-espace de $\mathbb{R}^n$ de dimension q. $\lambda \in \mathscr{M}_{q,1}(\mathbb{R})$ est un vecteur dit « parcimonieux », traduisant les composantes du dictionnaire sélectionnées. Le terme ε est un terme d'erreur de taille n ; c'est le bruit inhérent à l'observation, dont on cherche à se débarrasser. Le produit $D\lambda$ est alors le signal « débruité » que nous cherchons à obtenir.

On suppose les hypothèses suivantes satisfaites :

1. *Moyenne conditionnelle nulle :* $\mathbb{E}[\epsilon|D] = 0$.
2. *Erreur quadratique :* $\mathbb{V}[\epsilon|D] = \sigma^2 I_n$, qui implique les deux sous-hypothèses suivantes :
   – *Homoscédasticité :* $\forall i \in [1:n]\ \mathbb{E}[\epsilon_i^2|D] = \sigma^2$.
   – *Absence d'auto-corrélation :* $\mathbb{E}[\epsilon_i\epsilon_j|D] = 0, i \neq j$.
3. *Bruit blanc gaussien :* $\epsilon_{t_i} \hookrightarrow \mathscr{N}(0,\sigma^2)$

Une des hypothèses sous-jacentes de notre modèle est que les champs mesurés au cours de la première phase sont des perturbations, et donc assimilables à du bruit, dont il s'agit d'estimer les caractéristiques - ici, la variance, estimée à 6.25×10^{-25}. Une fois ce bruit caractérisé, nous pourrons implémenter notre algorithme du *matching pursuit* nous permettant de séparer effectivement corps du signal et bruit.

0.4 Matching Pursuit

0.4.1 Fonctionnement de l'algorithme

Le *Matching Pursuit (MP)*, introduit par Mallat et Zhang [1] - et sa variante orthogonale, le *Matching Pursuit] Orthogonal (OMP)*, notamment implémentée par Gabriel Peyré [2] - sont deux méthodes du traitement du signal reposant sur le problème suivant : disposant d'un signal Y et d'un dictionnaire de signaux $D = (D_i)$, on souhaite approcher Y par une combinaison linéaire de colonnes de D - i.e. déterminer quels atomes « significatifs » retenir.

Le *Matching Pursuit* propose une approche simple : on estime l'atome D_{i_0} le plus corrélé à Y ; on soustrait à Y sa projection sur D_{i_0} et on recommence l'opération sur le résidu $Y - \lambda D_{i_0}$, jusqu'à ce que la combinaison linéaire créée approche suffisamment bien le signal.
Dans l'*OMP*, on sélectionne également les atomes D_i un à un, mais à chaque nouvelle sélection d'atome, on recalcule la projection de Y sur l'espace vectoriel engendré par les vecteurs sélectionnés.
Quand Y s'écrit réellement comme une combinaison linéaire des atomes, l'*OMP* peut rapidement converger vers une erreur nulle ; l'erreur commise par la version classique sera, elle, rarement nulle.

0.4.2 Critère d'arrêt

Si un critère d'arrêt arbitraire peut convenir en premier lieu pour comprendre le mécanisme en jeu, débruiter un signal à l'aide du *Matching Pursuit* demande un critère statistique valide permettant de s'affranchir du bruit.

Le procédé est le suivant : on ajoute une nouvel élément au dictionnaire temporaire formé des atomes déjà retenus ; on teste alors, selon des méthodes classiques, l'hypothèse selon laquelle le nouveau vecteur parcimonieux ne comporte aucune composante selon cet élément - auquel cas ce dernier n'a aucune contribution significative au signal débruité, et est relégué au rang de bruit.

Etat du modèle temporaire
• A l'étape $p \leq q$, nous avons sélectionné p atomes $X_i \in \mathscr{M}_{n,1}(\mathbb{R})$. Ces vecteurs X_i sont les colonnes de notre dictionnaire temporaire X.
· Au rang p, l'équation associée s'écrit alors, avec β vecteur parcimonieux de taille p :

$$Y = X\beta + \epsilon_p.$$

· Dans le cadre du *Matching Pursuit* , on suppose que le nombre de régresseurs p est strictement inférieur au nombre d'observations n. Puisque les colonnes de X sont choisies parmi les colonnes

de D, elle-même injective, X est injective et $X^tX \in \mathscr{M}_p(\mathbb{R})$ devient inversible.
· L'estimateur de la variance du bruit, σ^2, s'écrit :

$$\hat{\sigma}^2 = \frac{1}{n-p}(Y - X\hat{\beta})^T(Y - X\hat{\beta}) \tag{2}$$

Inclusion du nouvel élément
• On s'intéresse aux $q - p$ atomes restants du dictionnaire D original ; soit x l'un d'entre eux. Nous cherchons alors à déterminer si l'ajout d'un atome $x \in \mathscr{M}_{n,1}(\mathbb{R})$ contribuera de manière significative au signal débruité : on note $Z = [X\ x]$ l'éventuel dictionnaire au rang $p + 1$. L'équation devient :

$$Y = Z \begin{pmatrix} \beta \\ \gamma \end{pmatrix} + \epsilon_{p+1}. \tag{3}$$

où ϵ_{p+1} est le nouveau résidu au rang $p + 1$.
L'estimateur des moindres carrés du nouveau vecteur parcimonieux s'écrit alors :

$$\begin{pmatrix} \hat{\beta} \\ \hat{\gamma} \end{pmatrix} = (Z^TZ)^{-1}Z^TY = (Z^TZ)^{-1} \begin{pmatrix} X^TY \\ x^TY \end{pmatrix} \tag{4}$$

Afin de déterminer la pertinence de x, on teste la significativité de γ. On teste alors l'hypothèse $\mathscr{H}_0 : \gamma = 0$ contre l'alternative $\mathscr{H}_1 : \gamma \neq 0$. Notre statistique de test s'écrit :

$$T_{n,p+1} = \frac{|x^T(Y - HY)|}{\sigma\sqrt{x^T(x - Hx)}}$$

avec $H = X(X^TX)^{-1}X^T$ un projecteur orthogonal.

• Si la variance est connue, la zone de rejet de l'hypothèse $\mathscr{H}_0$ contre l'hypothèse $\mathscr{H}_1$ s'écrit alors : $\{|T_{n,p+1}| > q_{n-p-1}^{1-\frac{\alpha}{2}}\}$, où $q_{n-p-1}^{1-\frac{\alpha}{2}}$ est le quantile d'ordre $\frac{\alpha}{2}$ de la loi de Student.
Dans le cas contraire, on remplace σ^2 par son estimation $\hat{\sigma}^2$, et l'on reprend ce qui précède.

0.5 Choix du dictionnaire

Le dernier point auquel nous nous sommes attelés fut le choix du dictionnaire.
Si le document de Gabriel Peyré introduisait un dictionnaire gaussien aléatoire, celui-ci n'avait pour but que d'illustrer le fonctionnement du *Matching Pursuit* ; dans notre étude, il produisit des résultats médiocres en termes de débruitage.
Aussi avons-nous essayé différents dictionnaires supposés fournir de meilleurs résultats, comme ceux obtenus à l'aide de la Transformée de Fourier Rapide (FFT) ou encore des ondelettes (Coiflets) d'Ingrid Daubechies ; le critère retenu pour la comparaison entre deux dictionnaires potentiels fut la décroissance de la variance des résidus en fonction du nombre d'atomes inclus dans le modèle.
A titre d'exemple, voici les signaux originaux et débruités obtenus pour des ondelettes de Daubechies à 1 et 3 moments nuls - i.e. *2-tap* et *6-tap* :

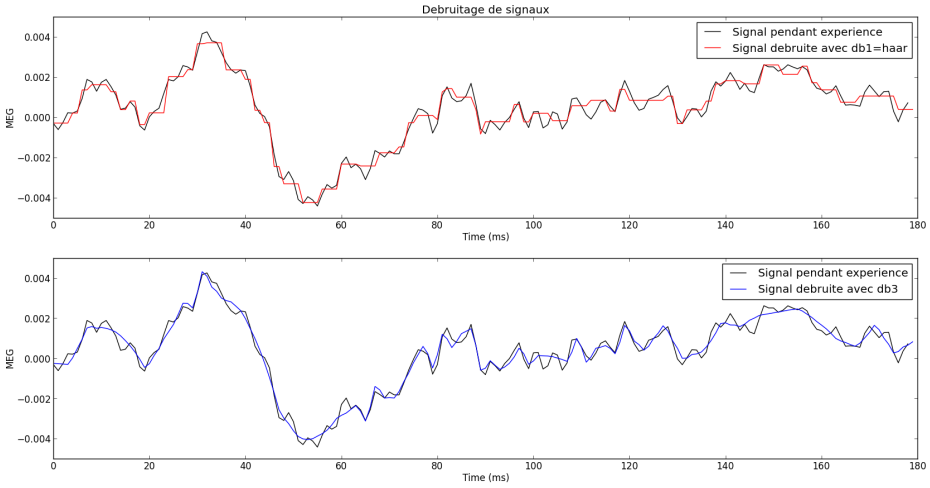

FIGURE 1 – *Débruitage à l'aide d'ondelettes de Daubechies.*

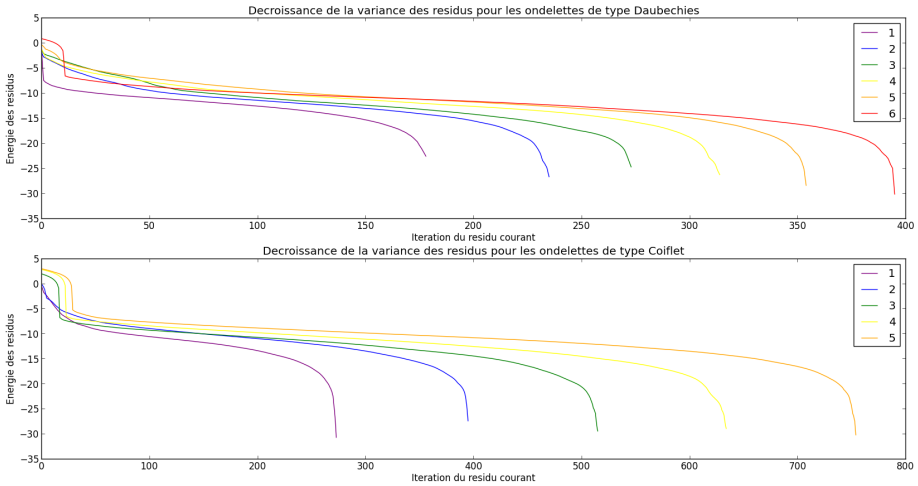

FIGURE 2 – *Décroissance des résidus selon le nombre d'atomes retenus.*

0.6 Conclusion

Les résultats majeurs obtenus jusqu'à présent furent l'obtention d'un critère d'arrêt « pertinent » à variance fixée - qui permette de ne retenir que les atomes significatifs et d'exclure le bruit du modèle - et d'un critère de comparaison entre différents dictionnaires compatibles avec nos données.
Nous envisageons par la suite de comparer notre critère d'arrêt avec un second autre, suggéré par nos encadrants - de voir dans quelle mesure peuvent-ils être équivalents - ainsi que de raffiner les hypothèses émises sur le terme d'erreur - jusqu'ici compris comme un bruit blanc gaussien - à partir du cours de Séries Temporelles, et de voir si nos résultats en sont modifiés.

Bibliographie

[1] Stéphane Mallat et Zhifeng Zhang *Matching Pursuit with time-Frequency Dictionaries*, IEEE Transactions On signal Processing,Vol.41. No.12,December 1993

[2] Gabriel Peyré *Sparse Regularization with Matching Pursuits*, Ceremade