Dictionary Learning for Multidimensional Data

Christos Papageorgakis, Sebastian Hitziger, Théodore Papadopoulo

To cite this version:

Christos Papageorgakis, Sebastian Hitziger, Théodore Papadopoulo. Dictionary Learning for Multidimensional Data. Proceedings of GRETSI 2017, Sep 2017, Juan-les-Pins, France. hal-01575263

HAL Id: hal-01575263
https://hal.inria.fr/hal-01575263

Submitted on 18 Aug 2017

Dictionary Learning for multidimensional data

Christos PAPAGEORGAKIS, Sebastian HITZIGER, Théodore PAPADOPOULO,

Université Côte d'Azur, Inria, France
2004, route des Lucioles, BP 93, 06902 Sophia Antipolis Cedex, France
christos.papageorgakis@inria.fr, sebastian.hitziger@gmx.de,
theodore.papadopoulo@inria.fr

Résumé – L'électro-encéphalographie (EEG) et la magnéto-encéphalographie mesurent les activités respectivement électriques et magnétiques dues au fonctionnement du cerveau à partir de capteurs (électrodes et magnétométres) placés à la surface du scalp. Les signaux EEG ou MEG (conjointement M/EEG) mesurent la même activité biologique: celle des neurones pyramidaux situés dans le cortex. Ces signaux sont fondamentalement vectoriels (à chaque instant on dispose d'un vecteur de mesures). Ils sont, par contre, trés bruités. Pour pallier à cela, les mesures sont traditionnellement acquises de façon répétée (plusieurs "essais") puis moyennées. Ce procédé n'est pas optimal à cause de la variabilité des signaux d'un essai à l'autre. La méthode "jitter-adaptive dictionary learning" (JADL) [1] permit découverte via les données des composantes prototypiques qui se répétent à travers les essais. Cette méthode est, à l'heure actuelle, limitée à un seul signal et n'exploite pas la dimension vectorielle des signaux mesurés. Ce papier propose donc une adaptation vectorielle de la méthode JADL. Cette nouvelle méthode est testée sur des données synthétiques comme sur des données MEG réelles. Les résultats obtenus semblent prometteurs et indiquent que de meilleures performances peuvent être obtenues par rapport à la méthode mono-dimensionnelle.

Abstract – Electroencephalography(EEG) and magnetoencephalography (MEG) measure the electrical activity of the functioning brain using a set of sensors placed on the scalp (electrodes and magnetometers). Magneto- or electroencephalography (M/EEG) have the same biological origin, the activity of the pyramidal neurones within the cortex. The signals obtained from M/EEG are very noisy and inherently multi-dimensional, i.e. provide a vector of measurements at each single time instant. To cope with the noise, researchers, traditionally acquire measurements over multiple repetitions (trials) and average them to classify various patterns of activity. This is not optimal because of trial to trial variability. The jitter-adaptive dictionary learning method (JADL) [1] has been developed to better handle for this variability. JADL is a data-based method that learns a dictionary from a set of signals, but is currently limited to a single channel, which restricts its capacity with very noisy data such as M/EEG. In this paper, we propose an extension to the jitter-adaptive dictionary learning method, in order to handle multidimensional measurements such as M/EEG. A modified model is developed and tested using synthetically generated data set as well as real M/EEG signals. The results obtained using our model look promising, and show superior performance compared to the original single-channel JADL framework.

1 Introduction

Biological signals show important variability and diversity. Furthermore, they are often very noisy. This is particularly true of magnetic or electrical signals such as those obtained from magneto- or electro-encephalography (M/EEG). Both MEG and EEG have the same biological origin, the activity generated by displacement of charges of billions of synchronously active cells existing within the brain, called pyramidal neurons. The sets of synchronously active neurons, generating the magnetic fields and electric potentials recorded by M/EEG measures, are called generators or sources of the brain activity.

While the source activity is spread spontaneously within the head, the electric potentials and magnetic fields that reach the surface of the head, are the result of the linear mixture of the individual sources' activities. The propagations of the source activity is instantaneous, leading to M/EEG measurements that reflect the sources synchronously, in means of well-aligned features in the recorded data. The mixture of the sources in the recorded data is explained by the following matrix expression:

$$\mathscr{M} = \mathbf{GS} \,, \tag{1}$$

where $\mathscr{M} \in \mathbb{R}^{C\times T}$ is the measurement matrix either MEG or EEG, $\mathbf{G} \in \mathbb{R}^{C\times S}$ is the lead-field matrix (or gain matrix), $\mathbf{S} \in \mathbb{R}^{S\times T}$ is the sources matrix. C, S and T are the numbers of channels, sources and time samples respectively. Practically, the lead field matrix is a linear operator that maps source activations, to the estimated M/EEG measurements at sensors locations.

The M/EEG measurements are very noisy and inherently multi-dimensional, i.e. provide a vector of measurements at each single time instant. The components of each vector correspond to M/EEG channels. Traditionally, researchers, acquire the M/EEG measurements over multiple repetitions, called trials, and average them to improve the signal to noise ratio. But this degrades the shapes and timings of the activities and hides their inherent variability. To cope with this problem, various methods have been developed such as the differentially Variable Component Analysis method (dVCA) [3], that relies on trial-to-trial variability in response amplitude and latency to identify multiple components from multi-channel recordings. Other methods are the Multichannel matching pursuit (MMP)[6] and the Consensus Matching Pursuit (CMP)[2] which aim at extracting signal patterns from raw (i.e. non averaged) signals while still accounting for repetitions (events of interest over all the trials). However, these methods need a previously defined dictionary.

To remove this constraint, the JADL method [1] has been developed. It is a data-based method, that learns a dictionary from a set of example signals. This method has the advantage of better modelling prototypical brain activity, and also allows some variability (time jitter). Yet, JADL method is currently limited to a single channel, which restricts its capacity with very noisy data such as M/EEG.

The goal of this work is to extend the jitter adaptive dictionary learning method to handle multi-dimensional measurements such as M/EEG and to study the improvements this brings to the detection of some brain activity.

1.1 Dictionary learning and the JADL model

The dictionary learning problem generally aims at decomposing a given set of signals as a weighted sum of basic elements, called atoms. The method finds the atom shapes and weights assuming that the signal is represented as a sparse combination of atoms. The set of the decomposed basic elements (atoms) is the learned dictionary over the given signals, while the coefficients can be used to reconstruct the given set of signals using the learned dictionary and weights.

Jitter-adaptive dictionary learning (JADL) [1] is a dictionary learning framework that is designed to compensate for variations in latency and phase of atoms. The JADL model supposes that atoms present in a signal can suffer from unknown time delays, which will be referred to as jitter. Atoms learned by JADL are defined on the entire signal domain and are supposed, but not restricted, to shift only up to a small fraction of the signal length and adapt their positions across trials. This is typically the case of multi-trial M/EEG signals. It is expected that any independent source in the measurements is learned by an atom.

The model hypothesizes that multiple recordings $\{\mathbf{x}_j\}_{j=1}^{M}$ of one electrode (a line of matrix $\mathscr{M}$ in Eq. 1), can be generated by a dictionary $D = \{\mathbf{d}_i\}_{i=1}^{K}$ with few atoms K in the following way: Given a finite set of time shift operations Δ, for every j there exist coefficients $a_{ij} \in \mathbb{R}$ and shift operators $\delta_{ij} \in \Delta$, such that

$$\mathbf{x}_j = \sum_{i=1}^{K} a_{ij}\delta_{ij}(\mathbf{d}_i) \ . \tag{2}$$

One also assumes that Δ contains only small shifts relative to the size of the time window. The JADL problem can be formulated as:

$$\min_{\mathbf{d}_i, a_{ij}, \delta_{ij}} \sum_{j=1}^{M} \left(\frac{1}{2} \left\| \mathbf{x}_j - \sum_{i=1}^{K} a_{ij}\delta_{ij}(\mathbf{d}_i) \right\|_2^2 + \lambda\|\mathbf{a}_j\|_1 \right) , \tag{3}$$

$$s.t. \ \|\mathbf{d}_i\|_2 = 1, \ \delta_{ij} \in \Delta, \ i = 1, \dots, K, \ j = 1, \dots, M. \tag{4}$$

The algorithm solving Eq. 3, is based on an implementation in [5] for common dictionary learning adopted to deal with jitters [1], which iteratively alternates between (i) sparse coding (finding the coefficients $\{a_{ij}\}$ and the jitters $\{\delta_{ij}\}$) and (ii) dictionary update (finding the shapes $\{\mathbf{d}_i\}$).

(i) Sparse coding This part is solved by rewriting the problem into a form similar to the Lasso problem, which allows to solve it using a modification of least angle regression (LARS) [4]. First, the "unrolled" version of the dictionary has to be defined. An "unrolled" version of the dictionary contains all allowed shifts of all its atoms; is given by $D^s = \{\delta(d) : d \in D, \delta \in \Delta\}$. It can be represented as a matrix $\mathbf{D}^s$ of dimension $T \times KL$, where $L = |\Delta|$ is the number of allowed shifts. The problem is then written as follows:

$$\mathbf{a}_j^s \leftarrow \operatorname{argmin} \frac{1}{2}\|\mathbf{x}_j - \mathbf{D}^s\mathbf{a}_j^s\|_2^2 + \lambda\|\mathbf{a}_j^s\|_1 \ , \tag{5}$$

$$s.t. \quad \|\mathbf{a}_j^{s,i}\|_0 \leq 1, \quad i = 1, \dots, K. \tag{6}$$

where $\mathbf{a}_j^s$ is the corresponding vector of coefficients for x_j. The constraint ensures that once an atom of D^s is chosen, all its shifts are forbidden. This choice selects the proper shift as it is encoded in D^s.

(ii) Dictionary update Block coordinate descent is used to iteratively solve the constrained minimization problem

$$\mathbf{d}_k = \underset{\mathbf{d}_k}{\operatorname{argmin}} \sum_{j=1}^{M} \frac{1}{2} \left\| \mathbf{x}_j - \sum_{i=1}^{K} a_{ij}\delta_{ij}(\mathbf{d}_i) \right\|_2^2 , \ s.t. \ \|\mathbf{d}_k\|_2 = 1 \tag{7}$$

for each atom $\mathbf{d}_k$. This can be solved in two steps, the solution of the unconstrained problem followed by normalization. This is summarized by:

$$\widetilde{\mathbf{d}_k} \ = \ \sum_{j=1}^{M} a_{kj}\delta_{kj}^{-1} \left(\mathbf{x}_j - \sum_{i=1, i \neq k}^{K} a_{ij}\delta_{ij}(\mathbf{d}_i) \right) , \tag{8}$$

$$\mathbf{d}_k = \frac{\widetilde{\mathbf{d}_k}}{\|\widetilde{\mathbf{d}_k}\|_2} \ . \tag{9}$$

δ^{-1} is the opposite shift of δ.

2 Our modified model

A simple way to extend the usage of the JADL implementation would be, to reshape the input multi-dimensional data to a shape compatible with the JADL implementation. Such a data transformation could be achieved by simply stacking the recordings from the different channels as trials or multiplexing the signals from different channels at the same time point, leading to one mixed signal per trial. As a consequence, in the first case (stacking), the dictionary would be learned over measurements that are assumed to have the same waveforms but can have different jitters and coefficients over the different channel, while in the second case (multiplexing), the dictionary would be learned over measurements that are assumed to have the same jitters, but can have different waveforms and coefficients over the different channel measurements. Those techniques would not be optimal as the multi-dimensional signals should share the same waveform and jitters, but have different coefficients over the channels.

To better account the nature of the M/EEG signals, we developed a novel model that learns a single dictionary over multi-dimensional recording that have the same waveforms and jitters, but different coefficients over the channels. Significant modifications are applied to the original JADL framework, especially in the (i) sparse coding step in order to handle multi-dimensional data.

The least angle regression algorithm (LARS) used for solving (i) is modified as:

1. *Atom Selection*: The best shifted versions of the atoms contained in the extended dictionary D^s are selected, over all the channels, leading to a compressed dictionary D. The selection of the atoms is performed using the following criterion:

$$\mathbf{d}_j^s = \underset{\mathbf{d}_j^s \in D^s}{\operatorname{argmax}} \sum_{c=1}^{C} \left\| \mathbf{s}_c \cdot \mathbf{d}_j^s \right\| , \qquad (10)$$

   where $\mathbf{s}_c$ is the signal of channel c and $\mathbf{d}_j^s$ is the j-th atom of the extended dictionary D^s.

2. *Standard LARS sparse coding over the channels for the current atom set*: During this step the multi-dimensional coefficients a_{ijc} are computed using the compressed dictionary D selected at step 1 and the multi-channel signals for the given trial.

The (ii) dictionary update problem is also slightly modified to treat the measurements corresponding to the different channels as additional trial. The dictionary update problem is then written similarly to Eq. (8) as:

$$\widetilde{\mathbf{d}_k} = \sum_{j=1}^{M} \sum_{c=1}^{C} a_{kjc} \delta_{kj}^{-1} \left(\mathbf{x}_{jc} - \sum_{i=1, i \neq k}^{K} a_{ijc} \delta_{ij}(\mathbf{d}_i) \right) , \qquad (11)$$

with δ^{-1} the opposite shift of δ and the normalization of Eq. (9).

In the implementation of our model, the initial dictionary contains atoms generated by random values independent from the signals. As a consequence, an atom learned by the dictionary update process can appears in any latency l in the dictionary. In order to be able to account for all the allowed latencies of the window Δ, a centering of the window should be applied, to realign Δ with respect to the latency l.

3 Results on lead field synthetic data

We created synthetic measurements using a dictionary of $K = 3$ synthetic atoms (a spike, and two oscillatory signals, see Fig. 1). We then selected K places as active brain locations (each place extends over 3 source points). Each location is associated with a specific atom of the dictionary. Each source point in a location receives a signal generated by introducing a random jitter (shift) to the atom, drawn from the set Δ of contiguous allowed shifts (a window of $[-51, 51]$ shifts). Therefore, 9 source signals are generated for the K groups of 3 sources. These signals are then combined with a leadfield matrix $\mathbf{G}$ computed from real EEG measurements [7] as in Eq. (1).

Performing the above procedure for several trials corresponding to new random jitters to the dictionary of $K = 3$ synthetic atoms, leads to the multidimensional M/EEG measurements of size $(T \times M \times C)$. The generated data contain measurements from $C = 6$ channels, $M = 200$ trials and $T = 515$ time samples. Additional data are also generated with various amplitudes of white Gaussian noise.

3.1 A comparison between the original and our multi-dimensional JADL model

In order to assess on the developed model's performance (multi-channel approach) and the improvements that it brings compared to the original JADL framework (single-channel approach), a comparison between the two implementations is made. Both algorithms are executed with the same signals, initial random dictionary and latency parameters (the same range of shifts that was used to generate the M/EEG measurements).

The multi-channel algorithm is executed using all the channels from the input data, while the single-channel algorithm is executed several times, each time using a different channel. Note that the results using the single-channel algorithm depend on how the atoms are represented in that channel.

A goodness of fit metric, is used to evaluate the quality of the learned dictionary computing the correlation between the generated and reconstructed atoms taking into account possible shifts of the atoms.

The results of our multi-channel algorithm (Fig. 1), look promising as the learned dictionary fits very well to the one used to generated the synthetic data. As the input signals are generated by introducing random jitters to the generated dictionary's atoms, the learned atoms can suffer by small jitters depending on the distribution of the picked jitters.

The results of our multi-channel approach look similar to the one obtained by the single-channel algorithm (Fig. 2) when the best channel is used, but when a medium or the worse channel is used, the results become worst and there are cases where the algorithm is unable to recover correctly all the atoms of the dictionary used to generate the signals. In addition, the goodness of fitness metric, showed a small but superior performance for the multi-channel approach giving the coefficients vector of 0.995, 0.996 and 0.995 instead of 0.992, 0.977 and 0.964 for the single-channel approach using the best channel and 0.939, 0.512, 0.512 using the worst channel. In general, finding the best channels is 1) source dependent and 2) highly non-trivial as it depends on both actual sources and leadfields.

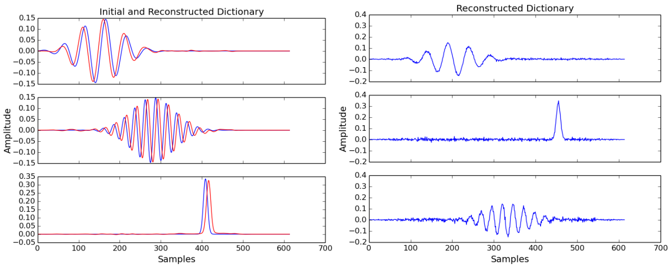

Figure 1 – The generated (blue) and learned (red) dictonary using our model with no noise (left). The learned dictionary on contaminated signals by noise of $SNR: 0.021$ (right).

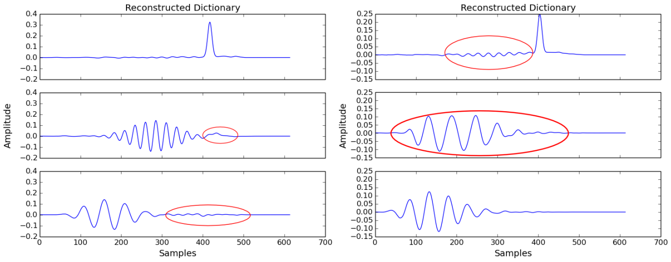

Figure 2 – The learned dictionaries by the single-channel method: using the best (left) and the worst (right) channel. Wrong recovered components are marked by the red ellipses.

The robustness of our algorithm was also tested on the synthetic data contaminated by several noise levels. The algorithm was able to recover correctly the shape of all the atoms even with an SNR of 0.001, whereas it failed to recover all the atoms with SNR of 0.0002 and smaller values.

Table 1 – Robustness to various noise levels

SNR	SNR_{dB}	Atom1	Atom2	Atom3
0.804	-0.944	0.998	0.999	0.997
0.021	-16.700	0.993	0.973	0.983
0.001	-29.240	0.954	0.821	0.892
0.0002	-36.107	0.826	0.585	0.462

4 Results on real data

The performance of the multi-dimensional approach is tested using real MEG and EEG data of 200 channels, 63 trials and 541 time samples. An input parameter of 103 latencies has been provided to the algorithm. Note that, with real data, there is no groundtruth to compare to the obtained results. Most of the input parameters (the number of atoms and the range of jitters) are tuned by “trial and error”.

The learned atoms by the multi-dimensional approach (Fig. 3) appear less noisy compared to the single-channel approach, with waveforms that seem to reveal more information for the underling brain activity (last row of atoms in Fig. 3).

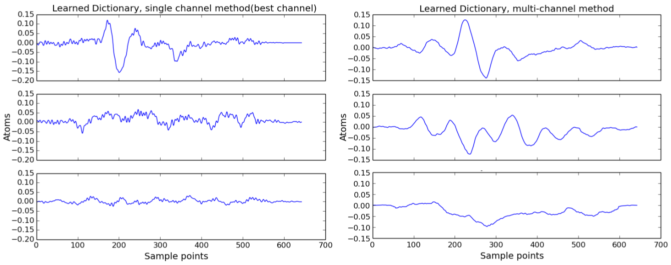

Figure 3 – The single-channel JADL method (left) and the proposed multi-channel method (right).

5 Conclusions

The results obtained using the proposed model, look promising, showing superior performance compared to the original single-channel JADL framework. In particular, the learned atoms appear less noisy.

As the multi-dimensional JADL approach uses all the channels from the input data, it relies on more information. There is also no need for a prior selection of the “best” channel, which gives an additional advantage to the multi-channel approach, compared to the original JADL implementation.

References

[1] S. Hitziger, M. Clerc, A. Gramfort, S. Saillet, C. Bénar, and T. Papadopoulo. *Jitter-adaptive dictionary learning-application to multi-trial neuroelectric signals*. arXiv preprint arXiv:1301.3611, 2013.

[2] C. Bénar, T. Papadopoulo, B. Torrésani, and M. Clerc. *Consensus matching pursuit for multi-trial EEG signals*. Journal of Neuroscience Methods, 180(1):161 – 170, 2009.

[3] K. Knuth, A. Shah, W. Truccolo, M. Ding, S. Bressler, and C. Schroeder. *Differentially variable component analysis: Identifying multiple evoked components using trial-to-trial variability*. Journal of Neurophysiology, 95 (5):3257–3276, 2006.

[4] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. *Least angle regression* . The Annals of Statistics, 32(2):407–499, 04 2004.

[5] J. Mairal, F. Bach, J. Ponce, and G. Sapiro. *Online learning for matrix factorization and sparse coding*. J. Mach. Learn. Res., 11:19–60, March 2010.

[6] P. Durka, A. Matysiak,E. Martinez-Montes,P. Sosa,K. Blinowska. *Multi-channel matching pursuit and EEG inverse solutions*. J Neurosci Methods 148:49–59, 2005.

[7] R. Henson, D. Wakeman, C. Phillips, J. Rowe. *Effective Connectivity between OFA and FFA during face perception: DCM of evoked MEG, EEG and fMRI*. Hum Brain Mapp, 2013.