Multivariate Temporal Dictionary Learning for EEG

Quentin Barthélemy, Cedric Gouy-Pailler, Yoann Isaac, Antoine Souloumiac, Anthony Larue, Jerome I. Mars

Multivariate Temporal Dictionary Learning for EEG

Q. Barthélemy[a,b], C. Gouy-Pailler[a], Y. Isaac[a], A. Souloumiac[a], A. Larue[a,*], J.I. Mars[a,b]

[a] *CEA, LIST, Data Analysis Tools Laboratory, Gif-sur-Yvette Cedex, 91191, France*
[b] *GIPSA-lab, DIS, UMR 5216 CNRS, Grenoble INP, Grenoble, 38402, France*

Abstract

This article addresses the issue of representing electroencephalographic (EEG) signals in an efficient way. While classical approaches use a fixed Gabor dictionary to analyze EEG signals, this article proposes a data-driven method to obtain an adapted dictionary. To reach an efficient dictionary learning, appropriate spatial and temporal modeling is required. Inter-channels links are taken into account in the spatial multivariate model, and shift-invariance is used for the temporal model. Multivariate learned kernels are informative (a few atoms code plentiful energy) and interpretable (the atoms can have a physiological meaning). Using real EEG data, the proposed method is shown to outperform the classical multichannel matching pursuit used with a Gabor dictionary, as measured by the representative power of the learned dictionary and its spatial flexibility. Moreover, dictionary learning can capture interpretable patterns: this ability is illustrated on real data, learning a P300 evoked potential.

Keywords: Dictionary learning, orthogonal matching pursuit, multivariate, shift-invariance, EEG, evoked potentials, P300.

1. Introduction

Scalp electroencephalography (EEG) measures electrical activity produced by post-synaptic potentials of large neuronal assemblies. Although this old medical imaging technique suffers from poor spatial resolution, EEG is still widely used in medical contexts (sleep analysis, anesthesia and coma monitoring, encephalopathies) as well as entertainment and rehabilitation contexts (Brain-Computer Interfaces – BCI). EEG devices are relatively cheap compared to other imaging techniques (*e.g.* MEG, fMRI, PET), and they offer both high temporal resolution (a short period of time between two acquisitions) and very low latency (a delay between the mental task and the recording on the electrodes).

These features are of particular concern for the practitioner interested in (Sanei and Chambers, 2007):

- Event-related potentials (ERPs) or evoked potentials: transient electrical activity that results from external sensory stimulation (*e.g.* P300);
- Steady-state evoked potentials: oscillatory brain activity that results from repetitive external sensory stimulation;
- Event-related synchronization/desynchronization (ERS/ERD): oscillatory activity that results from involvement of a specialized part of the brain; for example, activation of the primary motor area, known as μ (8 – 13 Hz) or β (13 – 30 Hz) bandpower synchronization or desynchronization, have been widely studied;
- Epileptic activity: transient electrical bursts of parts of the brain.

While EEG devices are known to be able to record such aforementioned activities through the wide areas of the sensors, methodologists are usually necessary in EEG experiments. Indeed, they have to provide the practitioners with tools that can capture the temporal, frequential and spatial content of an EEG. Consequently, signal interpretation usually yields a *representation* problem: which dictionary is able to best represent the information recorded in the EEG?

Fourier and wavelets dictionaries allow spectral analysis of the signals through well-defined mathematical bases (Durka, 2007; Mallat, 2009), although they show a lack of flexibility to represent the shape diversity of EEG patterns. The Gabor dictionary has also attracted high interest due to its temporal shift-invariance property. Nevertheless, it also suffers from a lack of flexibility to represent evoked potentials and EEG bursts (Niedermeyer and da Silva, 2004). For example widely studied sleep activities such as spindles (centroparietal or frontal areas) consist of a complex EEG shape; as same epileptic activities such as inter-epileptic peaks are another examples of repeatable and complexely shaped cerebral activities (Niedermeyer and da Silva, 2004). In these two cases practitioners should probably benefit from a custom-based dictionary approach over Fourier or wavelets dictionaries. While these approaches are based on *a priori* models of the data, recent methodological developments focus on data-driven representations: *dictionary learning* algorithms (Tošić and Frossard, 2011).

In EEG analysis, the spatial modeling consists of taking into account inter-channels links, and this has been done in several studies searching for more spatial flexibility (Durka, 2007). The EEG temporal modeling is more difficult. Some approaches use a hypothesis of temporal stationarity and treat only the spa-

*Corresponding author. Tel: + 33 1 69 08 84 39.
Email address: quentin.barthelemy@cea.fr, cedric.gouy-pailler@cea.fr, yoann.isaac@cea.fr, antoine.souloumiac@cea.fr, anthony.larue@cea.fr, jerome.mars@gipsa-lab.grenoble-inp.fr (J.I. Mars)

tial aspect, but this brings about loss of information. Other approaches use the generic Gabor dictionary which is shift-invariant. But it remains difficult to learn an EEG dictionary that integrates these two aspects (Jost et al., 2005; Hamner et al., 2011).

In this article, time-frequency analysis tools that are used to deal with EEG are first reviewed in Section 2. Then, temporal modeling is proposed based on the shift-invariance, and a spatial model called multivariate to provide an efficient dictionary learning in Section 3. As validation, the multivariate methods are applied to real EEG data, and then compared to other methods in Section 4. Finally, to show the interpretability of the learned kernels, methods are applied for learning the P300 evoked potential.

2. EEG analysis

In this section, some of the classical signal processing tools that are applied to EEG data for time-frequency analysis are reviewed: monochannnel and multichannel sparse approximations, shift-invariant dictionaries, and dictionary learning algorithms.

2.1. Monochannel sparse approximation

In this paragraph, the EEG analysis is provided independently for the different electrodes / channels, so it is called monochannel. A single channel signal $y \in \mathbb{R}^N$ of N temporal samples and a normalized dictionary $\Phi \in \mathbb{R}^{N\times M}$ composed of M time-frequency atoms $\{\phi_m\}_{m=1}^M$ are considered. The monochannel decomposition of the signal y is carried out on the dictionary Φ such that:

$$y = \Phi x + \epsilon\,, \tag{1}$$

assuming $x \in \mathbb{R}^M$ for the coding coefficients, and $\epsilon \in \mathbb{R}^N$ for the residual error. The approximation $\hat{y}$ is Φx. The dictionary is redundant since $M > N$, and thus the linear system of Eq. (1) is under-determined. Consequently, sparsity, smoothness or another constraint is needed to regularize the solution x. Considering the constant $K \ll M$, the sparse approximation is written as:

$$\min_x \| y - \Phi x \|^2 \text{ s.t. } \|x\|_0 \leq K\,, \tag{2}$$

with $\|.\|$ for the Frobenius norm, and $\|x\|_0$ for the ℓ_0 pseudo-norm, defined as the number of nonzero elements of vector x. The well-known Matching Pursuit (MP) (Mallat and Zhang, 1993) tackles this difficult problem iteratively, but in a suboptimal way. At iteration k, it iteratively selects the atom that is the most correlated to the residue ϵ^{k-1} as[1] :

$$m^k = \arg\max_m \left| \left\langle \epsilon^{k-1}, \phi_m \right\rangle \right|\,. \tag{3}$$

To carry out time-frequency analysis for EEG, MP is applied (Durka and Blinowska, 2001) with the Gabor parametric dictionary, which is a generic dictionary that is widely used to study EEG signals.

[1] $\langle A, B\rangle = trace(B^H A)$ is the matrix inner product.

2.2. Multichannel sparse approximations

Hereafter, the EEG signal $y \in \mathbb{R}^{N\times C}$ composed of several channels $c = 1..C$ are considered. The c^{th} channel of the signal y is denoted by $y[c]$. The following reviewed methods link these channels spatially with the multichannel model illustrated in Fig. 1(left).

A multichannel MP (Gribonval, 2003) was set up using a spatial (or topographic) prior based on structured sparsity. This model is an extension of Eq. (1), with $y \in \mathbb{R}^{N\times C}$, $\Phi \in \mathbb{R}^{N\times M}$ and $x \in \mathbb{R}^{M\times C}$. The multichannel model linearly mixes an atom in the channels, each channel being characterized by a coefficient. The underlying assumption is that few EEG events are spatially spread over in all of the different channels.

In (Durka, 2007), different multichannel selections are enumerated:

- the original multichannel MP (Gribonval, 2003), called MMP_1 by Durka, selects the maximal energy such that:

$$m^k = \arg\max_m \sum_{c=1}^{C} \left| \left\langle \epsilon^{k-1}[c], \phi_m \right\rangle \right|^2\,; \tag{4}$$

- the multichannel MP (Durka et al., 2005) called MMP_2 selects the maximal multichannel inner product:

$$m^k = \arg\max_m \left| \sum_{c=1}^{C} \left\langle \epsilon^{k-1}[c], \phi_m \right\rangle \right|\,; \tag{5}$$

- the multichannel MP (Matysiak et al., 2005) called MMP_3 selects the maximal energy as Eq. (4), but with complex coefficients that allows it to have a varying phase. In (Matysiak et al., 2005), the channels have the same varying phase φ whereas, in (Gratkowski et al., 2007), each channel c has its own varying phase φ_c, which is the version that is used hereafter;
- the multichannel MP (Sielużycki et al., 2009b), which will call MMP_4 selects the maximal multichannel inner product as Eq. (5), with complex coefficients that allows it to have a phase φ_c for each channel.

Note that other algorithms deal with EEG decomposition. For example, a multichannel decomposition was proposed in (Koenig et al., 2001), but it was based on the method of frames (Daubechies, 1988).

2.3. Shift-invariant dictionaries

The dictionary Φ used in the decomposition can have a particular form. In the shift (also called translation or temporal)-invariant model (Jost et al., 2005; Barthélemy et al., 2012), the signal y is coded as a sum of a few structures, named kernels, that are characterized independently of their positions. The L shiftable kernels of the compact Ψ dictionary are replicated at all positions, to provide the M atoms of the Φ dictionary. Kernels $\{\psi_l\}_{l=1}^L$ can have different lengths T_l, so they are zero-padded. The N samples of the signal y, the residue ϵ, the atoms ϕ_m and the kernels ψ_l are indexed by t. The subset σ_l collects

the translations τ of the kernel $\psi_l(t)$. For the few L kernels that generate all of the atoms, Eq. (1) becomes:

$$y(t) = \sum_{m=1}^{M} x_m \, \phi_m(t) + \epsilon(t) \tag{6}$$

$$= \sum_{l=1}^{L} \sum_{\tau \in \sigma_l} x_{l,\tau} \, \psi_l(t-\tau) + \epsilon(t) \, . \tag{7}$$

This model is also called convolutional, and as a result, the signal y is approximated as a weighted sum of a few shiftable kernels ψ_l. It is thus adapted to overcome the latency variability (also called the jitter) of the events studied.

Algorithms described in Section 2.2 are widely used with a Gabor dictionary for EEG (Durka, 2007). Its generic atoms ϕ_{Gabor} are parameterized as (Mallat, 2009):

$$\phi_{\text{Gabor}}(t) = \frac{1}{\sqrt{s}} \cdot g\left(\frac{t-\alpha\tau}{s}\right) \cdot \cos\left(\, 2\pi f t + \varphi \,\right) \, , \tag{8}$$

where $g(t) = \beta \cdot e^{-\pi t^2}$ is a Gaussian window, β is a normalization factor, s is the scale, τ is the shift parameter, α is the shift factor, f is the frequency and φ is the phase used in MMP_3 and MMP_4 algorithms. Note that a multiscale Gabor dictionary is not properly shift-invariant since the shift factor α, which depends on the dyadic scale s, is not equal to 1 (Mallat, 2009). The drawback of such a dictionary is that generic atoms introduce *a priori* for data analysis.

2.4. Dictionary learning

Recently, dictionary learning algorithms (DLAs) have allowed the learning of dictionary atoms in a data-driven and unsupervised way (Lewicki and Sejnowski, 1998; Tošić and Frossard, 2011). A set of iterations between sparse approximation and dictionary update provides learned atoms, which are no more generic but are adapted to the studied data. Thus, learned dictionaries overcome generic ones, showing better qualities for processing (Tošić and Frossard, 2011; Barthélemy et al., 2012). Different algorithms deal with this problem: the method of optimal directions (MOD) (Engan et al., 2000) generalized under the name iterative least-squares DLA (ILS-DLA) (Engan et al., 2007), the K-SVD (Aharon et al., 2006), the online DLA (Mairal et al., 2010) and others (Tošić and Frossard, 2011).

Only two studies have proposed to include dictionary learning for EEG data. In (Jost et al., 2005), the MoTIF algorithm, which is a shift-invariant DLA, is applied to EEG. It thus learns a kernels dictionary, but only in a monochannel case, which does not consider the spatial aspect. In (Hamner et al., 2011), the well-known K-SVD algorithm (Aharon et al., 2006) is used to carry out spatial or temporal EEG dictionary learning. The spatial learning is efficient and can be viewed as a generalization of the N-Microstate algorithm (Pascual-Marqui et al., 1995). Conversely, results of the temporal learning are not convincing, mainly because the shift-invariant model is not used.

The dictionary redundancy gives a more efficient representation than learning methods based on Principal Component Analysis or Independent Component Analysis (Lewicki and Sejnowski, 1998). Moreover, such methods provide only a base (with $M = N$), and are not adapted to cope with the shift-invariance required by the temporal variability of the EEG.

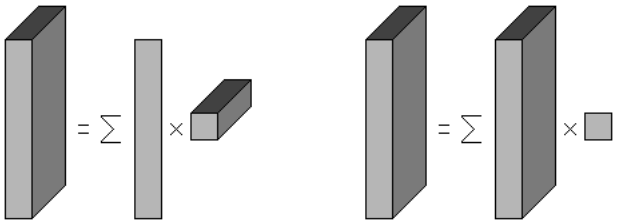

Figure 1: Illustration of the multichannel (left) and the multivariate (right) models.

2.5. Summary of the state of the art

To sum up, the previous paragraph on dictionary learning has already shown the relevance of the shift-invariance to learn the EEG temporal atom. That is also why the parametric Gabor dictionary, which is quasi shift-invariant, is well-adapted for such data and is widely used with the multichannel MPs (Durka, 2007). In multichannel decompositions, the different MPs try to be flexible to match the spatial variability. The use of complex Gabor atoms adds a degree of freedom that improves the quality of the representation (or reconstruction) (Matysiak et al., 2005).

The proposed multivariate approach takes into account these two aspects in a dictionary learning approach: a relevant shift-invariant temporal model and a spatial flexibility which considers all of the channels.

3. The multivariate approach

In this section, the general multivariate approach introduced in (Barthélemy et al., 2012) is adapted to the context of the EEG. The underlying model is first detailed, and then the methods are explained: multivariate orthogonal MP (M-OMP) and multivariate dictionary learning algorithm (M-DLA).
Note that the name *multivariate* is used in (Sielużycki et al., 2009a) to designate MMP_2 (Durka et al., 2005) applied to MEG data, and in (Sielużycki et al., 2009b) for its complex extension MMP_4; they are totally different from the methods explained in (Barthélemy et al., 2012).

3.1. Multivariate model

In the multivariate model, Eq. (1) is kept, but with $y \in \mathbb{R}^{N \times C}$, $\Phi \in \mathbb{R}^{N \times M \times C}$ and $x \in \mathbb{R}^{M}$, and now considering the multiplication Φx as an element-wise product along the dimension M. In the multichannel model, a same monochannel atom is linearly diffused in the different channels (imposing a rank-1 matrix). Whereas in the multivariate model illustrated in Fig. 1(right), each component has its own atom, forming a flexible multicomponent atom, multiplied by one coefficient. The differences between multichannel and multivariate models are detailed in (Barthélemy et al., 2012).

3.2. Multivariate methods

A brief description of the methods is given in this paragraph, as all of the computational details can be found in (Barthélemy et al., 2012). First, note that the OMP (Pati et al., 1993) is an optimal version of the MP, as the provided coefficients vector x is the least-squares solution of Eq. (2), contrary to the MP. The multivariate OMP is the extension of the OMP to the multivariate model described previously. The selection step is defined as:

$$m^k = \arg\max_m \left| \sum_{c=1}^{C} \left\langle \epsilon^{k-1}[c], \phi_m[c] \right\rangle \right| , \quad (9)$$

$$= \arg\max_m \left| \left\langle \epsilon^{k-1}, \phi_m \right\rangle \right| . \quad (10)$$

Concerning the M-DLA, for more simplicity, a non-shift-invariant formalism is used, with the atoms dictionary Φ. A training set of multivariate signals $\left\{y_p\right\}_{p=1}^{P}$ is considered (the index p is added to the other variables). In M-DLA, each trial y_p is treated one at a time. This is an *online* alternation between two steps: a multivariate sparse approximation and a multivariate dictionary update. The multivariate sparse approximation is carried out by M-OMP:

$$x_p = \arg\min_x \left\| y_p - \Phi x \right\|^2 \text{ s.t. } \|x\|_0 \leq K , \quad (11)$$

and the multivariate dictionary update is based on maximum likelihood criterion (Olshausen and Field, 1997), on the assumption of Gaussian noise:

$$\Phi = \arg\min_\Phi \left\| y_p - \Phi x_p \right\|^2 \text{ s.t. } \forall\, m \in \mathbb{N}_M, \|\phi_m\| = 1. \quad (12)$$

This dictionary update step is solved by a stochastic gradient descent. At the end of the M-DLA, the learned dictionary is adapted to the training set. Note that these multivariate methods are already given in a shift-invariant way in (Barthélemy et al., 2012).

Besides applying the M-DLA on high-noised data, the evolution of the kernels length has been improved (Experiment 1 and 2) and specific EEG activities have been time-localized to favour the learning (Experiment 2). Moreover, only two hypotheses are followed in this approach: EEG noise is a Gaussian additive noise, and EEG events can be considered as statistically repeated following the same stimulus. There are no spatial or temporal assumptions made on the dictionary: the learning results are data-driven at most.

4. Experiment 1: dictionary learning and decompositions

The following two experiments aim at showing that multivariate learned kernels are informative (Experiment 1) and interpretable (Experiment 2). In this first experiment, the algorithms presented are applied to EEG data. They are compared to other decomposition methods to highlight their model novelty and their representative performances.

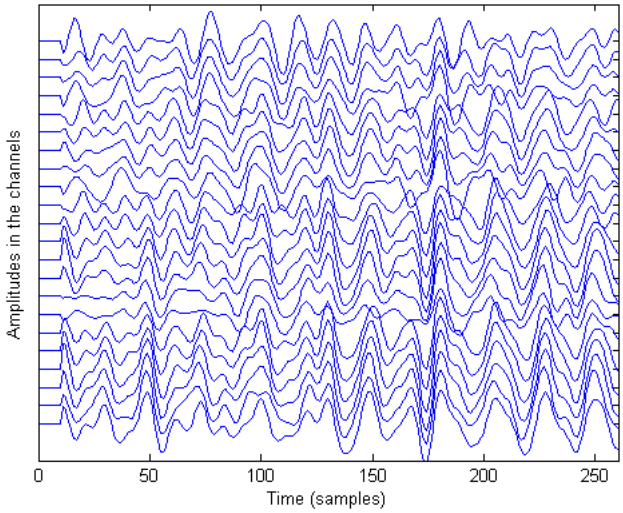

Figure 2: EEG signal $y_{p=1}$ sampled at 250 Hz with $C = 22$ channels.

4.1. EEG data

Real data are used in the following experiments and comparisons. Dataset 2a (Brunner et al., 2008) from BCI Competition IV is considered. There are four classes of motor tasks, but they are not taken into account in this paper. EEG signals are sampled at 250 Hz using $C = 22$ channels. Compliance to our model is natural, as signals are organized into trials. A trial consists of $N = 501$ temporal samples, during which subjects are asked to perform one among four specific motor tasks. Data come from 9 subjects, and the trials are divided in a training set and a testing set. Each set is composed of $P = 288$ signals. Raw data are filtered between 8 Hz and 30 Hz (motor imagery concerns μ and β bands) and zero-padded. Data resulting from this preprocessing are the inputs of the M-DLA. The first samples of the EEG signal $y_{p=1}$ are plotted in Fig. 2.

4.2. Models and Comparisons

M-DLA is applied to the training set of the first subject, and a dictionary of $L = 20$ kernels is learned with 100 iterations [2].

To show the novelty of the proposed multivariate model, the existing multicomponent dictionaries are compared, with $C = 22$ channels. In Fig. 3 and 4, at the top, amplitudes $x_m \times \phi_m$ (ordinate) of one atom are represented as a function of samples (abscissa), and on the bottom, spectrograms in reduced frequencies (ordinate) are represented as a function of samples (abscissa). A Gabor atom is plotted in Fig. 3, based on

[2] During the DLA, the control of the kernels length is tricky due to the low signal-to-noise ratio of the data. For the original M-DLA, the kernels were first initialized on an arbitrary length T^i, and they were then lengthened or shortened during the update step, depending on the energy presence in their edges. Nevertheless, with these rough data, kernels tend to lengthened without stopping. So, a new control method is set up: a limit length T^m borders the kernels over the first 2/3 of the iterations, and then, the border is fixed to $T^m + 40$ for the last iterations. This allows kernels to begin to converge, and to then have the possibility to obtain quasi-null edges, which avoids discontinuities in the latter decompositions using this dictionary.

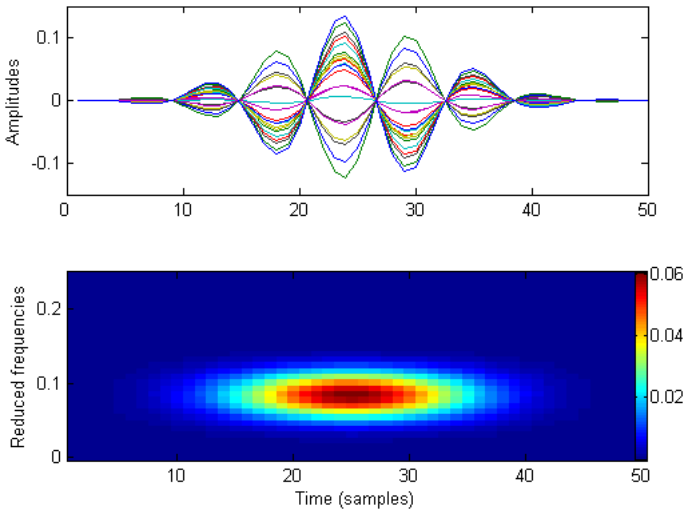

Figure 3: Gabor atom used with the multichannel MP: the temporal profiles of each channel (top) and the time-frequency visualization (bottom).

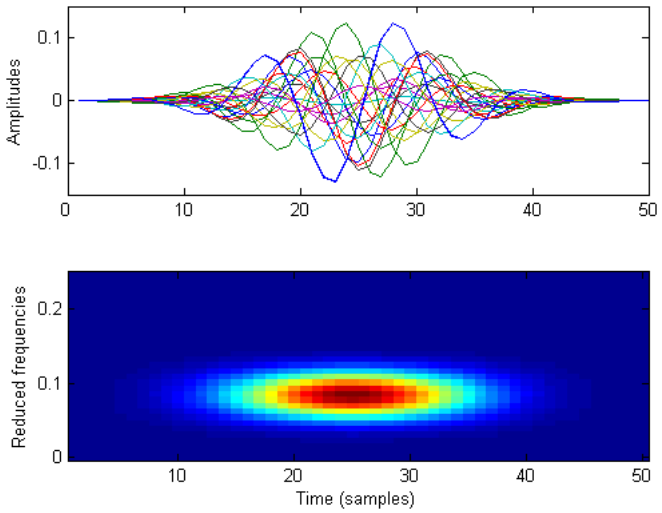

Figure 4: Gabor atom used with the topographic MP case: the temporal profiles of each channel (top) and the time-frequency visualization (bottom). Only the phase of the different channels varies, not the spectral content.

MMP_1 (Gribonval, 2003) or on MMP_2 (Durka et al., 2005), which gives the same kind of atoms. Gabor atom parameters are randomly chosen. In Fig. 4, a Gabor atom is plotted based on MMP_3 (Gratkowski et al., 2007) or on MMP_4 (Sielużycki et al., 2009b). Since this atom has a specific phase for each channel, it is more adaptive than the first one. Nevertheless, in these two cases, the spectral content is identical in each channel.

In Fig. 5, two learned multivariate kernels are plotted, $l = 9$ (top) and $l = 17$ (bottom). Components of Fig. 5(top) are similar, that is adapted to fit data structured like signal y_1 around samples $t = 175$ of Fig. 2. Components of Fig. 5(bottom) are not so different: they appear to be continuously and smoothly distorted, which corresponds exactly to data structured like signal y_1 around samples $t = 75$ of Fig. 2. Moreover, they have various spectral contents, as seen in Fig. 6, contrary to Gabor atoms, which look like monochannel filter banks (Fig. 3(bottom) and 4(bottom)). In the multivariate model, each channel has its own profile, and so its own spectral content, which gives an excellent spatial adaptability.

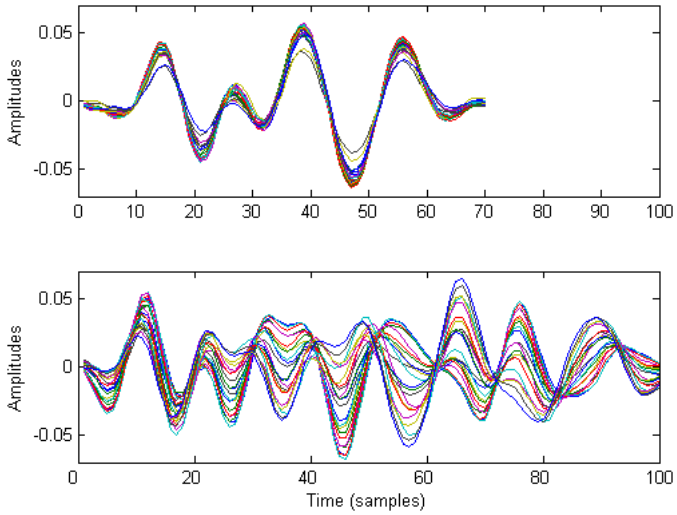

Figure 5: Kernels of the learned multivariate dictionary: kernel $l = 9$ (top) and kernel $l = 17$ (bottom).

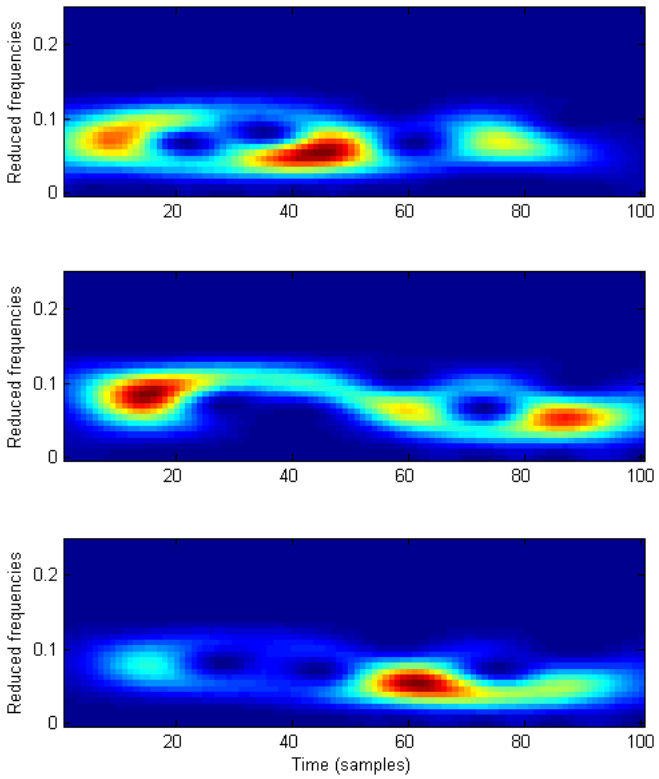

Figure 6: Spectrograms of three components of the kernel $l = 17$: component $c = 10$ (top), component $c = 14$ (middle), component $c = 20$ (bottom).

4.3. Decompositions and Comparisons

In this paragraph, the reconstructive power of the different dictionaries and multichannel sparse algorithms is evaluated.

In a first round, the training set of the first subject is considered. Learned dictionaries (LD) used with M-OMP, and a Gabor dictionary used with MMP_1, MMP_2, MMP_3 and MMP_4, are compared. The Gabor dictionary has $M = 30720$ atoms. Two learned dictionaries are used: one with $L = 20$ kernels (learned in Section 4.2), which gives $M \approx 10000$ atoms; and one with $L = 60$ which gives $M \approx 30000$ atoms, which is a size similar to the Gabor dictionary. For each case, K-sparse approximations are computed on the training set, and the reconstruction rate ρ is then computed. This is defined as:

$$\rho = 1 - \frac{1}{P} \sum_{p=1}^{P} \frac{\|\epsilon_p\|}{\|y_p\|} . \tag{13}$$

The rate ρ is represented as a function of K in Fig. 7. First, MMP_1 (blue dash-dot line) is better than MMP_2 (blue dotted line), and MMP_3 (green solid line) is better than MMP_4 (green dashed line), because the selection of Eq.(5) is more constraining than selection of Eq.(4) as well explained in (Barthélemy et al., 2012). Then, MMP_3 and MMP_4 are better than the other MMP_1 and MMP_2 due to their spatial flexibility on phases atoms. Finally, learned dictionaries (black solid lines) are better than other approaches, even with three times fewer atoms. These two representations (LD with $L = 20$ and $L = 60$) are more compact since they are adapted to the studied signals. If learned kernels code more energy than Gabor atoms, the learned dictionary takes more memory (for storage or transmission) than the parametric Gabor dictionary. Identical results are observed in (Barthélemy et al., 2012), but only in a monochannel case: the learned dictionary overcomes generic ones for sparse reconstructive power.

The generalization is tested in a second round, determining whether if the adapted representation that was learned on the training set of the subject 1, remains efficient for other acquisitions and other subjects. Thus, the testing sets of the 9 subjects are now considered. In the same way, K-sparse approximations are computed with the LD ($L = 60$) on the different testing sets, and the rate ρ is plotted as a function of K in Fig. 8. The different curves are very similar and look like the curve of LD (black solid line with stars) in Fig. 7, that shows the intra-user and inter-user robustness of the learned representation.

Since it has good representation properties, the learned dictionary can be useful for EEG data simulation. Moreover, as noted in (Tošić and Frossard, 2011), learned dictionaries overcome classical approaches for processing such as denoising, etc.

5. Experiment 2: Evoked Potentials Learning

The previous experiment has shown the performances and the relevance of such adaptive representations. The question is to know if these learned kernels can capture behavioral structures with physiological interpretations. To answer, we will focus on the P300 evoked potential.

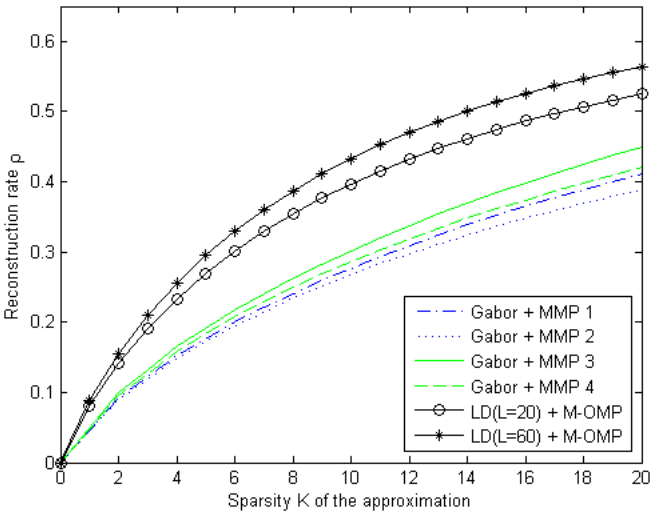

Figure 7: Reconstruction rate ρ on the training set as a function of the sparsity K of the approximation of the different dictionaries and algorithms.

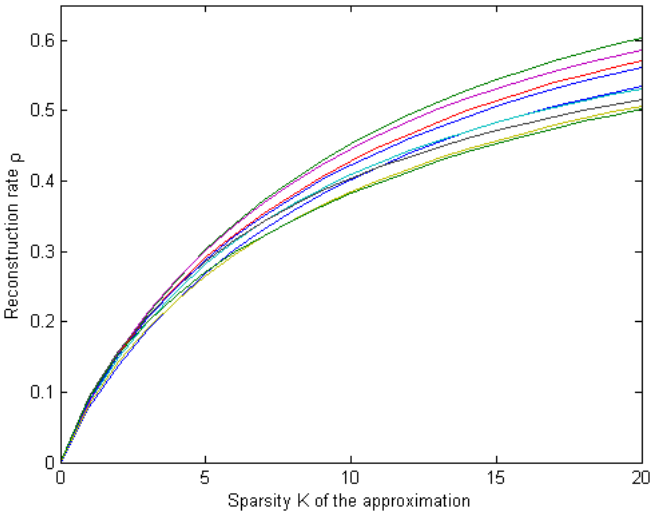

Figure 8: Reconstruction rate ρ of the M-OMP used with the LD ($L = 60$), as a function of the sparsity K of the approximation on the 9 testing sets.

5.1. P300 data

In this section, the experiment is carried out on dataset 2b of the P300 speller paradigm (Blankertz and Schalk, 2002), from BCI Competition II. To sum up, a subject is exposed to visual stimuli. When it is a target stimulus, a P300 evoked potential is provoked in the brain of the subject 300 ms after, contrary to a non-target stimulus. The difficulty is the low SNR (signal-to-noise ratio) of the evoked potentials.

In this dataset, $P = 1261$ target stimuli are carried out. Considering the complete acquisition $Y \in \mathbb{R}^{\mathcal{N} \times C}$ of $\mathcal{N}$ samples, it is parsed in P epoched signals $\left\{y_p\right\}_{p=1}^{P}$ of N samples. $D \in \mathbb{R}^{\mathcal{N} \times N}$ is a Toeplitz matrix, the first column of which is defined such that $D(t^p, 1) = 1$, where t^p is the onset of the pth target stimulus. Acquired signals have $C = 64$ channels and are sampled at 240 Hz. They are filtered between 1 Hz and 20 Hz by a 3rd order

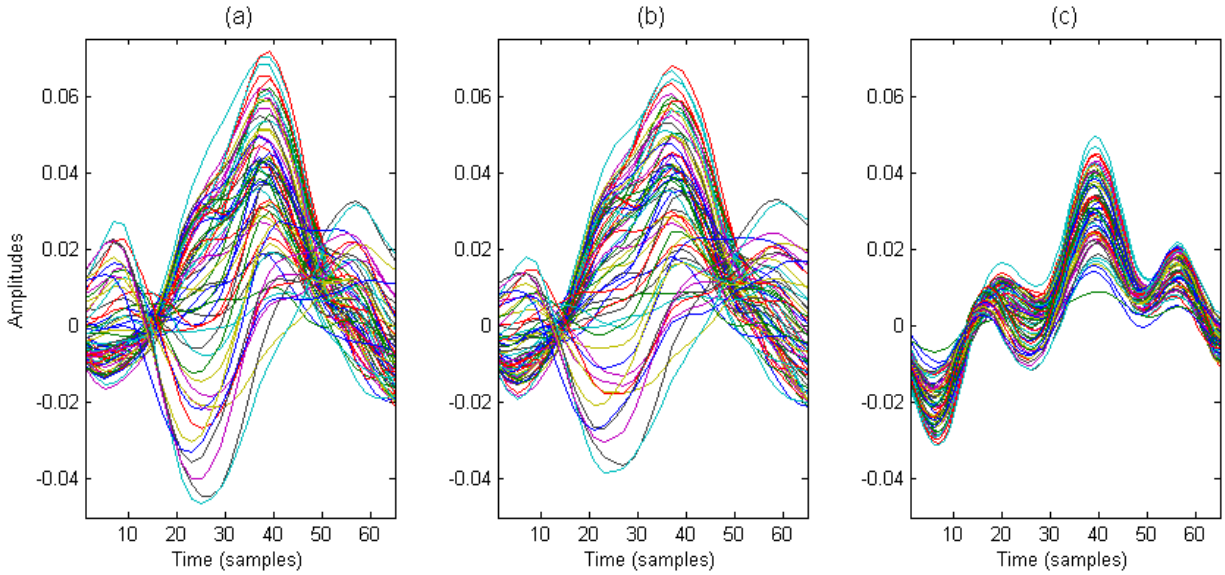

Figure 9: Multivariate temporal patterns of the P300 computed by the grand average (a), by least-squares (b) and by multivariate dictionary learning (c). Sampled at 240 Hz, the amplitudes are given as a function of the temporal samples.

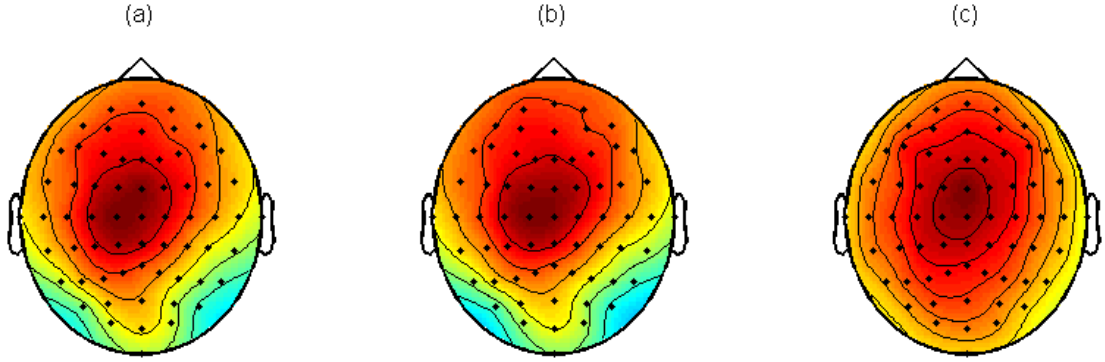

Figure 10: Spatial patterns of the P300 computed by the grand average (a), by least-squares (b) and by multivariate dictionary learning (c).

Butterworth filter.

5.2. *Review of models and methods*

There are two models for the P300 waveform. With $\phi_{\text{P300}} \in \mathbb{R}^{N\times C}$, the classical additive model can be written as:

$$y = \phi_{\text{P300}} + \epsilon \,, \tag{14}$$

and, with shift-invariance and an amplitude, it gives a more flexible model (Jaśkowski and Verleger, 1999):

$$y(t) = x\, \psi_{\text{P300}}(t-\tau) + \epsilon(t) \,. \tag{15}$$

Note that Eq. (6) is retrieved, but reduced to one kernel ($L = 1$) and in a multivariate case.

One classical way for working on P300 is the *dictionary design*, which uses a template, *i.e.* a pre-determined pattern designed to fit a P300 waveform. An *a priori* is thus injected through this prototype, generally with the shift-invariant model. For example, monochannel patterns as Gaussian functions (Lange et al., 1997), time-limited sinosoids (Jaśkowski and Verleger, 1999), generic mass potentials (Melkonian et al., 2003), Gamma functions (Li et al., 2009) and Gabor functions (Jörn et al., 2011) have been used to match the P300 and other evoked potentials.

Another way is *dictionary learning*, which learns EEG patterns in a data-driven way. Different recent methods allow the learning of evoked potentials, but only with monochannel patterns (D'Avanzo et al., 2011; Wu and Gao, 2011; Nonclercq et al., 2012). Here, we are interested in learning multivariate patterns. Based on Eq. (14), (Rivet et al., 2009) gives a least-squares (LS) estimation that takes into account overlaps of consecutive target stimuli, defined as:

$$\begin{aligned}\hat{\phi}_{\text{P300}} &= \arg\min_{\phi} \| Y - D\phi \|^2 \\ &= (D^T D)^{-1} D^T Y \,. \end{aligned} \tag{16}$$

This estimation is optimal if there is no variability in latencies and amplitudes. Furthermore, without overlap, this estimation

is equivalent to the grand average (GA) carried out on epoched signals $\{y_p\}_{p=1}^{P}$:

$$\bar{\phi}_{\text{P300}} = \frac{1}{P} D^T Y = \frac{1}{P} \sum_{p=1}^{P} y_p \,. \tag{17}$$

However, these two multivariate estimations made strong assumptions on latencies and amplitudes, which is a lack of flexibility. Based on model (15), we propose to use dictionary learning, which can be viewed as an iterative online least-squares estimation:

$$\min_{\psi} \sum_{p=1}^{P} \min_{x_p,\tau_p} \left\| y_p - x_p\, \psi(t-\tau_p) \right\|^2 \text{ s.t. } \|\psi\| = 1 \,, \tag{18}$$

with the variable τ restricted to an interval around 300 ms. The estimated kernel is denoted by $\tilde{\psi}_{\text{P300}}$.

Note that spatial filtering is not considered, which provides enhanced but projected signals (for example, the second part of (Rivet et al., 2009)).

5.3. Learning and qualitative comparisons

M-DLA is applied to the training set $\{y_p\}_{p=1}^{P}$, with $K = 1$. The grand average estimation $\bar{\phi}_{\text{P300}}$ is used for initialization, to provide a warm start, and the M-DLA is used on 20 iterations on the training set. The kernel length is $T = 65$ samples, which represents 270 ms, and it is constant during the learning. The optimal parameter τ is searched only on an interval of 9 points centered around 300 ms after the target stimulus [3] , which gives a latency tolerance of ± 16.7 ms.

To be compared to the kernel $\tilde{\psi}_{\text{P300}}$, estimations $\bar{\phi}_{\text{P300}}$ and $\hat{\phi}_{\text{P300}}$ are firstly limited to 65 samples and then normalized. Note that considered patterns have $C = 64$ channels and they are not spatially filtered to be enhanced. Multivariate temporal patterns are plotted in Fig. 9, with $\bar{\phi}_{\text{P300}}$ estimated by grand average (a), $\hat{\phi}_{\text{P300}}$ estimated by least-squares (b) and $\tilde{\psi}_{\text{P300}}$ by multivariate dictionary learning (c). The amplitude is given in ordinate and the samples in abscissa. Patterns (a) and (b) are very similar, whereas kernel (c) is thinner and the components are in-phase.

Associated spatial patterns are composed of the amplitudes of the temporal maximum of the patterns. They are then plotted in Fig. 10, where (a) is the pattern estimated by the grand average, (b) by least-squares, and (c) by multivariate dictionary learning. We observe that, similarly to the temporal comparison, patterns (a) and (b) are quite similar. The topographic scalp (c) is smoother than the others and does not exhibit a supplementary component behind the head. But, as the true P300 reference pattern is unknown, it is difficult to have a quantitative comparison between these patterns.

[3] However, an edge effect is observed during the learning: temporal shifts τ_p of plentiful signals are localized on the interval edges. It means that the global maximum of the correlation has not be found in this interval, and τ_p is a value by default. This can be due to the high level of noise that prevents the correlation from detecting the P300 position. Such signals will damage the $\tilde{\psi}_{\text{P300}}$ if they are used for the dictionary update, since the shift parameters are not optimal. To avoid this, the kernel update is not carried out for such signals.

5.4. Quantitative comparisons by analogy

A solution consists in validating the previous case by analogy with a simulation case. A P300 pattern previously learned with $C = 64$ channels is chosen to be the reference P300 of the simulation. $P = 1000$ signals are created using this reference P300 with shift parameters drawn from a Gaussian distribution. Spatially correlated noise, reproduced with a FIR filter learned on EEG data (Anderson et al., 1998), is added to signals to give a signal-to-noise ratio of −10 dB. First, the GA estimation is computed for this dataset. Secondly, this estimation is used for the initialization of M-DLA, which is then applied to the dataset. This experimentation is carried out 50 times for different standard deviations of the shift parameters. The patterns estimated from these two approaches are compared quantitatively, computing their correlations with the reference pattern. Since the patterns are normalized, the correlations absolute values are between 0 and 1.

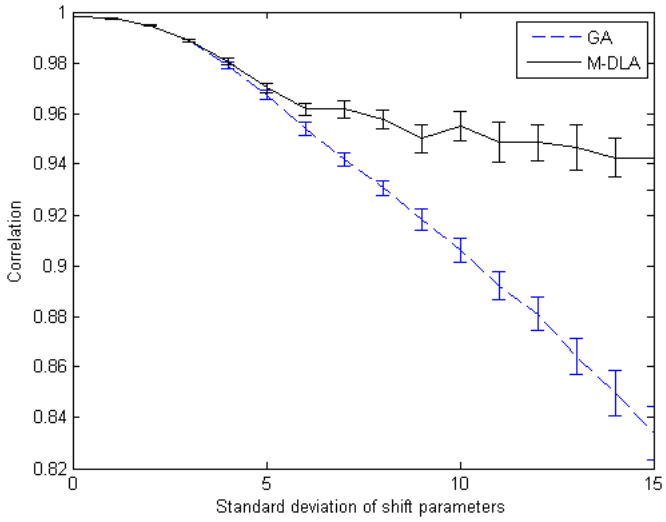

Figure 11: Correlations averaged over 50 experiments as a function of the standard deviation of shift parameters, for the grand average (GA) and the multivariate dictionary learning (M-DLA).

Correlations are plotted in Fig. 11 as a function of the standard deviation of the shift parameters. If the recovery performances of GA and M-DLA are similar for small standard deviations, the more patterns are shifted, and the more the M-DLA outperforms GA. This shows quantitatively that the shift-invariant dictionary learning is better than the grand average approach (equivalent to the LS estimation in this case, since there is no overlap between signals).

Temporal patterns from one experiment are plotted in Fig. 12, with a standard deviation of $\sigma = 6$. Note that the reference P300 pattern (a) comes from learning of Section 5.3, but with an interval of 1. It is thus estimated with signals giving their maximum correlation at 300 ms exactly. First, we observe that averaging shifted patterns gives a spread estimated pattern (b) compared to the reference one (a). Then, the M-DLA pattern (c) is thinner than (b). This confirms the results obtained for the correlations. By analogy, we can assume that the reference

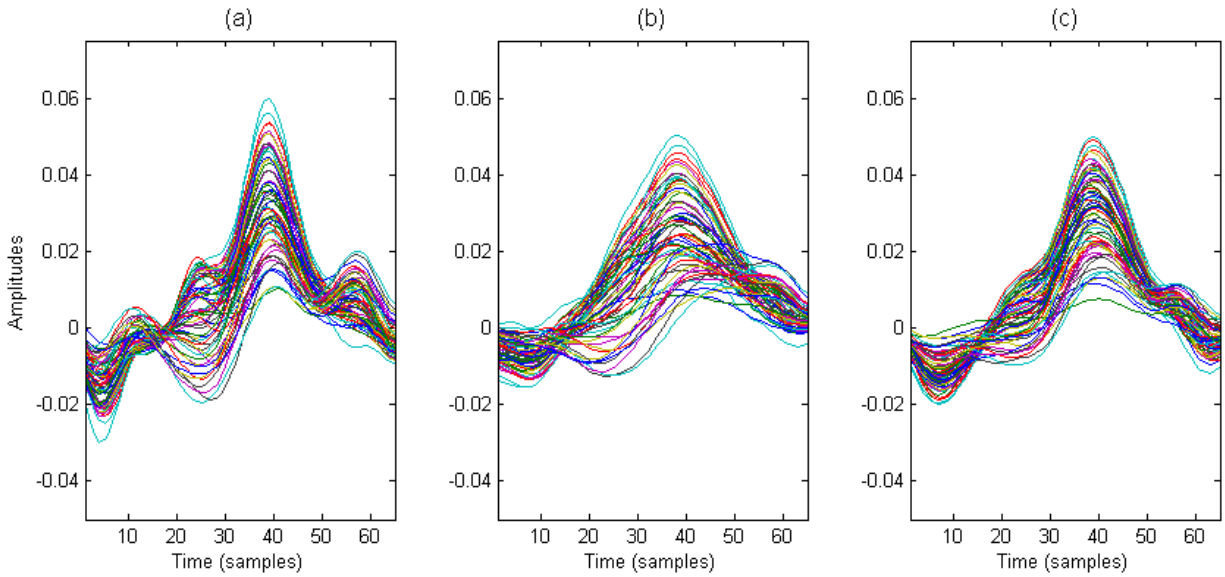

Figure 12: Temporal patterns of the P300: the reference (a), the grand average (b) and the multivariate dictionary learning (c).

P300 is a thin pattern, spread by the average of the shifted occurrences. The M-DLA allows a thinner pattern to be extracted due to its shift-invariance flexibility.

As shift-invariance is not easily integrated into EEG processing, the hypothesis of temporal stationarity is often carried out through the covariance matrix (Blankertz et al., 2008) or the grand average (Hamner et al., 2011). This experiment shows that this hypothesis is rough and provides a loss of temporal information. Although the shift-invariance model and the spatial flexibility of our approach is an obvious improvement, the goal is not to say that the learned pattern is better than the others[4], but to present a new estimation method, with the prospect to move forward with the knowledge of the P300, and to improve the processings based on the P300 estimation. Moreover, we observe that components of learned kernels are rather in-phase contrary to GA and LS estimations. This data driven result opens a question: are repeatable and energy EEG activities rather in-phase or rather with opposite phases?

This experiment shows that learned kernels are interpretable. However, due to the high noise, in order to be interpreted with a physiological meaning, the dictionary learning algorithm has to time-localize activities of interest as presented in Experiment 2, contrary to Experiment 1. Note that the M-DLA can be applied to other kinds of evoked potentials, such as mismatch negativity and the N200, among others.

6. Discussion and Conclusions

After reviewing the classical time-frequency approaches for representing EEG signals, our dictionary-based method has been described. It is characterized by a spatial model, referred to as multivariate, that is very flexible, with one specific profile in each component, and with a shift-invariance used for the temporal model that is obviously pertinent for EEG data. This provides multivariate and shift-invariant temporal dictionary learning. Our approach has been shown to outperform the Gabor dictionaries in terms of their sparse representative power; *i.e.* the number of atoms necessary to represent a fixed percentage of the EEG signals. Specifically high-energy and repeated patterns have been learned and the resulting dictionary has been shown to be robust to intra-user and inter-user variability. Interestingly, the proposed approach has also been able to extract custom patterns in a very low signal-to-noise ratio context. This property is here demonstrated in the particular context of the P300 signals, which are repeated and approximately time-localized. In the context of the EEG, the results obtained can be interpreted according to two distinct points of view.

First, EEG signal interpretation entails the analysis of huge amounts of multicomponent signals in the temporal domain. Consequently the best representation domain for neurophysiologists would be the ability to efficiently concentrate the information using a small number of active and informative components. In this sense, our sparse approach is shown to outperform classical approaches based on Gabor atoms. In other words, at a low and fixed number of active atoms, our method is able to better render the information available in initial EEG signals. This is also interesting for simulating multivariate EEG data. The issue of generating realistic multivariate EEG signals has indeed become recurrent over the past few years, to provide experimentally validated algorithms in a tightly controlled context. As shown in our first experiment, our approach can efficiently represent the diversity of EEG signals. Consequently, we believe that it represents a relevant and competitive candidate for realistic EEG generation.

[4]Moreover, between patterns plotted in Fig. 9(c) and 12(a), both estimated by M-DLA, we are not able to say which is the best.

Secondly, it should be kept in mind that strong *a-priori* conditions are considered by methodologists when they are considering pre-defined models based on generic dictionaries. While these assumptions can be accurate enough in the case of oscillatory activities (*e.g.*, Fourier, wavelets or Gabor), various EEG patterns cannot be efficiently represented through these dictionaries. The flexibility of our approach relies on the fact that shiftable kernels are learned directly from data. This point is of particular interest for evoked potentials, or event-related potentials. To conclude, multivariate learned kernels are informative and interpretable, which is excellent for EEG analysis.

For the propects relating to a brain-computer interface (BCI), on the one hand, classical BCI methods can be improved taking into account the shift flexibility. Then, on the other hand, as noted in (Tošić and Frossard, 2011), dictionaries learned with discriminative constraints are efficient for classification. The future prospect will thus be to modify the proposed multivariate temporal method to give a spatio-temporal approach for BCI. Finally, wavelet parameters learning methods as (Yger and Rakotomamonjy, 2011) can be extended to the multicomponent case.

Acknowledgments

The authors would like to thank V. Crunelli and anonymous reviewers for their fruitful comments, and C. Berrie for his help about English usage.

References

Aharon M, Elad M, Bruckstein A. K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. IEEE Trans Signal Process 2006;54:4311–22.

Anderson C, Stolz E, Shamsunder S. Multivariate autoregressive models for classification of spontaneous electroencephalographic signals during mental tasks. IEEE Trans Biomed Eng 1998;45:277–86.

Barthélemy Q, Larue A, Mayoue A, Mercier D, Mars J. Shift & 2D rotation invariant sparse coding for multivariate signals. IEEE Trans Signal Process 2012;60:1597–611.

Blankertz B, Schalk G. 2nd Wadsworth BCI Dataset (P300 Evoked Potentials). 2002.

Blankertz B, Tomioka R, Lemm S, Kawanabe M, Müller K. Optimizing spatial filters for robust EEG single-trial analysis. IEEE Signal Process Mag 2008;41:581–607.

Brunner C, Leeb R, Müller-Putz G, Schlögl A, Pfurtscheller G. BCI Competition 2008 - Graz data set A. 2008.

Daubechies I. Time-frequency localization operators: a geometric phase space approach. IEEE Trans Inf Theory 1988;34:605–12.

D'Avanzo C, Schiff S, Amodio P, Sparacino G. A bayesian method to estimate single-trial event-related potentials with application to the study of the P300 variability. J Neurosci Methods 2011;198:114–24.

Durka P. Matching Pursuit and Unification in EEG Analysis. Artech House, 2007.

Durka P, Blinowska K. A unified time-frequency parametrization of EEGs. IEEE Eng Med Biol Mag 2001;20:47–53.

Durka P, Matysiak A, Martínez-Montes E, Valdés-Sosa P, Blinowska K. Multichannel matching pursuit and EEG inverse solutions. J Neurosci Methods 2005;148:49–59.

Engan K, Aase S, Husøy J. Multi-frame compression: theory and design. Signal Process 2000;80:2121–40.

Engan K, Skretting K, Husøy J. Family of iterative ls-based dictionary learning algorithms, ILS-DLA, for sparse signal representation. Digit Signal Process 2007;17:32–49.

Gratkowski M, Haueisen J, Arendt-Nielsen L, Zanow F. Topographic matching pursuit of spatio-temporal bioelectromagnetic data. Prz Elektrotech 2007;83:138–41.

Gribonval R. Piecewise linear source separation. In: Proc. SPIE 5207. 2003. p. 297–310.

Hamner B, Chavarriaga R, del R. Millán J. Learning dictionaries of spatial and temporal EEG primitives for brain-computer interfaces. In: Workshop on structured sparsity: learning and inference ; ICML 2011. 2011. .

Jaśkowski P, Verleger R. Amplitudes and latencies of single-trial ERP's estimated by a maximum-likelihood method. IEEE Trans Biomed Eng 1999;46:987–93.

Jörn M, Sielużycki C, Matysiak M, Żygierewicz J, Scheich H, Durka P, König R. Single-trial reconstruction of auditory evoked magnetic fields by means of template matching pursuit. J Neurosci Methods 2011;199:119–28.

Jost P, Vandergheynst P, Lesage S, Gribonval R. Learning redundant dictionaries with translation invariance property: the MoTIF algorithm. In: Workshop Structure et parcimonie pour la représentation adaptative de signaux SPARS '05. 2005. .

Koenig T, Marti-Lopez F, Valdés-Sosa P. Topographic time-frequency decomposition of the EEG. NeuroImage 2001;14:383–90.

Lange D, Pratt H, Inbar G. Modeling and estimation of single evoked brain potential components. IEEE Trans Biomed Eng 1997;44:791–9.

Lewicki M, Sejnowski T. Learning overcomplete representations. Neural Comput 1998;12:337–65.

Li R, Keil A, Principe J. Single-trial P300 estimation with a spatiotemporal filtering method. J Neurosci Methods 2009;177:488–96.

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

Mallat S. A Wavelet Tour of signal processing. 3rd edition, New-York : Academic, 2009.

Mallat S, Zhang Z. Matching pursuits with time-frequency dictionaries. IEEE Trans Signal Process 1993;41:3397–415.

Matysiak A, Durka P, Martínez-Montes E, Barwiński M, Zwoliński P, Roszkowski M, Blinowska K. Time-frequency-space localization of epileptic EEG oscillations. Acta Neurobiol Exp 2005;65:435–42.

Melkonian D, Blumenthal T, Meares R. High-resolution fragmentary decomposition - a model-based method of non-stationary electrophysiological signal analysis. J Neurosci Methods 2003;131:149–59.

Niedermeyer E, da Silva F. Electroencephalography: Basic principles, clinical applications and related fields. 5th edition, 2004.

Nonclercq A, Foulon M, Verheulpen D, De Cock C, Buzatu M, Mathys P, Van Bogaert P. Cluster-based spike detection algorithm adapts to interpatient and intrapatient variation in spike morphology. J Neurosci Methods 2012;210:259–65.

Olshausen B, Field D. Sparse coding with an overcomplete basis set: a strategy employed by V1? Vision Res 1997;37:3311–25.

Pascual-Marqui R, Michel C, Lehmann D. Segmentation of brain electrical activity into microstates: Model estimation and validation. IEEE Trans Biomed Eng 1995;42:658–65.

Pati Y, Rezaiifar R, Krishnaprasad P. Orthogonal Matching Pursuit: recursive function approximation with applications to wavelet decomposition. In: Proc. Asilomar Conf. on Signals, Systems and Comput. 1993. p. 40–4.

Rivet B, Souloumiac A, Attina V, Gibert G. xDAWN algorithm to enhance evoked potentials: Application to brain-computer interface. IEEE Trans Biomed Eng 2009;56:2035–43.

Sanei S, Chambers J. EEG signal processing. John Wiley & Sons, 2007.

Sielużycki C, König R, Matysiak A, Kús R, Ircha D, Durka P. Single-trial evoked brain responses modeled by multivariate matching pursuit. IEEE Trans Biomed Eng 2009a;56:74–82.

Sielużycki C, Kús R, Matysiak A, Durka P, König R. Multivariate matching pursuit in the analysis of single-trial latency of the auditory M100 acquired with MEG. Int J Bioelectromagn 2009b;11:155–60.

Tošić I, Frossard P. Dictionary learning. IEEE Signal Process Mag 2011;28:27–38.

Wu W, Gao S. Learning event-related potentials (ERPs) from multichannel EEG recordings: a spatio-temporal modeling framework with a fast estimation algorithm. In: Int. Conf. IEEE EMBS. 2011. p. 6959–62.

Yger F, Rakotomamonjy A. Wavelet kernel learning. Pattern Recognit 2011;44:2614–29.