ℓ_1-NORM REGULARIZATION FOR SPARSE REPRESENTATION AND P300 WAVE DETECTION IN BRAIN-COMPUTER INTERFACES

Victoria Peterson[♭], Hugo L. Rufiner[♭, †] and Rubén D. Spies[*]

[♭] *Instituto de Investigación en Señales, Sistemas e Inteligencia Computacional, Facultad de Ingeniería y Ciencias Hídricas, Universidad Nacional del Litoral, Consejo Nacional de Investigaciones Científicas y Técnicas (sinc(i)-FICH-UNL-CONICET), Ruta Nac. 168, km 472.4 (3000), Santa Fe - Argentina, vpeterson@santafe-conicet.gov.ar*

[†] *Facultad de Ingeniería, Universidad Nacional de Entre Ríos (FI-UNER), Ruta Prov. 11, km 10 (3100), Oro Verde - Entre Ríos - Argentina, lrufiner@fich.unl.edu.ar*

[*] *Instituto de Matemática Aplicada del Litoral (IMAL-CONICET-UNL), Ruta Nac. 168, Paraje El Pozo (3000), Santa Fe - Argentina and Facultad de Ingeniería Química, Universidad Nacional del Litoral, Consejo Nacional de Investigaciones Científicas y Técnicas (FIQ-UNL), Santiago del Estero 2829 (3000), Santa Fe - Argentina, rspies@santafe-conicet.gov.ar*

Abstract: A Brain-Computer Interface (BCI) is a system which provides direct communication between the mind of a person and the outside world by using only brain activity (EEG). A common EEG-BCI paradigm is based on the so called Event-Related Potentials (ERP) which are responses of the brain to some external stimuli. One of the main components of ERP signals is an enhanced positive-going component called P300 wave. The ℓ_1-norm minimization has been widely used due to its sparsity-inducing property, convenient convexity and great success in several applications. In this work we propose a sparse representation and posterior classification of ERPs signals by means of an ad-hoc spatio-temporal dictionary composed of bidimensional Gaussian elements. The classification is based on minimizing the residual between a test sample and its estimation.

Keywords: *Brain-Computer Interface, Sparse Representation, ℓ_1-minimization.*
2000 AMS Subject Classification: primary: 92C55, 92C20, secondary: 65F22, 65J20

1 INTRODUCTION

Sparse representations have received great interest in recent years due to their success in many applications in signal and image processing [1], [2]. Given a $m \times n$ matrix A, called dictionary, an unknown signal $\mathbf{x} \in \mathbb{R}^n$ and a measurement vector $\mathbf{y} \in \mathbb{R}^m$, we seek to find the sparsest coefficient vector $\mathbf{x}$ such that $\mathbf{y} = A\mathbf{x}$. Recovering $\mathbf{x}$ given A and $\mathbf{y}$ is a non trivial inversion problem since in general the size of $\mathbf{x}$ is greater than the size of $\mathbf{y}$, the measurement $\mathbf{y}$ is contaminated by noise and the problem is usually severally ill-posed. Hence, regularization is required.

Most of the existing works on sparse learning are based on variants of the ℓ_1-norm regularization due to its sparsity-inducing property, convenient convexity, wide and strong theoretical support, and great success in several applications [3]. In practice $\mathbf{y}$ is contaminated by noise, thus the regularized sparse solution is obtained by solving the following minimization problem:

$$(P_{1,2}): \quad \min_{\mathbf{x} \in \mathbb{R}^n} ||\mathbf{x}||_1 \quad s.t. \quad ||\mathbf{y} - A\mathbf{x}||_2 \leq \epsilon, \tag{1}$$

where ϵ is a noise level.

Problem $(P_{1,2})$, known as the *basis pursuit denoising* problem (BPDN), is equivalent to the following unconstrained minimization problem:

$$(P_\lambda): \quad \min_{\mathbf{x} \in \mathbb{R}^n} \frac{1}{2}||\mathbf{y} - A\mathbf{x}||_2^2 + \lambda||\mathbf{x}||_1, \tag{2}$$

which can be viewed as a generalized Tikhonov-Phillips regularization method [4] or in a statistical context as a Least Absolute Shrinkage and Selection Operator (LASSO) [5].

In this article we use the sparse representation for posterior classification based on the ideas proposed in [6] in the context of Event Related Potential (ERP) recognition.

The Brain-Computer Interfaces (BCI) can significantly improve the quality of life of a person who cannot control his/her own body or even is not able to communicate. By using only brain activity, BCI provides a person a new way of communication and control without needing any peripheral nerves or muscles [7]. When a person is stimulated with some external and "rare" item (which can be auditory, visual or somatosensorial) an ERP is elicited. One of the main components of such ERP signals is an enhanced positive-going component with a latency of about 300 ms (called P300 wave) [8], [9]. Unfortunately, detecting a P300 wave (which means detecting the ERP signal) is not an easy task, mainly due to the fact that SNR between ERP and EEG signals is very low (about -50 db) and also due to the large variation present in P300 wave records among different trails.

In order to use all the information that can be found in EEG records we shall construct an ad-hoc dictionary whose elements are composed by two bidimensional Gaussian atoms representing the main spatio-temporal variation of EEG records with and without P300.

2 Material and Methods

2.1 Database

An Open-Access P300 speller database was used [10]. EEG records from 3 subjects were acquired by 10 electrodes at a rate of 256 Hz. Each subject participated in 4 sessions. In this work we used the first and second sessions as training and testing sets, respectively. In both sets there are records belonging to the target class (with ERP signals) and belonging to the non-target class (without ERP signals). The training set consisted of 2880 EEG epochs per channel (480 target records and 2400 non-target records) while the testing set consisted of 900 EEG epochs per channel (150 of them being target). In this context, an epoch is a EEG record of one second duration extracted at the beginning of a stimulus.

2.2 Ad-hoc Dictionary Generation

In order to use all the information available in an epoch, we used not only the information time-to-time but also the spatial information given by the electrode's positions. In this way, we constructed one image per trial in the time-channel plane. Because the trial size was 256 and the number of channels was ten, our images consisted of 256×10 pixels.

Since the registered amplitude of the P300 wave differs according to the sensors's position, the channels were re-ordered based on decreasing signal energy. Next we filtered the images with a median filter. Due to the fact that each re-ordered and filtered target image contained one or two notable peaks between 0.2s and 0.6s and because the 0.2-0.6 range is in agreement with the latency window of the P300 wave, we cropped all images between 0.2s and 0.6s, resulting in images of 104×10 pixels. In the sequel, a template will referred to the re-ordered, filtered and cropped image.

Next, we generated a dictionary by means of variations of an appropriate "mother" element. In our case we shall consider one mother element per class and per subject, given by a linear affine combination of two Gaussian functions (atoms). More precisely, let $\mathcal{P} = \{(p_1, p_2, ..., p_{11}) \in \mathbb{R}^{11} : p_5, p_6, p_{10}, p_{11} > 0\}$ and for $\mathbf{p} \in \mathcal{P}$ define:

$$z(t, c; \mathbf{p}) = p_1 + p_2 \exp\left(\frac{(t-p_3)^2}{p_5^2} + \frac{(c-p_4)^2}{p_6^2}\right) + p_7 \exp\left(\frac{(t-p_8)^2}{p_{10}^2} + \frac{(c-p_9)^2}{p_{11}^2}\right). \tag{3}$$

In the sequel we shall always identify the scalar field $z(t, c; \mathbf{p})$ with the vector $z(\mathbf{p}) \in \mathbb{R}^{1040}$ obtained after stacking in a column vector the matrix resulting of the evaluation of $z(t, c; \mathbf{p})$ over 104×10 image grid.

Given the template $f(t, c)$ (or simply $f \in \mathbb{R}^{1040}$) we formulate the corresponding non-linear fitting problem associated to f and the mother element $z(\mathbf{p})$ as:

$$(F_{\mathbf{p}}): \quad \min_{\mathbf{p} \in \mathcal{P}} ||f - z(\mathbf{p})||_2^2. \tag{4}$$

The Levenberg-Marquardt (LM) method [11], [12], which is a standard and efficient technique to solve nonlinear least squares problems, was used to find $\mathbf{p}$ in (4).

As the dictionary must capture the main variation of the P300 wave for each subject, in order to construct an appropriate dictionary we analyzed the sensitivity of the parameter vector $\mathbf{p}$ in each one of both classes. This process was done by solving (4) with templates generated as the average of n epoch (we took $n = 5$ for the target class and $n = 25$ for the non-target class). At the end, we obtained a range variation for each one of the 11 components of $\mathbf{p}$. For the dictionary generation itself we varied one parameter at the time while the others were kept constant. The variations were made in order to cover the whole range in each one of the components of $\mathbf{p}$ in one hundred equal increments[1].

From the sparse representation point of view, it is highly desirable not to have dictionary elements which are "too similar". The Mutual Coherence (MC) of a dictionary A, denoted by $\mu(A)$, measures the similarity between dictionary elements. It is defined as the maximal absolute scalar product between two different ℓ_2-normalized elements of the dictionary A [1]. In the dictionary generation algorithm we discarded an element if its MC was greater than some predefined number κ. To avoid classification bias, the dictionary sizes were required to be the same for both classes. With this objective in mind, we randomly eliminated the necessary number of elements from the larger dictionary whose MCs were greater than another predefined value (e.g. $\kappa - 0.5$).

2.3 Classification Based on Sparse Representation

Let us define a new matrix A as the concatenation of the n elements from both target and non-target dictionaries, A_1 and A_2, respectively, that is:

$$A \doteq [A_1 \; A_2] = [\mathbf{a}_{1,1}, \mathbf{a}_{1,2}, ..., \mathbf{a}_{1,n}, \mathbf{a}_{2,1}, ..., \mathbf{a}_{2,n}]. \tag{5}$$

For given ϵ, A and $\mathbf{y}$, let $\hat{\mathbf{x}}$ be the solution of the problem $(P_{1,2})$ in (1). For each class i, let $\delta_i : \mathbb{R}^n \to \mathbb{R}^n$ be the lifting function that selects the coefficients associated with the i^{th} class, that for $\mathbf{x} \in \mathbb{R}^n$, $\delta_i(\mathbf{x}) \in \mathbb{R}^n$ is a vector whose only nonzero entries are the elements in $\mathbf{x}$ that are associated to class i. The representation of a given test sample $\mathbf{y}$ in class i is then $\hat{\mathbf{y}}_i = A\delta_i(\hat{\mathbf{x}})$ [6]. The classification of $\mathbf{y}$ proceeds by assigning it to the class that minimizes the residual, i.e., we associate to $\mathbf{y}$ the class given by:

$$\underset{i=1,2}{\operatorname{argmin}} \; r_i(\mathbf{y}) \doteq ||\mathbf{y} - A\delta_i(\hat{\mathbf{x}})||_2. \tag{6}$$

The unweighted accuracy rate (UAR) was used as the performance classification measure. For a binary classification problem, this index is defined as: $UAR = \frac{1}{2}\left(\frac{TP}{P} + \frac{TN}{N}\right)$, where TP and TN are the number of *true positive* and *true negative*, while P and N indicate the amount of samples in each class, respectively.

3 Results

The solution of problem (P_λ) was obtained by the application of the SLEP 4.1 toolbox [13]. More precisely, we solved the ℓ_1-norm regularized least squares problem (LeastR), i.e. (2), and the non-negative ℓ_1-norm regularized least squares problem (NNLeastR), where the latter imposes the additional non-negative constrain $\mathbf{x} \geq 0$ in (2).

The training set was used to fix the "optimal" value of λ. Per each observation $\mathbf{y}$ we varied λ from 0 to 10^{-6} in a log-scale and analyzed the residue for λ in this range. Minimizing the residual and the sparsity are both desired properties, thus we fixed the range of nonzero entries in $\mathbf{x}$ between 1 and 10 and defined a new λ range in agreement with this nonzero entries range. We finally chose the λ value that produced the best classification result for the training set and used it for the posterior classification with the testing set.

In order to analyze the impact of the choice of the MC value on the generation of the dictionary and, consequently, on the classification performance, we varied the MC values from 0.90 to 0.95 in increments of 0.005.

Table 1 summarizes the best UAR results for each subject with the different optimization problems. The corresponding MC value is shown between parenthesis. An analysis of Table 1 seems to suggest that better

[1] The variation in p_1, corresponding to the offset parameter in (3) was negligible and therefore it was kept constant for the dictionary generation.

classification result are obtained by imposing $\mathbf{x} \geq 0$. Although further analysis is required, we strongly believe that this is due to the fact that the elements in the target dictionary present mainly positive peaks.

Table 1: Best UAR results for the different optimization problems.

Subject N^{o}	**LeastR**	**NNLeastR**
1	0.567 (0.985)	0.627 (0.935)
2	0.550 (0.965)	0.576 (0.905)
3	0.495 (0.960)	0.499 (0.955)

4 Conclusions

In the present work we explored the sparse properties for classification purposes of ERP signals. Although the classification performances are far for being optimal, we find important to point out that the classification method used is very simple and moreover noisy single trail epochs were used. We are currently devoting efforts in this direction. In particular it is of great interest to find alternative and/or complementary ways to the ℓ_1-minimization approach, which could allow classification improvement.

We constructed an ad-hoc Gaussian dictionary for representing the P300 wave in a channel-time space. We focused our work in constructing a subject-dependent dictionary in order to represent the variation of the P300 wave of that particular subject. We have good reasons to believe that better classification result could be obtained by improving the representation of non-target signals.

Acknowledgments

This work was supported in part by Consejo Nacional de Investigaciones Científicas y Técnicas, CONICET, through PIP 2014-2016 Nro. 11220130100216-CO, the Air Force Office of Scientific Research, AFOSR/SOARD, through Grant FA9550-14-1-0130 and by CAID-UNL 2011.

References

[1] M. Elad. *Sparse and Redundant Representation, from theory to applications in signal and image processing.* Springer, 2010, ISBN 978-1-4419-7010-7.

[2] J.M, Bruckstein, D. Donoho, and M. Elad. *From Sparse Solutions of Systems of Equations to Sparse Modeling of Signals and Images.* Society for Industrial and Applied Mathematics, Vol. 51 (2009), pp.34-81.

[3] D. Donoho. *For Most Large Underdetermined Systems of Equations the Minimal l1-norm Near-Solution Approximates the Sparsest Near-Solution.* Communications on Pure and Applied Mathematics, Vol. 59 (2006), pp.907-934.

[4] H.W Engl, M. Hanke and A. Neubauer. *Regularization of inverse problems.* Kluwer Academic Publishers, 2010, ISBN 0-7923-4157-0.

[5] R. Tibshirani. *Regression, Shrinkage and Selection via the lasso.* Journal of the Royal Statistical Society. Series B (Methodological), Vol. 58 (1996), pp.267-288.

[6] J. Wright, A.Y Yang, A. Ganesh, S.S Sastry and Y. Ma. *Robust Face Recognition via Sparse Representation.* IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 31 (1996), pp.1-18.

[7] J.R Wolpaw, N. Birbaumer, D.J McFarland, G. Pfurtscheller and T.M. Vaugha. *Brain computer interfaces for communication and control.* Clinical Neurophysiology, Vol. 113 (2002), pp.767-791.

[8] S.A Hiyard and M. Kutas. *Electrophysiology of congnitive processing.* Annual Reviews Phychol, Vol. 34 (1983), pp.33-61.

[9] L.A Farwell and E. Donchin, Talking off the top of your head: toward a metal prosthesis utilizing event-related brain potentials *E*lectroencephalography and clinical neurophysiology. Vol.70 (1988), pp.510-523.

[10] C. Ledesma-Ramirez, E. Bojorges-Valdez, o. Yañez-Suarez, C. Saavedra, L. Boygrain and G. Gentiletti. *An Open-Access P300 Speller Database.* Fourth international BCI meeting, Monterrey, USA, California, 2010.

[11] K. Levenberg. *A method for the solution of certain problems in least squares.* Quart. Appl. Math., Vol. 2 (1944), pp.164-168.

[12] D. Marquardt. *An algorithm for least-squares estimation of nonlinear parameters.* SIAM J. Appl. Math., Vol. 11 (1963), pp.431-441.

[13] J. Liu, S. Ji and J. Ye. *SLEP: Sparse Learning with Efficient Projections.* Arizona State University, 2009. http://www.public.asu.edu/ jye02/Software/SLEP.