32nd Annual International Conference of the IEEE EMBS
Buenos Aires, Argentina, August 31 - September 4, 2010

Nonlinear Feature Extraction Using Kernel Principal Component Analysis With Non-negative Pre-image

Maya Kallas, Paul Honeine, Cédric Richard, Hassan Amoud and Clovis Francis

***Abstract*— The inherent physical characteristics of many real-life phenomena, including biological and physiological aspects, require adapted nonlinear tools. Moreover, the additive nature in some situations involve solutions expressed as positive combinations of data. In this paper, we propose a nonlinear feature extraction method, with a non-negativity constraint. To this end, the kernel principal component analysis is considered to define the most relevant features in the reproducing kernel Hilbert space. These features are the nonlinear principal components with high-order correlations between input variables. A pre-image technique is required to get back to the input space. With a non-negative constraint, we show that one can solve the pre-image problem efficiently, using a simple iterative scheme. Furthermore, the constrained solution contributes to the stability of the algorithm. Experimental results on event-related potentials (ERP) illustrate the efficiency of the proposed method.**

***Index Terms*— Kernel-PCA, pre-image problem, non-negativity constraint, additive weight algorithm**

I. INTRODUCTION

There has been an ever-increasing interest of engineers and scientists in nonlinear feature extraction since, unfortunately, most natural systems exhibit nonlinear behavior. Furthermore, with some prior information on the system under investigation, a constrained solution is often required in many situations, in order to illustrate some physical characteristics such as the non-negativity.

Consider for instance an electroencephalographic (EEG) recording, which corresponds to a summation of individual contributions in the brain. A measure of the brain activity should always be positive, since the brain is always in activity. In practice, the recordings are zero-meaned by comparing them to some reference, resulting into positive and negative components. Nevertheless, to understand the underlying structure of theses recordings, and thus the brain activity, one should *keep in mind* the non-negative additivity of contributions. Non-negativity is a desirable property in many research areas. Independent component analysis impose a non-negative factorization of the data [1], i.e. for blind source separation with positive sources. In [2], a non-negative principal component analysis (PCA) is proposed. A more general approach is studied in [3] for signal and image restoration with a non-negativity constraint.

Kernel-based methods provide a breakthrough in both statistical learning theory and low computational cost for nonlinear algorithms. The main idea behind these algorithms is the *kernel trick* [4]. It gives a mean to transform conventional linear algorithms into nonlinear ones, under the only condition of expressing the algorithm in terms of pairwise inner products between data. By substituting the inner product operator with a (positive semi-definite) kernel function, this is equivalent to mapping the data from the input space into a feature space via some nonlinear map, and then apply the linear algorithm in the feature space. The resulting feature space is the so-called reproducing kernel Hilbert space (RKHS). For instance, in [5] the authors reported the superiority of nonlinear kernel-PCA over conventional linear PCA, combined with a discrimination scheme in order to classify event-related potentials (ERP).

While the mapping from input space to feature space is of primary importance in kernel methods, the reverse mapping from feature space back to input space is often very useful, as studied in this paper. Unfortunately, getting back from the RKHS to the input space is not obvious in general, as most features of the former may not have an exact pre-image in the latter. This is the pre-image problem, as one seeks an approximate solution. Furthermore, this is also non-trivial as the dimensionality of the feature space can even be infinite. In [6], Mika *et al.* studied this highly nonlinear optimization problem, and proposed a fixed-point iterative method. In [7], a technique based on multidimensional-scaling is considered, and recently a more adapted method is derived in [8]. While these techniques are applied in a de-noising scheme, we propose in this paper a feature extraction approach, incorporating a non-negativity constraint. The resulting algorithm is based on an iterative gradient descent scheme. The proposed method is general, and can be applied on any data, as long as kernel-PCA can be applied. In this paper, we illustrate its performance on a ERP problem in order to extract nonlinear features from EEG.

The paper is organized as follows: In Section II, we review the kernel-PCA technique. The problem of nonlinear feature extraction is presented in Section III, in the light of the pre-image problem. The non-negativity constraint is studied in Section IV, while in Section V experimental results are given.

Manuscript received April 26, 2010.

M. Kallas and P. Honeine are with Institut Charles Delaunay (UMR CNRS 6279), LM2S, Université de technologie de Troyes, France maya.kallas@utt.fr, paul.honeine@utt.fr

M. Kallas, H. Amoud and C. Francis are with the Université Libanaise, Lebanon hassan.amoud@gmail.com, cfrancis@ul.edu.lb

C. Richard is with Laboratoire Fizeau (UMR CNRS 6525), Observatoire de la Côte d'Azur, Université de Nice Sophia-Antipolis, France cedric.richard@unice.fr

II. Kernel-PCA

Principal Component Analysis (PCA) is a widely used technique for representing data, by extracting a small number of features from the data itself. This approach is regarded as a global approach, as opposed to methods such as parametric models and wavelet decomposition, where extracted features highly depend on the model or wavelet type under investigation. In PCA, features are obtained by diagonalizing the correlation matrix of the data, conserving only the most relevant eigenvectors. Without loss of generality, we assume zero-mean data, given column-wise in $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n \in \mathbb{R}^d$. PCA technique seeks the m features $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_m \in \mathbb{R}^d$, as the eigenvectors in the eigen-problem $\lambda \boldsymbol{v} = \boldsymbol{C}\,\boldsymbol{v}$, with $\boldsymbol{C} = \frac{1}{n}\sum_{j=1}^n \boldsymbol{x}_j \boldsymbol{x}_j^\top$ the correlation matrix. The relevance of each eigenvector $\boldsymbol{v}$ is given by its corresponding eigenvalue λ, which measures the amount of captured variance of the data. From the linearity property of the operations, the eigenvectors lie in the span of the data, taking the form $\boldsymbol{v} = \sum_{i=1}^n \alpha_i\, \boldsymbol{x}_i$.

Unlike conventional PCA which is restricted to learn only linear structures within data, kernel-PCA is a popular generalization to discover nonlinearities. To recognize nonlinear features, a common strategy consists in mapping the data into some feature space, with $\Phi\colon \mathbb{R}^d \mapsto \mathcal{H}$, and then compute PCA on mapped data, $\Phi(\boldsymbol{x}_1), \Phi(\boldsymbol{x}_2), \ldots, \Phi(\boldsymbol{x}_n) \in \mathcal{H}$. While eigenvectors are linear in the transformed data, they are nonlinear in the original data. Without the need to evaluate explicitly the map, it turns out that one can efficiently compute such nonlinear PCA, for a broad class of nonlinearities, using the concept of the *kernel trick*. It corresponds to writing the algorithm using only pairwise inner products between data, thus substituting these proximity measurements with nonlinear ones, defined by a kernel function. This widespread principle is illustrated here on kernel-PCA [9].

First, we write PCA algorithm in terms of inner products in the feature space, $\langle \Phi(\boldsymbol{x}_i), \Phi(\boldsymbol{x}_j)\rangle_{\mathcal{H}}$, for $i, j = 1, 2, \ldots, n$. Each extracted feature $\varphi \in \mathcal{H}$ satisfies the expression

$$\lambda\,\varphi = \boldsymbol{C}^\Phi\,\varphi, \tag{1}$$

where $\boldsymbol{C}^\Phi$ represents the correlation between mapped data, expressed in a finite-dimensional space as $\boldsymbol{C}^\Phi = \frac{1}{n}\sum_{j=1}^n \Phi(\boldsymbol{x}_j)\,\Phi(\boldsymbol{x}_j)^\top$. By analogy with the linear case, all solutions φ lie in the span of the Φ-images of the data. This means that there exists coefficients $\alpha_1, \alpha_2, \ldots, \alpha_n$ such that

$$\varphi = \sum_{i=1}^n \alpha_i\, \Phi(\boldsymbol{x}_i). \tag{2}$$

Substituting $\boldsymbol{C}^\Phi$ and the expansion (2) into the eigen-problem (1), and defining a $n \times n$ matrix $\boldsymbol{K}$ whose (i,j)-th entry is $\langle \Phi(\boldsymbol{x}_i), \Phi(\boldsymbol{x}_j)\rangle_{\mathcal{H}}$, we get the eigen-problem in terms of inner product matrix

$$n\,\lambda\,\boldsymbol{\alpha} = \boldsymbol{K}\,\boldsymbol{\alpha}, \tag{3}$$

where $\boldsymbol{\alpha} = [\alpha_1\ \alpha_2\ \cdots\ \alpha_n]^\top$. In order to get the normalization as in PCA[1], i.e. $\langle \varphi, \varphi\rangle_{\mathcal{H}} = 1$, one operates a normalization on the resulting solution $\boldsymbol{\alpha}$, with $\|\boldsymbol{\alpha}\|^2 = 1/\lambda$.

Substituting the inner product operator with a kernel function, $\kappa\colon \mathbb{R}^d \times \mathbb{R}^d \mapsto \mathbb{R}$, provides a nonlinear extension to PCA, the so-called kernel-PCA. Kernels with a positive semi-definite property correspond to an implicit mapping, and thus can be written as $\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \langle \Phi(x_i), \Phi(x_j)\rangle_{\mathcal{H}}$, in a feature space $\mathcal{H}$, the so-called RKHS. Examples of admissible kernels include the polynomial kernel $\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = (1 + \langle \boldsymbol{x}_i, \boldsymbol{x}_j\rangle)^p$, and the Gaussian kernel $\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp(\frac{1}{\sigma^2}\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2)$, the latter implicitly maps data into an infinite-dimensional space.

III. Feature extraction as a pre-image problem

As illustrated above, it is easy to compute the coefficients in (2), thanks to the kernel trick. When a supervised learning is required, the resulting features are only used in a pre-processing scheme, for dimensionality reduction purpose, before applying a discrimination machine such as Support Vector Machines. In such cases, the features need not to be explicated since, for any given $\boldsymbol{x}$, the projection of $\Phi(\boldsymbol{x})$ onto any $\varphi \in \mathcal{H}$ can be given by $\langle \varphi, \Phi(\boldsymbol{x})\rangle_{\mathcal{H}} = \sum_{i=1}^n \alpha_i\, \kappa(\boldsymbol{x}_i, \boldsymbol{x})$. When an unsupervised learning is desired, such as in pattern recognition, it is not sufficient to know the weighting coefficients. One is often interested in the feature itself, as defined in (2), or more precisely in its counterpart in the input space, i.e. a $\boldsymbol{x}^*$ such that its map is equivalent to $\varphi = \sum_{i=1}^n \alpha_i\, \Phi(\boldsymbol{x}_i)$. However, very few elements of a RKHS satisfy this property. In general, one seeks an approximate solution, i.e. $\boldsymbol{x}^*$ in $\mathbb{R}^d$ whose map $\Phi(\boldsymbol{x}^*)$ is as close as possible to φ.

This is the pre-image problem in machine learning, where one seeks to map back elements from the RKHS to the input space. This optimization problem was originally studied by Mika *et al.* in [6]. It consists of minimizing the distance in the RKHS between both elements, with

$$\boldsymbol{x}^* = \arg\min_{\boldsymbol{x}\in\mathbb{R}^d} \|\varphi - \Phi(\boldsymbol{x})\|^2_{\mathcal{H}}, \tag{4}$$

where $\|\cdot\|_{\mathcal{H}}$ denotes the norm in the RKHS. Worth noting that this is a non-convex and highly nonlinear optimization problem. In [6], the authors propose a fixed-point iterative method to solve this problem. Unfortunately, this technique tends to be unstable and suffers from local minima. In [7], a technique based on the multidimensional-scaling is proposed, while in [10] the authors illustrate the connection of this problem with other dimensionality reduction methods. More recently, two of the authors propose a more adapted method to solve the pre-image problem [8], [11]. Interestingly, all these methods suggest that the resulting pre-image lies in the span of the original data, namely

$$\boldsymbol{x}^* = \sum_{i=1}^n \gamma_i\, \boldsymbol{x}_i. \tag{5}$$

[1] Furthermore, data should be centered in the feature space, a task efficiently operated by replacing the matrix $\boldsymbol{K}$ in (3) with the modified matrix $(\boldsymbol{1} - \boldsymbol{1}_n)\boldsymbol{K}(\boldsymbol{1} - \boldsymbol{1}_n)$, with $\boldsymbol{1}_n$ the n-by-n matrix of entries $1/n$ and $\boldsymbol{1}$ the identity matrix.

While this is a linear system, it is computed on the basis of closeness to the nonlinear feature, where distance is computed in the feature space.

All these techniques have been proposed for de-noising purpose, i.e. any new data is mapped, projected into the most relevant subspace, and then mapped back to the input space. To our knowledge, mapping the features back to the input space was not considered in the literature, yet feature extraction is as (if not more) important as de-noising data. Moreover, in many physical phenomena, one may require a constrained solution (see for instance [12] for an application regarding positive temperatures). Next, we show that one may provide an easy, yet efficient, scheme to solve this optimization problem with a non-negativity constraint, i.e. $\gamma_1, \gamma_2, \ldots, \gamma_n \geq 0$ in the expansion (5). Worth noting that including a constraint contributes to the stability of the solution, and provides often sparsity [13].

IV. The pre-image with non-negativity constraint

We begin by injecting the expansion in (5) into the optimization problem (4), reducing the problem into finding the coefficients vector $\boldsymbol{\gamma}^* = [\gamma_1^* \; \gamma_2^* \; \cdots \; \gamma_n^*]^\top$. Let $J(\boldsymbol{\gamma}^*)$ be the resulting cost function. Next, we consider solving the pre-image problem, independently of the used kernel, by writing the pre-image problem in the general form [3]

$$\boldsymbol{\gamma}^* = \arg\min_{\boldsymbol{\gamma}} J(\boldsymbol{\gamma})$$
$$\text{subject to } \boldsymbol{\gamma} \geq 0$$

with ≥ 0 denotes element-wise non-negativity. The corresponding Lagrangian can be described as $J(\boldsymbol{\gamma}) - \boldsymbol{\mu}^\top \boldsymbol{\gamma}$, where $\boldsymbol{\mu}$ is the vector of the non-negative Lagrange multipliers. The Kuhn-Tucker conditions must be satisfied at the optimum, with the expressions

$$\nabla_{\boldsymbol{\gamma}} \left[J(\boldsymbol{\gamma}^*) - \boldsymbol{\mu}^{*\top} \boldsymbol{\gamma}^* \right] = 0$$
$$\mu_i^* \, \gamma_i^* = 0 \quad \forall i$$

where γ_i^* (resp. μ_i^*) is the i-th component de $\boldsymbol{\gamma}^*$ (resp. $\boldsymbol{\mu}^*$). Thus the resulting problem to be solved is

$$\gamma_i^* [-\nabla_{\boldsymbol{\gamma}} J(\boldsymbol{\gamma})]_i = 0$$

with $[\nabla_{\boldsymbol{\gamma}} J(\boldsymbol{\gamma})]_i$ is the i-th component of $\nabla_{\boldsymbol{\gamma}} J(\boldsymbol{\gamma})$ and the minus sign is used to explicitly describe the gradient descent of $J(\boldsymbol{\gamma})$.

To solve this problem iteratively, we consider the fixed-point approach, leading to the element-wise gradient descent algorithm [13], [14]

$$\gamma_i(k+1) = \gamma_i(k) + \eta_i(k) f_i(\boldsymbol{\gamma}(k)) \gamma_i(k) [-\nabla_\gamma J(\boldsymbol{\gamma})]_i$$

where $\eta_i(k)$ is a step size factor used to control convergence, and $f_i(\boldsymbol{\gamma}(k))$ is a function having positive values. To guarantee the non-negativity of $\gamma_i(k+1)$, updated from the previously estimated one, $\gamma_i(k)$, the following condition should be satisfied: if $[\nabla_\gamma J(\boldsymbol{\gamma})]_i > 0$,

$$\eta_i(k) \leq \frac{1}{f_i(\boldsymbol{\gamma}(k)) [\nabla_\gamma J(\boldsymbol{\gamma})]_i};$$

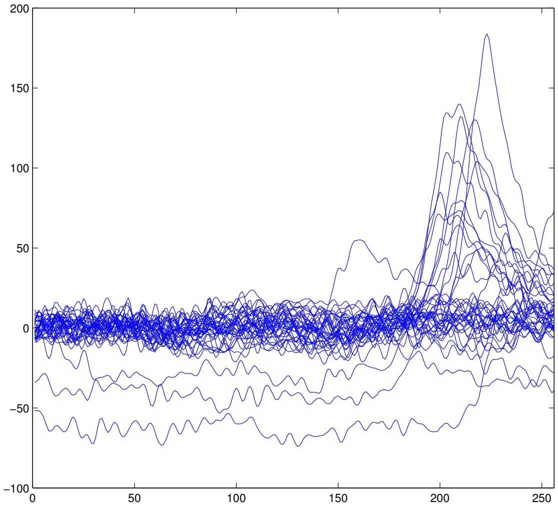

Fig. 1. The 40 trials from electrode FP1.

Otherwise, when $[\nabla_\gamma J(\boldsymbol{\gamma})]_i \leq 0$, no restriction related to the positivity is imposed on this step size factor. Finally, we deduct that the general expression of the algorithm, in matrix form, is

$$\boldsymbol{\gamma}(k+1) = \boldsymbol{\gamma}(k) + \eta(k) \, \boldsymbol{d}(k),$$

where $\boldsymbol{d}(k)$ defines the direction of descent, with

$$\boldsymbol{d}(k) = -\text{diag}[f_i(\boldsymbol{\gamma}(k)) \gamma_i(k)] \, \nabla_\gamma J(\boldsymbol{\gamma}).$$

The optimal step size $\eta(k)$ can be computed eventually from a linear search algorithm in the interval $]0, \eta_{max}(k)]$ with

$$\eta_{max}(k) = \min_i \frac{1}{f_i(\boldsymbol{\gamma}(k)) [\nabla_\gamma J(\boldsymbol{\gamma})]_i}.$$

V. Experiments

The proposed method for solving the pre-image problem with non-negative constraint, provides a general technique for feature extraction, and can be applied in any feature extraction problem. In this section, we illustrate it on a set of event-related potentials (ERP) from EEG recordings. The experimental signals are taken from a large study on selected sets of people, with a genetic predisposition to alcoholism [15]. The acquisition system is composed of 64 electrodes, positioned on the scalps, taking the measurements sampled at 256 Hz, for 1 second. There were 122 subjects, each one has completed 120 trials where different visual stimuli were shown to them: the subject was exposed either to one stimulus (S1), or two stimuli (S1 and S2). We have considered only one subject with ERP resulting from one stimulus, and chosen one electrode, FP1[2]. The number of trials is 40, resulting into 40 signals of 256 samples each, illustrated in Fig. 1.

[2] The considered EEG signals can be downloaded from http://archive.ics.uci.edu/ml/databases/eeg/eeg.data.html.

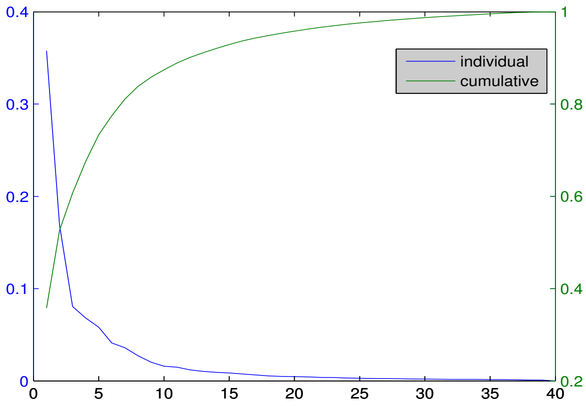

Fig. 2. The individual and the cumulative captured variance of the 40 available features. The first 4 features capture 67% of overall variance

To perform the non-negative coefficients pre-image, the kernel applied on the signals was the Gaussian kernel, with bandwidth set to $\sigma = 300$. The kernel-PCA algorithm was applied using this kernel, with the overall captured variance illustrated in Fig. 2 with individual $\lambda_k / \sum_{i=1}^{40} \lambda_i$ (left axis) and cumulative $\sum_{j=1}^{k} \lambda_j / \sum_{i=1}^{40} \lambda_i$ (right axis) eigenvalues for each of the k eigenvectors. In the following, we consider the four most relevant features, capturing 67 % of the data variance, and consider the proposed method to get back from the (infinite dimensional) RKHS to the input space of 256-sample signals. The additive weight update algorithm is applied to pre-image the four features. For this purpose, the following parameters were considered: the step size factor was set to $\eta = 0.9$, and the number of iterations to 200. The resulting pre-imaged features are given in Fig. 3, and compared to an arbitrary less-relevant feature, the 21st extracted feature, which exhibits *less structure* within data. We can easily verify that all the coefficients are nonnegative.

VI. CONCLUSIONS

Real-life phenomena, such as some biological characteristics, impose constraints on the extracted features. In this paper, we have shown that nonlinear features can be extracted by jointly applying a kernel-PCA algorithm and a pre-image technique. The pre-image problem is solved under the non-negative constraint, using an additive fixed-point iterative algorithm. The utility of the method was demonstrated on real EEG data.

REFERENCES

[1] A. Hyvärinen and E. Oja, “Independent component analysis: algorithms and applications,” *Neural Netw.*, vol. 13, no. 4-5, pp. 411–430, 2000.

[2] E. Oja and M. Plumbley, “Blind separation of positive sources using non-negative PCA,” in *In 4th International Symposium on Independent Component Analysis and Blind Signal Separation*, 2003, pp. 11–16.

[3] H. Lantéri, M. Roche, O. Cuevas, and C. Aime, “A general method to devise maximum-likelihood signal restoration multiplicative algorithms with non-negativity constraints,” *Signal Processing*, vol. 81, no. 5, pp. 945–974, May 2001.

[4] M. A. Aizerman, E. A. Braverman, and L. Rozonoer, “Theoretical foundations of the potential function method in pattern recognition learning.” in *Automation and Remote Control,*, no. 25, 1964, pp. 821–837.

[5] R. Rosipal, M. Girolami, and L. Trejo, “Kernel PCA feature extraction of event-related potentials for human signal detection task,” *Artificial Neural Networks in Medicine and Biology*, pp. 321–326, May 2000.

[6] S. Mika, B. Schölkopf, A. Smola, K.-R. Müller, M. Scholz, and G. Rätsch, “Kernel PCA and de-noising in feature spaces,” in *Proceedings of the 1998 conference on Advances in neural information processing systems II*. Cambridge, MA, USA: MIT Press, 1999, pp. 536–542.

[7] J. T. Kwok and I. W. Tsang, “The pre-image problem in kernel methods,” in *ICML*, T. Fawcett and N. Mishra, Eds. AAAI Press, 2003, pp. 408–415.

[8] P. Honeine and C. Richard, “Solving the pre-image problem in kernel machines: a direct method,” in *Proc. IEEE Workshop on Machine Learning for Signal Processing (MLSP)*, Grenoble, France, September 2009.

[9] B. Schölkopf, A. Smola, and K.-R. Müller, “Nonlinear component analysis as a kernel eigenvalue problem,” *Neural Comput.*, vol. 10, no. 5, pp. 1299–1319, 1998.

[10] P. Arias, G. Randall, and G. Sapiro, “Connecting the out-of-sample and pre-image problems in kernel methods,” *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, vol. 0, pp. 1–8, 2007.

[11] P. Honeine and C. Richard, “A closed-form solution for the pre-image problem in kernel-based machines,” *Journal of Signal Processing Systems*, in press 2010.

[12] C. Richard, P. Honeine, H. Snoussi, A. Ferrari, and C. Theys, “Distributed learning with kernels in wireless sensor networks for physical phenomena modeling and tracking,” in *Proc. 30th IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, Honolulu (Hawaii), USA, 25 - 30 July 2010.

[13] J. Chen, C. Richard, P. Honeine, H. Lanteri, and C. Theys, “System identification under non-negativity constraints,” in *Proc. 18th European Conference on Signal Processing (EUSIPCO)*, Aalborg, Denmark, 23 - 27 Aug. 2010.

[14] J. Chen, C. Richard, P. Honeine, H. Snoussi, H. Lantéri, and C. Theys, “Techniques d’apprentissage non-linéaires en ligne avec contraintes de positivite,” in *Actes de la VI^{ème} Conférence Internationale Francophone d’Automatique*, Nancy, France, 2 - 4 Juin 2010.

[15] L. Ingber, “Statistical mechanics of neocortical interactions: Training and testing canonical momenta indicators of eeg,” *Mathematical Computer Modelling*, vol. 27, no. 3, pp. 33–64, 1998.

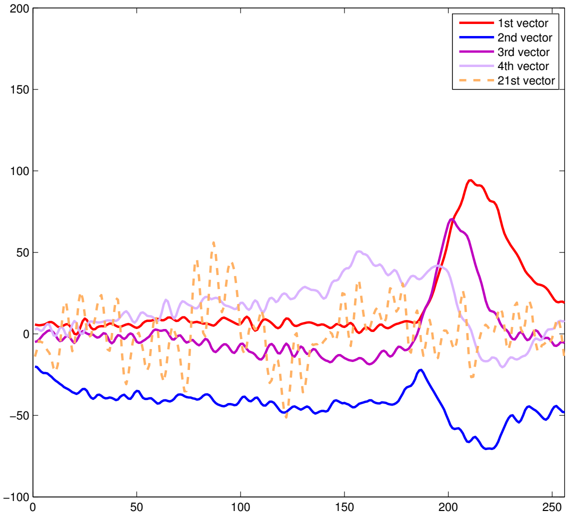

Fig. 3. The four most relevant features as well as a less relevant one.