Proceedings of the 28th IEEE
EMBS Annual International Conference
New York City, USA, Aug 30-Sept 3, 2006

Analysis of Large-Scale Brain Data for Brain-Computer Interfaces

Koel Das, Joerg Meyer, and Zoran Nenadic

Abstract— **We present a systematic technique for extraction of useful information from large-scale neural data in the context of brain-computer interfaces. The technique is based on a direct linear discriminant analysis, recently developed for face recognition problems. We show that this technique is capable of extracting useful information from brain data in a systematic fashion and can serve as a general analytical tool for other types of biomedical data, such as images and collections of images (movies). The performance of the method is tested on intracranial electroencephalographic data recorded from the human brain.**

I. INTRODUCTION

A. Brain-Computer Interfaces

A brain-computer interface (BCI) is a set of communication and control devices and algorithms that allow the interaction of a patient with its environment without generating any motor output [1], [2]. While the major application of BCIs is to assist disabled individuals by using neural activity from the brain to control prosthetic devices (computers, robots, autonomous vehicles, etc.), BCIs play an increasingly important role as a tool for studying brain mechanisms and testing new hypotheses about brain function.

Depending on the type of neural signals and underlying recording technology, BCIs can be classified as invasive and non-invasive. Invasive BCIs use arrays of electrodes implanted in the brain to record action potentials of single cortical neurons [3], [4], [5], or local field potentials (i.e. the composite extracellular potential from hundreds or thousands of neurons) [6], [7]. Non-invasive BCI techniques rely on electroencephalographic (EEG) signals, recorded from the surface of the scalp [2], [8].

The ultimate goal of any BCI is to decode in real time information contained in neural signals, and to use this information for the control of assistive devices, such as computers or robots. Typically, the decoding is facilitated by recording neural correlates of movement intentions or hand trajectories, and by accumulating sufficient data over multiple trials into a *training database*. Future data can then be compared against the training database, and the intended movements or trajectories are decoded.

B. Feature Extraction

Brain signals arising in BCI applications are typically spatio-temporal and large-dimensional. While abundance of neural data is ultimately important for decoding, not all data samples carry useful information for the task at hand. Ideally, relevant data samples should be combined into meaningful features, while irrelevant data should be discarded as noise. This dimensionality reduction procedure is commonly referred to as *feature extraction*. An obvious benefit of feature extraction is that data becomes computationally more manageable. More importantly, the statistical properties of data can be more accurately estimated in the low-dimensional feature space. This is especially important under the *small sample size* conditions [9], common to many BCI applications, where the dimension of data, p, exceeds the number of examples, n_t, in the training database.

Current approaches to feature extraction from neural data are largely heuristic and rely heavily on off-the-shelf signal processing tools. For example, the use of spectral power of EEG signals in various frequency bands, e.g. μ-band or β-band [10], [11], can be viewed as a heuristic way of extracting features from data. While spectral features have clear physical interpretation, there is no evidence that they are optimal features for decoding. A common approach to processing of spatio-temporal neural signals is a two-step procedure: data is first processed spatially, typically by applying off-the-shelf tools such as the Laplacian filter [8], [10], followed by temporal processing such as autoregressive frequency analysis [8], [11]. However, the assumption of space-time separability is not justified and may be responsible for suboptimal performance. In addition, many of the heuristic strategies attempt to rank individual (scalar) features according to some criterion, and then construct the feature vector by a concatenation of the several most significant features. The major shortcoming of this approach is that the joint statistical properties of the features are not accounted for, which may produce suboptimal feature vectors.

In this article, we present a feature extraction technique suitable for BCI applications under the small sample size conditions. The method is based on a variant of linear discriminant analysis (LDA), called direct LDA [12]. The technique is computationally efficient (linear), and extracts features by utilizing their joint statistical properties. In addition, when applied to spatio-temporal signals, the technique does not need the space-time separability assumption.

II. DIRECT LDA APPROACH

A. Background

LDA is perhaps the most widely used feature extraction technique. The objective of LDA is to perform dimensionality reduction while enhancing class separability, normally by

K. Das is with the Department of Electrical Engineering and Computer Science,University of California,Irvine,CA 92697,USA kdas@uci.edu
J. Meyer is with the Department of Biomedical Engineering and the Department of Electrical Engineering and Computer Science,University of California,Irvine,CA 92697,USA jmeyer@uci.edu
Z. Nenadic is with the Department of Biomedical Engineering,University of California,Irvine,CA 92697,USA znenadic@uci.edu

1-4244-0033-3/06/$20.00 ©2006 IEEE.

maximizing an objective function. The most popular form of LDA relies on the maximization of the Fisher criterion [13]:

$$J(T) = \arg \max_{T \in \mathbb{R}^{m \times p}} \frac{|T S_b T^T|}{|T S_w T^T|} \quad (1)$$

where S_b and S_w represent between-class and within-class scatter matrix [14]. However, for large-scale data, the traditional LDA approach faces a couple of challenges. Firstly, large-scale data is associated with large covariance matrices which are difficult to store and manipulate. Secondly, large-scale data inevitably leads to the small sample size conditions ($p > n_t$) and singular scatter matrix S_w, and so the classical solution to (1) based on the eigenvalue/eigenvector decomposition of $S_w^{-1} S_b$ [14], is not directly applicable.

There have been several research efforts to solve the small sample size problem using variants of LDA (see [15] for review). Recently, Yu and Yang [12] developed a novel direct LDA (DLDA) algorithm for high-dimensional data. The algorithm discards the null space of S_b by means of eigenanalysis, followed by a simultaneous diagonalization of S_w and S_b in the remaining subspace. Since $\text{rank}(S_b) \leq c-1$, where c is the number of classes, it follows that the dimension of the feature space, m, is at most $c-1$. Formally, the algorithm finds a matrix $A \in \mathbb{R}^{q \times p}$ ($q \ll p$) such that

$$A S_b A^T = I, \quad A S_w A^T = \Lambda_w, \quad (2)$$

where $I \in \mathbb{R}^{q \times q}$ is an identity matrix and $\Lambda_w \in \mathbb{R}^{q \times q}$ is a positive (semi)definite diagonal matrix with elements sorted in ascending order. The feature extraction matrix T is chosen as the first m ($1 \leq m \leq q \leq c-1$) rows of A. Yu and Yang also show how to efficiently handle the calculation of A for large-scale data, and the details can be found in [12]. If analyzed data represent spatio-temporal signals (written in a vector form), each row of the feature extraction matrix T can be viewed as a *spatio-temporal filter* that maps data into one component of the feature vector. Note that in this case no space-time separability assumption is necessary. Also note that when $m \geq 2$, the features are extracted simultaneously, thereby utilizing their joint statistical properties.

B. Modifications of the Direct LDA Approach

While Yu and Yang report very good performance of DLDA to face recognition problems, the direct application of their technique to brain data for BCI (see Section III), generally yields poor results. The discrepancy in performances could be explained by relatively low noise levels in face recognition problems, compared to very noisy brain signals. In general, under the small sample size conditions the empirical covariance matrix is not a good estimate of the true covariance matrix, as shown in [16]. Inaccurate estimates of large scale covariance matrices typically lead to erroneous eigenvalue calculations used in LDA or any variants thereof. One way to reduce the noise is to improve the estimates of the covariance matrix by use of shrinkage [16]. In this article we apply a thresholding approach to get rid of the noise in the data. Thresholding is used extensively as a signal denoising technique, especially in conjunction with wavelet basis functions [17]. The near-zero coefficients of the spatio-temporal filter T, obtained from DLDA are likely to be fluctuations due to noise, and conversely, large filter coefficients are likely to correspond to data attributes that are informative. Consequently, the near-zero coefficients of T are shrunk to 0, while the large coefficients are preserved. This type of thresholding is often referred to as the *hard thresholding* rule [17].

For large-scale datasets suffering from "small n_t, large p" problem, where estimated class-covariance matrices are highly singular, we propose a novel two-pass algorithm which systematically reduces the singularity of the estimated covariance matrices by the use of hard thresholding. In the first pass, we run DLDA [12] on the training dataset and get the spatio-temporal filter F. The training filter F gives a good initial guess as to which attributes are prominent in the p-dimensional data space. In the second pass, we use hard thresholding on the coefficients of F to get the feature extraction matrix T which extracts the spatio-temporal samples of relevance, effectively selecting a p'-dimensional data subspace which contains most relevant information. Features are then extracted by DLDA on the selected subspace, determined by hard thresholding. Both the optimal threshold value and the performance of a classifier in the feature space are determined through a leave-one-out cross-validation (CV).

In applications where prior information is available, the training filter can be constrained by imposing conditions on both spatial and temporal domain. For example, if we believe that only a certain brain area is important for the task at hand, we can set to 0 the coefficients of the training filter outside of this area. Examples of spatially and/or temporally constrained filters will be presented in Section III.

In summary, the first pass of the algorithm gives a global picture of the data space, which can be treated as a *saliency map* [18], [19] for the second pass of the algorithm. Thus using the two-pass algorithm, the inherent error associated with the estimated covariance matrices under the small size conditions is reduced by the systematic selection of the most informative data subspace.

III. EXPERIMENTAL RESULTS

We illustrate the performances of the DLDA and modified DLDA methods on a data set adopted from Rizzuto *et al.* [20]. The data represents intracranial encephalographic (iEEG) recordings from the human brain during a standard memory reach task. At the beginning of each trial a red fixation stimulus is presented in the middle of a touchscreen which marks a *fixation* period. Next, a green target is flashed on the screen, marking the onset of a *target* period. After the target disappears, the subject is asked to plan a reach movement to the target location without making any eye or arm movements. This stage of the experiment is referred to as a *delay* period. After the *delay* period, the fixation stimulus is extinguished, which signals the participant to reach to the target location. The duration of *fixation*, *target* and *delay* periods varied uniformly between 1 and 1.3 s.

The subject had several electrodes implanted into each of the following target brain areas: orbital frontal (OF) cortex, amygdala (A), hippocampus (H), anterior cingulate (AC) cortex, supplementary motor (SM) cortex and parietal (P) cortex. The total number of electrodes in both hemispheres was 91. The targets were presented at six different locations: right, top right, top left, left, bottom left and bottom right position, with respect to the fixation stimulus. The number of trials per stimulus varied, yielding a total of $n_t = 438$ trials. The electrode signals were amplified, sampled at 200 Hz and band-pass filtered. Note that 1 s of data can be represented as a vector in 18200-dimensional space ($p = 91 \times 200$). The goal of our analysis is to decode the target location based on the brain data alone. Such a method could be used to decode a person's motor intentions in real-time, supporting BCI applications. It should be noted that the iEEG signals are essentially local field potentials (see Section I-A). All decoding results are based on the linear classifier. The choice of classifier did not affect the results significantly, i.e. similar results were obtained with a quadratic classifier.

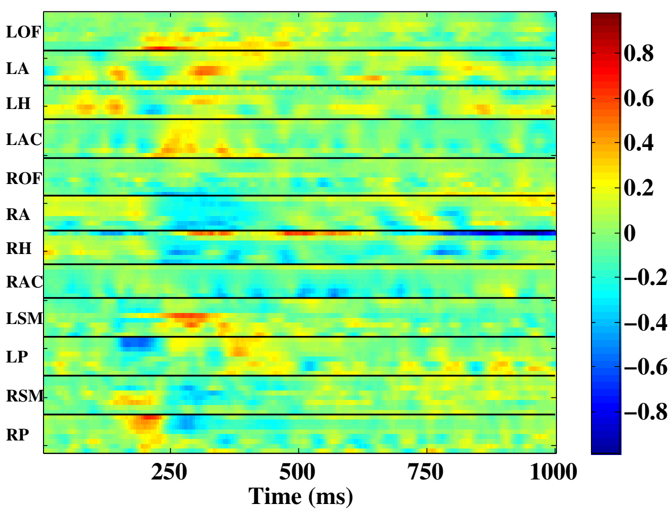

Fig. 1. Training filter $F \in \mathbb{R}^{1\times 18200}$ corresponding to the *target* period on one of the folds of CV, shown as an image with the coefficients normalized between -1 and 1.

In the interest of clarity we will concentrate on a subset of data corresponding to two (out of 6) target locations: left and right. Consequently, our analysis is confined to a one-dimensional (1-D) feature space ($c = 2$, $m = 1$, see Section II-A). Our method readily generalizes to multicategory cases ($c > 2$). The total number of trials for these two stimulus conditions is $n_t = 162$.

Fig. 1 shows the training filter, F, obtained through the DLDA method, and corresponding to 1 s of data during the *target* period on one of the folds of CV. The prefix L or R is used to distinguish between the left and the right hemisphere. Note that the spatial component of F changes in time, thus confirming that space and time are not separable. Large positive (dark red) and large negative (dark blue) coefficients correspond to the brain areas and time samples that are most informative for predicting the target location. As expected, the majority of these large coefficients appear ~150 to 200 ms poststimulus (onset of the target is at 0), consistent with the latency of visual information processing [21]. A straightforward application of DLDA to this data yields a performance ~70% (see Table I), assessed through CV as follows. A single trial (out of 162) was designated for testing, with the remaining trials designated for training. The feature extraction filter T was obtained by DLDA of training data, and 1-D features were extracted. A linear classifier was then designed in the feature space, where test data was projected and classified. This procedure was repeated 162 times, each time selecting different trial as a test data. The overall performance was estimated by dividing the number of correctly classified trials by 162.

TABLE I

THE PERFORMANCES (%) OF DLDA AND HARD THRESHOLDING (HT) WITH DLDA, FOR THE *target*, *delay* AND *reach* PERIOD. (TOP) UNCONSTRAINED, (BOTTOM) CONSTRAINED FILTERS.

Period	Brain area	Time	DLDA	HT+DLDA	p'
target	all	all	70.37	72.22	2513
delay	all	all	58.02	58.64	8269
target	all	150:750	72.22	80.25	918
	P	150:750	86.42	87.06	13872
	SM	all	83.95	87.65	222
	SM,P	all	85.80	92.59	706
delay	P	all	59.86	62.96	1015
	SM, P, OF	all	65.43	71.60	2271
reach	OF	all	75.31	85.18	157

Fig. 1 shows that the filter F provides a sparse representation [17], [22], that is useful discriminatory information is localized in time and space, while the distribution of noise is broad, as illustrated by large areas of near-zero coefficients. The goal of hard thresholding is to extract the most informative data subspace of dimension p' ($p' < p$) by setting the near-zero coefficients to 0. The data samples corresponding to the coefficients not affected by thresholding are used for feature extraction and the performance is evaluated through CV. In general $p' > n_t$, therefore DLDA must be used. The threshold is systematically varied as a multiple of the standard deviation of the coefficients of F and the value that provides the best performance is selected. Note that DLDA can be obtained from above by setting the threshold to 0. Table I shows that this technique improves significantly upon DLDA (for the *reach* period in the orbital frontal (OF) cortex region, the improvement is ~13%). Fig. 2 shows a feature extraction filter T, corresponding to the optimal choice of threshold, in the dataspace of dimension p' determined by the hard thresholding of the training filter in Fig. 1 . Note that the improvement in the performance is more prominent for datasets where p' is sufficiently smaller than p. This result seems consistent with [22], where a sparse representation is indeed a necessary condition for efficient denoising by thresholding. Note that the optimal threshold value obtained through CV gives an idea of the achievable error rates but does not have a generalization power, since validatory data is used to determine the optimal threshold.

From Fig. 1 it follows that some important coefficients appear 100 ms within the onset of the stimulus, which stands in contradiction to the latency of visual information processing ($\sim$150 ms [21]). Clearly, no discriminatory information is expected this early in the trial, and these coefficients likely represent noise artifacts, such as biological noise, recording hardware noise or ambient noise. A constraint can be placed on the training filter to reflect this fact, e.g. $F(t) = 0, \forall t < 150$. Table I also gives the performances based on training filters which are constrained in time and/or space, followed by DLDA. Further gain in performance can be obtained by applying hard thresholding to these constrained problems.

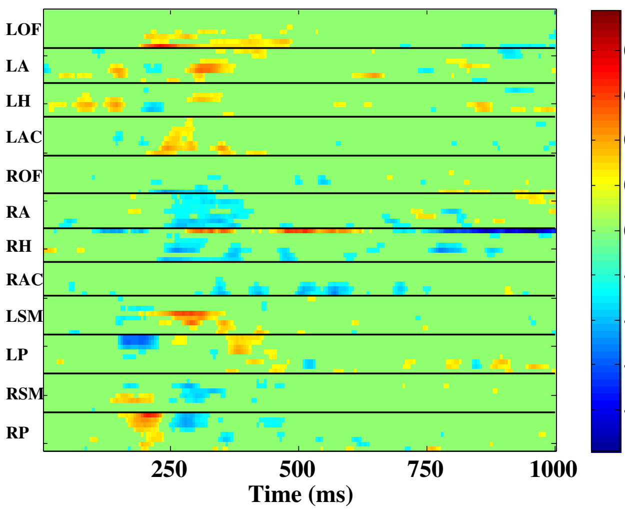

Fig. 2. Feature extraction filter $T \in \mathbb{R}^{1\times 18200}$, corresponding to the optimal threshold value for the *target* period. The locations of zero coefficients are determined by applying optimal thresholding to the training filter F in Fig. 1. Nonzero coefficients are obtained by performing DLDA on the remaining data subspace.

IV. CONCLUSIONS AND FUTURE WORKS

A. Conclusions

We have presented a feature extraction technique for large-scale brain data under the small sample size conditions. Using a thresholding approach in conjunction with DLDA, we have developed a fast and improved feature extraction technique. The technique does not require the space-time separability assumption, common in the analysis of large-scale spatio-temporal signals. For multidimensional features ($m \geq 2$), our method utilizes joint statistical properties of the feature vector, thereby avoiding heuristic feature selection strategies and computationally expensive search algorithms. The technique may have applications beyond BCIs, e.g. epileptic seizure localization. Also, the technique can be applied to any type of spatial, temporal, or spatio-temporal signals.

B. Future Works

Our future research efforts are directed toward testing of this technique in multicategory environment ($c > 2$), and possibly for other types of brain signals. Also, potential advantages (disadvantages) of this technique will be compared against various covariance shrinkage approaches.

V. ACKNOWLEDGMENTS

The authors would like to thank Daniel S. Rizzuto and Richard A. Andersen for providing iEEG data.

REFERENCES

[1] J.R. Wolpaw, N. Birbaumer, D.J McFarland, G. Pfurtscheller, and T.M. Vaughan. Brain-computer interfaces for communication and control. *Clin. Neurophysiol.*, 6(113):767–791, 2002.

[2] G. Pfurtscheller, C. Neuper, G. R. Muller, B. Obermaier, G. Krausz, A. Schlogl, R. Scherer, B. Graimann, C. Keinrath, D. Skliris, G. Supp M. Wortz, and C. Schrank. Graz-BCI: State of the art and clinical applications. *IEEE T. Neur. Sys. Reh.*, 11(2):177–180, 2003.

[3] J. Wessberg, C.R. Stambaugh, J.D. Kralik, P.D. Beck, M. Laubach, J.K. Chapin, J. Kim, S.J. Biggs, M. A. Srinivasan, and M. A. Nicolelis. Real-time prediction of hand trajectory by ensembles of cortical neurons in primates. *Nature*, 408(6810):361–365, 2000.

[4] D.M. Taylor, S.I. Tillery, and A.B. Schwartz. Direct cortical control of 3D neuroprosthetic devices. *Science*, 296(5574):1829–1832, 2002.

[5] S. Musallam, B. D. Corneil, B. Greger, H. Scherberger, and R. A. Andersen. Cognitive control signals for neural prosthetics. *Science*, 305(5681):258–262, 2004.

[6] B. Pesaran, J. S. Pezaris, M. Sahani, P. P. Mitra, and R. A. Andersen. Temporal structure in neuronal activity during working memory in macaque parietal cortex. *Nat. Neurosci.*, 5(8):805–811, 2002.

[7] H. Scherberger, M. R. Jarvis, and R. A. Andersen. Cortical local field potential encodes movement intentions. *Neuron*, 46:347–354, 2005.

[8] J. R. Wolpaw and D. J. McFarland. Control of a two-dimensional movement signal by a noninvasive brain-computer interface in humans. *P. Natl. Acad. Sci. USA*, 101(51):17849–17854, 2004.

[9] K. Fukunaga. *Inroduction to Statistical Pattern Recognition*. Academic Press, 1990.

[10] D. J. McFarland, L. M. McCane, S. V. David, and J. R. Wolpaw. Spatial filter selection for EEG-based communication. *Electroen. Clin. Neuro.*, 103(3):386–394, 1997.

[11] G. Pfurtscheller, C. Neuper, D. Flotzinger, and M. Pregenzer. EEG-based discrimination between imagination of right and left hand movement. *Electroen. Clin. Neuro.*, 103(6):642–651, 1997.

[12] H. Yu and H. Yang. A direct lda algorithm for high-dimensional data - with application to face recognition. *Patter Recognition Letters*, 34(10):2067–2070, 2001.

[13] R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7:179–188, 1936.

[14] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. Wiley-Interscience, 2001.

[15] L.-F. Chen, H.-Y. Mark Liao, M.-T. Ko, J.-C. Lin, and G.-J. Yu. A new lda-based face recognition system which can solve the small sample size problem. *Pattern Recognition*, 33(10):1713–1726, 2000, Pages.

[16] J. Schäfer and K. Strimmer. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology*, 4(1):Article 32, 2005.

[17] D. L. Donoho. De-noising by soft thresholding. *IEEE Transactions on Information Theory*, 41(3):613–627, 1995.

[18] F. Nielsen. Visualization and analysis of 3d functional brain images. Master's thesis, Department of Mathematical Modelling, Technical University of Denmark, Lyngby, Denmark., 1996.

[19] S. Baluja and D. A. Pommerleau. Using a saliency map for active spatial selective attention: Implementation & initial results. In *Advances in Neural Information Processing Systems*, volume 7, pages 451–458. The MIT Press, 1995.

[20] D.S. Rizzuto, A.N. Mamelak, W.W. Sutherling, I. Fineman, and R.A. Andersen. Spatial selectivity in human ventrolateral prefrontal cortex. *Nat. Neurosci.*, 8:415–417, 2005.

[21] S. Thorpe, D. Fize, and C. Marlot. Speed of processing in the human visual system. *Nature*, 381(6582):520–522, 1995.

[22] S. Mallat. *A Wavelet Tour of Signal Processing, 2nd edition*. Academic Press, Inc., 1999.