sinc(*i*) Research Institute for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
V. Peterson, H. L. Rufiner & R. Spies; "ANALYSIS OF DIFFERENT DISCRIMINANT MEASURES ON A PENALIZED MIX-NORM CLASSIFICATION METHOD FOR ERP DETECTION"
VI Congreso de Matemática Aplicada, Computacional e Industrial (VI MACI 2017), may, 2017.

ANALYSIS OF DIFFERENT DISCRIMINANT MEASURES ON A PENALIZED MIX-NORM CLASSIFICATION METHOD FOR ERP DETECTION

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

[♭]*Instituto de Investigación en Señales, Sistemas e Inteligencia Computacional, UNL, CONICET, FICH, Ruta Nac. 168, km 472.4, 3000, Santa Fe, Argentina, vpeterson@sinc.unl.edu.ar*
[†]*Facultad de Ingeniería, Universidad Nacional de Entre Ríos, Ruta Prov. 11, km 10, 3100, Oro Verde, Argentina,*
[*]*Instituto de Matemática Aplicada del Litoral IMAL, CONICET-UNL, Predio CCT-CONICET-Santa Fe, Ruta Nac. 168, km. 472, 3000, Santa Fe, Argentina,*

Abstract: A brain-computer interface (BCI) system based on event related potentials (ERPs) consists mainly of solving a binary classification problem. Although the linear discriminant analysis (LDA) method is widely used for this type of problems, it does not yield satisfactory performances when the number of features is large relative to the number of observations. In this article we present a generalized sparse discriminant analysis method and analyze the impact of six different discriminant measures (used in the construction of certain anisotropy matrices) in classification performance. Numerical results indicate that the best measures for this type of ERP classification problems are those belonging to the Shannon-Entropy family.

Keywords: *Brain-Computer Interface, Discriminant Measures, Sparse Discriminant Analysis, Mixed Penalization*
2000 AMS Subject Classification: primary: 92C55, 92C20, secondary: 65F22, 65J20

1 INTRODUCTION

A Brain-Computer Interface (BCI) is a system aimed to establish an alternative way of communication between the brain of a disabled person and the outside world [4]. In particular, by using the well-known "oddball" paradigm, a BCI based on brain signals (EEGs) can decode the subject's desire by detecting what is called an "event related potential" (ERP) [5]. One of the main components of such ERPs is the so called P300 wave, which is a positive deflection occurring in the scalp-recorded EEG approximately 300 ms after an infrequent stimulus has been applied. The detection of those ERPs on the background EEG conforms a binary classification problem: EEG record with ERP (target class) and EEG record without ERP (non-target class).

The Linear Discriminant Analysis (LDA) criterion is a well-known dimensionality reduction tool in the context of supervised classification [1]. Let $\mathbf{X}$ be a $n \times p$ data matrix and let $\mathbf{y} \in \{1,2\}^n$ be a categorical vector accounting for class membership, i.e. each row of $\mathbf{X}$ (known as a pattern) $\mathbf{x}_i$ belongs to one and only one of the two aforementioned classes. The LDA method seeks to find the discriminant direction vector $\boldsymbol{\beta}$ of maximal separability between classes. Thus, it is defined as:

$$\boldsymbol{\beta}^* = \hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2), \tag{1}$$

where $\hat{\boldsymbol{\Sigma}} \doteq \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^T$ is the common covariance matrix, $\boldsymbol{\mu}_1 \doteq \frac{1}{n_1}\sum_{i\in I_1}\mathbf{x}_i$, $\boldsymbol{\mu}_2 \doteq \frac{1}{n_2}\sum_{i\in I_2}\mathbf{x}_i$ are the sample means for classes 1 and 2, respectively, in which I_1, n_1 and I_2, n_2 are the set of indices and the number of patterns belonging to classes 1 and 2, respectively, and $\boldsymbol{\mu} \doteq \frac{n_1\mu_1+n_2\mu_2}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is the common sample mean.

It has been shown that an effective LDA training requires of a number of samples between five and ten times the dimensionality of the patterns [2]. If that is not the case, then the covariance matrix $\hat{\Sigma}$ is highly ill-conditioned. Regularization may help to overcome this issue, since it limits the influence of outliers, avoids over-fitting and improves the estimation of the ill-conditioned covariance matrix.

In high dimensional data problems, in order to reduce the dimensionality, it is desirable to select a subset of features that be most relevant for the classification problem. Different "metrics" from the statistical literature have been used as measures of discrepancy between classes [3]. We have developed a regularized

version of LDA which performs feature selection and classification by simultaneously using ℓ_1 and ℓ_2 norm penalizers. Two anisotropy matrices are also added in order to include pointwise heterogeneously weighted penalization as dictated by the a-priori discriminative information provided by the discrepancy measure being used. This method can be thought of as a penalized version of the sparse discriminant analysis (SDA) [6], reason for which we call it "generalized sparse discriminant analysis" (GSDA) [7]. This method has been proved to outperform not only the standard LDA, but also several other state-of-the-art regularized versions of LDA (including SDA). However, since GSDA has only been tested with Kullback-Leibler divergence (see [7]), the aim of this work is to find out, from a number of discriminant measures, which one of them is the best for ERP classification. To shead light on this issue, experiments with a real EEG-based BCI database are carried out in small training size scenarios.

2 Material and Methods

2.1 Generalized Sparse Discriminant Analysis

Let K be the number of classes (in our case $K = 2$), $\mathbf{X}$ as before and $\mathbf{Y}$ an $n \times K$ matrix of binary variables such that y_{ij} is an indicator variable of whether the i^{th} observation belongs to the j^{th} class. Let $\boldsymbol{\theta} \in \mathbb{R}^K$ be a *score* vector and $\boldsymbol{\beta} \in \mathbb{R}^p$ the discriminant vector. Then the GSDA scheme consists of solving the following regularized constrained least squares problem:

$$\left(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}\right) = \underset{\boldsymbol{\beta}\in\mathbb{R}^p, \boldsymbol{\theta}\in\mathbb{R}^K}{\arg\min} \{\|\mathbf{Y}\boldsymbol{\theta} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda_1\|\mathbf{D}_1\boldsymbol{\beta}\|_1 + \lambda_2\|\mathbf{D}_2\boldsymbol{\beta}\|_2^2\}, \quad \text{s.t.} \quad \frac{1}{n}\boldsymbol{\theta}^T\mathbf{Y}^T\mathbf{Y}\boldsymbol{\theta} = 1, \tag{2}$$

where λ_1 and λ_2 are positive regularization parameters and $\mathbf{D}_1$ and $\mathbf{D}_2$ are appropriately defined $p \times p$ positive definite matrices which incorporate the a-priori discriminative information about the classes.

The solution of problem (2) is iteratively approximated by alternating two steps (with an adequate initialization): first keep $\boldsymbol{\theta}$ fixed and find $\hat{\boldsymbol{\beta}}$, and then keep $\boldsymbol{\beta}$ fixed and find $\hat{\boldsymbol{\theta}}$. The former step is a generalized version of the well-known elastic-net (e-net) problem [8]: $\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p}\{\|\mathbf{Y}\boldsymbol{\theta} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda_1\|\mathbf{D}_1\boldsymbol{\beta}\|_1 + \lambda_2\|\mathbf{D}_2\boldsymbol{\beta}\|_2^2\}$. This generalized e-net problem can be re-written by means of lasso (least absolute shrinkage selection operator) [9] if $\mathbf{D}_1$ is invertible. The LARS-EN algorithm presented in [10] has been appropriately modified to find the solution vector $\hat{\boldsymbol{\beta}}$ over which the classes of the projected data $\boldsymbol{X}\hat{\boldsymbol{\beta}} \in \mathbb{R}^n$ can be well-separated by a simple linear classifier. In a previous work, we have found that GSDA outperforms other well-known classification methods for ERP detection in small training size scenarios.

2.2 Discriminant measures

Discriminative information can be incorporated into GSDA by appropriately quantifying the "distance" between classes or, more precisely, between their probability distributions. A wide variety of "metrics" is available for comparing probability distributions [3]. Let $f_1(n)$ and $f_2(n)$ with $n \in \mathcal{N}$ be two (discrete) probability functions (here, think of them as being those associated to the target and non-target classes, respectively). Different discriminant measures d can be used to compare those two probability distributions and hence, to quantify the difference between classes. For instance, let us consider the Kullback-Leibler Divergence [11] defined as:

$$d_{\text{KL}}(f_1||f_2) \doteq \sum_{n\in\mathcal{N}} f_1(n)\log\left(\frac{f_1(n)}{f_2(n)}\right), \tag{3}$$

with the convention that $0.\log 0 \doteq 0$. A symmetric d_{KL} version, called the J-divergence [12], is defined as follows:

$$d_{\text{J}}(f_1, f_2) \doteq d_{KL}(f_1||f_2) + d_{KL}(f_2||f_1). \tag{4}$$

Another measure within the Shannon-Entropy family is the Jensen-Shannon divergence [13], defined as:

$$d_{\text{JS}}(f_1, f_2) \doteq \frac{1}{2}d_{KL}(f_1||f_3) + \frac{1}{2}d_{KL}(f_2||f_3), \tag{5}$$

where $f_3 \doteq \frac{f_1+f_2}{2}$. Note that d_{JS} is symmetric and its square root $\hat{d_{\text{JS}}} \doteq \sqrt{d_{\text{JS}}}$ is a metric in the rigorous mathematical sense.

Another discrepancy measure, widely used in machine learning, is the Fisher information distance which can be obtained as the second derivative of d_{KL}. In particular, as described in [14] the Fisher distance between two univariate normal distribution $f_1(\mu_1, \sigma_1)$ and $f_2(\mu_2, \sigma_2)$ can be found in terms of the first two moments as:

$$d_{\mathrm{F}}(f_1, f_2) = \sqrt{2}\ \ln\left(\frac{\mathcal{F}((\mu_1,\sigma_1),(\mu_2,\sigma_2)) + (\mu_1-\mu_2)^2 + 2(\sigma_1^2+\sigma_2^2)}{4\sigma_1\sigma_2}\right), \tag{6}$$

where $\mathcal{F}((\mu_1,\sigma_1),(\mu_2,\sigma_2)) = \sqrt{((\mu_1-\mu_2)^2+2(\sigma_1-\sigma_2)^2)((\mu_1-\mu_2)^2+2(\sigma_1+\sigma_2)^2)}$.

Finally, the discriminant information can also be evaluated by the squared point-wise bi-serial correlation coefficients with sign, simply called "signed r^2-value". The r-value is defined as: $r = \frac{\sqrt{n_1 n_2}}{n_1+n_2}\left(\frac{\mu_1-\mu_2}{\sigma}\right)$,

where $\sigma = std\{\mathbf{x}_i | i \in I_1, I_2\}$ is the joint standard deviation. Then, the signed r^2-value is defined as

$$d_{\mathrm{r}}(f_1, f_2) = sgn(r)\, r^2. \tag{7}$$

In the present work we are particularly interested in comparing the impact of using these different measures in highlighting the ERPs from the background EEG. Independently of which discriminant measure d is used, we can define the measure of discrepancy between classes at the i^{th} feature, $d(i)$, for $i = 1, 2, \ldots, p$, as $d(i) \doteq d\left(\left\{f_1^i, f_2^i\right\}\right)$, where f_1^i, f_2^i denote the probability distributions of classes 1 and 2, respectively, at feature i. The anisotropy matrices $\mathbf{D}_1$ and $\mathbf{D}_2$ are then constructed by using the a-priori discriminant information $d(\cdot)$ as follows: $\mathbf{D_1} \doteq diag\,(1-\alpha_i+\alpha_i c_i)$ and $\mathbf{D_2} \doteq diag(c_i)$, where $c_i \doteq \dfrac{\left(\prod_{j=1}^{p} d(j)\right)^{1/p}}{d(i)}$, $\alpha_i \doteq \dfrac{\max_{1\le j\le p}\{c_j\} - c_i}{\max_{1\le j\le p}\{c_j\} - \min_{1\le j\le p}\{c_j\}}$, for $i = 1, \ldots, p$. Note that with $\mathbf{D_1}$ and $\mathbf{D_2}$ so defined, c_i is large where $d(i)$ is small, and vice-versa. The parameter α_i (observe that $0 \le \alpha_i \le 1,\ \forall i = 1, \ldots, p$) weights the discriminant information proportionally to its relevance.

2.3 DATABASE

The database used for this work is an open-access P300 speller database from the "Laboratorio de Investigación en Neuroimagenología de la Universidad Autónoma Metropolitana", Mexico D.F., described in [15]. This database consists of EEG records acquired from 25 healthy subjects, recorded by 10 channels (Fz, C3, Cz, C4, P3, Pz, P4, PO7, PO8, Oz) at 256 Hz sampling rate. In this work we used the first two copy-spelling sessions as our dataset, in which each subject had to spell 21 characters. In the preprocessing stage, the EEG records were filtered from 0.1 Hz to 12 Hz by a 4^{th} order forward-backward Butterworth band-pass filter. A 1000 ms data segment (trial) was extracted from the EEG records at the beginning of each stimulus, and then they were downsampled to 32 Hz. A total of 3780 EEG trials (630 of them being target) of dimension of $10 \times 32 = 320$, conforms each subject's database.

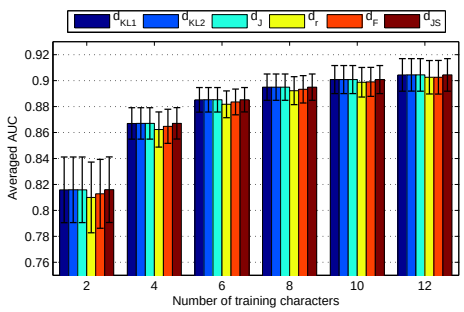
We simulated small training size scenarios by randomly selecting patterns for spelling different given number of characters (2, 4, 6, 8, 10 y 12). This selection procedure was repeated 100 times.

3 RESULTS AND DISCUSSIONS

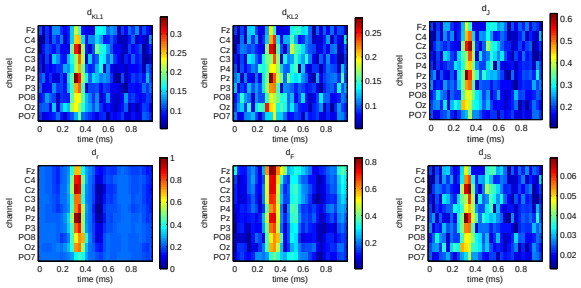
The classification performance was measured by the area under the receiver operator characteristics curve (AUC) [16]. Figure 1a shows the average result over the 25 subjects in each small training size scenarios delivered by GSDA using different discriminant measures. Note that d_{KL1} and d_{KL2} in Figure 1a refers to $d_{\mathrm{KL}}(f_1||f_2)$ and $d_{\mathrm{KL}}(f_2||f_1)$, respectively. Observe that the measures belonging to the Shannon-Entropy family yielded the best performances. Figure 1b show the time-channel plots of $d(\cdot)$ by using the aforementioned discrepancy measures for one subject in the 10 character scenario. Note how the P300 wave is well highlighted in all cases.

4 CONCLUSIONS

In the present work we have analyzed the impact of using different discriminant measures for constructing the anisotropy matrices $\mathbf{D}_1$ and $\mathbf{D}_2$ in the classification performances achieved by the GSDA method.



(a) Average AUC on test data by using GSDA.



(b) Discrepancy measure d in each sample i.

Figure 1: Classification results and discrepancy measure plots by using different discriminant measures.

We have found that the best classification results are achieved by both the symmetric Kullback-Leibler divergence and by its smoothed-out version, the Jensen-Shannon divergence. In regard to the discriminant measure plots, it seems that those measures that "better" depict the P300 wave do not yield the best classification results. This observation leads as to conjecture that the assumption of normal distribution (which conducts to measuring the distance by using only the first two statistical moments), results in the discriminant information being focused exclusively on the most prominent wave (P300), neglecting other pieces of important discriminant information.

ACKNOWLEDGMENTS

This work was supported in part by Consejo Nacional de Investigaciones Científicas y Técnicas, CONICET, through PIP 2014-2016 No. 11220130100216-CO, the Air Force Office of Scientific Research, AFOSR/SOARD, through Grant FA9550-14-1-0130 and and by Universidad Nacional del Litoral, UNL, through CAID-UNL 2011 Project No.525 within PACT "Señales, Sistemas e Inteligencia Computacional".

REFERENCES

[1] R. DUDA, P. HART AND D.G STORK, *Pattern classification.* JOHN WILEY & SONS, 2012.

[2] T. HASTIE, A. BUJA AND R. TIBSHIRANI, *Penalized Discriminant Analysis.* The Annals of Statistics, (1995), pp.73-102.

[3] M. BASSEVILLE, *Distance measures for signal processing and pattern recognition*. Signal Proc., Vol. 18 (1989), pp.349-369.

[4] J. WOLPAW AND E.W WOLPAW, *Brain-Computer Interfaces: principles and practice.* OXFORD UNIV. PRESS, USA, 2012.

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

[6] L. CLEMMENSEN, T. HASTIE, D. WITTEN AND B. ERSBØLL, *Sparse Discriminant Analysis*. Technometrics, Vol. 53 (2012), pp. 406-413.

[7] V. PETERSON, H.L RUFINER AND R.D SPIES, *Generalized Sparse Discriminant Analysis for Event-Related Potential Classification*. Biomedical Signal Processing and control, Vol. 35 (2017), pp. 70-78.

[8] H. ZOU AND T. HASTIE, *Regularization and variable selection via the elastic net*. Journal of the Royal Statistical Society: Series B (Methodological), Vol. 67 (2005), pp.301-320.

[9] H. ZOU AND T. HASTIE, *Regression shrinkage and selection via the lasso*. Journal of the Royal Statistical Society: Series B (Methodological), Vol. 58 (1996), pp. 267-288.

[10] K. SJÖSTRAND, L. CLEMMENSEN, R. LARSEN AND B. ERSBØLL, *SpaSM: A matlab toolbox for sparse statistical modeling*. Journal of Statistical Software Accepted for publication), (2012).

[11] S. KULLBACK, AND R.A LEIBLER, *On information and sufficiency*. The annals of math. statistics, Vol. 22 (1951), pp. 79-86.

[12] H. JEFFREYS, *An invariant form for the prior probability in estimation problems.* Proc. of the Royal Society of London A: Mathematical, Physical and Engineering Sciences, Vol. 186 (1946), pp. 453-461.

[13] J. LIN, *Divergence measures based on the Shannon entropy*. IEEE Trans. on Information theory, Vol. 37 (1991), pp. 145-151.

[14] S.I.R COSTA, S.A SANTOS AND J.E STRAPASSON, *Fisher information distance: a geometrical reading*. Discrete Applied Mathematics, Vol. 197 (2015), pp. 59-69.

[15] 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.

[16] A.P BRADLEY. *The use of the area under the ROC curve in the evaluation of machine learning algorithms.* Pattern recognition, Vol. 30 (1997), pp. 1145-1159.