REGULARIZED LDA BASED ON SEPARABLE SCATTER MATRICES FOR CLASSIFICATION OF SPATIO-SPECTRAL EEG PATTERNS

Mohammad Shahin Mahanta, Amirhossein S. Aghaei, Konstantinos N. Plataniotis

Email: {mahanta, aghaei, kostas} @comm.utoronto.ca
The Edward S. Rogers Sr. Department of Electrical and Computer Engineering
University of Toronto

ABSTRACT

Linear discriminant analysis (LDA) is a commonly-used feature extraction technique. For matrix-variate data such as spatio-spectral electroencephalogram (EEG), matrix-variate LDA formulations have been proposed. Compared to the standard vector-variate LDA, these formulations assume a separable structure for the within-class and between-class scatter matrices; these structured parameters can be estimated more accurately with a limited number of training samples. However, separable scatters do not fit some data, resulting in aggravated performance for matrix-variate methods. This paper first proposes a common framework for the vector-variate LDA with non-separable scatters and our previously proposed solution with separable scatters. Then, a regularization of the non-separable scatter estimates toward the separable estimates is introduced. This novel regularized framework integrates vector-variate and matrix-variate approaches, and allows the estimated scatter matrices to adapt to the data characteristics. Experiments on data set V from BCI competition III demonstrate that the proposed framework achieves a considerable classification performance gain.

Index Terms— regularization, separable covariance, matrix-variate Gaussian, linear discriminant analysis, 2DLDA.

1. INTRODUCTION

Electroencephalogram (EEG) is a record of electrical activities of the brain captured through electrodes mounted on the scalp [1]. Classification of EEG signals is desired in many applications, including medical diagnosis [2] and design of brain computer interface (BCI) systems [3]. A BCI system provides an interface to control external devices. It can be used in prosthetics, to perform highly demanding tasks, or for navigation in virtual environments. This paper considers a spontaneous BCI system using noninvasive multichannel EEG to classify motor imagery tasks [4].

Motor imagery tasks affect both spatial and spectral characteristics of EEG signals [5, 6]. Therefore, BCI systems operate on the power spectra of different EEG channels which form a matrix-variate sample. These samples are high-dimensional and consist of highly correlated components. In classifying these signals, the discriminant data components need to be extracted by techniques such as the commonly used linear discriminant analysis (LDA) [7]. A trivial one-directional LDA (1DLDA) feature extraction approach applies the vector-variate LDA on vectorized multichannel EEG samples; whereas matrix-variate LDA methods [8, 9, 10, 11, 12] utilize the inherent matrix-variate structure of the data to facilitate the estimation of within-class and between-class scatter matrices.

Among matrix-variate methods, Ye's two-directional LDA (Y2DLDA) [8] is widely used in the literature but does not provide Bayes optimal features [13, 14, 15]. We have previously proposed matrix-to-vector LDA (MVLDA) [16, 17] as a Bayes optimal matrix-variate LDA. Similar to other matrix-variate methods, MVLDA assumes a *separable* structure for scatter matrices of the data.

In the current work, in Section 3, MVLDA is presented in a common framework with 1DLDA, but with its distinct separable scatter matrix estimates. Based on this framework, Section 4 introduces a regularized scatter matrix estimate as a trade-off between MVLDA and 1DLDA. Compared to vector-variate regularized LDA solutions [18], this novel approach integrates the vector-variate and matrix-variate solutions, and provides a generally superior performance as demonstrated in Section 5.

2. PROBLEM DEFINITION

An overview of the target classification problem is shown in Fig. 1. Preprocessed spatio-spectral EEG samples are denoted as matrices $\mathbf{X}_{m\times n}$. [1] In the training stage, N_i training samples $\mathbf{X}_{ij}, 1 \leq j \leq N_i$, are used to estimate parameters of each class Ω_i. Then, in the testing stage, the BCI system classifies each spatio-spectral pattern $\mathbf{X}_{m\times n}$ into one of the classes

[1] In this paper, scalars, vectors, and matrices are respectively shown in regular lowercase/uppercase (e.g. a or A), boldface lowercase (e.g. $\mathbf{a}$), and boldface uppercase (e.g. $\mathbf{A}$). The transpose of $\mathbf{A}$, trace of $\mathbf{A}$, null (kernel) space of $\mathbf{A}$, and Kronecker product of $\mathbf{A}$ and $\mathbf{B}$ are respectively denoted by $\mathbf{A}^T$, $\mathrm{tr}(\mathbf{A})$, $\mathrm{Null}(\mathbf{A})$, and $\mathbf{A}\otimes\mathbf{B}$. The vectorized representation of a matrix $\mathbf{A}$ through concatenation of its columns is shown as $\mathrm{vec}(\mathbf{A})$.

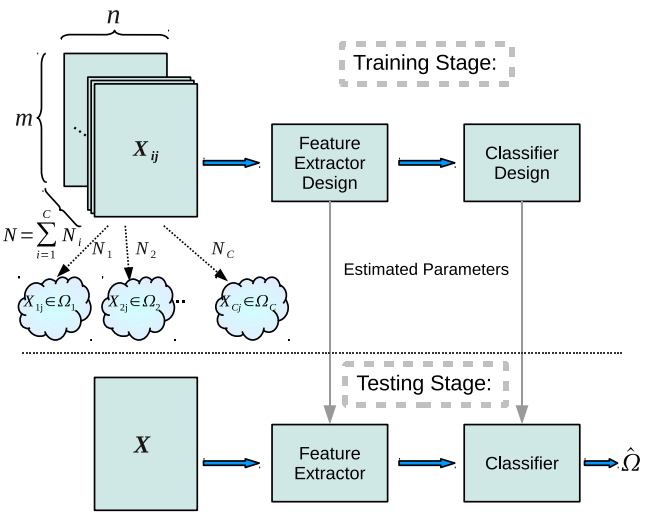

Fig. 1: Outline of the classification system.

Ω_i, $1 \leq i \leq C$, corresponding to different BCI tasks. The target is to maximize the probability of correct classification.

This paper focuses on the design of the feature extractor so that the features with the most discriminatory information are extracted. A regularized feature extractor is proposed which integrates the existing 1DLDA and MVLDA methods. This integrated approach is based on a common framework for 1DLDA and MVLDA as described in the next section.

3. 1DLDA VS. MVLDA: A COMMON FRAMEWORK

This section describes 1DLDA and the previously proposed matrix-variate MVLDA method [16, 17] based on a common framework consisting of a linear operation:

$$\mathbf{y}_{d\times 1} = \mathbf{T}_{d\times mn}\mathbf{x}_{mn\times 1}, \tag{1}$$

where $\mathbf{x}_{mn\times 1} = \mathrm{vec}(\mathbf{X})$ denotes the (column-wise) vectorized data. The 1DLDA method uses a vector-variate approach to find $\mathbf{T}$ in contrast to the matrix-variate formulation of MVLDA; although both methods provide a set of Bayes optimal features according to their corresponding assumptions.

3.1. 1DLDA

The 1DLDA method operates on the (column-wise) vectorized data $\mathbf{x}_{mn\times 1} = \mathrm{vec}(\mathbf{X})$. Starting from the matrix-variate mean $\mathbf{M}_i$ for each class Ω_i, $\mathbf{M}_i = \frac{1}{N_i}\sum_{j=1}^{N_i}\mathbf{X}_{ij}$, and the overall mean $\mathbf{M} = \sum_{i=1}^{C}\frac{N_i}{N}\mathbf{M}_i$, the corresponding mean vectors for $\mathbf{x}$ consist of $\boldsymbol{\mu}_i = \mathrm{vec}(\mathbf{M}_i)$ and $\boldsymbol{\mu} = \mathrm{vec}(\mathbf{M})$. Then, 1DLDA estimates the within-class scatter $\mathbf{S}_W$ and the between-class scatter $\mathbf{S}_B$ as

$$\mathbf{S}_W = \sum_{i=1}^{C}\sum_{j=1}^{N_i}(\mathbf{x}_{ij}-\boldsymbol{\mu}_i)(\mathbf{x}_{ij}-\boldsymbol{\mu}_i)^T, \tag{2}$$

$$\mathbf{S}_B = \sum_{i=1}^{C} N_i(\boldsymbol{\mu}_i-\boldsymbol{\mu})(\boldsymbol{\mu}_i-\boldsymbol{\mu})^T. \tag{3}$$

The 1DLDA operator $\mathbf{T}_{d\times mn}$ in (1) is constructed with its rows as the d eigenvectors of $\mathbf{S}_W^{-1}\mathbf{S}_B$ corresponding to the largest eigenvalues.

The 1DLDA features are Bayes optimal if the data in each class follow a Gaussian distribution with a common covariance matrix among the classes, i.e., the data are homoscedastic Gaussian, and the corresponding parameters are accurately known [19]. In practice, the data parameters need to be estimated. For most data sets, the dimensionality of 1DLDA scatter matrices $\mathbf{S}_W$ and $\mathbf{S}_B$, i.e., $mn \times mn$, is so large that the available training samples do not suffice for accurate estimation of these parameters. Thus, a matrix-variate method utilizing the inherent structure of these data is preferable.

3.2. MVLDA

When the original data are matrix-variate, the inherent structure of the data can be utilized in estimation of $\mathbf{S}_W$ and $\mathbf{S}_B$. In [16, 17] we proposed the MVLDA feature extractor using separable estimates for the scatter matrices: $\mathbf{S}_W = \mathbf{S}_{WR}\otimes\mathbf{S}_{WL}$, and $\mathbf{S}_B = \mathbf{S}_{BR}\otimes\mathbf{S}_{BL}$. These scatters are estimated as a Kronecker product of column-wise (left) and row-wise (right) components. Maximum-likelihood estimates for left and right within-class scatters $\mathbf{S}_{WL}$ and $\mathbf{S}_{WR}$ were used, which are calculated through iteration on the following two steps [16, 20]:

$$\mathbf{S}_{WL} = \frac{1}{Nn}\sum_{i=1}^{C}\sum_{j=1}^{N_i}(\mathbf{X}_{ij}-\mathbf{M}_i)\mathbf{S}_{WR}{}^{-1}(\mathbf{X}_{ij}-\mathbf{M}_i)^T,$$

$$\mathbf{S}_{WR} = \frac{1}{Nm}\sum_{i=1}^{C}\sum_{j=1}^{N_i}(\mathbf{X}_{ij}-\mathbf{M}_i)^T\mathbf{S}_{WL}{}^{-1}(\mathbf{X}_{ij}-\mathbf{M}_i). \tag{4}$$

The left and right between-class scatters are also calculated as [16, 17]

$$\mathbf{S}_{BL} = \sum_{i=1}^{C} N_i(\mathbf{M}_i-\mathbf{M})(\mathbf{M}_i-\mathbf{M})^T,$$

$$\mathbf{S}_{BR} = \frac{1}{\mathrm{tr}(\mathbf{S}_{BL})}\sum_{i=1}^{C} N_i(\mathbf{M}_i-\mathbf{M})^T(\mathbf{M}_i-\mathbf{M}). \tag{5}$$

Then, it can be shown that the previously proposed MVLDA can be written in the format of (1), with the rows of $\mathbf{T}_{d\times mn}$ as the d eigenvectors of $(\mathbf{S}_{WR}\otimes\mathbf{S}_{WL})^{-1}(\mathbf{S}_{BR}\otimes\mathbf{S}_{BL})$ corresponding to the largest eigenvalues.

The MVLDA method provides Bayes optimal features if the data are homoscedastic Gaussian with accurately known parameters, and $\mathbf{S}_W$ and $\mathbf{S}_B$ follow a separable structure as above. The assumption of separability of $\mathbf{S}_W$ and $\mathbf{S}_B$ is practical for most matrix-variate data [16, 17]. However, specific data sets may deviate from this assumption, and this deviation

reduces the effectiveness of MSLDA extracted features [17]. To obviate these challenges of 1DLDA and MVLDA, the next section proposes a flexible regularized estimation approach.

4. PROPOSED REGULARIZED MVLDA

Based on the previous section, the essential difference between 1DLDA and MVLDA lies in their corresponding procedure for estimating within-class and between-class scatters. The separable structure adopted by MVLDA provides a consistent estimate for most practical matrix-variate data sets. However, for data sets with deviation from the assumed separability, i.e, if $\mathbf{S}_W \neq \mathbf{S}_{WR} \otimes \mathbf{S}_{WL}$ or $\mathbf{S}_B \neq \mathbf{S}_{BR} \otimes \mathbf{S}_{BL}$, this structure may provide an oversimplified description. Therefore, we need a flexible trade-off between the non-separable 1DLDA estimates and the separable estimates of MVLDA. Parameter regularization [18] can trade-off between a general estimate with less bias and a constrained estimate with less variance. Using this technique in the framework of Section 3, we propose a novel regularization of the generally non-separable estimates $\mathbf{S}_W$ and $\mathbf{S}_B$ in (2) and (3) toward the separable estimates $\mathbf{S}_W^s = \mathbf{S}_{WR} \otimes \mathbf{S}_{WL}$ and $\mathbf{S}_B^s = \mathbf{S}_{BR} \otimes \mathbf{S}_{BL}$:

$$\begin{aligned} \mathbf{S}_W^r &= (1-\gamma_w)\,\mathbf{S}_W + \gamma_w\,\mathbf{S}_W^s, \\ \mathbf{S}_B^r &= (1-\gamma_b)\,\mathbf{S}_B + \gamma_b\,\mathbf{S}_B^s. \end{aligned} \tag{6}$$

It should be noted that from (5), $\mathrm{tr}(\mathbf{S}_B^s) = \mathrm{tr}(\mathbf{S}_B)$, and due to the convergence of the iteration in (4), $\mathrm{tr}(\mathbf{S}_W^s) = \mathrm{tr}(\mathbf{S}_W)$ [20]. Therefore, the regularization coefficients $0 \leq \gamma_w \leq 1$ and $0 \leq \gamma_b \leq 1$ determine the actual weighting between the non-separable and separable estimates.

The proposed regularized MVLDA (R-MVLDA) method uses a linear operator $\mathbf{T}_{d\times mn}$ as in (1) whose rows are selected as the d eigenvectors of $(\mathbf{S}_W^r)^{-1}\mathbf{S}_B^r$ corresponding to the largest eigenvalues. The regularization coefficients can be estimated using procedures such as cross-validation [7], or through prior knowledge of the characteristics of the data.

Setting $\gamma_w = \gamma_b = 0$ simplifies R-MVLDA to 1DLDA, while $\gamma_w = \gamma_b = 1$ leads to MVLDA. Therefore, R-MVLDA integrates the two methods into a common framework. The regularization coefficients γ_w and γ_b respectively represent the degree of separability of $\mathbf{S}_W$ and $\mathbf{S}_B$ for the data, with $\gamma_w = \gamma_b = 0$ corresponding to the non-separable extreme case and $\gamma_w = \gamma_b = 1$ denoting the fully separable case. With γ_w and γ_b selected based on the nature of the given data set, R-MVLDA adapts to the data characteristics and, as shown in the next section, outperforms both 1DLDA and MVLDA.

5. EXPERIMENTAL EVALUATION

In this section, we evaluate the effectiveness of R-MVLDA on the actual spatio-spectral EEG patterns in a BCI scenario. In order to focus on the performance of the feature extraction stage, a simple linear Gaussian classifier is used in the system of Fig. 1. The resulting correct classification rate (CCR) performance using R-MVLDA is compared with that of 1DLDA, Y2DLDA [8], and MVLDA [16, 17].

5.1. EEG Data Set and Preprocessing Procedure

Data set V from BCI competition III [4] is used which contains EEG data of three normal subjects collected in four sessions. The data are recorded using 32-electrode Biosemi system at 512Hz sampling rate. Each record consists of sequential 15-second trials of three possible mental imagery tasks: left-hand movement, right-hand movement, and generation of words beginning with a random letter. The last session is used for testing and the rest for training. The goal of the competition is to classify the mental task every 0.5 second using only the last second of the data. The highest CCR in this competition without post-processing was %62.72 [21].

The raw EEG data are spatially filtered using a surface Laplacian filter computed using spherical splines of order 2 and regularization parameter of 0.01. Then, a short-time Fourier transform with a Hamming window of length one second and overlapping factor of $\frac{15}{16}$ is applied. The resulting power spectral components from $8-30$Hz with resolution of 2Hz are averaged every 0.5 seconds, leading to 12 spectral components per EEG channel. This procedure results in 12×32 matrix-variate spatio-spectral EEG samples. Furthermore, the 8 centro-parietal channels (C3, Cz, C4, CP1, CP2, P3, Pz, and P4) which are highly correlated to motor imagery tasks are selected to form a new set of 12×8 EEG samples with a lower dimensionality (ref. Tab. 1).

5.2. Experimental Results

The CCR values for different feature extractors are reported in Tab. 1. The number of features d and the regularization parameters are chosen so that CCR is maximized. Practically, these parameters need to be estimated through methods such as cross-validation. However, in this study, we use the optimal values in order to investigate the essential performance limitations for different methods. For Y2DLDA, 10 iterations are performed [8]. For MVLDA and R-MVLDA, a threshold of 10^{-5} on the incremental change in the Frobenius norm of $\mathbf{S}_{WL}$ and $\mathbf{S}_{WR}$ is used to terminate the iteration in (4); this choice leads to an average of 18 or 14 iterations respectively for the data with all or selected EEG channels.

From Tab. 1, R-MVLDA provides a considerable performance gain compared to 1DLDA and MVLDA, whose performances are in turn superior to that of the non-Bayes-optimal Y2DLDA. When all the EEG channels are used, the data dimensionality is higher and MVLDA significantly outperforms 1DLDA. Therefore, γ_w for R-MVLDA is much higher in this case, so that R-MVLDA leans toward the preferable MVLDA method. On the other hand, γ_b is almost always large, which signifies that for this data set, between-class scatter is better estimated as a separable matrix.

Table 1: Classification results for different feature extractors.

Data Size	Method	Subject a				Subject b				Subject c				Avg.
$(m \times n)$		%CCR	d	γ_w	γ_b	%CCR	d	γ_w	γ_b	%CCR	d	γ_w	γ_b	%CCR
(12×8)														
	1DLDA	74.89	2	_	_	63.71	2	_	_	51.26	1	_	_	63.29
	Y2DLDA	36.38	36	_	_	45.15	1	_	_	38.45	15	_	_	39.99
	MVLDA	74.04	11	_	_	62.24	36	_	_	53.57	48	_	_	63.28
	R-MVLDA	77.66	43	0.10	1.00	68.57	37	0.00	1.00	55.88	4	0.65	1.00	67.37
(12×32)														
	1DLDA	69.15	2	_	_	58.23	2	_	_	50.21	1	_	_	59.20
	Y2DLDA	37.66	32	_	_	43.25	60	_	_	40.97	20	_	_	40.62
	MVLDA	76.81	206	_	_	65.19	2	_	_	57.35	51	_	_	66.45
	R-MVLDA	78.72	4	0.80	0.80	68.78	4	0.95	0.90	62.82	16	0.75	0.95	70.10

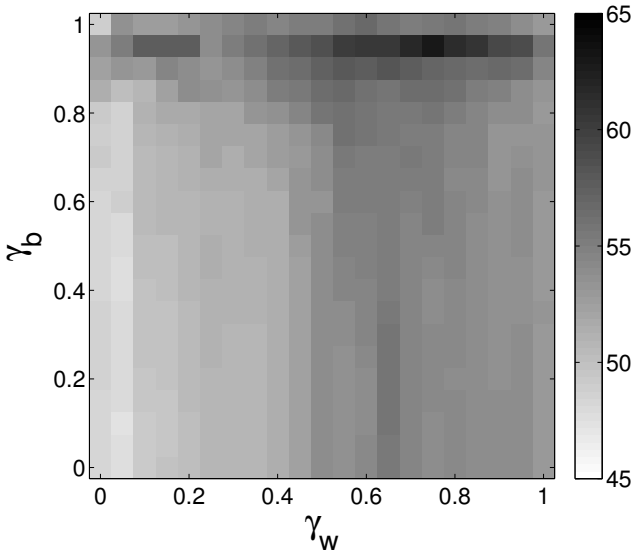

Fig. 2: CCR of R-MVLDA versus the regularization parameters. Darkness of the pixels denote the %CCR for subject 'c' with data from all the 32 channels and $d = 16$.

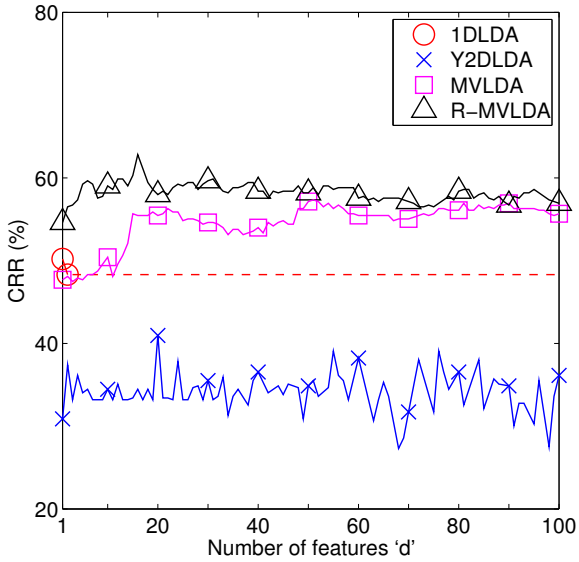

Fig. 3: CCR of different methods versus number of features d. CCR values belong to subject 'c' with data from all the 32 channels.

The performance gain achieved by R-MVLDA in Tab. 1 and its γ_w and γ_b values vary for different subjects. When using all the channels, R-MVLDA's gain is most significant for subject 'c'. For this subject, R-MVLDA's CCR profile versus γ_w and γ_b at the optimal d value is shown in Fig. 2. In this plot, $(0,0)$ and $(1,1)$ corners respectively correspond to 1DLDA and MVLDA as extreme cases. For $\gamma_b \lesssim 0.8$, the low-rank $\mathbf{S}_B$ matrix dominates the largest eigenvalues of $\mathbf{S}_B^r$, and thus γ_b does not affect the CCR significantly.

In Fig. 3, CCR of different methods is plotted versus $1 \leq d \leq 100$ for subject 'c'. It should be noted that since the number of 1DLDA features is limited to $C - 1 = 2$, a horizontal dashed line is drawn from its last CCR to the end of range of d. Again, it is demonstrated that R-MVLDA can significantly improve the performance of both 1DLDA and MVLDA.

6. CONCLUSIONS AND FUTURE WORKS

In this paper, for the first time, a regularized LDA formulation as a trade-off between the existing matrix-variate and vector-variate approaches was proposed. Although this generalized approach does not provide the computational efficiency of the matrix-variate approach, it provides a superior performance compared to the solutions on both sides of the trade-off.

In the experiments on spatio-spectral EEG patterns, the optimal regularized scatters were closer to the separable estimates in most cases. This result demonstrates that for most EEG data sets, separable scatter matrix estimates are preferable to non-separable estimates. This finding and its physical interpretation can be further studied in a future work.

7. REFERENCES

[1] Saeid Sanei and Jonathon A. Chambers, *EEG Singal Processing*, John Wiley and Sons Ltd, fifth edition, 2007.

[2] Abdulhamit Subasi and M. Ismail Gursoy, "EEG signal classification using PCA, ICA, LDA and support vector machines," *Expert Systems with Applications*, vol. 37, no. 12, pp. 8659 – 8666, 2010.

[3] F. Lotte, M. Congedo, A. Lécuyer, F. Lamarche, and B. Arnaldi, "A review of classification algorithms for EEG-based brain-computer interfaces.," *Journal of neural engineering*, vol. 4, no. 2, June 2007.

[4] Jdel.R. Millan, "On the need for on-line learning in brain-computer interfaces," in *Neural Networks, 2004. Proceedings. 2004 IEEE International Joint Conference on*, July 2004, vol. 4, pp. 2877–2882.

[5] B. Blankertz, R. Tomioka, S. Lemm, M. Kawanabe, and K.-R. Muller, "Optimizing spatial filters for robust EEG single-trial analysis," *Signal Processing Magazine, IEEE*, vol. 25, no. 1, pp. 41–56, 2008.

[6] C. Neuper and G. Pfurtscheller, "Event-related dynamics of cortical rhythms: Frequency-specific features and functional correlates," *International Journal of Psychophysiology*, vol. 43, no. 1, pp. 41–58, 2001.

[7] A.K. Jain, R.P.W. Duin, and Jianchang Mao, "Statistical pattern recognition: a review," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 22, no. 1, pp. 4–37, 2000.

[8] Jieping Ye, Ravi Janardan, and Qi Li, "Two-dimensional linear discriminant analysis.," in *Advances in Neural Information Processing Systems, NIPS 2004*, 2004.

[9] K. Inoue and K. Urahama, "Non-iterative two-dimensional linear discriminant analysis," in *18th International Conference on Pattern Recognition, ICPR 2006*, 2006, vol. 2, pp. 540 –543.

[10] Shiladitya Chowdhury, Jamuna Kanta Sing, Dipak Kumar Basu, and Mita Nasipuri, "Face recognition by generalized two-dimensional FLD method and multi-class support vector machines," *Applied Soft Computing*, vol. 11, no. 7, pp. 4282 – 4292, 2011.

[11] S. Noushath, G. Hemantha Kumar, and P. Shivakumara, "(2D)2LDA: An efficient approach for face recognition," *Pattern Recognition*, vol. 39, no. 7, pp. 1396 – 1400, 2006.

[12] J. Zhao, P.L.H. Yu, L. Shi, and S. Li, "Separable linear discriminant analysis," *Computational Statistics and Data Analysis*, vol. 56, no. 12, pp. 4290–4300, 2012.

[13] Bo Yang and Yingyong Bu, "A comparative study on vector-based and matrix-based linear discriminant analysis," *Journal of Computers*, vol. 6, no. 4, pp. 818–824, 2011.

[14] Zhizheng Liang, Youfu Li, and Pengfei Shi, "A note on two-dimensional linear discriminant analysis," *Pattern Recognition Letters*, vol. 29, no. 16, pp. 2122 – 2128, 2008.

[15] Wei-Shi Zheng, J.H. Lai, and Stan Z. Li, "1D-LDA vs. 2D-LDA: When is vector-based linear discriminant analysis better than matrix-based?," *Pattern Recognition*, vol. 41, no. 7, pp. 2156 – 2172, 2008.

[16] Mohammad Shahin Mahanta, Amirhossein S. Aghaei, and Konstantinos N. Plataniotis, "A Bayes optimal matrix-variate LDA for extraction of spatio-spectral features from EEG signals," in *Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE*, 2012, pp. 3955 –3958.

[17] Mohammad Shahin Mahanta, Amirhossein S. Aghaei, and Konstantinos N. Plataniotis, "A Bayes optimal matrix-to-vector approach to 2DLDA," *submitted to Pattern Recognition*.

[18] J.H. Friedman, "Regularized discriminant analysis," *Journal of the American Statistical Association*, vol. 84, no. 405, pp. 165–175, March 1989.

[19] O.C. Hamsici and A.M. Martinez, "Bayes optimality in linear discriminant analysis," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 30, no. 4, pp. 647 –657, april 2008.

[20] Pierre Dutilleul, "The MLE algorithm for the matrix normal distribution," *Journal of Statistical Computation and Simulation*, vol. 64, no. 2, pp. 105–123, 1999.

[21] Ferran Galán, Francesc Oliva, and Joan Guàrdia, "Using mental tasks transitions detection to improve spontaneous mental activity classification," *Medical & Biological Engineering & Computing*, vol. 45, no. 6, pp. 603–609, 2007.