CSLT TECHNICAL REPORT-20160013 [Tuesday 27th December, 2016]

Sparse Discriminative Analysis and Its Application in Distraction Classification

Dong Wang

Correspondence: wangdong99@mails.tsinghua.edu.cn
Center for Speech and Language Technology, Research Institute of Information Technology, Tsinghua University, ROOM 1-303, BLDG FIT, 100084 Beijing, China
Center for Speech and Language Technologies, Division of Technical Innovation and Development, Tsinghua National Laboratory for Information Science and Technology, ROOM 1-303, BLDG FIT, 100084 Beijing, China
Full list of author information is available at the end of the article

Abstract

Sparse discriminant analysis (SDA) imposes l-1 regularization to encourage sparse coefficients in linear discriminant transform. This approach has found a broad range of machine learning tasks, due to its capability of identifying the most promising features so that the feature dimensionality can be significantly reduced, leading to most robust and generalizable models. This paper reviews the development of SDA from linear discriminative analysis (LDA), and presents its application to the driving distraction detection task.

Keywords: sparse discriminative analysis; distraction detection

1 Introduction

Machine learning methods suffer from curse of dimensionality, which means that the amount of data required to train a reliable model is exponentially increased with respect to the dimensionality of the feature. In the case where the the dimensionality is high and the training data is limited, the trained model tends to be severely over-fitting, i.e., performs very good on training data but degrades on test data substantially [1]. To prevent over-fitting, it is often desirable to to reduce the feature dimensionality. An attractive feature dimensionality reduction approach is to select the important dimensions and ignore unimportant ones. A clear advantage of this approach is that those dimensions that are easy to be contaminated by noise can be removed, leading to more noise-robust models. In summary, machine learning requires dimension reduction, keeping the most prominent dimensions (features) and removing trivial and task-unrelated ones.

Linear discriminative analysis (LDA) is probably the simplest dimension reduction approach. The objective of LDA is to discover a low-rank linear transform, by which the training data are projected to a low-dimensional space where the intra-class variance is minimized and the inter-class variance is maximized. This approach, however, fails in situations where (1) the dimensionality of the data is very high; (2) the classes can not be well described by a single Gaussian; (3) the classes can not be well separated by linear boundaries [2]. Penalized discriminant analysis (PDA) [3] and LDA based on Gaussian mixtures [4] were presented to address these difficulties. The sparse discriminative analysis (SDA) [2] extended PDA and multi-Gaussian LDA by introducing an l-1 norm, which encourages discovering sparse discriminant directions, e.g., the non-zero elements of these directions are sparse. Almost at the same time, [5] presented a similar l-1 penalization for LDA. Shao et al. [6] proposed a sparse LDA that is asymptotically optimal under some sparsity conditions on the parameters. SDA, and other sparse version of discriminant analysis were thereafter extensively studied and found a broad range of applications, e.g. [7, 8, 9, 7, 10, 11, 12, 13]. A number of toolkits have been designed to support SDA and its variants, e.g., Spasm from Sjöstrand et al. [14].

This paper is organized as follows: in Section 2 we start from the classical LDA and review three perspectives that can lead to LDA. Section 3 discusses some extensions based on different perspectives, including heterogeneous LDA, penalized LDA, sparse LDA. Section 4 focuses on the sparse LDA, particularly the SDA model based on the optimal scoring framework. Section 5 presents an experiment to demonstrate the capability of SDA, where SDA is used to select important features from the EEG data and use the selected features to detect drivers' mental distraction. The entire paper is concluded in Section 6.

2 Linear discriminant analysis (LDA): Three perspectives

LDA can be derived from three perspectives [5]: Fisher's discriminant, probabilistic modeling and optimal scoring. These three perspectives are correlated and formulate the learning task from different angles: Fisher's discriminant seeks for the best discriminative directions on which the projected training samples are best classified; the probabilistic modeling approach formulates the classification problem as a maximum likelihood estimation problem with a probabilistic model; the optimal scoring approach reformulates the classification problem into a regression problem.

2.1 Fisher discriminant view

let $X = [x_1 x_2 ... x_N]^T \in R^{N \times P}$ denote a data matrix containing N observations of P dimensions; further let $Y \in \{0,1\}^{N \times K}$ represent the class of the N observations, where Y_{nk} indicates whether the n-th observation belongs to the k-th class. The projected image of x_n is $z_n = \beta^T x_n$, where $\beta = [\beta_1, \beta_2, ..., \beta_D] \in R^{P \times D}$ is the projection matrix that projects a P-dimensional data sample x to a D-dimensional image z. The definition of LDA is an optimal β by which the training data in different classes are mostly separated in the projected space, where the 'separateness' is measured by the Fisher criterion, defined as follows:

$$J(\beta) = Tr\{(\beta S_W \beta^T)^{-1} (\beta S_B \beta^T)\} \tag{1}$$

where S_W and S_B are within-class variance and between-class variance of the training data, respectively, and β projects the variance in the original data space to the projected space. The between class S_B is defined by:

$$S_B = \sum_{k=1}^{K} N_k (\bar{x}_k - \bar{x})(\bar{x}_k - \bar{x})^T$$

where N_k is the number of training samples in the k-th class, and $\bar{x}^k$ is the mean vector of the training data of the k-th class, and $\bar{x}$ is the mean vector of all the training data:

$$\bar{x}^k = \frac{1}{N_k} \sum_{C(x_n)=k} x_n \qquad \bar{x} = \frac{1}{N} \sum_{n=1}^{N} x_n$$

where $C(x)$ returns the class label of x. The within-class S_W is defined as the sum of the variance of individual classes:

$$S_W = \sum_{k=1}^{K} \sum_{n \in C_k} (x_n - \bar{x}^k)(x_n - \bar{x}^k)^T.$$

It was shown that for a particular dimension D, $J(\beta)$ is maximized when the columns of β (the discriminant directions) coincide with the D eigenvectors of $S_W^{-1} S_B$ corresponding to the D largest eigenvalues [15]. This means the LDA problem under the Fisher discriminant view can be solved by simple eigen analysis.

2.2 Probabilistic model view

LDA can be also derived from a maximum likelihood estimation of a constrained Gaussian model [16]. The definitions of X, β, $\bar{x}$ and $\bar{x}^k$ as as in the Fisher discriminant view. The difference is that the training data of a particular class k now is assumed to follow a Gaussian distribution $N(\mu_k, \Sigma_k)$. The projected image $z = \beta^T x$ therefore follows a Gaussian distribution as well, denoted by $N(m_k, V_k)$, where $m_k = \beta^T \mu_k$, and $V_k = \beta^T \Sigma_k \beta$.

Define the between-class and within-class variances as follows:

$$S_B = \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})(x_i - \bar{x})^T$$

$$S_W^k = \frac{1}{N_k} \sum_{C(x)=k} (x - \bar{x}_k)(x - \bar{x}_k)^T$$

$$S_W = \frac{1}{N} \sum_{k=1}^{K} N_k S_W^k.$$

Note that the definitions of S_B and S_W here are slightly different from the ones in the Fisher discriminant view, but the basic idea is totally the same. Besides the Gaussian assumption, LDA also assumes that only the first D dimensions of m_k are distinct for different k, and the covariance matrices of all the classes are identical. These conditions are formally written as:

$$m_k = \begin{bmatrix} m_{k,1} \\ m_{k,2} \\ ... \\ m_{k,D} \\ m_{0,D+1} \\ ..m_{0,P} \end{bmatrix} = \begin{bmatrix} m_D^k \\ m_0 \end{bmatrix}$$

$$V_k = \begin{bmatrix} V_D & 0 \\ 0 & V_0 \end{bmatrix}$$

Under these constraints, LDA optimizes β such that the log likelihood function $J(\beta) = \sum_x logP(x;\beta)$ is maximized. Since $|\frac{\partial z}{\partial x}| = |\beta|$, we have :

$$P(x) = |\beta|P(z) = \frac{|\beta|}{\sqrt{(2\pi)^P|V_{C(x)}|}} exp\{-\frac{(\beta^T x - m_{C(x)})^T V_{C(x)}^{-1}(\beta^T x - m_{C(x)})}{2}\}$$

where $C(x)$ denotes the class of x. The likelihood function is then written by:

$$\begin{aligned} J(\beta) &= \sum_x logP(x;\beta) \qquad (2) \\ &= Nlog|\beta| - \frac{N}{2}log|V| - \frac{1}{2}\sum_x\{(\beta^T x - m_{C(x)})^T V^{-1}(\beta^T x - m_{C(x)})\} + const \end{aligned}$$

where *const* is a constant value unrelated to β. Maximizing $J(\beta)$ with respect to β leads to the optimal projection. Firstly treat β as fixed and optimize m_k and V. A simple computation shows:

$$\begin{aligned} \hat{m}_k^D &= \beta_D^T \bar{x}_k \quad k = 1, 2, .., K \\ \hat{m}_0 &= \beta_0^T \bar{x} \\ \hat{V}^D &= Diag(\beta_D^T S_W \beta_D) \\ \hat{V}_0 &= Diag(\beta_0^T S_B \beta_0) \end{aligned}$$

where β_D is a submatrix that involves the first D columns of β, and β_0 involves the rest $P - D$ columns of β. Substituting these results into Eq. 2 leads to:

$$J(\beta) = Nlog|\beta| - \frac{1}{2}log|Diag(\beta_0^T S_B \beta_0^T)| - \frac{N}{2}log|Diag(\beta_D^T S_W \beta_D)|.$$

Kumar [17] proved that taking the eigenvectors of $S_W^{-1}S_B$ corresponding to the D largest eigenvalues as β_D and the rest eigenvectors as β_0 also maximizes the above $J(\beta)$. Note that these eigenvectors are also the solution of the conventional LDA based on the Fisher discriminant, thus confirming that LDA can be regarded as a maximum likelihood parameter estimation of a constrained Gaussian model, where (1) some dimensions of the class-dependent Gaussians are distinguishable, and (2) the covariance matrices of all the classes are identical.

The probabilistic model that LDA assumes can be written as a neat form:

$$x_{k,i} = \mu + Bz_k + \sigma_{k,i} \tag{3}$$

where $x_{k,i}$ is the i-th sample of the k-th class, μ the global mean vector, and z_k is the mean vector of the k-th class in the projected space. B is the projection matrix, and $\sigma_{k,i} \sim N(0, \Sigma)$ is the residual Gaussian noise. It is easy to verify that this formulation holds the same assumptions as LDA: all the classes are Gaussians with shared covariance. Note that this is a linear Gaussian model and similar to principle component analysis (PCA) and factor analysis, except that the data are labelled in class.

2.3 Optimal scoring view

The third approach that can lead to LDA is to cast the classification problem to a linear regression problem. The main difficulty, obviously, is that the regression target is a categorical variable, which often leads to high sensitivity with outliers in the training data, if the regression is simply conducted on the categorical targets. The optimal scoring approach [3] tackles the problem by turning the one-hot categorical target (class label) to a continuous vector θ_k. The objective function takes the following form [18]:

$$min_{\beta_k,\theta_k}\{||Y\theta_k - X\beta_k||_2^2\} \quad s.t. \quad \frac{1}{N}\theta_k^T Y^T Y \theta_k = 1, \;\; \theta_k^T Y^T Y \theta_l = 0 \;\; \forall l < k, \tag{4}$$

where θ_k is the score vector of the k-th class, and β_k is the k-th discriminative direction vector. Note that this is a sequential optimization problem where the 'discriminative directions' $\{\beta_k\}$ are attained one by one, and the new derived discriminant direction is orthogonal to all the existing discriminant directions.

3 LDA extension based on different perspectives

A couple of extensions have been proposed to improve LDA, based on different perspectives of the LDA formulation. We review several typical ones and discuss their correlation.

3.1 Penalized LDA based on the Fisher discriminant view

Based on the Fisher discriminant formulation, Witten and Tibshirani [5] proposed to add a convex penalty to the objective function. According to the Fisher discriminant view, the original LDA problem can be formulated as:

$$max_{\beta_k}(\beta_k^T S_B \beta_k) \quad s.t. \;\; \beta_k^T S_W \beta_k \le 1, \beta_k^T S_W \beta_i = 0 \;\; \forall i < k$$

Note that this criterion looks different from the one in Section 2.1, but it can be easily verified that they state the same thing, if we adapt the length of β_k to meet $\beta_k^T S_W \beta_k = 1$ in Eq.1. Adding a penalty on β leads to:

$$max_{\beta_k}(\beta_k^T S_B \beta_k - P(\beta_k)) \quad s.t. \;\; \beta_k^T S_W \beta_k \leq 1, \beta_k^T S_W \beta_i = 0 \;\; \forall i < k$$

where $P(\beta_k)$ is the added regulation item. In Witten's work [5], two types of penalty terms are investigated: the first one is the l-1 norm $\sum_{j=1}^{P} |\sigma_j \beta_{kj}|$ where σ_j is the within-class standard deviation of the j-th feature. This penalty pushes unimportant features to zero, leading to a sparse β_k. The second penalty encourages smoothness between adjacent dimensions, formulated by $\sum_{j=2}^{P} |\sigma_j \beta_{k,j} - \sigma_{j-1}\beta_{k,j-1}|$.

3.2 Heterogeneous LDA based on the probabilistic model view

According to the probabilistic model view, the assumption behind LDA is that all the classes are Gaussian and the covariances are shared. If the covariance matrices are not shared, we reach the heterogeneous LDA (HLDA), as proposed by Kumar [16].

Following the notation in Section 2.2, the variance V_k of the k-th class can be written by:

$$V_k = \begin{bmatrix} V_k^D & 0 \\ 0 & V_0 \end{bmatrix}$$

The likelihood function is $J(\beta) = \sum_x logP(x;\beta)$, where

$$P(x) = |\beta| P(z) = \frac{|\beta|}{\sqrt{(2\pi)^P |V_{C(x)}|}} exp\{-\frac{(\beta^T x - m_{C(x)})^T V_{C(x)}^{-1} (\beta^T x - m_{C(x)})}{2}\}.$$

Note the covariance matrices of different classes are different. The objective function is then derived as:

$$\begin{aligned} J(\beta) &= \sum_x logP(x;\beta) \qquad (5) \\ &= Nlog|\beta| - \frac{N}{2} log|V_{C(x)}| - \frac{1}{2}\sum_x \{(\beta^T x - m_{C(x)})^T V^{-1} (\beta^T x - m_{C(x)})\} + const \end{aligned}$$

where *const* is a constant value unrelated to β. Fixing β and optimizing m_k and V_k leads to:

$$\begin{aligned} \hat{m}_k^D &= \beta_D^T \bar{x}_k \quad k = 1, 2, .., K \\ \hat{m}_0 &= \beta_0^T \bar{x} \\ \hat{V}_k^D &= Diag(\beta_D^T S_W^k \beta_D) \quad k = 1, 2, .., K \\ \hat{V}_0 &= Diag(\beta_0^T S_B \beta_0) \end{aligned}$$

Substituting these results into Eq. 5 leads to:

$$J(\beta) = Nlog|\beta| - \frac{1}{2}log|Diag(\beta_0^T S_B \beta_0^T)| - \sum_{k=1}^{K} \frac{N_k}{2} log|Diag(\beta_D^T S_W^k \beta_D)|.$$

Formulating HLDA as a linear Gaussian model gives:

$$x_{k,i} = \mu + Bz_k + \sigma_{k,i}.$$

where $\sigma_{k,i} \sim N(0, \Sigma_k)$, i.e., the variance of the residual noise is different for different classes.

3.3 Probabilistic LDA based on the probabilistic model view

A key note of the linear Gaussian model of LDA (Eq. (3)) is that z_k, the mean vector of class k in the latent space, is deterministic rather than probabilistic, i.e., z_k is a parameter that should be learned during model training. Once the model has been trained, the values of $\{z_k\}$ are fixed. This leads to several disadvantages: firstly, z_k is 'almost purely' estimated from the data of the k-th class. For classes with limited training data, z_k tends to be weakly estimated. Secondly, it is impossible to introduce new classes once the model has been trained, preventing it from dealing with new classes at inference (test) time.

To overcome this problem, probabilistic LDA was proposed by several authors [19, 20]. The basic idea is to change z_k from a deterministic parameter to a random variable, so that z_k can be inferred from data during the test phase. A simple setting is to assume a normal distribution for z_k, i.e., $z_k \sim N(0, 1)$, although more complex settings are possible. With the prior distribution, it would be possible to infer the projected image $z_{k,i}$ even if the class has never been seen in the training data.

3.4 Regularized LDA based on the optimal scoring view

The optimal scoring view casts the classification problem to a regression problem, opening rich opportunities to involve various regularizations. For example, Clemmensen et al. [5, 2] appended an l-2 term and an l-1 term to the cost function Eq.(4), leading to:

$$\begin{aligned} &min_{\beta_k,\theta_k}\{||Y\theta_k - X\beta_k||_2^2 + \gamma\beta_k^T\Omega\beta_k + \lambda||\beta_k||_1\} \\ &\quad s.t. \quad \frac{1}{N}\theta_k^T Y^T Y \theta_k = 1, \quad \theta_k^T Y^T Y \theta_l = 0 \ \ \forall l < k, \end{aligned} \tag{6}$$

where Ω is a positive definite matrix to avoid singularity when the observations are mutually dependent or when the dimension is large, i.e., $P > N$, and λ and γ are non-negative hyperparameters. Note that the l-1 penalty introduced by the third term in the above equation enforces sparsity on β_k, and more dimensions of β_k are driven to zeros with a larger λ [21]. More discussion about this model can be seen in the next section.

3.5 Some discussions

We have reviewed a number of extensions for LDA. The underlying idea of these extensions can be categorized into two fold: regularization and assumption relaxation. Regularization imposes additional terms to the objective functions to encourage particular properties, e.g., parameter shrinkage or sparsity [3, 5, 2]. This regularization is also a principle approach to solving the singularity problem caused by high feature dimensionality or limited training data. Assumption relaxation, on the other hand, tries to substitute the strong shared-covariance Gaussian assumption with more realistic assumptions, e.g., the distinct-covariance assumption in HLDA [16] and the multiple Gaussian assumption in the Gaussian mixture LDA [4].

Compared to the three perspectives, the Fisher discriminant view is the most straightforward and the inference is easy to interpret. The standard solutions based on eigen analysis makes it easy to implement by using off-the-shelf eigen analysis tools. It is also easy to add regularization terms to the cost function, but the additional terms may cause difficulty in optimization as it may not be simply an egien problem. The optimal scoring view is particularly easy to add regularization, but it is not simple to change the model structure, i.e., involving more realistic assumptions. The probabilistic model is a more principle way to deal with both model change and regularization introduction. Since the probabilistic model describes the underlying assumptions in a simple and explicit way, and many regularizations can be derived from a prior distribution, the probabilistic view provides a theoretical framework to enhance the model structure, leading to an elegant way to involve human knowledge, plausible assumptions, and trade-off between performance and complexity.

4 Sparse linear discriminant analysis (SDA)

In this section, we focus on a particular regularization for LDA: the sparsity constraint. The importance of this regularization can be attributed to several aspects. Firstly, it can address the singularity problem associated with very high dimensional features. Secondly, it helps identify the most discriminant feature group, leading to a natural dimension reduction approach. Thirdly, the sparsity constraint promotes the most discriminative information and ignores unimportant nuances, leading to more noise-robust features in the projected space. This section reviews a typical sparsity-oriented LDA, named as 'sparse discriminative analysis', or SDA.

SDA resorts to the l-1 penalty to achieve sparse projections (discriminative directions). Imposing an l-1 penalty to achieve sparsity has been extensively studied in both regression [21, 22] and classification [18, 5, 23, 24]. By adding to the objective function an l-1 regularization term with respect to the model parameters, unimportant parameters tend to be driven to zeros. For a linear model $\sigma(\beta^T x)$ where $\sigma(\cdot)$ denotes the activation function, x the input data and β the parameters, an l-1 regularized objective function encourages the coefficient β_i to zero if x_i is less important, leading to a natural and efficient approach for feature selection.

The SDA approach described in [18] belongs to such l-1 derived sparse model. It is based on the optimal scoring formulation of LDA, and adds an l-2 term and an l-1 term to the cost function, leading to:

$$\begin{aligned} & min_{\beta_k,\theta_k}\{||Y\theta_k - X\beta_k||_2^2 + \gamma\beta_k^T\Omega\beta_k + \lambda||\beta_k||_1\} \\ & \quad s.t. \quad \frac{1}{N}\theta_k^T Y^T Y\theta_k = 1, \quad \theta_k^T Y^T Y\theta_l = 0 \ \ \forall l < k, \end{aligned} \tag{7}$$

where Ω is a positive definite matrix to avoid singularity, and λ and γ are non-negative hyperparameters. Note that the l-1 penalty introduced by the third term in the above equation enforces sparsity on β_k, and more dimensions of β_k are driven to zeros with a larger λ [21].

In the case of a two-class classification problem, there is only one discriminative direction β. The optimization problem is then reduced to the following:

$$\begin{aligned} & min_{\beta,\theta}\{||Y\theta - X\beta||_2^2 + \gamma\beta^T\Omega\beta + \lambda||\beta||_1\} \\ & \quad s.t. \quad \frac{1}{N}\theta^T Y^T Y\theta = 1. \end{aligned} \tag{8}$$

Eliminating θ by a simple variable substitution leads to:

$$min_{\beta}\{||\hat{Y} - X\beta||_2^2 + \gamma\beta^T\Omega\beta + \lambda||\beta||_1\}, \tag{9}$$

where $\hat{Y}$ is the normalized class indicator matrix whose elements are given by:

$$\hat{Y}_{n,k} = \sqrt{\frac{N}{N_k}},$$

where N_k is the number of observations of the k-th class. We see that the optimization problem for the classification task equals to the optimization problem of a regression task in the case of two classes, which has been stated in [1]. Furthermore, notice that Eq. (9) is an elastic net problem if $\Omega = I$, and a generalized elastic net problem for an arbitrary symmetric positive definite matrix Ω. This elastic net problem can be solved by the algorithm proposed by [22].

The mental distraction detection task that we will study in this paper is a two-class problem. In this case, the discriminative direction β actually plays a role of feature selection, i.e., the dimensions whose corresponding elements of β are zero are simply discarded. Note that Eq. (9) coincides with the elastic net regression proposed by [22]. For multiple classification tasks, the SDA model is a general framework to derive sparse coefficients $\{\beta_k\}$. In this case, the non-zero dimensions of different β_k are usually different, requiring task-dependent treatment.

5 SDA for driving distraction detection

Mental distraction in car driving is very dangerous and is one of the major causes of traffic disasters. It would be highly valuable if we can monitor the mental status of drivers and produce some alarms when they are in distraction. Mental distraction can be caused by various factors, e.g., answering calls, tuning air conditions, looking at people besides or behind, etc. These ‘explicit distraction’ can be detected by either eye fixation or driving behavior [25, 26]. However, for distraction caused by psychological overload, e.g., thinking some puzzling things, it would be hard to detect by just monitoring drivers’ face or behavior. One way to detect such ‘implicit distraction’ is to monitor the activity of drivers’ brain. In this study, we investigate using electroencephalography (EEG) data to detect mental distraction. Although it is still far from practice (it would be not easy to persuade drivers to wear EEG helmets), the study at least can show a possible way of detecting psychological distraction. More importantly, we hope to discover more evidence for the connection between psychological load and brain activities.

5.0.1 Data

The EEG data was provided by Prof. Guozhen Zhao from the psychological research institute of Chinese academy of social science (CASS). The data were collected from 8 subjects using a driving simulation. The collection was divided into 6 phases, where two phases are normal driving, while the other four phases are mental distracted, by increasing the drivers’ psychological load with some disturbing tasks. The data were collected using 34 electrodes (channels), with the sampling rate of 10 Hz. Three features were selected from each channel according to previous psychological studies, lead to a feature of 102 dimensions. Note that this high-dimensional feature involves much redundance and noise, as neighbouring channels tend to record similar patterns, and some channels provide noise only.

5.0.2 SVM results on all channels

In the first set of experiments, we use an off-the-shell machine learning tool to discriminate normal and distracted mental status. Due to its simplicity and high performance, the SVM model is selected as the classifier, for which the LibSVM package is used for model training and inference.[1] Positive samples (distracted) and negative samples (normal) are balanced by random sampling some negative data.

Two experiments are conducted: the first ‘Single-subject’ experiment trains a single SVM for each subject, and the second ‘Multi-subject’ experiment trains a single SVM with the data from all the subjects, though the test is still conducted on individual subjects.

The results are shown in Table 1, where the performance is evaluated in terms of frame error rate (FER), i.e., the proportion of the frames that are classified incorrectly. It can be seen that with either approach, the performance on the training set is very good (almost 100% correct), while the performance on the test set is pretty low. This suggests that using all the features leads to severe over-fitting. It is not surprising as the training data is very limited (about 900 positive samples and 300 negative samples per subject), and the dimensionality is pretty high (102). We will see the over-fitting problem exists with linear models as well.

[1] http://www.csie.ntu.edu.tw/ cjlin/libsvm/

Table 1 **Results of SVM, based on features collected on all channels.**

Training Approach	Training subject	Test subject	Training FER	Test FER
Single	1	1	0.00	48.54
Single	2	2	0.00	48.85
Single	3	3	0.00	38.16
Single	4	4	0.00	49.70
Single	5	5	0.00	41.54
Single	6	6	0.00	49.40
Single	7	7	0.00	49.70
Single	8	8	0.00	47.43
Multi	1-8	1	-	48.55
Multi	1-8	2	-	49.00
Multi	1-8	3	-	49.47
Multi	1-8	4	-	49.70
Multi	1-8	5	-	49.08
Multi	1-8	6	-	50.30
Multi	1-8	7	-	49.70
Multi	1-8	8	-	48.93
Multi	1-8	1-8	0.02	48.93

5.0.3 SVM approach on selected channels

To alleviate the over-fitting problem, the feature dimensionality needs to be reduced. From previous studies, it has been shown that some channels are more related to mental status. By this knowledge, we choose the most prominent 5 channels, leading to 15 features. This feature selection is purely psychologically driven.

Again, a 'Single-subject' experiment that trains a single SVM for each subject, and a 'Multi-subject' experiment that trains a single SVM for all subjects are conducted. The results are shown in Table 2. It can be seen that the over-fitting is alleviated a little bit, but the performance is still unacceptable. This indicates that the psychology-based feature dimension reduction is not ideal.

Table 2 **Results of SVM, based on features collected from selected channels.**

Training Approach	Training subject	Test subject	Training FER	Test FER
Single	1	1	0.00	43.80
Single	2	2	0.15	49.92
Single	3	3	0.40	47.36
Single	4	4	0.00	45.35
Single	5	5	0.05	49.23
Single	6	6	6.90	57.06
Single	7	7	0.05	50.30
Single	8	8	0.15	49.24
Multi	1-8	1	-	42.42
Multi	1-8	2	-	51.77
Multi	1-8	3	-	46.46
Multi	1-8	4	-	47.45
Multi	1-8	5	-	48.92
Multi	1-8	6	-	53.45
Multi	1-8	7	-	54.29
Multi	1-8	8	-	48.64
Multi	1-8	1-8	9.70	49.29

5.0.4 SDA approach: Multi-subject model

In this experiment, we employ SDA to select the most representative features from the 102 dimensions. Once the features are selected, a simple linear model (logistic regression) is applied to conduct the classification. The data from all the 8 subjects are used to train the SDA and the classifier (so it is a 'Multi-subject' experiment), and then test on each subject as well as the entire test data. The results on the entire training set and test set are reported in Fig. 1 and Fig. 2 respectively, where the sparsity of SDA is set in different values so that the dimensionality of the selected features changes from 1 to 102. It can be seen that on the training set, more dimensions lead to better performance, while on the test set, there is an optimal dimensionality that leads to the best test performance. This confirms the existence of over-fitting, and demonstrates that SDA can help select the most prominent features so that the over-fitting problem can be alleviated. Compared to the SVM results in the previous section, it can be seen that the best performance with SDA is much better than with the simple SVM.

The results on each subject are presented in Fig. 3, where each subject is represented by a curve. We observe that there is a large variance among subjects: some subjects can obtain very good performance, while others' performances are rather poor.

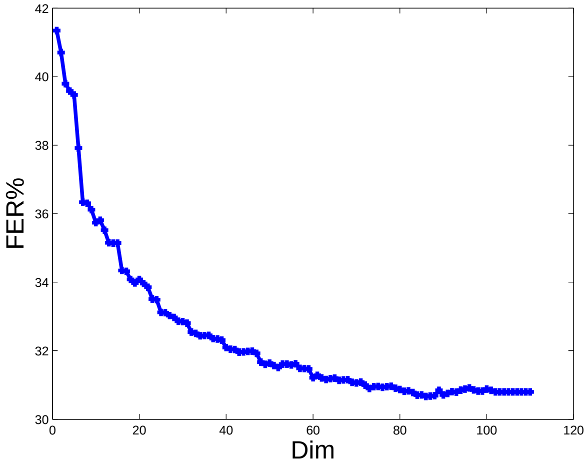

Figure 1 FER on the entire train set, with SDA based on features of all channels. Models are trained on data of all subjects.

5.0.5 SDA approach: Single-subject model

Motivated by the great variance among subjects, we train specific SDAs for individual subjects in this experiment. The results on the training sets and test sets are presented in Fig. 4 and Fig. 5 respectively.

From the results on the training data, it can be observed that the single models can learn each subject very well and obtain rather good performance, compared to the multi-subject model shown in Fig. 1. The results on test sets do not shown much advantage compared to those obtained with the multi-subject model as shown in Fig. 3; however, there are no subjects that perform very bad as in Fig. 3. This again suggests that subject variability is an important factor, and training subject-specific models is necessary. However, the subject-specific model suffers from data sparsity more seriously, leading to more serious over-fitting. This is why the highest performance obtained by the single-subject models is even worse than the one obtained with the multi-subject model.

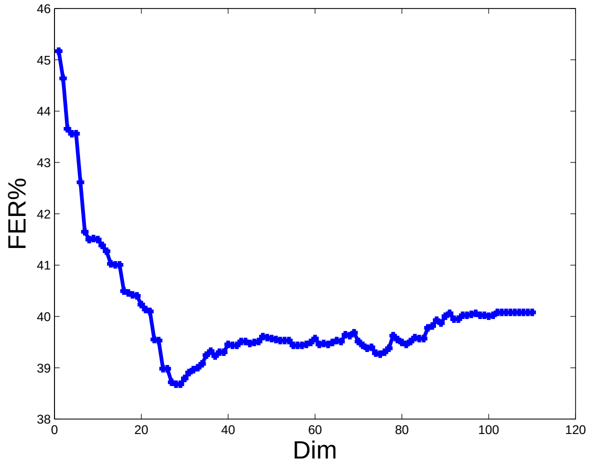

Figure 2 FER on the entire test set, with SDA based on features of all channels. Models are trained on data of all subjects.

6 Conclusion

In this paper, we review three perspectives of LDA and several extensions based on different perspectives. We particularly focus on a typical sparsity-oriented LDA approach, SDA, and present an experiment that applies this technique to the task of mental distraction detection based on EEG data. Our experiments show that the SDA-based feature selection works pretty good and can deliver even better performance than the human selection approach based on psychological knowledge.

Acknowledgements

This research was supported by the National Science Foundation of China (NSFC) under the project No. 61371136, and the MESTDC PhD Foundation Project No. 20130002120011. Thanks to Prof. Guozhen Zhao from Chinese Academy of Social Sciences to provide the EEG data.

References

1. Christopher M Bishop, "Pattern recognition," *Machine Learning*, vol. 128, 2006.
2. Line Clemmensen, Trevor Hastie, Daniela Witten, and Bjarne Ersbøll, "Sparse discriminant analysis," *Technometrics*, 2012.
3. Trevor Hastie, Andreas Buja, and Robert Tibshirani, "Penalized discriminant analysis," *The Annals of Statistics*, pp. 73–102, 1995.
4. Trevor Hastie and Robert Tibshirani, "Discriminant analysis by gaussian mixtures," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 155–176, 1996.
5. Daniela M Witten and Robert Tibshirani, "Penalized classification using fisher's linear discriminant," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 73, no. 5, pp. 753–772, 2011.
6. Jun Shao, Yazhen Wang, Xinwei Deng, Sijian Wang, et al., "Sparse linear discriminant analysis by thresholding for high dimensional data," *The Annals of statistics*, vol. 39, no. 2, pp. 1241–1265, 2011.
7. Guosheng Cheng, Xingxiang Li, Peng Lai, Fengli Song, and Jun Yu, "Robust rank screening for ultrahigh dimensional discriminant analysis," *Statistics and Computing*, pp. 1–11, 2016.
8. Xixuan Han and Line Clemmensen, "Regularized generalized eigen-decomposition with applications to sparse supervised feature extraction and sparse discriminant analysis," *Pattern Recognition*, vol. 49, pp. 43–54, 2016.
9. Zhihui Lai, Wai Keung Wong, Yong Xu, Jian Yang, and David Zhang, "Approximate orthogonal sparse embedding for dimensionality reduction," *IEEE transactions on neural networks and learning systems*, vol. 27, no. 4, pp. 723–735, 2016.
10. Hoai An Le Thi and Duy Nhat Phan, "Dc programming and dca for sparse optimal scoring problem," *Neurocomputing*, vol. 186, pp. 170–181, 2016.
11. Zhe Bai, Steven L Brunton, Bingni W Brunton, J Nathan Kutz, Eurika Kaiser, Andreas Spohn, and Bernd R Noack, "Data-driven methods in fluid dynamics: Sparse classification from experimental data," in *Whither Turbulence and Big Data in the 21st Century?*, pp. 323–342. Springer, 2017.

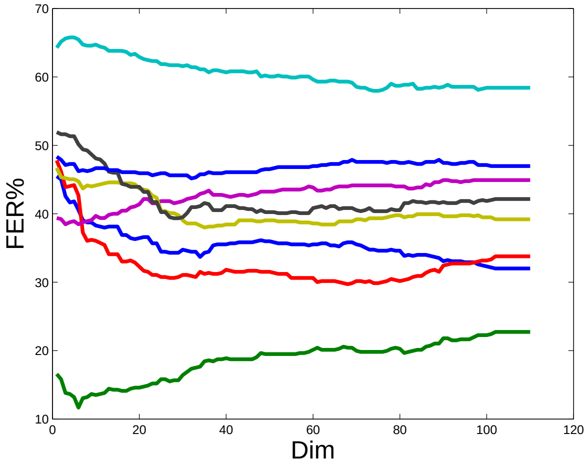

Figure 3 FER on each subject, with SDA based on features of all channels. Models are trained on data of all subjects.

12. Victoria Peterson, Hugo Leonardo Rufiner, and Ruben Daniel Spies, "Kullback-leibler penalized sparse discriminant analysis for event-related potential classification," *arXiv preprint arXiv:1608.06863*, 2016.
13. Qiang Wu, Yu Zhang, Ju Liu, Jiande Sun, and Jie Li, "Sparse optimal score based on generalized elastic net model for brain computer interface," in *Information Science and Technology (ICIST), 2016 Sixth International Conference on*. IEEE, 2016, pp. 66–71.
14. Karl Sjöstrand, Line Harder Clemmensen, Rasmus Larsen, and Bjarne Ersbøll, "Spasm: A matlab toolbox for sparse statistical modeling," *Journal of Statistical Software Accepted for publication*, 2012.
15. Keinosuke Fukunaga, *Introduction to statistical pattern recognition*, Academic press, 1990.
16. Nagendra Kumar and Andreas G Andreou, "Heteroscedastic discriminant analysis and reduced rank hmms for improved speech recognition," *Speech communication*, vol. 26, no. 4, pp. 283–297, 1998.
17. Nagendra Kumar and Andreas G Andreou, *Investigation of silicon auditory models and generalization of linear discriminant analysis for improved speech recognition*, Ph.D. thesis, Johns Hopkins University, 1997.
18. Line Clemmensen, Trevor Hastie, Daniela Witten, and Bjarne Ersbøll, "Sparse discriminant analysis," *Technometrics*, vol. 53, no. 4, pp. 406–413, 2011.
19. Simon JD Prince and James H Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *2007 IEEE 11th International Conference on Computer Vision*. IEEE, 2007, pp. 1–8.
20. Sergey Ioffe, "Probabilistic linear discriminant analysis," in *European Conference on Computer Vision*. Springer, 2006, pp. 531–542.
21. Robert Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B*, pp. 267–288, 1996.
22. Hui Zou and Trevor Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B*, vol. 67, no. 2, pp. 301–320, 2005.
23. Jinbo Bi, Kristin P. Bennett, Mark Embrechts, Curt Breneman, and Minghu Song, "Dimensionality reduction via sparse support vector machines," *The Journal of Machine Learning Research*, vol. 3, pp. 1229–1243, 2003.
24. Mingkui Tan, Li Wang, and Ivor W Tsang, "Learning sparse SVM for feature selection on very high dimensional datasets," in *Proceedings of the International Conference on Machine Learning*, 2010, pp. 1047–1054.
25. Yulan Liang, John D. Lee, and Michelle L. Reyes, "Nonintrusive detection of driver cognitive distraction in real time using bayesian networks," *Transportation Research Record Journal of the Transportation Research Board*, vol. 2018, no. 2018, pp. 1–8, 2007.
26. Yulan Liang, M. L. Reyes, and J. D. Lee, "Real-time detection of driver cognitive distraction using support vector machines," *Intelligent Transportation Systems IEEE Transactions on*, vol. 8, no. 2, pp. 340–350, 2007.

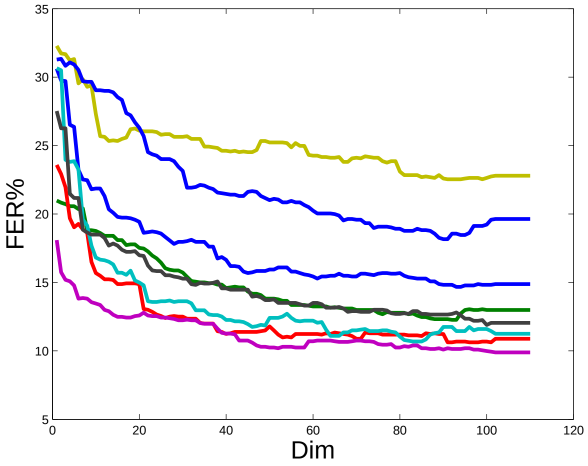

Figure 4 FER on training data of each subject, with SDA based on features of all channels. Models are trained on data of each subject.

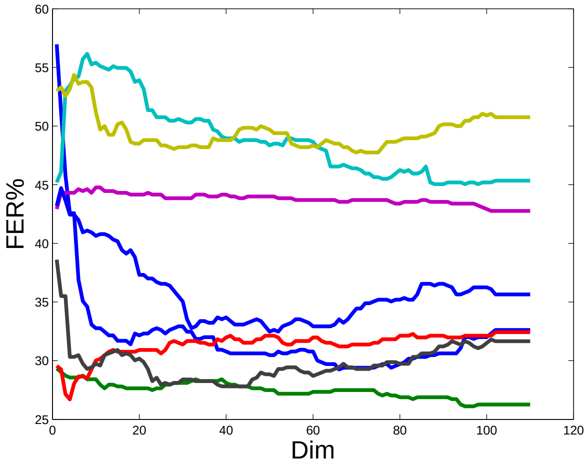

Figure 5 FER on test data of each subject, with SDA based on features of all channels. Models are trained on data of each subject.