Two-Step Linear Discriminant Analysis for Classification of EEG Data

Nguyen Hoang Huy, Stefan Frenzel, and Christoph Bandt

Abstract We introduce a multi-step machine learning approach and use it to classify electroencephalogram (EEG) data. This approach works very well for high-dimensional spatio-temporal data with separable covariance matrix. At first all features are divided into subgroups and linear discriminant analysis (LDA) is used to obtain a score for each subgroup. Then LDA is applied to these scores, producing the overall score used for classification. In this way we avoid estimation of the high-dimensional covariance matrix of all spatio-temporal features. We investigate the classification performance with special attention to the small sample size case. We also present a theoretical error bound for the normal model with separable covariance matrix, which results in a recommendation on how subgroups should be formed for the data.

1 Introduction

Fisher's classical linear discriminant analysis (LDA) is still one of the most widely used techniques for data classification. For two normal distributions with common covariance matrix Σ and different means μ_1 and μ_2, LDA classifier achieves the minimum classification error rate. The LDA score or discriminant function δ of an observation X is given by

$$\delta(X) = (X-\mu)^T \Sigma^{-1} \alpha \;\text{ with }\; \alpha = \mu_1 - \mu_2 \;\text{ and }\; \mu = \frac{1}{2}(\mu_1 + \mu_2)\,.$$

In practice we do not know Σ and μ_i, and have to estimate them from training data. This worked well for low-dimensional examples, but estimation of Σ for high-dimensional data turned out to be really difficult. When the sample size n of the

Nguyen Hoang Huy · Stefan Frenzel · Christoph Bandt
Department of Mathematics and Computer Science, University of Greifswald, Germany
e-mail: nhhuy@hua.edu.vn

training data is smaller than the number d of features, then the empirical $d \times d$ covariance matrix $\hat{\Sigma}$ is not invertible. The pseudoinverse can be used but this will impair the classification. Even when n is larger, but of the same magnitude as d, the aggregated estimation error over many entries of the sample covariance matrix will significantly increase the error rate of LDA. For our EEG data, where $200 \leq n \leq 3500$ and $160 \leq d \leq 1280$, these facts will be discussed below in Sect. 5 and Fig. 3.

One possible solution of the estimation problem is regularized LDA, where a multiple of the unity matrix I is added to the empirical covariance. $\hat{\Sigma} + rI$ is invertible for each $r > 0$. The most useful regularization parameter r has to be determined by time-consuming optimization, however.

Bickel and Levina (2004) recommended a simpler solution: to neglect all correlations of the features and use the diagonal matrix D_Σ of Σ instead of Σ. This is called the independence rule. Its discriminant function δ_I is defined by

$$\delta_I(X) = (X - \mu)^T D_\Sigma^{-1} \alpha.$$

In this paper, we present another solution, which uses some but not all correlations of the features and which worked very well for the case of spatio-temporal data, in the context of an experiment with a brain-computer interface.

2 Two-step Linear Discriminant Analysis

We introduce multi-step linear discriminant analysis which applies LDA in several steps instead of applying it to all features at one time. Here we consider the case of two steps (two-step LDA). All d features of an observation $X \in \mathbb{R}^d$ are divided into disjoint subgroups

$$X = \left[X_1^T, \cdots, X_q^T\right]^T,$$

where $X_j \in \mathbb{R}^p$, and $pq = d$. LDA is applied to obtain a score for each subgroup of features. In the second step, LDA is again applied to these scores which gives the overall score used for classification. Thus the discriminant function of two-step LDA is

$$\delta^\star(X) = \delta(\delta(X_1), \cdots, \delta(X_q)),$$

where δ denotes the LDA function. Fig. 1a illustrates the two-step LDA procedure. The assumption of normality which is needed for LDA will be fulfilled in the second step. The distribution of scores can be calculated applying basic linear algebra and the properties of the multivariate normal distribution.

Proposition 1. *Suppose X is normally distributed with known μ_1, μ_2 and Σ. Let $\mu_2 - \mu_1 = (\alpha_1^T, \ldots, \alpha_q^T)^T$ and $\Sigma_{ij} \in \mathbb{R}^{p \times p}$ denote the submatrix of Σ corresponding to subgroups i and j such that $\Sigma = (\Sigma_{ij})_{i,j=1}^q$. The scores $(\delta(X_1), \ldots, \delta(X_q))^T$ are then normally distributed with common covariance matrix Θ and means $\pm(1/2)m$ given by*

$$\Theta_{ij} = \alpha_i^T \Sigma_{ii}^{-1} \Sigma_{ij} \Sigma_{jj}^{-1} \alpha_j, \;\; m_i = \Theta_{ii}, \;\; \text{with} \;\; i, j = 1, \ldots, q.$$

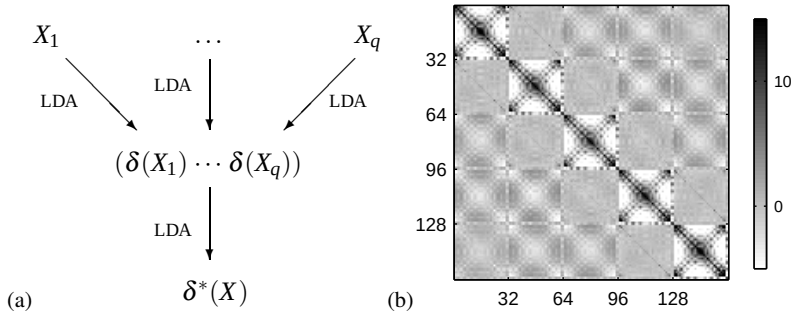

Fig. 1 (a) Schematic illustration of two-step LDA. (b) Sample covariance matrix of a single dataset estimated from 5 time points and 32 locations.

3 Separable Models

Statistical modelling of spatio-temporal data often is based on separable models which assume that the covariance matrix of the data is a product of spatial and temporal covariance matrices. This greatly reduces the number of parameters in contrast to unstructured models. Genton (2007) argues that separable approximations can be useful even when dealing with non-separable covariance matrices.

A spatio-temporal random process $X(\cdot,\cdot) : S \times T \to \mathbb{R}$ with time domain $T \subset \mathbb{R}$ and space domain $S \subset \mathbb{R}^3$ is said to have a separable covariance function if, for all $s_1, s_2 \in S$ and $t_1, t_2 \in T$, it holds

$$\mathrm{Cov}(X(s_1,t_1),X(s_2,t_2)) = u(t_1,t_2) \cdot v(s_1,s_2), \tag{1}$$

where u and v is the temporal and spatial covariance function, respectively. Suppose that the data from $X(\cdot,\cdot)$ is only selected at a finite set of locations $s_1,\ldots,s_p$ and time points $t_1,\ldots,t_q$. An observation for classification is obtained by concatenation of spatial data vectors at times $\{t_1,\cdots,t_q\}$

$$X = [X(s_1;t_1)\cdots X(s_p;t_1)\cdots X(s_1;t_q)\cdots X(s_p;t_q)]^T. \tag{2}$$

Equation (1) says that the covariance matrix of X can be written as Kronecker product of the spatial covariance matrix V with entries $v_{ij} \equiv v(s_i,s_j)$ and the temporal covariance matrix U with $u_{ij} \equiv u(t_i,t_j)$,

$$\Sigma = U \otimes V.$$

In the context of EEG, the locations $s_1,\ldots,s_p$ are defined by the electrode positions on the scalp. Huizenga et al. (2002) demonstrated that separability is a proper assumption for this kind of data. Fig. 1b visualizes the Kronecker product structure of the covariance matrix of one of our data sets. There are $p = 32$ electrodes and $q = 5$ time points. Each of the five blocks on the diagonal represents the covariance

between the electrodes for a single time point. The other blocks represent covariance for different time points.

4 An Error Bound for Two-step LDA

In this section we derive a theoretical error estimate for two-step LDA in the case of separable models. The following theorem, illustrated in Fig. 2a, shows that the loss in efficiency of two-step LDA in comparison to ordinary LDA even in the worst case is not very large when the condition number of the temporal correlation matrix is moderate. The assumption that the means and covariance matrices are known may seem a bit unrealistic, but it is good to have such a general theorem. The numerical results in Sect. 5 will show that the actual performance of two-step LDA for finite samples is much better. To compare the error rate of δ and $\delta^\star$, we use the technique of Bickel and Levina (2004) who compared independence rule and LDA in a similar way.

Theorem 1. *Suppose that mean vectors μ_1, μ_2 and common separable covariance matrix $\Sigma = U \otimes V$ are known. Then the error rate e_2 of the two-step LDA fulfils*

$$e_1 \le e_2 \le \Phi\left(\frac{2\sqrt{\kappa}}{1+\kappa}\,\Phi^{-1}(e_1)\right), \tag{3}$$

where e_1 is the LDA error rate, $\kappa = \kappa(U_0)$ denotes the condition number of the temporal correlation matrix $U_0 = D_U^{-1/2} U D_U^{-1/2}$, $D_U = diag(u_{11}, \cdots, u_{qq})$, and Φ is the Gaussian cumulative distribution function.

Proof. $e_1 \le e_2$ follows from the optimality of LDA. To show the other inequality, we consider the error $\bar{e}$ of the two-step discriminant function $\bar{\delta}$ defined by

$$\bar{\delta}(X) = \delta_I(\delta(X_1), \cdots, \delta(X_q)),$$

where δ_I is the discriminant function of the independence rule. The relation $e_2 \le \bar{e}$ again follows from the optimality of LDA and Proposition 1. We complete the proof by showing that $\bar{e}$ is bounded by the right-hand side of (3), by the technique of Bickel and Levena (2004). We repeat their argument in our context, demonstrating how U_0 comes up in the calculation. We rewrite the two-step discriminant function $\bar{\delta}$ applied to the spatio-temporal features X with $\alpha = \mu_1 - \mu_2$ and $\mu = (\mu_1 + \mu_2)/2$

$$\bar{\delta}(X) = (X-\mu)^T \bar{\Sigma}^{-1} \alpha, \text{ where } \bar{\Sigma} = D_U \otimes V = \begin{bmatrix} u_{11}V & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & u_{qq}V \end{bmatrix}.$$

The errors e_1 of $\delta(x)$ and $\bar{e}$ of $\bar{\delta}(x)$ are known, see Bickel and Levina (2004):

$$e_1 = \Phi\left(\frac{-(\alpha^T\Sigma^{-1}\alpha)^{1/2}}{2}\right), \quad \bar{e} = \Phi\left(\frac{-\alpha^T\bar{\Sigma}^{-1}\alpha}{2(\alpha^T\bar{\Sigma}^{-1}\Sigma\bar{\Sigma}^{-1}\alpha)^{1/2}}\right).$$

Writing $\alpha_0 = \bar{\Sigma}^{-1/2}\alpha$, we determine the ratio

$$r = \frac{\Phi^{-1}(\bar{e})}{\Phi^{-1}(e_1)} = \frac{(\alpha_0^T\alpha_0)}{[(\alpha_0^T\widetilde{\Sigma}\alpha_0)(\alpha_0^T\widetilde{\Sigma}^{-1}\alpha_0)]^{1/2}}, \tag{4}$$

where

$$\begin{aligned}\widetilde{\Sigma} = \bar{\Sigma}^{-1/2}\Sigma\bar{\Sigma}^{-1/2} &= (D_U^{-1/2}\otimes V^{-1/2})(U\otimes V)(D_U^{-1/2}\otimes V^{-1/2}) \\ &= (D_U^{-1/2}UD_U^{-1/2})\otimes(V^{-1/2}VV^{-1/2}) = U_0\otimes I.\end{aligned}$$

Clearly $\widetilde{\Sigma}$ is a positive definite symmetric matrix and its condition number $\kappa(\widetilde{\Sigma})$ is equal to the condition number $\kappa = \kappa(U_0)$ of the temporal correlation matrix U_0. In the same way as Bickel and Levina we obtain from (4) by use of the Kantorovich inequality $r \geq 2\sqrt{\kappa}/(1+\kappa)$. With (4) and $\Phi^{-1}(e_1) < 0$ this implies

$$\bar{e} \leq \Phi\left(\frac{2\sqrt{\kappa}}{1+\kappa}\,\Phi^{-1}(e_1)\right), \text{ which completes the proof.}$$

5 Classification of EEG Data

To check the performance of two-step LDA, we use the data of a brain-computer interface experiment by Frenzel et al. (2011). A mental typewriter was established using 32-electrode EEG. Users sit in front of a screen which presents a matrix of characters. They are instructed to concentrate on one target character by performing a mental count. Then characters are highlighted many times in random order. About 300 ms after highlighting the target character, a so-called event-related potential should appear in the EEG signal. This potential should not appear for the other characters.

The experiment intended to control the effect of eye view. Users were told to concentrate their eyes on a specific character. When this is the target character, the condition is described as overt attention, and the expected potential is fairly easy to identify. However, most of the time users looked at a different character and counted the target character in their visual periphery. This is referred to as covert attention, see Treder and Blankertz (2010). Controlling a brain-computer interface by covert attention is particularly difficult.

Detection of target characters from the EEG data is a binary classification problem. Each time interval where a character was highlighted is considered as a sample. Class labels are defined according to whether the target character is presented or not. Our data consists of nine datasets of $m = 7290$ samples measured with $p = 32$ electrodes. For each sample, data of the time interval of about 600 ms were downsam-

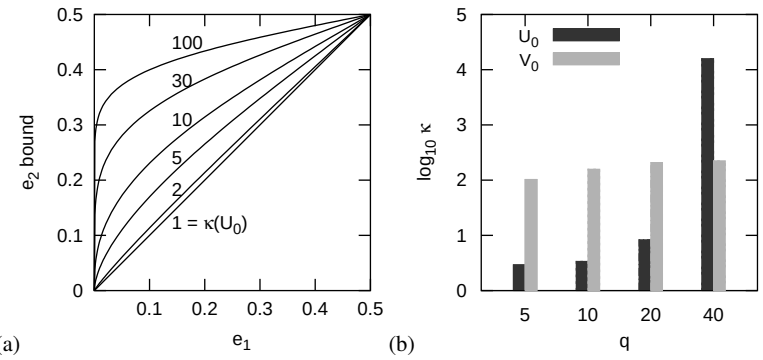

Fig. 2 (a) The error bound of two-step LDA as function of the LDA error rate. (b) Condition numbers of U_0 and V_0 estimated from a single dataset for different number of time points q.

pled from the acquisition rate of the hardware to a predefined sampling frequency. For typical values of 8, 16, 32, 64 Hz one obtains $q = 5, 10, 20, 40$ time points and thus $d = pq = 160, 320, 640, 1280$ spatio-temporal features in total.

Defining the Feature Subgroups of Two-step LDA

LDA is invariant with respect to reordering of features whereas two-step LDA is only when reordering is performed within the subgroups. For the latter we saw that it is preferable to define the subgroups such that the statistical dependencies between them are smaller than within. This is reflected in the influence of condition number of U_0 in the bound of the error rate (3).

In Sect. 3 we defined the features to be ordered according to their time index, see (2), and X_i to contain all features at time point i. In other words, in the first step LDA was applied to the spatial features. However, it also seems natural to order the features according to their spatial index and to assign all features from electrode i to X_i, thus interchanging the role of space and time. In this case the covariance matrix becomes $V \otimes U \neq U \otimes V$ and we have to replace $\kappa(U_0)$ by $\kappa(V_0)$ in (3). We argue that the decision between both approaches should be based on a comparison of both condition numbers using the data. This done in the following.

Our EEG data is, rather typical, normalized such that the means over all time points and all electrodes are zero. This implies U and V both to have a single zero eigenvalue. Maximum-likelihood estimation of both in general requires their inverses to exist, see Mitchell et al. (2006). We bypassed this problem by using the simple average-based estimator

$$\hat{V} = \frac{1}{q} \sum_{i=1}^{q} \hat{\Sigma}_{ii} \ ,$$

where $\hat{\Sigma}_{ii}$ is the sample covariance matrix of the i-th subgroup. It can be shown that $\hat{V}$ is an unbiased and consistent estimator of $\bar{\lambda}V$ with $\bar{\lambda}$ being the average eigenvalue of U. Since the correlation matrix corresponding to $\bar{\lambda}V$ is V_0 we estimated $\kappa(V_0)$ by $\kappa(\hat{V}_0)$, ignoring the single zero eigenvalue. Estimation of $\kappa(U_0)$ was done in the same way.

Fig. 2b shows the condition numbers estimated from a single dataset for different number of time points q. Except for $q = 40$ the condition numbers of U_0 were much smaller than those of V_0. This also applied for the corresponding error bounds, see Fig. 2a. It is thus likely that the actual error rates are smaller. Indeed, we never encountered a single case where first applying LDA to the temporal features gave better results for our data. For $q = 40$ the upper bounds were too loose to draw any conclusions. This could be observed in all nine datasets and gives rise to the following recommendation.

Remark 1. When two-step LDA is applied to EEG data it is preferable to define the feature subgroups such that X_i contains all features at time point i.

Learning Curves

We investigated the classification performance using $p = 32$ electrodes and $q = 40$ time points and hence $d = 1280$ features in total. Two-step LDA was compared to

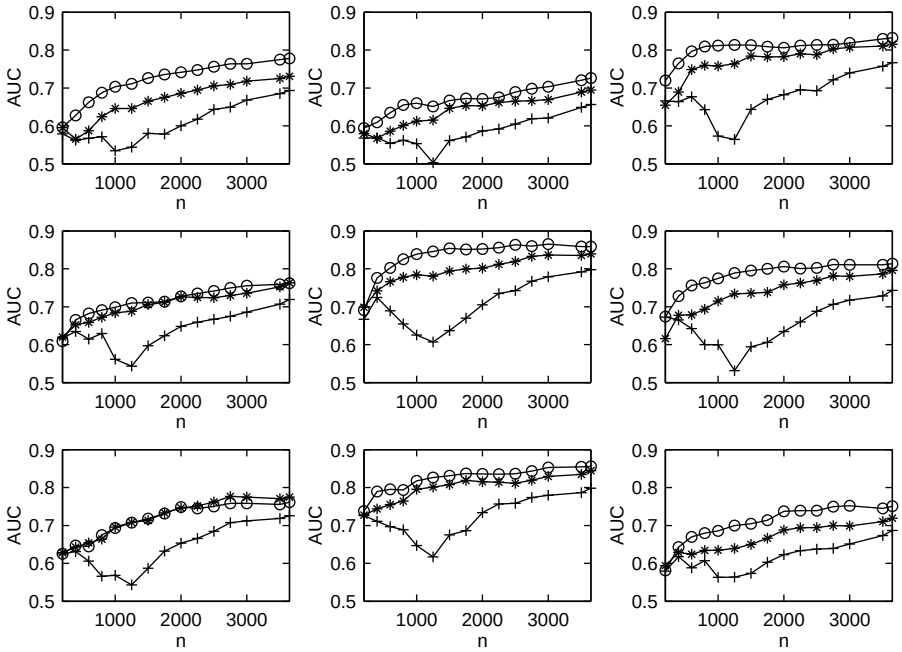

Fig. 3 Learning curves of two-step LDA (o), regularized LDA $(*)$ and LDA $(+)$ for all nine datasets.

ordinary and regularized LDA. For each dataset classifiers were trained using the first n samples, with $200 \leq n \leq 3500$. Scores of the remaining $m-n$ samples were calculated and classification performance was measured by the AUC value, i.e. the relative frequency of target trials having a larger score than non-target ones.

Fig. 3 shows the learning curves for all nine datasets. The prominent dip in the learning curves of LDA around d is due to the use of the pseudoinverse for $n < d+2$. Regularized LDA for $n \approx d$ performed much better than LDA, supporting the findings of Blankertz et al. (2011).

Two-step LDA showed similar or slightly better performance than both regularized and ordinary LDA. For large n the difference was rather small. For some datasets, however, it showed much faster convergence, i.e. it needed less training samples to achieve a certain classification performance. Sample size $n = 3500$ corresponds to a training period of approximately 20 minutes. Since two-step LDA gave reasonable performance even with short training periods, it might offer an practically relevant advantage. Although all three classifiers are computationally cheap, it should be noted that two-step LDA does not involve the inversion of the full sample covariance matrix.

6 Conclusion

When linear discriminant analysis is applied to high-dimensional data, it is difficult to estimate the covariance matrix. We introduced a method which avoids this problem by applying LDA in two steps for the case of spatio-temporal data. For our EEG data, the two-step LDA performed better than regularized LDA.

References

1. BICKEL, P. J. and LEVINA, E. (2004): Some theory of Fisher's linear discriminant function, 'naive Bayes', and some alternatives when there are many more variables than observations. *Bernoulli, 10, 989–1010*.
2. BLANKERTZ, B., LEMM, S., TREDER, M., HAUFE, S and MÜLLER K. R. (2011): Single-trial analysis and classification of ERP components – A tutorial. *NeuroImage, 56, 814–825*.
3. FRENZEL, S., NEUBERT, E. and BANDT, C. (2011): Two communication lines in a 3×3 matrix speller. *Journal of Neural Engineering, 8, 036021*.
4. GENTON, M. G. (2007): Separable approximation of space-time covariance matrices. *Environmetrics, 18, 681–695*.
5. HUIZENGA, H. M., DE MUNCK, J. C., WALDORP, L. J., and GRASMAN, R. P. P. P. (2002): Spatiotemporal EEG/MEG Source Analysis Based on a Parametric Noise Covariance Model. *IEEE Transactions on Biomedical Engineering, 49, 533–539*.
6. MITCHELL, M. W., GENTON, M. G. and GUMPERTZ, M. L. (2006): A likelihood ratio test for separability of covariances. *Journal of Multivariate Analysis, 97, 1025–1043*.
7. TREDER, M. S. and BLANKERTZ, B. (2010): (C)overt attention and visual speller design in an ERP-based brain-computer interface. *Behavioral and Brain Functions, 6, 28*.