Some Linear Classifiers for High-Dimensional Data

Nguyen Hoang Huy

Department of Applied Maths & Informatics, Hanoi University of Agriculture

January 6, 2014

Contents

Introduction

- Each object is characterized by its vector of measurements $\boldsymbol{x} = [x_1, \ldots, x_p]^T \in \mathbb{R}^p$ and response class $k \in \{1, 2\}$
  - $p =$ the number of variables
- Given **training data** $\mathfrak{G} = \{(\boldsymbol{x}^{(i)}, k_i),\ i = 1, \ldots, n\}$
  - $n =$ the training sample size

The Problem of Classification

- *Finding: a* **classification function** $g : \mathbb{R}^p \to \{1, 2\}$, *which can predict the unknown class* k *of new observation* $\boldsymbol{x} \in \mathbb{R}^p$ *using available training data as accurately as possible*

- **Assuming:** $\{(\boldsymbol{x}^{(i)}, k_i),\ (\boldsymbol{x}, k),\ i = 1, \ldots, n\}$ are independent and come from a certain distribution
- **The error rate** of a classification function g for a new observation $\boldsymbol{x}$ in class k is

$$\overline{W}(g) = \mathrm{P}(g(\boldsymbol{x}) \neq k)$$

Introduction

The Dimension p of $\boldsymbol{x}$

- *In classical applications, p is small (a few variables)*
- *Modern technologies: a large p (many variables)*
    - *data from brain computer interfaces*
    - *genetic and microarray data*
    - *high-frequency financial data*

Example: Brain-Computer Interface Data, Frenzel et al. (2011)

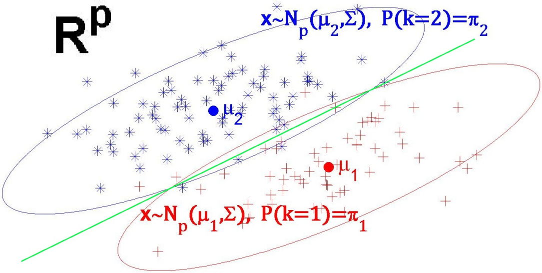
- *Distinguishing brain states into two classes*
- *Classification based on ElectroEncephaloGraphy (EEG) signals*
    - *over* 30 000 *scalp potential values*
    - $p = 1024$ *values after preprocessing*
- *As small training sample sizes as possible because of the online application requirements: n only in thousands or hundreds*

Introduction

Example: Classifying human acute leukemias into two types

- *Gene expression microarray, see Golub et al. (1999)*
- *Two types of human acute leukemias*
  - *acute myeloid leukemia (AML)*
  - *acute lymphoblastic leukemia (ALL)*
- *Distinguishing ALL from AML is crucial for successful treatment*
- *Classification based solely on gene expression monitoring*
  - *over* $7,000$ *genes*
  - $p = 1,714$ *genes after an initial screening*
- *A training data set:*
  - 47 *ALL*
  - 25 *AML*
- p *can be much larger than training sample sizes*

Wald's Approach (1944)



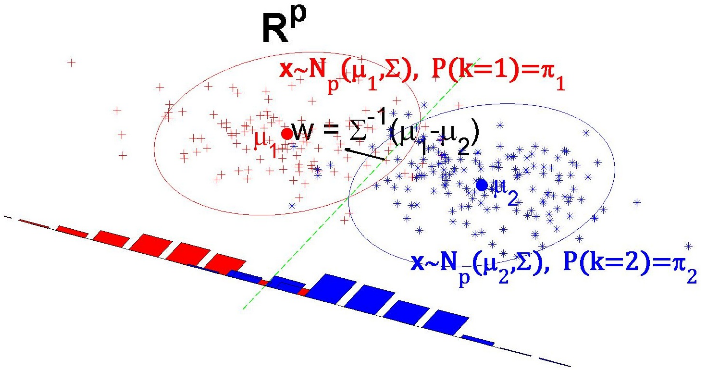
Finding the best classification function

Classify an object to class 1 *or class* 2 *based on its observed vector*

$$\boldsymbol{x} \sim \mathcal{N}_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}) \text{ or } \boldsymbol{x} \sim \mathcal{N}_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$$

$\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$: *the p-dimensional normal distribution with mean vector* $\boldsymbol{\mu}$ *and covariance matrix* $\boldsymbol{\Sigma}$

When $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ and $\boldsymbol{\Sigma}$ are known



- Fisher's discriminant function δ_F of observation $\boldsymbol{x}$ is given by

$$\delta_F(\mathbf{x}) = \boldsymbol{x}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\alpha},\ \boldsymbol{\alpha} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 \tag{1}$$

- δ_F projects $\boldsymbol{x}$ down to one dimension using $\delta_F(\mathbf{x}) = \mathbf{w}^T \mathbf{x}$

When $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ and $\boldsymbol{\Sigma}$ are known

- The classification function of Fisher's Linear Discriminant Analysis (LDA) is defined by

$$g(\boldsymbol{x}) = \begin{cases} 1, & \delta_F(\boldsymbol{x}) \geq w_0 \\ 2, & \delta_F(\boldsymbol{x}) < w_0, \end{cases}$$

  where $w_0 = \delta_F(\boldsymbol{\mu}) + \log(\pi_2/\pi_1)$, $\boldsymbol{\mu} = (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)/2$
- LDA is the Bayes classifier
- If $\pi_1 = \pi_2 = 1/2$ then $w_0 = \delta_F(\boldsymbol{\mu})$, the error rate of LDA is

$$\overline{W}(\delta_F) = \bar{\Phi}(d_p/2), \;\; d_p = [\boldsymbol{\alpha}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\alpha}]^{1/2}$$

  $\bar{\Phi}(t) = 1 - \Phi(t)$: the tail probability of the standard normal distribution, d_p : the Mahalanobis distance between two classes
- The larger dimension p is the better

$$\lim_{d_p \to \infty} \overline{W}(\delta_F) = 0$$

When $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ and $\boldsymbol{\Sigma}$ are known

Remark 1 (Cai and Liu (2011))

- *Write* $\boldsymbol{\alpha} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = \begin{bmatrix} \boldsymbol{\alpha}_0 \\ \mathbf{0} \end{bmatrix}$, $\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12}^T \\ \boldsymbol{\Sigma}_{12} & \boldsymbol{\Sigma}_{22} \end{bmatrix}$, *then* $d_p^2 = \boldsymbol{\alpha}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\alpha}$ *can be decomposed as follows:*

$$d_p^2 = \boldsymbol{\alpha}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\alpha} = \boldsymbol{\alpha}_0^T \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\alpha}_0 + (\boldsymbol{B}\boldsymbol{\alpha}_0)^T \boldsymbol{W}^{-1} (\boldsymbol{B}\boldsymbol{\alpha}_0), \quad (2)$$

where $\boldsymbol{B} = \boldsymbol{\Sigma}_{22}^{-1} \boldsymbol{\Sigma}_{12}$. *Since* $\boldsymbol{W} = \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{12} \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\Sigma}_{12}^T$ *is positive definite, if* $\boldsymbol{B}\boldsymbol{\alpha}_0 \neq \mathbf{0}$, *then the last term in (2) is positive and*

$$d_p^2 > \boldsymbol{\alpha}_0^T \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\alpha}_0$$

- Some components of $\boldsymbol{x}$ contribute to classification through their correlations with others although they have no mean effects

When $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ and $\boldsymbol{\Sigma}$ are unknown

Statistical Issue

- *We have training data of size n*

$$\mathfrak{G} = \{(\boldsymbol{x}^{(i)}, k_i),\ i = 1, \ldots, n\},\ \#\mathfrak{G} = n$$
$$\boldsymbol{x}^{(i)} \in \mathbb{R}^p,\ k_i \in \{1, 2\}$$

$$\boldsymbol{x}^{(i)} \overset{i.i.d.}{\sim} \begin{cases} \mathcal{N}_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}), & \text{for } k_i = 1 \\ \mathcal{N}_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}), & \text{for } k_i = 2 \end{cases}$$

- $\mathfrak{G}$ *is independent of* $\boldsymbol{x}$
- *How to use the training data* $\mathfrak{G}$ *to construct a classifier having the error rate close to the optimal error rate* $\overline{W}(\delta_F)$

- Checking assumption:
  - **How to test $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$?**

When $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ and $\boldsymbol{\Sigma}$ are unknown

Classical Applications: Fixed-p-large-n

- *Replacing unknown $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$, and $\boldsymbol{\Sigma}$ by*

$$\hat{\boldsymbol{\mu}}_k = \frac{1}{n_1} \sum_{k_i = k} \mathbf{x}^{(i)},\ n_k = \#\{i : k_i = k\},\ k = 1, 2$$

$$n = n_1 + n_2,\ \hat{\boldsymbol{\Sigma}} = \frac{1}{n-2} \sum_{k=1}^{2} \sum_{k_i = k} (\mathbf{x}^{(i)} - \hat{\boldsymbol{\mu}}_k)(\mathbf{x}^{(i)} - \hat{\boldsymbol{\mu}}_k)^T,$$

we obtain the sample Fisher's discriminant function

$$\hat{\delta}_F(\boldsymbol{x}) = \boldsymbol{x}^T \hat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\alpha}},\ \hat{\boldsymbol{\alpha}} = \hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2$$

- What kind of p (which may diverge to ∞) is LDA efficient?

Linear Discriminant Analysis and Asymptotic Results

Regularity Conditions

There is a constant c_0 (not depending on p or n) such that

- $c_0^{-1} \leq$ *all eigenvalues of* $\boldsymbol{\Sigma} \leq c_0$
- $c_0^{-1} \leq \max_{j \leq p} \alpha_j^2 \leq c_0$, α_j *is the j-th component of* $\boldsymbol{\alpha} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$

Asymptotic Setting

- $n = n_1 + n_2$, $n_1/n \to c \in (0,1)$ *as* $n \to \infty$
- p *is a function of* n, $p/n \to b \in [0, \infty]$ *as* $n \to \infty$

Asymptotic Optimality ($n \to \infty$)

- *The sample LDA is asymptotically optimal if* $\overline{W}(\hat{\delta}_F \mid \mathfrak{G}) / \overline{W}(\delta_F) \xrightarrow{\text{P}} 1$
- *The sample LDA is asymptotically sub-optimal if* $\overline{W}(\hat{\delta}_F \mid \mathfrak{G}) \xrightarrow{\text{P}} \overline{W}(\delta_F)$
- *The sample LDA is asymptotically worst if* $\overline{W}(\hat{\delta}_F \mid \mathfrak{G}) \xrightarrow{\text{P}} 1/2$

Linear Discriminant Analysis ($p < n$)

Theorem 1 (Shao et al. (2011))

Suppose that $s_n = p\sqrt{\log p}/\sqrt{n} \to 0$.

(i) *The conditional error rate of the sample LDA is equal to*

$$\overline{W}(\hat{\delta}_F \mid \mathfrak{G}) = \bar{\Phi}([1 + O_{\mathrm{P}}(s_n)]d_p/2).$$

(ii) *If* d_p *is bounded, then the sample LDA is asymptotically optimal and*

$$\frac{\overline{W}(\hat{\delta}_F \mid \mathfrak{G})}{\overline{W}(\delta_F)} - 1 = O_{\mathrm{P}}(s_n).$$

(iii) *If* $d_p \to \infty$, *then the sample LDA is asymptotically sub-optimal.*

(iv) *If* $d_p \to \infty$ *and* $s_n d_p^2 \to 0$, *then the sample LDA is asymptotically optimal.*

Modern Application: Large-p-not-so-large-n

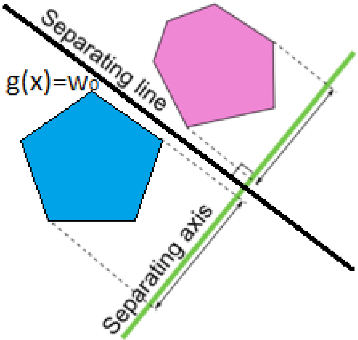
- "High-dimensional data are nowadays rule rather than exception in areas like information technology, bioinformatics or astronomy, to name just a few", see P. Bühlmann and S. van de Geer (2011)
- A large p results in more information, but produces more uncertainty when the distribution of $\boldsymbol{x}$ is unknown, see Shao et al. (2011)
- **Our problem:** Estimation of covariance matrix $\boldsymbol{\Sigma}$
  - with $p = 1024$, $\approx 10^6$ entry parameters
  - but sample size n only in thousands or hundreds
- Bickel and Levina (2004) showed that the sample LDA is as bad as random guessing when $\frac{p}{n} \to \infty$

Separating Hyperplane When p is Large

Is optimization enough to find separating hyperplanes for test data?

Theorem 2 (Hahn-Banach)

$K_1 = \text{conv}\{\boldsymbol{x}^{(i)} | k_i = 1\}, K_2 = \text{conv}\{\boldsymbol{x}^{(i)} | k_i = 2\}$. *If* $K_1 \cap K_2 = \emptyset$, *there exists a separating hyperplane* $g(\boldsymbol{x}) = w_0$ *for training data*



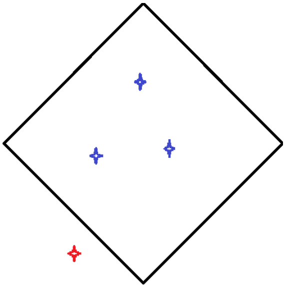
Separating Hyperplane When p is Large

Remark 2

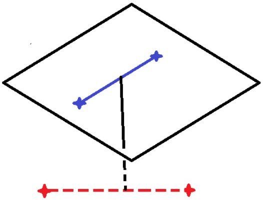
If $n > p + 1$, the problem usually cannot be solved because $K_1 \cap K_2 = \emptyset$. In high dimension $p \geq n - 1$ these sets are in general disjoint

Examples: $n = 4, p = 3$

$n_1 = 3,\ n_2 = 1$



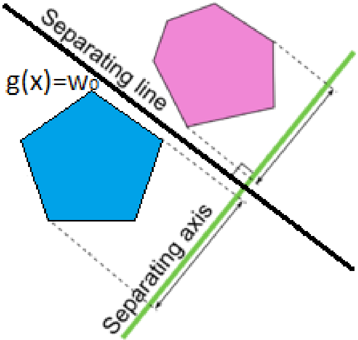
$n_1 = n_2 = 2$



Separating Hyperplane When p is Large

Remark 3

Determining separating hyperplanes by linear programming



LDA When p is Large

- Ignore the dependence
  - (i) Independence rule, see Bickel and Levina (2004)
  - (ii) Features annealed independence rule, see Fan and Fan (2008)
- Estimate $\mathbf{\Sigma}^{-1}$
  - (i) **$\mathbf{\Sigma}$ is sparse**: Estimation $\mathbf{\Sigma}^{-1}$ by the inverse of a thresholding estimate of $\mathbf{\Sigma}$, see Shao et al. (2011)
  - (ii) **$\mathbf{\Sigma}^{-1}$ is sparse**: Glasso estimator of $\mathbf{\Sigma}^{-1}$, see Rothman et al. (2008)
- See also Witten and Tibshirani (2009), Tibshirani et al. (2002), Guo et al. (2007), Wu et al. (2009), and Hall et al. (2009)

Regularized Linear Discriminant Analysis

- Regularized LDA replaces $\hat{\mathbf{\Sigma}}$ by

$$\widetilde{\mathbf{\Sigma}}(\gamma) := (1-\gamma)\widehat{\mathbf{\Sigma}} + \gamma\nu\boldsymbol{I}$$

  where $\nu = tr(\widehat{\mathbf{\Sigma}})/p$, $\gamma \in [0,1]$: a tuning parameter

- Extreme eigenvalues of $\hat{\mathbf{\Sigma}}$ are modified towards the average ν

- The optimal parameter $\gamma^\star$ can be calculated by the analytic formula as in Schäfer and Strimmer (2005)

$$\gamma^\star = \frac{n}{(n-1)} \frac{\sum_{j_1,j_2=1}^{p} \mathrm{var_k}(z_{j_1j_2}(k))}{\sum_{j_1\neq j_2} \hat{\sigma}_{j_1j_2}^2 + \sum_{j_1} (\hat{\sigma}_{j_1j_1} - \nu)^2}$$

  where $\boldsymbol{x}^{(k)} = [x_{kj}]$, common mean $\hat{\boldsymbol{\mu}} = [\hat{\mu}_j], \hat{\mathbf{\Sigma}} = [\hat{\sigma}_{j_1j_2}]$ and

$$z_{j_1j_2}(k) = (x_{kj_1} - \hat{\mu}_{j_1})(x_{kj_2} - \hat{\mu}_{j_2})$$

  or performing n-fold cross-validation, see Frenzel et al. (2010)

- The state-of-the-art classifier for Brain-Computer Interface data, see Blankertz et al. (2011)

Linear Discriminant Analysis ($p > n$)

Reason for bad performance of the LDA when $p > n$

Too many parameter in $\boldsymbol{\alpha}$, $\boldsymbol{\Sigma}$ to be estimated

Solutions?

A reasonable classifier can be obtained of both $\boldsymbol{\alpha}$ and $\boldsymbol{\Sigma}$ are sparse

Sparsity

- *Many elements of $\boldsymbol{\alpha}$ are 0 or very small*
- *Many off-diagonal elements of $\boldsymbol{\Sigma}$ are 0 or very small*
- *Both are true in many applications*

Sparse Covariance Matrices

Sparsity measure for $\boldsymbol{\Sigma}$

Bickel and Levena (2008) consider the following sparsity measure for $\boldsymbol{\Sigma}$

$$C_{h,p} = \max_{j \le p} \sum_{l=1}^{p} |\sigma_{jl}|^h$$

σ_{jl} *is the* (j,l)*th element of* $\boldsymbol{\Sigma}$, h *is a constant not depending on* p, $0 \le h < 1$

Special case of $h = 0$

$C_{0,p}$ *is the maximum of the numbers of nonzero elements of rows of* $\boldsymbol{\Sigma}$

Sparsity on $\boldsymbol{\Sigma}$

- *Not sparse:* $C_{h,p} = O(p)$
- *Sparse:* $C_{h,p} = O(\log p)$ *or* $C_{h,p} = O(n^\beta)$, $0 \le \beta < 1$

Sparse Covariance Matrices

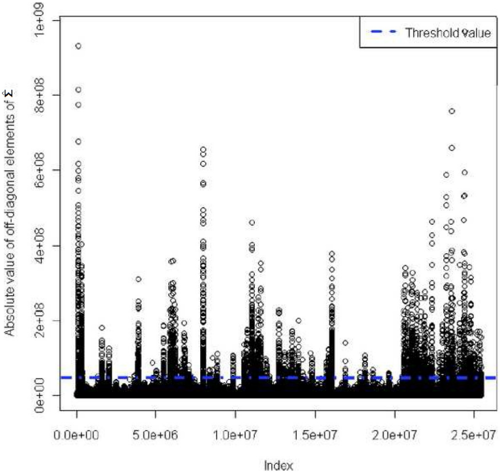


Figure : Plot of off-diagonal elements of $\hat{\mathbf{\Sigma}}$ (98.77% of off-diagonal elements of $\hat{\mathbf{\Sigma}}$ are smaller than the threshold value)

Sparse Covariance Matrices

Bickel and Levina's thresholding estimator of $\mathbf{\Sigma}$

$\hat{\mathbf{\Sigma}}$: *sample covariance matrix.* $\tilde{\mathbf{\Sigma}}$ *is* $\hat{\mathbf{\Sigma}}$ *thresholded at* $t_n = M_1\sqrt{\log p}/\sqrt{n}$ *(*M_1 *is a constant) i.e.,*

$$\textit{if } \tilde{\mathbf{\Sigma}} = [\tilde{\sigma}_{j,l}], \ \ \textit{then } \tilde{\sigma}_{j,l} = \hat{\sigma}_{j,l}\boldsymbol{I}(\hat{\sigma}_{j,l} > t_n)$$

$\hat{\sigma}_{j,l}$ *is the* (j,l)*th element of* $\hat{\mathbf{\Sigma}}$*, and* $\boldsymbol{I}$ *is the indicator function*

Consistency of $\tilde{\mathbf{\Sigma}}$

If $\log p/n \to 0$ *and* $f_n = C_{h,p}(\log p/n)^{(1-h)/2} \to 0$ *then*

$$\|\tilde{\mathbf{\Sigma}} - \mathbf{\Sigma}\| = O_{\mathrm{P}}(f_n) \ \ \textit{and} \ \ \|\tilde{\mathbf{\Sigma}}^{-1} - \mathbf{\Sigma}^{-1}\| = O_{\mathrm{P}}(f_n)$$

$\|\cdot\|$ *is the spectral norm*

Sparsity on $\boldsymbol{\alpha}$

- A large $\|\boldsymbol{\alpha}\|$ results in large difference between $\mathcal{N}_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ and $\mathcal{N}_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$
- But it also results in a more difficult task of constructing a good classification rule, since $\boldsymbol{\alpha}$ has to be estimated based on the training sample $\mathfrak{G}$

Sparsity measure for $\boldsymbol{\alpha}$

- *Shao et al (2011) consider the following sparsity measure for* $\boldsymbol{\alpha}$

$$D_{g,p} = \sum_{j=1}^{p} \alpha_j^{2g}$$

  - α_j *is the* j*th component of* $\boldsymbol{\alpha}$
  - g *is a constant not depending on* p, $0 \leq g < 1$
  - $\boldsymbol{\alpha}$ *is sparse if* $D_{g,p}$ *is much smaller than* p

Sparsity on $\boldsymbol{\alpha}$

Sparse estimator of $\boldsymbol{\alpha}$

$\tilde{\boldsymbol{\alpha}}$: $\hat{\boldsymbol{\alpha}}$ *thresholded at*

$$a_n = M_2(\log p/n)^{\xi} \;\; \textit{with constants} \;\; M_2 > 0 \;\; \textit{and} \;\; \xi \in (0, 1/2)$$

i.e., the j*th component of* $\tilde{\boldsymbol{\alpha}}$ *is* $\hat{\alpha}_j \boldsymbol{I}(|\hat{\alpha}_j| > a_n)$, $\hat{\alpha}_j$ *is the* j*th component of* $\hat{\boldsymbol{\alpha}}$

A useful result (Shao et al. (2011))

If $\log p/n \to 0$, *then*

$$\mathrm{P}(|\hat{\alpha}_j| \le a_n, j = 1, \ldots, p \;\; \textit{with} \;\; |\alpha_j| \le a_n/r) \to 1,$$

$$\mathrm{P}(|\hat{\alpha}_j| > a_n, j = 1, \ldots, p \;\; \textit{with} \;\; |\alpha_j| > a_n/r) \to 1$$

Sparsity on $\boldsymbol{\alpha}$

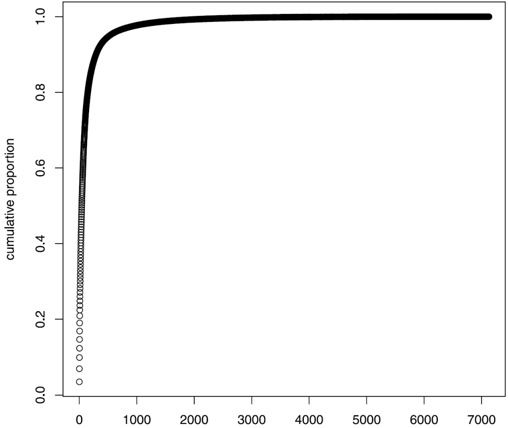


Figure : The cumulative proportions defined as $\sum_{j=1}^{l} \hat{\alpha}_{(j)}^2 / \|\hat{\boldsymbol{\alpha}}\|^2$, $l = 1, \ldots, p$ where $\hat{\alpha}_{(j)}^2$ is the jth largest value among the squared components of $\hat{\boldsymbol{\alpha}}$

Sparse linear discriminant analysis (SLDA)

Classify $\boldsymbol{x}$ to class 1 if and only if

$$\hat{\delta}_{SLDA}(\boldsymbol{x}) = \tilde{\boldsymbol{\alpha}}^T \tilde{\boldsymbol{\Sigma}}^{-1} \boldsymbol{x} \geq \hat{\delta}_{SLDA}(\boldsymbol{\mu})$$

Theorem 3 (Shao et al. (2011))

Assume $\log p/n \to 0$ *and*

$$b_n = \max\{f_n, \frac{a_n^{1-g}\sqrt{D_{g,p}}}{d_p}, \frac{\sqrt{C_{h,p}q_n}}{d_p\sqrt{n}}\} \to 0$$

$$d_p = \sqrt{\boldsymbol{\alpha}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\alpha}},\ a_n = (\log p/n)^{\xi},\ f_n = C_{h,p}(\log p/n)^{(1-h)/2}$$

$$C_{h,p} = \max_{j \leq p} \sum_{l=1}^{p} |\sigma_{jl}|^h, D_{g,p} = \sum_{j=1}^{p} \alpha_j^{2g}, q_n = \#\{j : |\alpha_j| > a_n/r\}$$

Sparse linear discriminant analysis (SLDA)

Theorem 3 (continued)

(i) *The conditional error rate of the SLDA is equal to*

$$\overline{W}(\hat{\delta}_{SLDA} \mid \mathfrak{G}) = \bar{\Phi}([1 + O_{\mathrm{P}}(b_n)]d_p/2)$$

(ii) *If d_p is bounded, then the SLDA is asymptotically optimal and*

$$\frac{\overline{W}(\hat{\delta}_{SLDA} \mid \mathfrak{G})}{\overline{W}(\delta_F)} - 1 = O_{\mathrm{P}}(b_n)$$

(iii) *If $d_p \to \infty$, then the SLDA is asymptotically sub-optimal*

(iv) *If $d_p \to \infty$ and $b_n d_p^2 \to 0$, then the SLDA is asymptotically optimal*

Applying the SLDA to human acute leukemias classification

- $p = 1,714$ genes
- $n_1 = 47,\ n_2 = 25,\ n = 72$

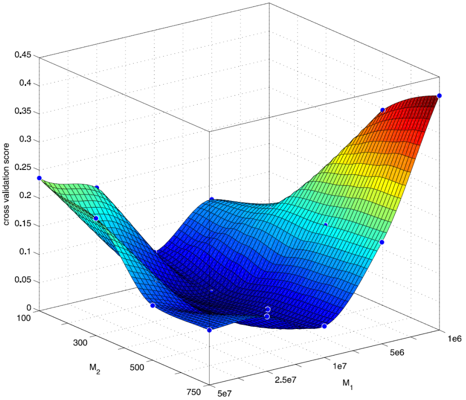


Figure : Cross-validation score vs (M_1, M_2) (Shao et al. (2011))

Cross validation estimates (Shao et al. (2011))

- Cross validation for SLDA
  - error rate is 0.0417
  - 2 of 47 cases in class 1 are misclassified
  - 1 of 25 cases in class 2 are misclassified
- Cross validation for LDA
  - error rate is 0.0694
  - 2 of 47 cases in class 1 are misclassified
  - 3 of 25 cases in class 2 are misclassified

Simulation (Shao et al. (2011))

Data are generated from $\mathcal{N}_p(\hat{\boldsymbol{\mu}}_1, \tilde{\boldsymbol{\Sigma}})$ *and* $\mathcal{N}_p(\hat{\boldsymbol{\mu}}_2, \tilde{\boldsymbol{\Sigma}})$, $n_1 = 47$, $n_2 = 25$, $p = 1,714$. *Error rates of*

- $LDA = 0.15$
- $SLDA = 0.07$
- $optimal\ rule = 0.03$

Simulation

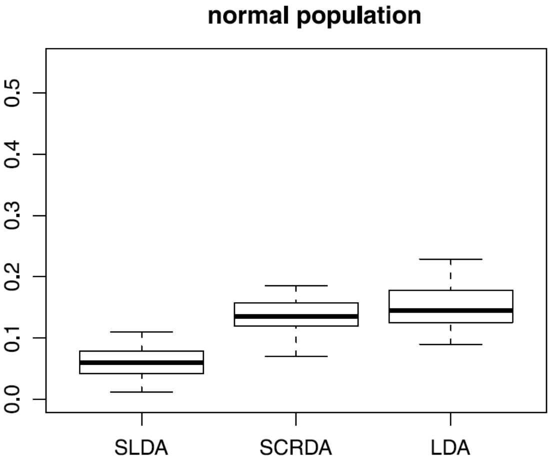


Figure : Boxplots of conditional error rates of LDA, the shrunken centroids regularized discriminant analysis (SCRDA), and SLDA

Independence Rule

- The discriminant function of Independence Rule is

$$\delta_I(\boldsymbol{x}) = \boldsymbol{x}^T \boldsymbol{D}_{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2), \text{ where } \boldsymbol{D}_{\boldsymbol{\Sigma}} = \mathrm{diag}(\boldsymbol{\Sigma})$$

- Independence Rule does not achieve the minimum error rate
- The sample version: $\hat{\delta}_I(\boldsymbol{x}) = \boldsymbol{x}^T \hat{\boldsymbol{D}}_{\hat{\boldsymbol{\Sigma}}}^{-1}(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2)$

Consider The Parameter Space

$$\Gamma = \{\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\Sigma}) : \boldsymbol{\alpha}^T \boldsymbol{D}_{\boldsymbol{\Sigma}}^{-1} \boldsymbol{\alpha} \geq C_p,\ \lambda_{\max}(\boldsymbol{R}) \leq b_0,\ \min_{1 \leq j \leq p} \sigma_j^2 > 0\},$$

where C_p is a deterministic positive sequence, $\boldsymbol{R} = \boldsymbol{D}_{\boldsymbol{\Sigma}}^{-1/2} \boldsymbol{\Sigma} \boldsymbol{D}_{\boldsymbol{\Sigma}}^{-1/2}$, b_0 is a positive constant, and σ_j^2 is the j-th diagonal element of $\boldsymbol{\Sigma}$

Impact of Dimensionality on Independence Rule

Let $\boldsymbol{x}$ be in class 1. Define the worst case posterior error rate as

$$W(\hat{\delta}_I) = \mathrm{P}(\hat{\delta}_I(\boldsymbol{x}) < \hat{\delta}_I(\hat{\boldsymbol{\mu}}) \mid \mathfrak{S}),\ \hat{\boldsymbol{\mu}} = (\hat{\boldsymbol{\mu}}_1 + \hat{\boldsymbol{\mu}}_2)/2$$

$$W_\Gamma(\hat{\delta}_I) = \max_{\boldsymbol{\theta}\in\Gamma} W(\hat{\delta}_I)$$

Theorem 4 (Fan and Fan (2008))

Suppose that $\log p = o(n)$, $n = o(p)$ *and* $nC_p \to \infty$.

(i) *The posterior error rate fulfils*

$$W(\hat{\delta}_I) \leq \bar{\Phi}\left(\frac{\sqrt{\frac{n_1 n_2}{pn}}\,\boldsymbol{\alpha}^T \boldsymbol{D}_{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\alpha}(1 + o_p(1)) + \sqrt{\frac{p}{nn_1 n_2}}\,(n_1 - n_2)}{2\sqrt{\lambda_{\max}(\boldsymbol{R})}\,[1 + \frac{n_1 n_2}{pn}\boldsymbol{\alpha}^T \boldsymbol{D}_{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\alpha}(1 + o_p(1))]^{1/2}}\right) \quad (3)$$

Features Annealed Independence Rule

Theorem 4 (Fan and Fan (2008))

(ii) *If $\sqrt{n_1 n_2/(np)}\, C_p \to C_0$ with C_0 some positive constant, then the worst case posterior error rate*

$$W_\Gamma(\hat{\delta}_I) \xrightarrow{P} \bar{\Phi}\left(\frac{C_0}{2\sqrt{b_0}}\right)$$

In particular, if $C_0 = 0$, then $W_\Gamma(\hat{\delta}_I) \xrightarrow{P} \frac{1}{2}$

- The inequality (3) is very useful
- If we only include the first m features $j = 1, \ldots, m$ in the independence rule, then (3) still holds with each term p replaced by m

Features Annealed Independence Rule

- The contribution of the j-th feature is evaluated by its utility value α_j^2/σ_j^2
- Assume that the importance of the features is already ranked in the descending order of $\{\alpha_j^2/\sigma_j^2, j = 1, \ldots, p\}$
- Then $\frac{1}{\sqrt{m}}\sum_{j=1}^{m}\alpha_j^2/\sigma_j^2$ will first increase and then decrease as we include more and more features, and thus the right hand side of (3) first decreases and then increases with m
- Minimizing the upper bound in (3) can help us to find the optimal number of features m

A Direct Estimation Approach

- Fisher's rule depends on $\mathbf{\Omega} = \mathbf{\Sigma}^{-1}$ and $\boldsymbol{\alpha} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ only through their product $\mathbf{\Omega}\boldsymbol{\alpha}$
- If there is a way to estimate the product $\mathbf{\Omega}\boldsymbol{\alpha}$ directly, then one does not need to estimate $\mathbf{\Omega}$ and $\boldsymbol{\alpha}$ separately
- Cai and Liu (2011) proposed a constrained ℓ_1 minimization method to directly estimate the product $\mathbf{\Omega}\boldsymbol{\alpha}$ by exploiting the (approximate) sparsity of $\mathbf{\Omega}\boldsymbol{\alpha}$

Linear Programming Discriminant Rule

- Estimate $\boldsymbol{\beta} = \boldsymbol{\Omega}\boldsymbol{\alpha}$ via constrained ℓ_1 minimization

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min} \|\boldsymbol{\beta}\|_1 \text{ subject to } \|\hat{\boldsymbol{\Sigma}}\boldsymbol{\beta} - \hat{\boldsymbol{\alpha}}\|_\infty \leq \lambda_n, \quad (4)$$

  where λ_n is a tuning parameter which will be specified later.
- Given the solution $\hat{\boldsymbol{\beta}}$ to (4), we classify $\boldsymbol{x}$ to class 1 if and only if

$$\hat{\delta}_{LPD}(\boldsymbol{x}) = \boldsymbol{x}^T\hat{\boldsymbol{\beta}} \geq \hat{\delta}_{LPD}(\hat{\boldsymbol{\mu}}) \quad (5)$$

- (4) can be cast as a linear program. We call (5) the *Linear Programming Discriminant (LPD)* rule
- The direct estimate leads to a classifier that is more effective, efficient than those based on estimating $\boldsymbol{\Omega}$ and $\boldsymbol{\alpha}$ separately
- In certain setting, $\boldsymbol{\Omega}\boldsymbol{\alpha}$ can be well estimated even when $\boldsymbol{\Omega}$ is not estimate consistently

Motivation

- Note that $\boldsymbol{\beta} = \boldsymbol{\Omega}\boldsymbol{\alpha}$ is the solution to the equation

$$\boldsymbol{\Sigma}\boldsymbol{\beta} - \boldsymbol{\alpha} = \mathbf{0}$$

- When $\boldsymbol{\Sigma}$ and $\boldsymbol{\alpha}$ are unknown, they are replaced by their respective sample version $\hat{\boldsymbol{\Sigma}}$ and $\hat{\alpha}$. We then seek the most sparse solution within the feasible set

$$\{\boldsymbol{\beta} : \|\hat{\boldsymbol{\Sigma}}\boldsymbol{\beta} - \hat{\boldsymbol{\alpha}}\|_\infty \leq \lambda_n\}$$

  to account for the variability in $\hat{\boldsymbol{\Sigma}}$ and $\hat{\boldsymbol{\delta}}$

- The convex relaxation of using ℓ_1 minimization in place of ℓ_0 minimization is a standard technique in sparse signal recovery

Remark

- Both $\mathbf{\Omega}$ and $\boldsymbol{\alpha}$ are sparse $\Rightarrow$ $\mathbf{\Omega}\boldsymbol{\alpha}$ is sparse: If $\boldsymbol{\alpha}$ is k_1-sparse and $\mathbf{\Omega}$ is k_2-sparse, then $\mathbf{\Omega}\boldsymbol{\alpha}$ is at most $k_1 k_2$-sparse
- The sparsity of $\mathbf{\Omega}\boldsymbol{\alpha}$ does not require $\mathbf{\Omega}$ being sparse. Suppose $\boldsymbol{\alpha}$ is k_1-sparse with $\boldsymbol{\alpha} = [\boldsymbol{\alpha}_1\, \mathbf{0}]^T$ where $\boldsymbol{\alpha}_1$ is a k_1-dimensional vector. Write
$$\mathbf{\Omega} = \begin{bmatrix} \mathbf{\Omega}_{11} & \mathbf{\Omega}_{21}^T \\ \mathbf{\Omega}_{21} & \mathbf{\Omega}_{22} \end{bmatrix}$$
$\mathbf{\Omega}\boldsymbol{\alpha} = [\mathbf{\Omega}_{11}\boldsymbol{\alpha}_1 \;\; \mathbf{\Omega}_{21}\boldsymbol{\alpha}_1]^T$ does not depend on $\mathbf{\Omega}_{22}$ at all. $\mathbf{\Omega}\boldsymbol{\alpha}$ is sparse if $\mathbf{\Omega}_{21}$ is sparse. In particular, if there are at most k_2 nonzero elements on each column of $\mathbf{\Omega}_{21}$, then $\mathbf{\Omega}\boldsymbol{\alpha}$ is $k_1(k_2+1)$ sparse
- In general, it is not possible to consistently estimate $\mathbf{\Omega}$ under the spectral norm without regularity conditions on $\mathbf{\Omega}_{22}$

Theoretical Analysis

- The error rate of LDA is

$$\overline{W}(\delta_F) = \bar{\Phi}(d_p/2), \text{ with } d_p = [\boldsymbol{\alpha}^T \boldsymbol{\Omega} \boldsymbol{\alpha}]^{1/2}$$

  which is the best possible performance in the ideal setting where $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ and $\boldsymbol{\Sigma}$ are known. This thus serves as an oracle benchmark

- Given the training samples $\mathfrak{G} = \{(\boldsymbol{x}^{(i)}, k_i), i = 1, \ldots, n\}$, the conditional error rate of the LPD rule is

$$\overline{W}(\hat{\delta}_{LPD} \mid \mathfrak{G}) = \frac{1}{2}\bar{\Phi}\left(\frac{(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_2)^T \hat{\boldsymbol{\beta}}}{(\hat{\boldsymbol{\beta}}^T \boldsymbol{\Sigma} \hat{\boldsymbol{\beta}})^{1/2}}\right) + \frac{1}{2}\bar{\Phi}\left(-\frac{(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_1)^T \hat{\boldsymbol{\beta}}}{(\hat{\boldsymbol{\beta}}^T \boldsymbol{\Sigma} \hat{\boldsymbol{\beta}})^{1/2}}\right)$$

  where $\hat{\boldsymbol{\beta}}$ is given in (4)

- How close is $\overline{W}(\hat{\delta}_{LPD} \mid \mathfrak{G})$ to $\overline{W}(\delta_F)$?

Consistency

(C1). $n_1 \asymp n_2, \log p = o(n), c_0^{-1} \leq \lambda_{\min}(\boldsymbol{\Sigma}) \leq \lambda_{\max}(\boldsymbol{\Sigma}) \leq c_0$ for some constant $c_0 > 0$ and $d_p \geq c_1$ for some $c_1 > 0$

Theorem 5 (Cai and Liu (2011))

Let $\lambda_n = C\sqrt{d_p \log p/n}$ with $C > 0$ being sufficiently large. Suppose (C1) holds and $\|\boldsymbol{\Omega\alpha}\|_0 = o\left(\sqrt{\frac{n}{\log p}}\right)$. Then as $n \to \infty$ and $p \to \infty$,

$$\overline{W}(\hat{\delta}_{LPD} \mid \mathfrak{G}) - \overline{W}(\delta_F) \xrightarrow{\text{P}} 0 \tag{6}$$

In practice, λ_n is chosen by cross-validation

Rate of Convergence

Theorem 6 (Cai and Liu (2011))

Let $\lambda_n = C\sqrt{d_p \log p/n}$ *with* $C > 0$ *being sufficiently large. Suppose (C1) holds and* $\|\mathbf{\Omega\alpha}\|_0 = o\left(d_p^{-1}\sqrt{\frac{n}{\log p}}\right)$*, then*

$$\frac{\overline{W}(\hat{\delta}_{LPD} \mid \mathfrak{G})}{\overline{W}(\delta_F)} - 1 = O\left(\|\mathbf{\Omega\alpha}\|_0 d_p \sqrt{\frac{\log p}{n}}\right) \tag{7}$$

with probability greater than $1 - O(p^{-1})$

Remark 4 (Cai and Liu (2011))

The results can be extended to non-Gaussian distributions

Numerical Performance (Cai and Liu (2011))

- The LPD classifier can be implemented efficiently using linear programming
- Simulation results show that the LPD rule significantly outperforms the alternative methods in terms of the average error rate
- The LPD rule is also applied to the analysis of two real datasets, one from a lung cancer study, see Gordon et al. (2002) and another from leukemia study, see Golub et al. (1999). It performs favorably in comparison to existing methods

Two-Step Linear Discriminant Analysis

Divide all components (features) $\boldsymbol{x} \in \mathbb{R}^p$ into q groups

$$\mathbf{x} = \begin{bmatrix} \boldsymbol{x}_1 \\ \vdots \\ \boldsymbol{x}_q \end{bmatrix}$$

where $\boldsymbol{x}_j \in \mathbb{R}^{\ddot{p}}$, $j = 1, \ldots, q$, and $\ddot{p}q = p$

Definition 1

Two-step LDA function is defined by

$$\delta^\star(\mathbf{x}) = \delta_F(\delta_F(\mathbf{x}_1), \cdots, \delta_F(\mathbf{x}_q)) \tag{8}$$

where δ_F denotes Fisher's discriminant function

Two-Step Linear Discriminant Analysis

$$
\begin{tikzcd}
\mathbf{x}_1 \arrow[dr, "\text{LDA}"'] & \cdots \arrow[d, "\text{LDA}"'] & \mathbf{x}_q \arrow[dl, "\text{LDA}"] \\
& {[\delta_F(\mathbf{x}_1), \cdots, \delta_F(\mathbf{x}_q)]^T} \arrow[d, "\text{LDA}"'] & \\
& \delta^\star &
\end{tikzcd}
$$

- Multi-step LDA is motivated by the recursiveness of the wavelet multiresolution decomposition, see Mallat (1989)
- Using different "atoms", given by LDA projection vectors as in Mallat and Zhang (1993) instead of a fixed low-pass filter
- Multi-step LDA does not neglect all correlations of the features as independence rule

When $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ and $\boldsymbol{\Sigma}$ are known

Theorem 7

Let $\boldsymbol{\Delta} = [\delta_F(\mathbf{x}_1) \cdots \delta_F(\mathbf{x}_q)]^T$. Suppose that $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ and $\boldsymbol{\Sigma}$ are known then $\boldsymbol{\Delta}$ has common covariance matrix $\boldsymbol{\Theta}$ and means $\pm\frac{1}{2}\boldsymbol{m}$ given by

$$\boldsymbol{\Theta} = \bigoplus_{j=1}^{q} \boldsymbol{\alpha}_j^T \cdot \bigoplus_{j=1}^{q} \boldsymbol{\Sigma}_j^{-1} \cdot \boldsymbol{\Sigma} \cdot \bigoplus_{j=1}^{q} \boldsymbol{\Sigma}_j^{-1} \cdot \bigoplus_{j=1}^{q} \boldsymbol{\alpha}_j \tag{9}$$

$$\boldsymbol{m} = (m_1, \cdots, m_q), \;\; m_j = \boldsymbol{\alpha}_j^T \boldsymbol{\Sigma}_j^{-1} \boldsymbol{\alpha}_j, \, j = 1, \ldots, q \tag{10}$$

where $\boldsymbol{\Sigma}_j \in \mathbb{R}^{\ddot{p} \times \ddot{p}}$: covariance matrix of $\mathbf{x}_j$, $\boldsymbol{\alpha} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$, $\boldsymbol{\alpha} = [\boldsymbol{\alpha}_1^T, \cdots, \boldsymbol{\alpha}_q^T]^T$, $\boldsymbol{\alpha}_j \in \mathbb{R}^{\ddot{p}}$, $j = 1, \ldots, q$.

The direct sum of q matrices $\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_q$ is

$$\bigoplus_{j=1}^{q} \boldsymbol{\Sigma}_j = \begin{bmatrix} \boldsymbol{\Sigma}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{\Sigma}_q \end{bmatrix}$$

When $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ and $\boldsymbol{\Sigma}$ are known

Remark 5

The theoretical error rate of two-step LDA discriminant function

$$\overline{W}(\delta^\star) = \bar{\Phi}(\frac{d_p^\star}{2}),$$

where $d_p^\star = [\boldsymbol{m}^T \boldsymbol{\Theta}^{-1} \boldsymbol{m}]^{1/2}$ is the Mahalanobis distance between two score classes

In the case of $\boldsymbol{\Sigma} = \bigoplus_{j=1}^{q} \boldsymbol{\Sigma}_j$, we have $d_p^\star = d_p = [\boldsymbol{\alpha}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\alpha}]^{1/2}$ and

$$\overline{W}(\delta^\star) = \overline{W}(\delta_F) = \bar{\Phi}(\frac{d_p}{2})$$

When $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ and $\boldsymbol{\Sigma}$ are known

Separability

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 $\boldsymbol{s}_1, \boldsymbol{s}_2 \in S$ and $t_1, t_2 \in T$, it holds

$$cov(x(\boldsymbol{s}_1; t_1), x(\boldsymbol{s}_2; t_2)) = C^{(s)}(\boldsymbol{s}_1, \boldsymbol{s}_2) \cdot C^{(t)}(t_1, t_2) \tag{11}$$

where $C^{(s)}$, $C^{(t)}$ are spatial, temporal covariance functions respectively

An observation for classification is selected at a finite set of locations $\boldsymbol{s}_1, \dots, \boldsymbol{s}_{\ddot{p}}$ and time points $t_1, \dots, t_q$

$$\boldsymbol{x} = \left[x(\boldsymbol{s}_1; t_1) \cdots x(\boldsymbol{s}_{\ddot{p}}; t_1) \cdots x(\boldsymbol{s}_1; t_q) \cdots x(\boldsymbol{s}_{\ddot{p}}; t_q)\right]^T \in \mathbb{R}^{\ddot{p}q}$$

From (11), the covariance of $\mathbf{x}$ is

$$\boldsymbol{\Sigma} = \mathbf{U} \otimes \mathbf{V},$$

where spatial covariance $\mathbf{V} = [C^{(s)}(\boldsymbol{s}_i, \boldsymbol{s}_j)]$ and temporal covariance $\mathbf{U} = [C^{(t)}(t_i, t_j)]$

Upper Bound of Error Rate

Theorem 8 (Huy et al. (2012))

Suppose that spatio-temporal observation $\mathbf{x}$ *are drawn from the normal distribution with known* $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ *and* $\boldsymbol{\Sigma}$, *moreover* $\boldsymbol{\Sigma} = \boldsymbol{U} \otimes \boldsymbol{V}$, *then the error rate* e_2 *of the two-step LDA fulfils*

$$e_1 \leq e_2 \leq \bar{\Phi}\left(\frac{2\sqrt{\kappa(\boldsymbol{U}_0)}}{1+\kappa(\boldsymbol{U}_0)}\bar{\Phi}^{-1}(e_1)\right) \tag{12}$$

where e_1 *is LDA error,* $\kappa(\boldsymbol{U}_0)$ *denotes the condition number of the temporal correlation matrix* $\boldsymbol{U}_0 = \boldsymbol{D}_{\boldsymbol{U}}^{-1/2}\boldsymbol{U}\boldsymbol{D}_{\boldsymbol{U}}^{-1/2}, \boldsymbol{D}_{\boldsymbol{U}} = diag(u_{11}, \cdots, u_{qq})$

If temporal features are independent then $K_0 = 1$ and $e_2 = e_1$

Upper Bound of Error Rate

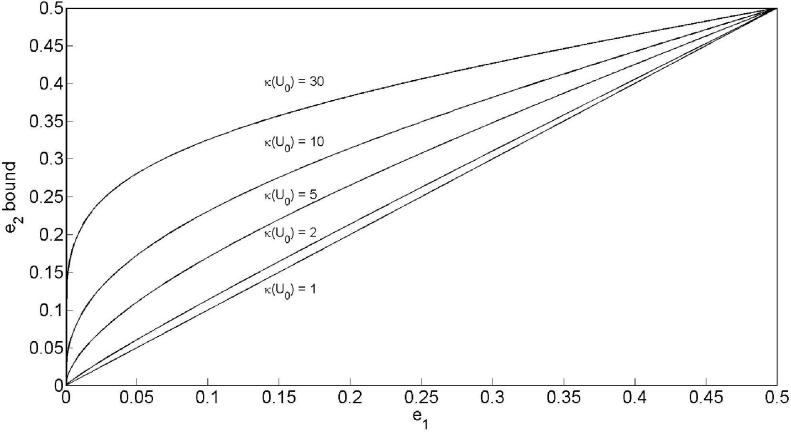


Figure : The error bound of two-step LDA as a function of the LDA error rate e_1 for several values of $\kappa(\boldsymbol{U}_0)$

When $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ and $\boldsymbol{\Sigma}$ are unknown

- Given training data $\mathfrak{G} = \{(\boldsymbol{x}^{(i)}, k_i),\ i = 1, \ldots, n\}$, $\#\mathfrak{G} = n$, $k_i \in \{1, 2\}$, $\boldsymbol{x}^{(i)} \in \mathbb{R}^p$, $p = \ddot{p}q$
- At the first step, we calculate all $\hat{\delta}_F(\boldsymbol{x}_j)$, $j = 1, \ldots, q$

$$\hat{\boldsymbol{\mu}}_{kj} = \frac{1}{n_k} \sum_{k_i = k} \boldsymbol{x}_j^{(i)},\ k = 1, 2,\ \hat{\boldsymbol{\alpha}}_j = \hat{\boldsymbol{\mu}}_{1j} - \hat{\boldsymbol{\mu}}_{2j},$$

$$\hat{\boldsymbol{\Sigma}}_j = \frac{1}{n-2} \sum_{k=1}^{2} \sum_{k_i = k} (\boldsymbol{x}_j^{(i)} - \hat{\boldsymbol{\mu}}_{kj})(\boldsymbol{x}_j^{(i)} - \hat{\boldsymbol{\mu}}_{kj})^T,\ \hat{\delta}_F(\boldsymbol{x}_j) = \boldsymbol{x}_j^T \hat{\boldsymbol{\Sigma}}_j^{-1} \hat{\boldsymbol{\alpha}}_j$$

- The conditional means and covariance matrix of the score $\hat{\boldsymbol{\Delta}} = [\hat{\delta}_F(\boldsymbol{x}_1), \cdots, \hat{\delta}_F(\boldsymbol{x}_q)]^T$ given by

$$\pm\frac{1}{2}\widetilde{\boldsymbol{m}} = \mathrm{E}(\hat{\boldsymbol{\Delta}} \mid \mathfrak{G}) = \pm\frac{1}{2}[\widetilde{m}_1, \cdots, \widetilde{m}_q]^T,\ \widetilde{m}_j = \boldsymbol{\alpha}_j^T \hat{\boldsymbol{\Sigma}}_j^{-1} \hat{\boldsymbol{\alpha}}_j,\ j = 1, \ldots, q,$$

$$\widetilde{\boldsymbol{\Theta}} = \mathrm{cov}(\hat{\boldsymbol{\Delta}}, \hat{\boldsymbol{\Delta}} \mid \mathfrak{G}) = \bigoplus_{j=1}^{q} \hat{\boldsymbol{\alpha}}_j^T \cdot \bigoplus_{j=1}^{q} \hat{\boldsymbol{\Sigma}}_j^{-1} \cdot \boldsymbol{\Sigma} \cdot \bigoplus_{j=1}^{q} \hat{\boldsymbol{\Sigma}}_j^{-1} \cdot \bigoplus_{j=1}^{q} \hat{\boldsymbol{\alpha}}_j$$

When $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ and $\boldsymbol{\Sigma}$ are unknown

- $\boldsymbol{\Sigma}_j \in \mathbb{R}^{\ddot{p}\times\ddot{p}}$

$$\bigoplus_{j=1}^{q} \hat{\boldsymbol{\Sigma}}_j^{-1} = \begin{bmatrix} \hat{\boldsymbol{\Sigma}}_1^{-1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \hat{\boldsymbol{\Sigma}}_2^{-1} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \hat{\boldsymbol{\Sigma}}_q^{-1} \end{bmatrix} \to \begin{bmatrix} \boldsymbol{\Sigma}_1^{-1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_2^{-1} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{\Sigma}_q^{-1} \end{bmatrix}$$

$$\|\bigoplus_{j=1}^{q} \hat{\boldsymbol{\Sigma}}_j^{-1} - \bigoplus_{j=1}^{q} \boldsymbol{\Sigma}_j^{-1}\| = O_{\mathrm{P}}(\ddot{p}\sqrt{\log p}/\sqrt{n})$$

- $\boldsymbol{\Sigma} \in \mathbb{R}^{p\times p}$

$$\|\hat{\boldsymbol{\Sigma}}^{-1} - \boldsymbol{\Sigma}^{-1}\| = O_{\mathrm{P}}(p\sqrt{\log p}/\sqrt{n})$$

When $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ and $\boldsymbol{\Sigma}$ are unknown

Theorem 9

Suppose that

$$c_0^{-1} \leq \textit{all eigenvalues of } \boldsymbol{\Sigma} \leq c_0, \tag{13}$$

$$\max_{j \leq q} \|\boldsymbol{\alpha}_j\|^2 \leq c_0, \tag{14}$$

where $\boldsymbol{\alpha}_j \in \mathbb{R}^{\ddot{p}}$, $\boldsymbol{\alpha} = [\boldsymbol{\alpha}_1^T, \cdots, \boldsymbol{\alpha}_q^T]^T$, *and* $\ddot{p}\sqrt{q \log p}/\sqrt{n} \to 0$*. Then*

$$\|\widetilde{\boldsymbol{\Theta}} - \boldsymbol{\Theta}\| = O_{\mathrm{P}}(\max[\sqrt{\ddot{p}}/n^{\beta}, \ddot{p}\sqrt{\log p}/\sqrt{n}]),$$

$$\|\widetilde{\boldsymbol{m}} - \boldsymbol{m}\| = O_{\mathrm{P}}(\ddot{p}\sqrt{q \log p}/\sqrt{n})$$

for every $\beta < \frac{1}{2}$, $\boldsymbol{m}$, *and* $\boldsymbol{\Theta}$ *as in Theorem 7. If* $\ddot{p} \geq n^{\gamma}$ *with any* $\gamma > 0$,

$$\|\widetilde{\boldsymbol{\Theta}} - \boldsymbol{\Theta}\| = O_{\mathrm{P}}(\ddot{p}\sqrt{\log p}/\sqrt{n}).$$

When $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ and $\boldsymbol{\Sigma}$ are unknown

Divide all training data into two parts $\mathfrak{G} = \mathfrak{G}_1 \cup \mathfrak{G}_2$, $\mathfrak{G}_1 \cap \mathfrak{G}_2 = \emptyset$ such that the sample size of every class in every part equals $\Omega(n)$

- Use $\mathfrak{G}_1$ to calculate $\hat{\delta}_F(\boldsymbol{x}_j) = \boldsymbol{x}_j^T \hat{\boldsymbol{\Sigma}}_j^{-1} \hat{\boldsymbol{\alpha}}_j$, $j = 1, \ldots, q$
- Use training scores $\{[\hat{\delta}_F(\boldsymbol{x}_1^{(i)}), \cdots, \hat{\delta}_F(\boldsymbol{x}_q^{(i)})]^T, \boldsymbol{x}^{(i)} \in \mathfrak{G}_2\}$ to estimate $\hat{\boldsymbol{\Theta}}$, $\pm\frac{1}{2}\hat{\boldsymbol{m}}$, and $\hat{\delta}^\star(\boldsymbol{x}) = [\hat{\delta}_F(\boldsymbol{x}_1), \cdots, \hat{\delta}_F(\boldsymbol{x}_q)]\hat{\boldsymbol{\Theta}}^{-1}\hat{\boldsymbol{m}}$

Regularity Conditions of Score $\boldsymbol{\Delta}$

There is a constant c_1 (not depending on q) such that

$$c_1^{-1} \leq \textit{all eigenvalues of } \boldsymbol{\Theta} \leq c_1, \tag{15}$$

$$c_1^{-1} \leq \max_{j \leq q} m_j^2 \leq c_1, \tag{16}$$

where $\boldsymbol{m} = [m_1, \cdots, m_q]^T$ and $\boldsymbol{\Theta}$ is given by Theorem 7

Impact of Dimensionality on Two-Step LDA

Corollary 1

If $\max\{\ddot{p}\sqrt{q\log p},\ q\sqrt{\log q}\}/\sqrt{n} \to 0$ *and* $\ddot{p} \geq n^{\gamma}$ *with any* $\gamma > 0$, *then the conditional error rate of* $\hat{\delta}^{\star}(\boldsymbol{x})$, *given* $\mathfrak{G}$ *satisfies*

$$\overline{W}(\hat{\delta}^{\star} \mid \mathfrak{G}) = \bar{\Phi}([1 + O_{\mathrm{P}}(\max\{\ddot{p}\sqrt{q\log p},\ q\sqrt{\log q}\}/\sqrt{n})]d_p^{\star}/2),$$

$d_p^{\star} = [\boldsymbol{m}^T\boldsymbol{\Theta}^{-1}\boldsymbol{m}]^{1/2}$: *Mahalanobis distance between two score classes*

Remark 6

If $\ddot{p} = O(p^{1/3}), q = O(p^{2/3})$ *then the conditional error rate of* $\hat{\delta}^{\star}$

$$\overline{W}(\hat{\delta}^{\star} \mid \mathfrak{G}) = \bar{\Phi}([1 + O_{\mathrm{P}}(p^{2/3}\sqrt{\log p}/\sqrt{n})]d_p^{\star}/2), \tag{17}$$

whereas the error rate of the ordinary LDA, see Shao el al. (2011)

$$\overline{W}(\hat{\delta}_F \mid \mathfrak{G}) = \bar{\Phi}([1 + O_{\mathrm{P}}(p\sqrt{\log p}/\sqrt{n})]d_p/2) \tag{18}$$

Impact of Dimensionality on Multi-Step LDA

- Multi-step LDA procedure divides all features or scores into consecutive disjoint subgroups at each step
- $\boldsymbol{t} = (\ddot{p}_1, \dots \ddot{p}_l)$, $\prod_{s=1}^{l} \ddot{p}_s = p$ where $\ddot{p}_s$: size of subgroups at step s, $s = 1, \dots, l$ is called the type of multi-step LDA

Remark 7

- *If $\ddot{p}_1 = O(p^{1/3})$, $\ddot{p}_2 = O(p^{2/9})$, $\ddot{p}_3 = O(p^{4/9})$ then error rate of three-step LDA with type $\boldsymbol{t} = (\ddot{p}_1, \ddot{p}_2, \ddot{p}_3)$ satisfies*

$$\overline{W}(\hat{\delta}^{\star\star} \mid \mathfrak{G}) = \bar{\Phi}([1 + O_{\mathrm{P}}(p^{4/9}\sqrt{\log p}/\sqrt{n})]d_p^{\star\star}/2) \tag{19}$$

where $d_p^{\star\star}$ is the Mahalanobis distance between two score classes at the third step

Impact of Dimensionality on Multi-Step LDA

Remark 8

- *The error rate of $\hat{\delta}^{l\star}$ with optimal type $\boldsymbol{t} = (\ddot{p}_1, \ldots \ddot{p}_l)$ satisfies*

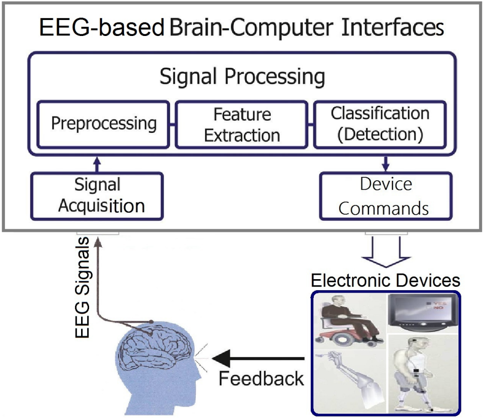
$$\overline{W}(\hat{\delta}^{l\star} \mid \mathfrak{G}) = \bar{\Phi}([1 + O_{\mathrm{P}}(p^{\frac{9}{6l+2+\frac{(-1)^{l+1}}{2^{l-1}}}}\sqrt{\log p}/\sqrt{n})]d_p^{l\star}/2)$$

where $d_p^{l\star}$ is the Mahalanobis distance between two score classes at the l-th step

- We can divide all training data into l parts using for l-step LDA such that the sample size of every class in every part equals $\Omega(n/\log p)$

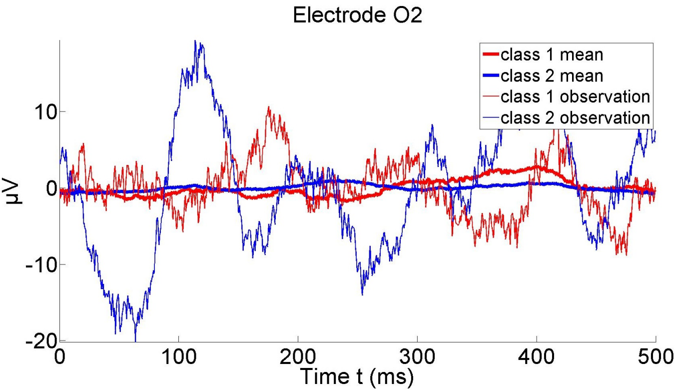
EEG-based Brain-Computer Interfaces

- Brain-Computer Interfaces (BCIs) enable users to control electronic devices or computers by using only their brain activity



- An EEG-based BCI classifies brain activity during differential tasks into differential classes based on their associated EEG signals

Binary Classification Problem



The assumption that observations being normally distributed with common covariance matrix holds well enough for BCI data, see Blankertz et al. (2011), Frenzel et al. (2011)

Separability

Separability is a proper assumption for EEG data, Huizenga et al. (2002)

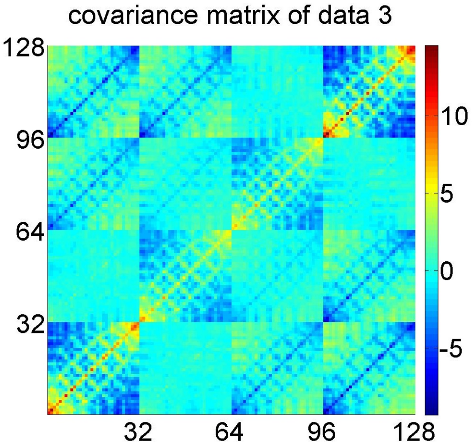


Figure : Covariance matrix estimated from 4 time points and 32 locations

Defining the Feature Subgroups of Two-Step LDA

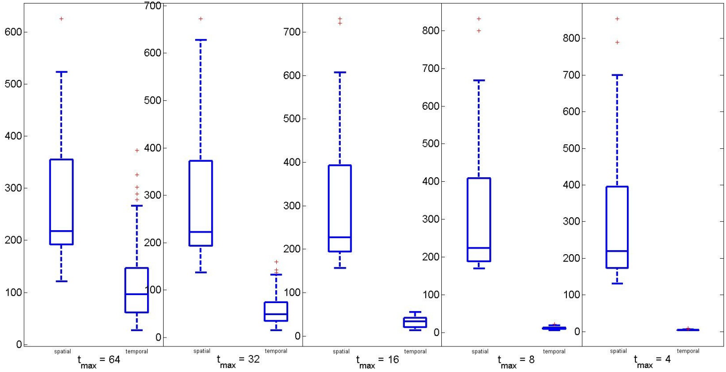


Figure : Comparison for condition numbers of correlation matrices between spatial and temporal features using 30 datasets, see Huy et al. (2012)

Remark 9

Feature subgroup $\mathbf{x}_j$ should contain all features at time point j

Learning Curves

- Brain-Computer Interface data from 9 subjects, see Frenzel et al. (2011)
  - Each dataset contain 7290 samples
- Train classifiers using the first n samples, with $200 \leq n \leq 3500$ and apply them to the remain ones
- Number of features $p = 1024$
- Since one of two classes are rare error rate is not a meaningful performance measure
- AUC value is often used as a standard measure of classification performance
  - AUC $= 1$ represents a perfect separation
  - AUC $= 0.5$ represents a worthless separation

Learning Curves

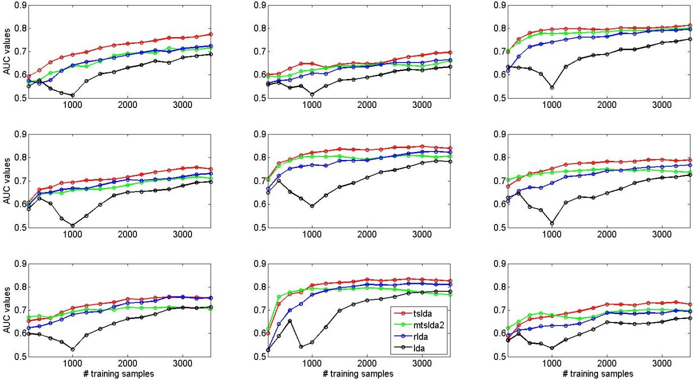


Figure : Performance of multi-step LDA with type (2, 2, 2, 2, 2, 2, 2, 2, 2, 2), two-step LDA, regularized LDA, LDA, Huy et al. (2012)

- Two-step LDA showed better performance than regularized LDA
- Multi-step LDA showed faster convergence than regularized LDA

Classification Performance

- Using 30 Brain-Computer Interface datasets, see Frenzel et al. (2011)
  - The number of samples of each dataset is from 450 to 477
- Train classifiers using the first n samples, with $100 \leq n \leq 250$ and apply them to the remain ones
- The dimension of $\boldsymbol{x}$: $p = 1024$
- The regularization parameter of regularized LDA was estimated by both the analytic formula (oprlda) as in Schäfer and Strimmer (2005) and cross-validation (cvrlda)

Classification Performance

$n =$	100($\approx 1min$)	125	150	175	200	225	250
lda	0.770	0.772	0.782	0.792	0.801	0.813	0.816
cvrlda	*0.789*	*0.802*	*0.810*	*0.813*	*0.822*	*0.835*	*0.839*
oprlda	0.782	0.792	0.803	0.809	0.816	0.826	0.830
tslda	0.747	0.777	0.796	0.810	0.822	0.837	0.847
mtslda1	0.806	0.822	0.836	0.843	0.856	0.862	0.865
mtslda2	0.808	0.819	0.828	0.837	0.841	0.844	0.850
mtslda3	0.808	0.819	0.833	0.841	0.846	0.852	0.856
mtslda4	**0.821**	**0.831**	**0.844**	**0.849**	**0.858**	**0.865**	**0.869**
mtslda5	0.787	0.808	0.824	0.831	0.842	0.850	0.853

Table : Average AUC values of LDA, regularized LDA, two-step LDA (tslda), multi-step LDA with type $(16, 2, 2, 2, 2, 2, 2)$ (mtslda1), $(2, 2, 2, 2, 2, 2, 2, 2, 2, 2)$ (mtslda2), $(4, 8, 2, 2, 2, 2, 2)$ (mtslda3), $(8, 4, 2, 2, 2, 2, 2)$ (mtslda 4), $(32, 2, 2, 2, 2, 2)$ (mtslda5)

- Except for type $(32, 2, 2, 2, 2, 2)$ (mtslda5) multi-step LDA showed better performance than regularized LDA
- The performance of two-step LDA, mtslda5 is worse for small $n\,(n = 100)$ but better for large $n\,(n = 250)$ due to the impact of dimensionality

Classification Performance

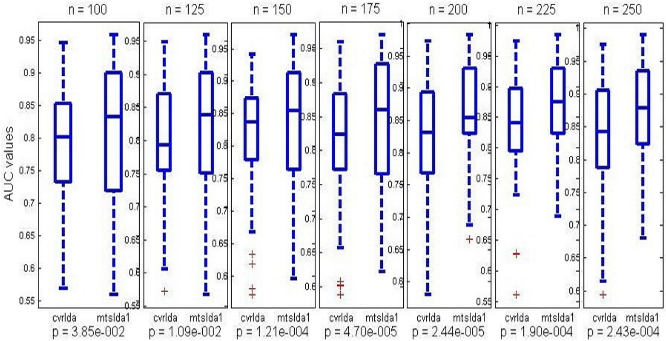


Figure : Performance comparison of multi-step LDA with type $(16, 2, 2, 2, 2, 2, 2)$ and regularized LDA for 30 Brain-Computer Interface datasets. Statistical significance p values were computed using a Wilcolxon signed rank test.

- The medians of AUCs of multi-step LDA are higher than regularized LDA
- p-value decreases until $n = 175$ and then increases since multi-step LDA achieves the faster convergence rate than regularized LDA

Conclusions

- Our method avoids estimation of the high-dimensional covariance matrix by applying LDA in several steps
- In the case of separable models, the theoretical loss in efficiency of two-step LDA in comparison to LDA is not very large
- For our EEG data, multi-step LDA performed better than regularized LDA which is the state-of-the-art classification method, see Blankertz et al. (2011)
- Multi-step LDA has faster convergence rate than LDA

Conjecture

- *The error rates of independence rule, two-step LDA, and LDA:*

$$\overline{W}(\delta_I) \leq \overline{W}(\delta^\star) \leq \overline{W}(\delta_F)$$

THANK YOU!

References

1. S. Frenzel, C. Bandt, N. H. Huy, and L. T. Kien. Single-trial classification of P300 speller data. Frontiers in Computational Neuroscience conference abstract: Bernstein Conference on Computational Neuroscience, 2010.
2. N. H. Huy, S. Frenzel, and C. Bandt. Two-Step Linear Discriminant Analysis for Classification of EEG Data. In *Studies in Classification, Data Analysis, and Knowledge Organization*, 2012. Accepted.

References

3. P. J. Bickel and E. Levina. Some theory of Fisher's linear discriminant function, 'naive Bayes' and some alternatives when there are many more variables than observations. *Bernoulli*, 10(6):989-1010, December 2004.
4. B. Blankertz, S. Lemm, M. Treder, S. Haufe, and K. -R. Müller. Single-trial analysis and classification of ERP components - A tutorial. *NeuroImage*, 56(2):814-825, May 2011.
5. P. Bühlmann and S. van de Geer. *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer Series in Statistics. Springer, 2011.
6. T. Cai and W. Liu. A direct estimation approach to sparse linear discriminant analysis. *Journal of the American Statistical Association*, 106(496):1566-1577, December 2011.
7. J. Fan and Y. Fan. High-Dimensional Classification Using Features Annealed Independence Rules. *The Annals of Statistics*, 36(6):2605-2637, December 2008.
8. S. Frenzel, E. Neubert, and C. Bandt. Two communication lines in a 3×3 matrix speller. *Journal of Neural Engineering*, 8(3):036021, May 2011.

References

9. H. M. Huizenga, J. C. De Munck, L. J. Waldorp, and R. P. P. P. Grasman. Spatiotemporal EEG/MEG Source Analysis Based on a Parametric Noise Covariance Model. *IEEE Transactions on Biomedical Engineering*, 49(6):533-539, June 2002.
10. D. J. Krusienski, E. W. Sellers, D. J. McFarland, T. M. Vaughan, and J. R. Wolpaw. Toward enhanced P300 speller performance. *Journal of Neuroscience Methods*, 167(1):15-21, January 2008.
11. S. G. Mallat. A Theory for Multiresolution Signal Decomposition: The Wavelet Representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(7):674-693, July 1989.
12. S. G. Mallat and Z. Zhang. Matching Pursuits With Time-Frequency Dictionaries. *IEEE Transactions on Signal Processing*, 41(12):3397-3415, December 1993.
13. E. Neubert. Untersuchung ereigniskorrelierter Potentiale im EEG. Diploma thesis, Ernst-Moritz-Arndt-Universität Greifswald, May 2010.
14. J. Schäfer and K. Strimmer. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Application in Genetics and Molecular Biology*, 4(1):1544-6115, November 2005.
15. J. Shao, Y. Wang, X. Deng, and S. Wang. Sparse Linear Discriminant Analysis by Thresholding for High Dimensional Data. *The Annals of Statistics*, 39(2):1241-1265, April 2011.

References

16. E. Candes, and T. Tao. The Dantzig selector: statistical estimation when p is much large than n. *The Annals of Statistics*, 35: 2313-2351, 2007.
17. D. Donoho, M. Elad, and V. Temlyakov. Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Transactions on Information Theory*, 52: 6-18, 2006.