Multi-Step Linear Discriminant Analysis for Classification of Event-Related Potentials

Nguyen Hoang Huy   Stefan Frenzel   Christoph Bandt

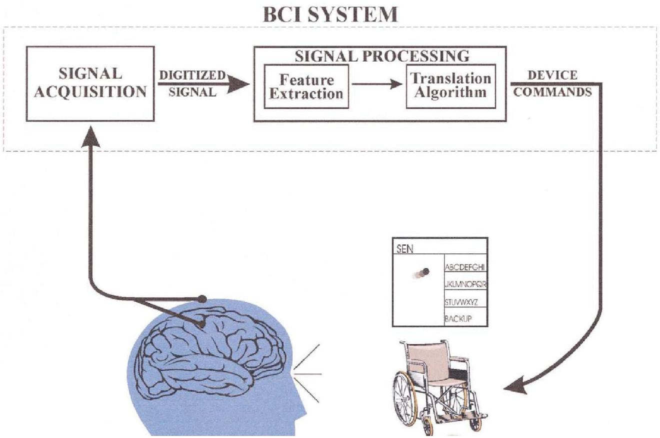
Institute of Mathematics and Informatics, Arndt University Greifswald, Germany

GfKI2012, August 1-3, 2012, Hildesheim, Germany

Contents

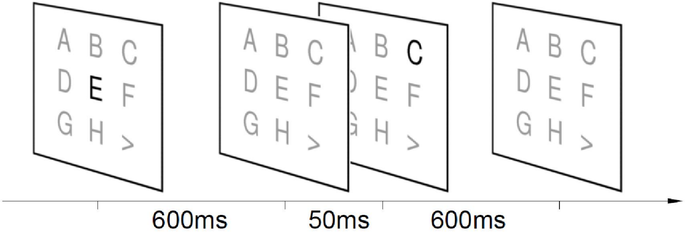
Brain-Computer Interfaces

- Brain-Computer Interfaces (BCI) are communication systems



Visual Speller

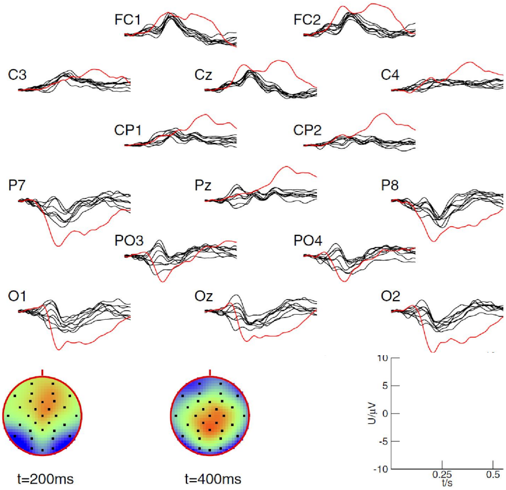
- Users sit in front of a screen
- Concentrate on a target letter
- One random letter highlighted in each single-trial



Frenzel et al. (2011)

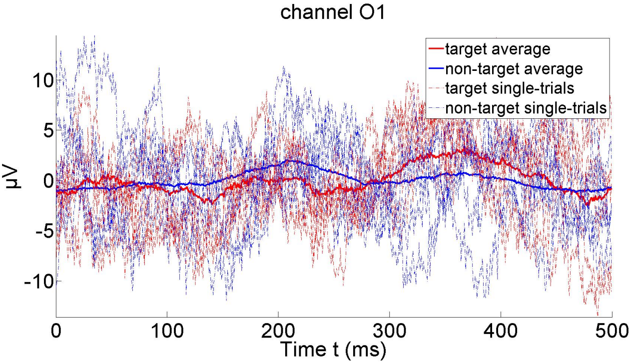
Event-Related Potentials

- Recognition of target letters induces characteristic EEG signals

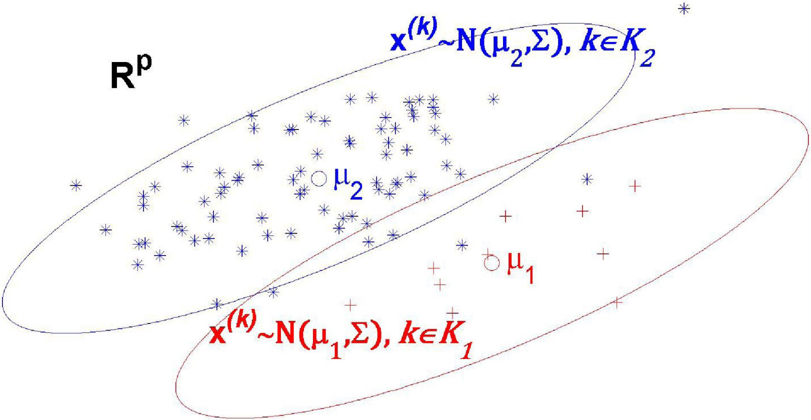


Binary Classification Problem

- Distinguish the target trials from the rest
- As few training samples as possible
- Utilize structure of Event-Related Potentials

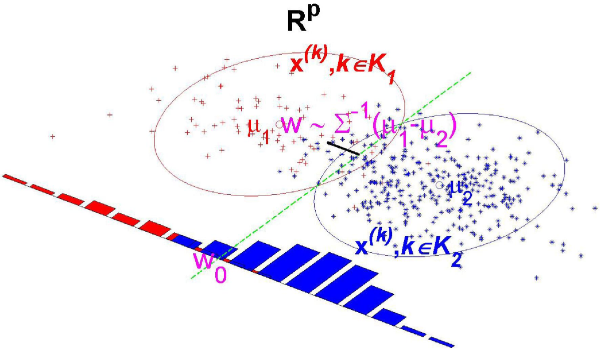


Homoscedastic Normal Model



- Appropriate model of event-related potential distributions

Linear Discriminant Analysis



- Fisher's linear discriminant (LDA) function: $\delta_F(\mathbf{x}) = \mathbf{w}^T\mathbf{x}$
- Maximum likelihood parameter estimation using training data
- Problem: Estimation of covariance matrix $\mathbf{\Sigma}$
  - with 32 channel, 40 time points, $\approx 10^6$ entries
  - but sample size n only in thousands

Linear Discriminant Analysis

- Bickel and Levina[1] showed that the sample LDA is as bad as random guessing when $\frac{p}{n} \to \infty$
- The asymptotic error rate $\widehat{e}_1$ of sample LDA given by[5]

$$\widehat{e}_1 \xrightarrow{P} \Phi(-J_1 \frac{\sqrt{1-y}}{2\sqrt{J_1 + y_1 + y_2}})$$

where $J_1 = \lim_{p\to\infty} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2), y = \lim_{p\to\infty} \frac{p}{n}, y_k = \lim_{p\to\infty} \frac{p}{n_k}, k = 1, 2, \Phi$ Gaussian cumulative function

- Regularization: Successful event-related potential classifier[2]

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

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

[1] Bickel et al. (2004)
[2] Blankertz et al. (2011)
[5] Serdobolskii (2000)

Two-Step Linear Discriminant Analysis

- Divide all features $\mathbf{x} \in \mathbb{R}^p$ into q groups

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

where $\mathbf{x}_j \in \mathbb{R}^{p_j}, \;\; j = 1, \cdots, q, \;\; p_1 + \cdots + p_q = p$

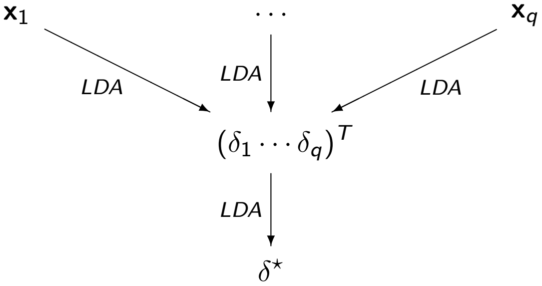
Definition 1

Two-step LDA function is defined by

$$\delta^\star(\mathbf{x}) = \delta_F(\delta_F(\mathbf{x}_1), \cdots, \delta_F(\mathbf{x}_q)) \tag{1}$$

where δ_F denotes the LDA function

Two-Step Linear Discriminant Analysis



Two-Step Linear Discriminant Analysis

Theorem 1

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

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

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

where $\mathbf{\Sigma}_j$: *covariance matrix of* $\mathbf{X}_j$, $\boldsymbol{\alpha} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$, $\boldsymbol{\alpha} = [\boldsymbol{\alpha}_1^T, \cdots, \boldsymbol{\alpha}_q^T]^T$, $\boldsymbol{\alpha}_j \in \mathbb{R}^{p_j}; j = 1, \cdots, q$.

Two-step LDA asymptotic setting

- $\max\limits_{1 \le j \le q} \frac{p_j}{n}, \frac{q}{n^{\beta}} \to 0, \frac{n_1}{n} \to c, \frac{p_1 + \cdots + p_1}{n} \to y \in [0, \infty), \beta, c \in (0, 1)$
- $\lim\limits_{p \to \infty} \mathbf{m}^T \mathbf{\Theta}^{-1} \mathbf{m} = J_2$

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

Regularity conditions

- *There are c_1, c_2 such that $c_1 \leq$ all eigenvalues of $\boldsymbol{\Sigma} \leq c_2$*
- *The mean $\boldsymbol{\mu} = [\boldsymbol{\mu}_1^T, \cdots, \boldsymbol{\mu}_q^T]^T$ satisfying that*

$$\boldsymbol{\mu}_j \in B = B_{\mathbf{a},d} = \left\{ \mathbf{z} = (z_1, z_2, \dots) \in l_2 : \sum_{j=1}^{\infty} a_j z_j^2 \leq d^2 \right\},$$

where $\mathbf{a} = (a_1, a_2, \cdots)$, and $a_j \to \infty$

Theorem 2

The error rate of sample Two-step LDA $\widehat{\delta}^{\star}(\mathbf{x})$ satisfies

$$\widehat{e}_2 \xrightarrow{P} \Phi(-\frac{J_2}{2})$$

Separability

Assumption

The spatio-temporal features $X(c;t)$ satisfy

$$cov(X(c_1;t_1), X(c_2;t_2)) = C^{(s)}(c_1, c_2) \cdot C^{(t)}(t_1, t_2) \qquad (2)$$

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

- From (2), the covariance of $\mathbf{X}$ is

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

  where spatial covariance $\mathbf{V} = [C^{(s)}(c_i, c_j)]$ and temporal covariance $\mathbf{U} = [C^{(t)}(t_i, t_j)]$
- EEG signals fit separability test[4]

[4] Mitchell et al. (2006)

Separability

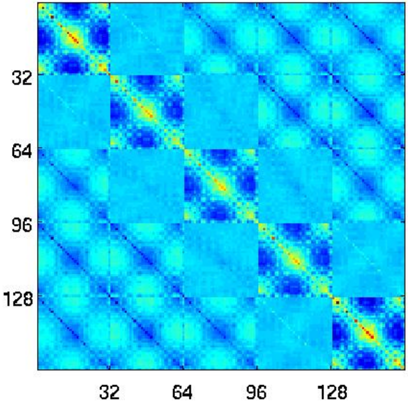


Figure: Pooled covariance matrix of all features

Upper Bound of Error Rate

Theorem 3

Suppose that spatio-temporal features $\mathbf{x}$ are drawn from the normal distribution with known $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ and $\boldsymbol{\Sigma}$, moreover $\boldsymbol{\Sigma} = \mathbf{U} \otimes \mathbf{V}$, then the error rate e_2 of the Two-step LDA can be bounded by:

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

e_1: LDA error, $\kappa(\mathbf{U}_0)$: condition number of temporal correlation matrix $\mathbf{U}_0 = \mathbf{D}_{\mathbf{U}}^{-1/2}\mathbf{U}\mathbf{D}_{\mathbf{U}}^{-1/2}, \mathbf{D}_{\mathbf{U}}^{-1/2} = diag(u_{11}^{-1/2}, \cdots, u_{\tau\tau}^{-1/2})$.

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

Upper Bound of Error Rate

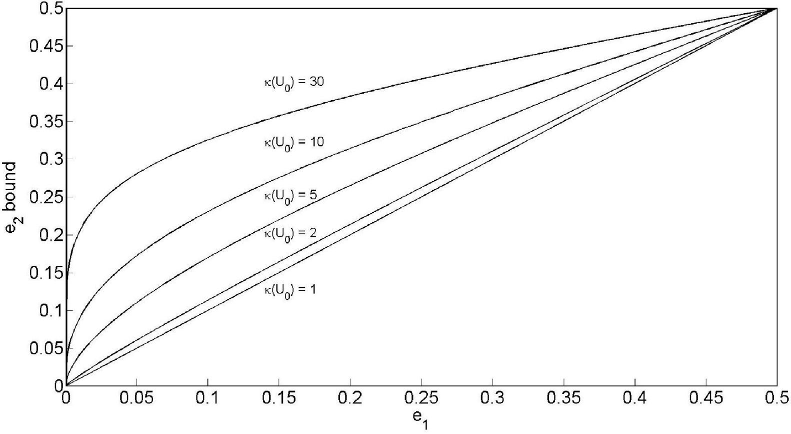
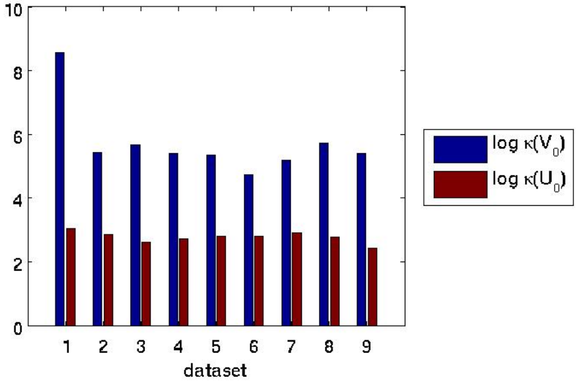


Figure: The bound on the error rate of Two-step LDA as a function of

How to Group Features



- Features should be ordered according to their time index

Learning Curves

- Visual speller data from 9 subjects[3]
- Train classifier using the first n samples and apply them to the remain ones
- Number of features $p = 1280$
- Since target trials are rare overall error rate is not a meaningful performance measure
- Instead we consider AUC value

[3] Frenzel et al. (2011)

Learning Curves

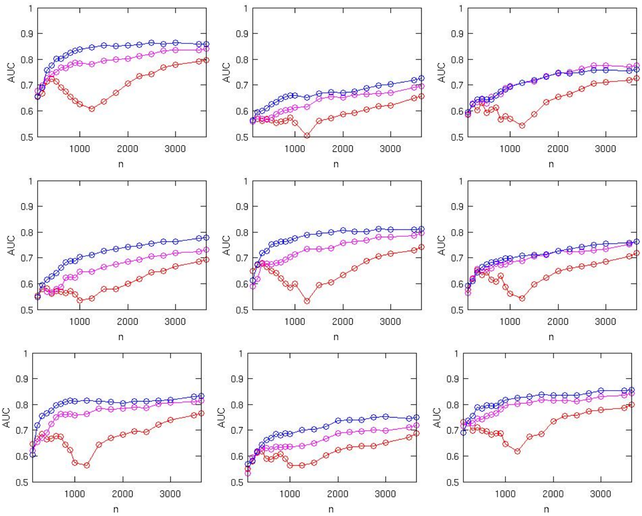


Figure: Two-step LDA, Regularized LDA, LDA

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. and LEMM, S. and TREDER, M. and HAUFE, S and MÜLLER K. R. (2011): Single-trial analysis and classification of ERP components – A tutorial. *NeuroImage, 56, 814-825*.
3. FRENZEL, S. and NEUBERT, E. and BANDT, C. (2011): Two communication lines in a 3×3 matrix speller. *Journal of Neural Engineering, 8, 036021*.
4. MITCHELL, M. W. and GENTON, M. G. and GUMPERTZ, M. L. (2006): A likelihood ratio test for separability of covariances. *Journal of Multivariate Analysis, 97, 1025–1043*.
5. SERDOBOLSKII, V. I. (2000): Multivariate Statistical Analysis: A High-Dimensional Approach. *Volume 41 of Mathematical and Statistical Methods, Kluwer Academic Publishers, Dordrecht, The Netherlands*.

Thank you for your attention