Brain-Computer Interfacing

WS 2018/2019 – Lecture #05

Benjamin Blankertz

Lehrstuhl für Neurotechnologie, TU Berlin

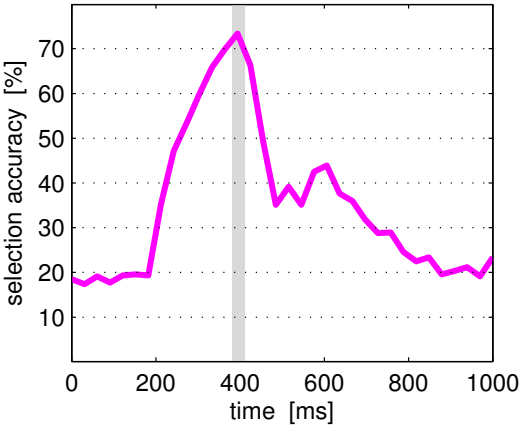
benjamin.blankertz@tu-berlin.de

14 · Nov · 2018

Today's Topics

- Classification with spatio-temporal features
- Estimation bias in the sample covariance matrix
- Shrinkage of the sample covariance matrix to counterbalance that bias

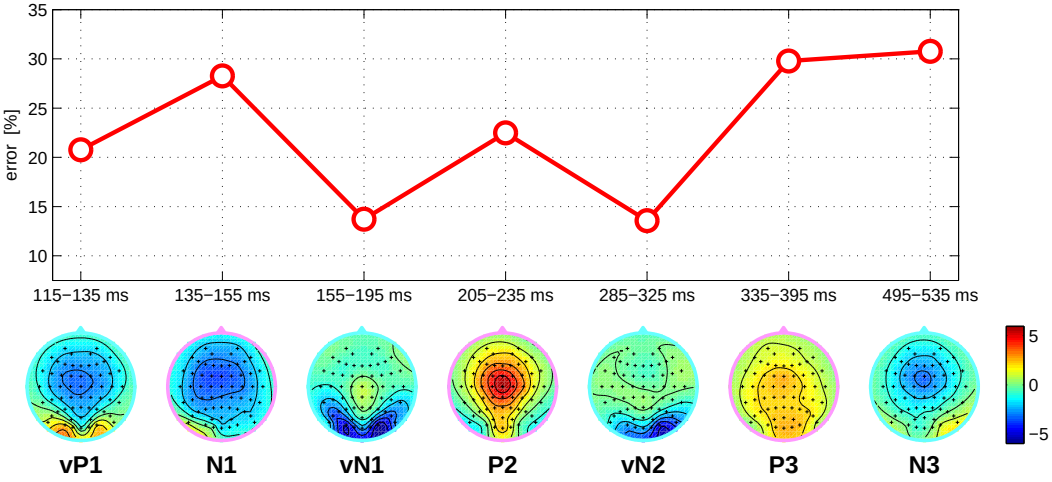
Recap: Classification of Spatial Features



The classification error of spatial features was determined for each time interval of 30 ms duration, shifted from 0 to 1000 ms. (Here, chance level is 16.6%).

Results of Classifying Spatial Features

Classifying on spatial features for various time intervals results in error rates between 14% and 31% in this example data set (visual speller):



Classification of Spatio-Temporal Features

Remember: purely temporal and purely spatial features are used to investigate discriminability (spatial and temporal profile). For an efficient classification spatio-temporal feature should be used.

Advancing from temporal or spatial features to spatio-temporal features means increasing the information.

Accordingly, a better classification performance is to be expected.

But in our example data set, the classification error **increases** from

- **14%** for the spatial feature at the best interval to
- **25%** for spatio-temporal features

Side Trip: Introduction to Pattern Recognition

Assumption

> *Objects have an underlying regularity, but individual observations are corrupted by random noise.*

Pattern Recognition / Machine Learning (ML) is about

- automatic discovery of regularities in data and
- using these regularities to take action such as classifying.

Some good text books on pattern recognition:

- Bishop: Pattern Recognition and Machine Learning.
- Hasties, Tibshirani & Friedman: The Elements of Statistical Learning, 2nd edition
- Duda, Hart & Storck: Pattern Classification, 2nd edition

Concept of Supervised Learning

We observe samples of noisy data $y = f(\mathbf{x}) + \varepsilon$ from an unknown fcn $f(\mathbf{x})$ with $\mathbf{x} \in \mathbb{R}^d$ and $y \in \mathbb{R}$ (or $y \in \{-1, +1\}$ for binary classification):

$$\langle (\mathbf{x}_k^{\text{tr}}, y_k^{\text{tr}}) \rangle_{k=1,\ldots,K_{\text{tr}}}$$

and the task is to learn an approximation of f.

General approach:

1. choose a class of functions $\mathcal{F}$ and
2. use ML method to select the best fit $\hat{f} \in \mathcal{F}$ on training data

$$y_k^{\text{tr}} \approx \hat{f}(\mathbf{x}_k^{\text{tr}}) \quad \text{for } k = 1, \ldots, K_{\text{tr}}$$

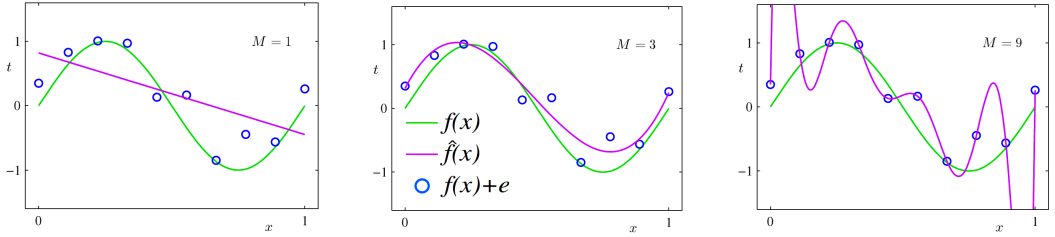
with good generalization to previously unseen test data

$$y_k^{\text{te}} \approx \hat{f}(\mathbf{x}_k^{\text{te}}) \quad \text{for } k = 1, \ldots, K_{\text{te}}$$

(quantified by a loss function).

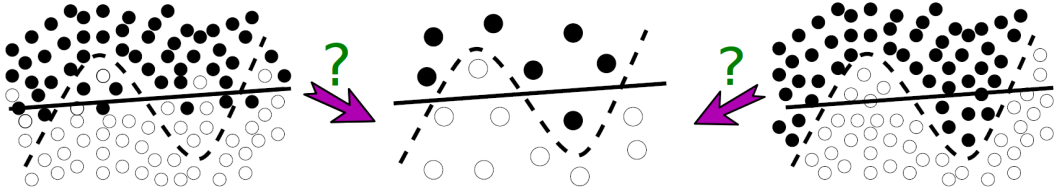
Complexity of the Function Class

Regression example: Fit 10 noisy samples of $f(x) = \sin(2\pi x)$ with a function from class $\mathcal{F}_M = \{\sum_{i=0}^{M} w_i x^i \mid \mathbf{w} \in \mathbb{R}^{M+1}\}$.



Classification example:

Linear or nonlinear separation? Undecidable with few samples.



Overfitting

Using a larger (more complex) function class allows a better fit with the training data.

However, despite a low training error, the selected function might not describe the regularity in the data well. It may also be overfitted to the noise that is present in the particular set of samples that is available as training data.

This overfitting becomes apparent in a cross-validation when the error on the training data deviates substantially from the error on the test data.

Back on the Main Path: Overfitting in LDA

After this short side trip, we return to the initial point:

When LDA was applied to high-dimensional (spatio-temporal) features, the performance broke down (result worse than on sub-features).

Even being aware of overfitting, given the optimality theorem, this should not happen, right?

So far, we did not discuss the third assumption:
The true distributions are known.

- This assumptions is always violated in non-artificial problems.
- Distribution parameters have to be estimated from given data.
- **Estimated** (empirical) distribution parameters necessarily deviate from the **true** ones.
- How much this deviation deteriorates performance is variable.

Bias in Estimating Covariance Matrices

For LDA we need estimates for the distribution parameters:

- $\hat{\boldsymbol{\mu}} = \frac{1}{K} \sum_{k=1}^{K} \mathbf{x}_k$ **empirical mean**
- $\hat{\boldsymbol{\Sigma}} = \frac{1}{K-1} \sum_{k=1}^{K} (\mathbf{x}_k - \hat{\boldsymbol{\mu}})(\mathbf{x}_k - \hat{\boldsymbol{\mu}})^\top$ **empirical covariance matrix**

But, if the number of samples K is not large relative to the dimension d ($\mathbf{x} \in \mathbb{R}^d$), the estimation, in particular $\hat{\boldsymbol{\Sigma}}$, is error-prone.

This may affect classification with LDA badly.

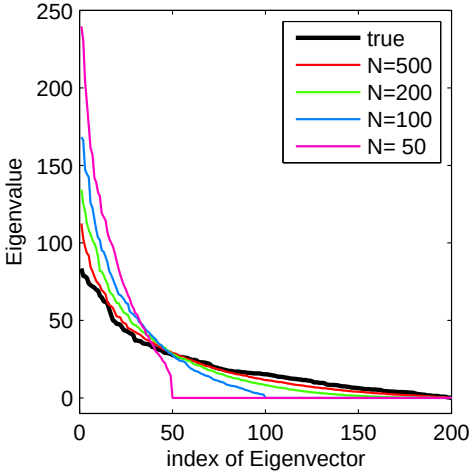
There is a systematical bias in the empirical covariance matrix:

- Large Eigenvalues of $\hat{\boldsymbol{\Sigma}}$ are too large and
- Small Eigenvalues of $\hat{\boldsymbol{\Sigma}}$ are too small

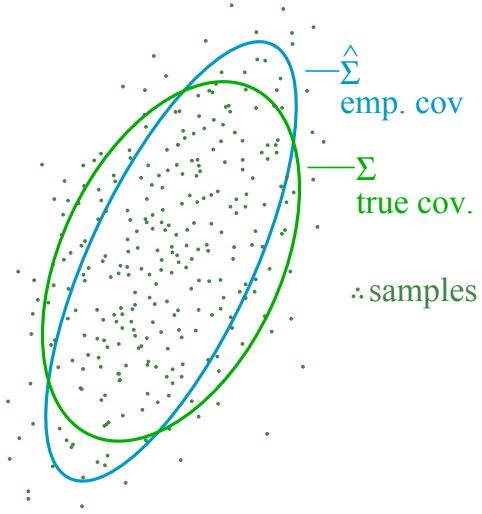
compared to those of $\boldsymbol{\Sigma}$ assuming $\mathbf{x}_1, \ldots, \mathbf{x}_K \in \mathbb{R}^d$ are drawn from a Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Bias in Estimating Covariances Visualized

Simulation for $d = 200$:



Cartoon in 2D:

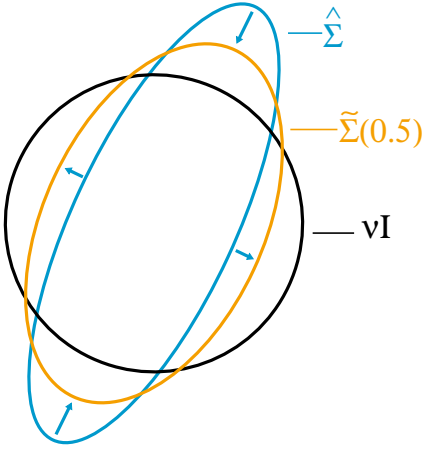


A Remedy for the Estimation Bias

A simple way that counteracts the bias is **shrinkage**:
The empirical covariance matrix $\hat{\mathbf{\Sigma}}$ is modified to be more spherical:

$$\tilde{\mathbf{\Sigma}}(\gamma) = (1-\gamma)\hat{\mathbf{\Sigma}} + \gamma\nu\mathbf{I}$$

for a $\gamma \in [0, 1]$ and ν defined as average Eigenvalue $\mathrm{trace}(\hat{\mathbf{\Sigma}})/d$.



Next, we check that shrinkage serves the intended purpose. Covariance matrices are described by their Eigenvectors and Eigenvalues.

So, we have to investigate, what happens to those, when we change over from the empirical covariance matrix $\hat{\mathbf{\Sigma}}$.

Properties of the Shrunk Covariance Matrix

From the Eigenvalue decomposition of the empirical covariance matrix $\hat{\mathbf{\Sigma}} = \mathbf{V}\mathbf{D}\mathbf{V}^\top$ with orthonormal $\mathbf{V}$ and diagonal $\mathbf{D}$, we get an Eigenvalue decomposition of $\tilde{\mathbf{\Sigma}}(\gamma) = (1-\gamma)\hat{\mathbf{\Sigma}} + \gamma\nu\mathbf{I}$ like this:

$$\begin{aligned}\tilde{\mathbf{\Sigma}}(\gamma) &= (1-\gamma)\mathbf{V}\mathbf{D}\mathbf{V}^\top + \gamma\nu\mathbf{I} \\ &= (1-\gamma)\mathbf{V}\mathbf{D}\mathbf{V}^\top + \gamma\nu\mathbf{V}\mathbf{I}\mathbf{V}^\top \\ &= \mathbf{V}\underbrace{\left((1-\gamma)\mathbf{D} + \gamma\nu\mathbf{I}\right)}_{\text{diagonal matrix}}\mathbf{V}^\top\end{aligned}$$

We see that (uniqueness of the EVD)

- $\hat{\mathbf{\Sigma}}$ and $\tilde{\mathbf{\Sigma}}(\gamma)$ have the same Eigenvectors (columns of $\mathbf{V}$)
- Extreme Eigenvalues (large/small) are shrunk/extended towards the average Eigenvalue ν as $d_i \mapsto (1-\gamma)d_i + \gamma\nu$
- $\gamma = 0$ means no shrinkage: $\tilde{\mathbf{\Sigma}}(0) = \hat{\mathbf{\Sigma}}$
- $\gamma = 1$ corresponds to spherical covariance matrices: $\tilde{\mathbf{\Sigma}}(1) = \nu\mathbf{I}$

Regularized Linear Discriminant Analysis

This technique can be used to enhance LDA to work better in the case of a low number-of-samples to dimensionality ratio. The empirical covariance matrix $\hat{\mathbf{\Sigma}}$ is replaced by a shrunk covariance matrix $\tilde{\mathbf{\Sigma}}(\gamma)$:

$$\mathbf{w}_\gamma := \tilde{\mathbf{\Sigma}}(\gamma)^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)$$

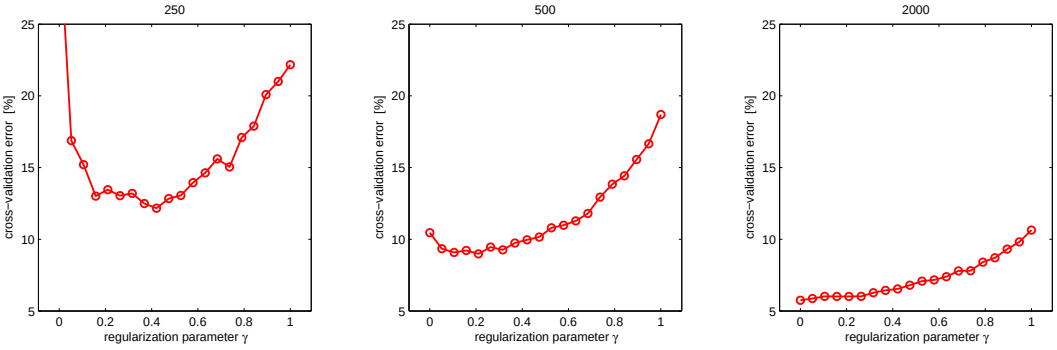
Here, γ is a hyperparameter that has to be selected between 0 and 1.

- $\gamma = 0$ yields $\mathbf{w}_0 = \hat{\mathbf{\Sigma}}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)$, i.e. unregularized LDA
- $\gamma = 1$ yields $\mathbf{w}_1 = \boldsymbol{\mu}_2 - \boldsymbol{\mu}_1$, i.e. NCC

But: There is no golden rule for setting γ. A pragmatic, however time-consuming way is to use **cross-validation** for the selection.

LDA with Different Shrinkage Parameters

Cross-validation results for different sizes of training data (250, 500, 2000) for different values of the shrinkage parameter γ (x-axis). Features vectors have 250 dimensions.



Optimal Selection of Shrinkage Parameter

As a (relatively) novel method for selecting the free parameter (γ) other than with cross-validation, there is an analytical method.

Let $\mathbf{x}_1, \ldots, \mathbf{x}_K \in \mathbb{R}^d$ be K feature vectors and let $\hat{\boldsymbol{\mu}} = \frac{1}{K} \sum_{k=1}^{K} \mathbf{x}_k$ be the empirical mean.

Aim: get a better estimate of the true covariance matrix $\boldsymbol{\Sigma}$ (especially in case $K < d$) than the sample covariance matrix

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{K-1} \sum_{k=1}^{K} (\mathbf{x}_k - \hat{\boldsymbol{\mu}})(\mathbf{x}_k - \hat{\boldsymbol{\mu}})^\top$$

by selecting a γ in

$$\tilde{\boldsymbol{\Sigma}}(\gamma) := (1-\gamma)\hat{\boldsymbol{\Sigma}} + \gamma \nu \mathbf{I}.$$

Optimal Selection of Shrinkage Parameter

The approach of [Ledoit & Wolf, J Multivar Anal, 2004] is to minimize

$$\|\tilde{\mathbf{\Sigma}}(\gamma) - \mathbf{\Sigma}\|_F^2 \qquad \text{with } \|\cdot\|_F^2 \text{ being the Frobenius norm.}$$

We define the covariance matrix **of trial** k:

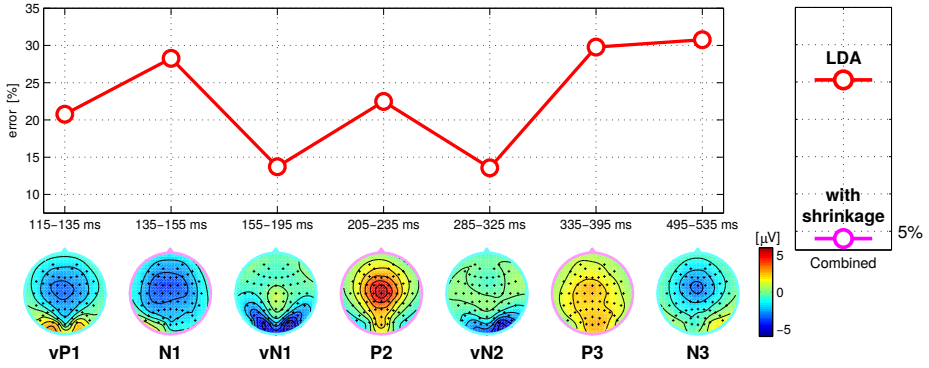
$$\mathbf{Z}^k = (\mathbf{x}_k - \hat{\boldsymbol{\mu}})\,(\mathbf{x}_k - \hat{\boldsymbol{\mu}})^\top$$

Denoting by s_{ij} the element in the i-th row and j-th column of the matrix $\hat{\mathbf{\Sigma}} - \nu\mathbf{I}$, the optimal shrinkage parameter $\gamma^\star = \text{argmin}_\gamma \|\tilde{\mathbf{\Sigma}}(\gamma) - \mathbf{\Sigma}\|_F^2$ can be analytically calculated as [Schäfer & Strimmer 2005]

$$\gamma^\star = \frac{K}{(K-1)^2} \frac{\sum_{i,j=1}^{d} \text{var}\,\langle (\mathbf{Z}^k)_{ij} \mid k = 1, \ldots, K \rangle}{\sum_{i,j=1}^{d} s_{ij}^2}.$$

Shrinkage-LDA: use $\tilde{\mathbf{\Sigma}}(\gamma^\star)$ instead of $\hat{\mathbf{\Sigma}}$.

Classification on Single Components and Combined



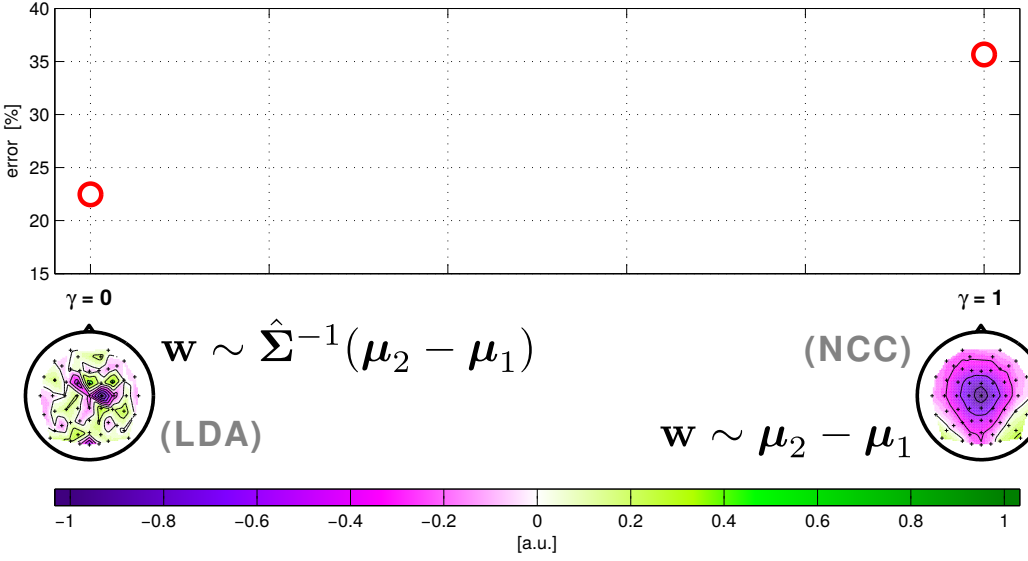
Classification (with $N = 750$ training samples) on seven different single components ($d = 55$ each) yields errors between **14%** and 31%.

LDA on the concatenated feature ($d = 7 \cdot 55 = 385$) performs with **25%** worse, although information is added: *overfitting*.

Shrinkage-LDA: only **4%** error. [Blankertz et al, NeuroImage 2011]

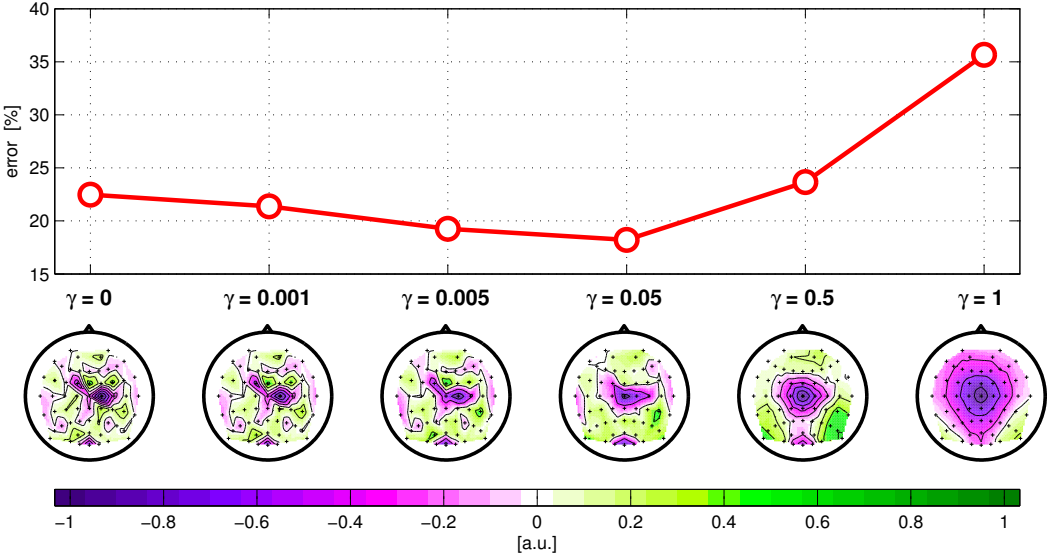
Impact of Shrinkage as Trade-off

LDA with shrinkage: $\mathbf{w} = \tilde{\Sigma}(\gamma)^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1); \quad \tilde{\Sigma}(\gamma) = (1-\gamma)\hat{\Sigma} + \gamma\nu\mathbf{I}$

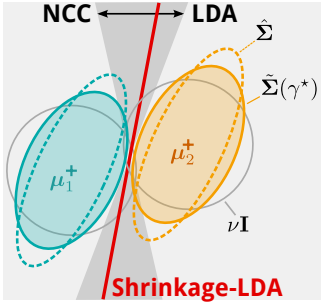


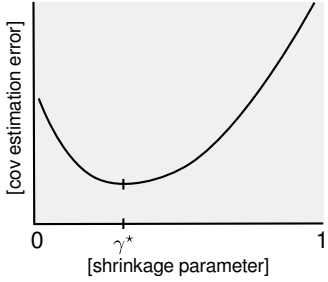
Impact of Shrinkage as Trade-off

With increasing shrinkage, the spatial filters (classifier) look smoother, but classification may degrade with too much shrinkage.



Classification with Shrinkage-LDA at a Glance





Shrinkage-LDA hyperplane is defined by:

$$\mathbf{w} := \tilde{\Sigma}(\gamma^\star)^{-1}\,(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)$$

$$\tilde{\Sigma}(\gamma) := (1-\gamma)\hat{\Sigma} + \gamma\nu\mathbf{I}$$

Selection of shrinkage parameter γ by calculating the 'optimal' $\gamma^\star$ analytically:

$$\gamma^\star = \operatorname*{argmin}_{\gamma} \|\tilde{\Sigma}(\gamma) - \Sigma\|_F^2$$

$$= \frac{K}{(K-1)^2}\,\frac{\sum_{i,j=1}^{d} \operatorname{var}_k\,(\mathbf{Z}^k)_{ij}}{\sum_{i,j=1}^{d} s_{ij}^2} \quad \text{with}$$

$$\mathbf{Z}^k = (\mathbf{x}_k - \hat{\boldsymbol{\mu}})\,(\mathbf{x}_k - \hat{\boldsymbol{\mu}})^\top$$

[Ledoit & Wolf 2004], [Schäfer & Strimmer 2005]

Lessons Learnt

After this lecture you should

- be aware of the problem when classifying high dimensional features and how it corresponds to the LDA optimality statement
- have an idea about the bias in the empirical covariance matrix
- know how to counterbalance the estimation bias with shrinkage
- be familiar with the Shrinkage-LDA

References I

- Blankertz, B., Lemm, S., Treder, M. S., Haufe, S., and Müller, K.-R. (2011).
  Single-trial analysis and classification of ERP components – a tutorial.
  NeuroImage, 56:814–825.
- Ledoit, O. and Wolf, M. (2004).
  A well-conditioned estimator for large-dimensional covariance matrices.
  J Multivar Anal, 88:365–411.
- Schäfer, J. and Strimmer, K. (2005).
  A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics.
  Stat Appl Genet Mol Biol, 4:Article32.