BME I5100: Biomedical Signal Processing

Linear Discrimination

Lucas C. Parra
Biomedical Engineering Department
City College of New York

CCNY

Schedule

Week 1: Introduction
Linear, stationary, normal - the stuff biology is **not** made of.

Week 1-4: Linear systems
Impulse response
Moving Average and Auto Regressive filters
Convolution
Discrete Fourier transform and z-transform
Sampling

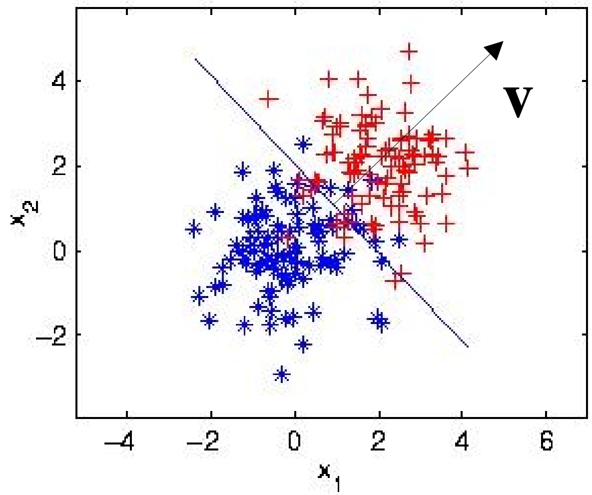
Week 5-8: Random variables and stochastic processes
Random variables
Moments and Cumulants
Multivariate distributions
Stochastic processes

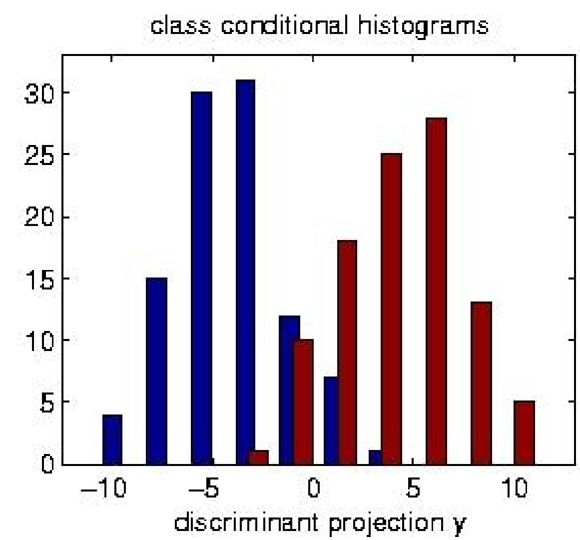
Week 9-14: Examples of biomedical signal processing
Probabilistic estimation
Harmonic analysis - **estimation** circadian rhythm and speech
Linear discrimination - **detection** of evoked responses in EEG/MEG
Independent components analysis - **analysis** of MEG signals
Dynamical Models - Kalman filter and Hidden Markov Models
Matched and Wiener filter - **filtering** in ultrasound

Linear Discrimination

Given samples $\boldsymbol{x}$ from two classes c_1 and c_2 find vector $\mathbf{v}$ so that the projection y separates the two classes:

$$y = \boldsymbol{v}^T \boldsymbol{x} + v_0$$





Redefine $\boldsymbol{x} = [1, x_1, x_2, \ldots, x_d]^T$ and $\boldsymbol{v} = [v_0, v_1, v_2, \ldots, v_d]^T$ to write in short:

$$y = \boldsymbol{v}^T \boldsymbol{x}$$

Linear Discrimination - Logistic Regression

The goal is to find a mapping from $\boldsymbol{x}$ to class label c or at least an expected value:

$$\hat{c} = f(y) = f(\boldsymbol{v}^T \boldsymbol{x})$$

However, since there is overlap, rather than making a fixed determination on the class label we will build a model that tells us what is the likelihood of a class c given input $\boldsymbol{x}$.

$$p(c|y) = p(c|\boldsymbol{v}^T \boldsymbol{x})$$

We will now derive an expression of the Likelihood of class labels c given projection y, which are given by input $\boldsymbol{x}$ and parameters $\boldsymbol{v}$. The optimal $\boldsymbol{v}$ results then from ML.

Linear Discrimination - Logistic Regression

Consider the posterior probability $p(c_1|\, y)$

$$p(c_1|y)=\frac{p(y|c_1)\,p(c_1)}{p(y|c_2)\,p(c_2)+p(y|c_1)\,p(c_1)}=\frac{1}{1+\exp(-l)}$$

We introduced here the **Likelihood ratio** l:

$$l=\ln\frac{p(y|c_1)\,p(c_1)}{p(y|c_2)\,p(c_2)}$$

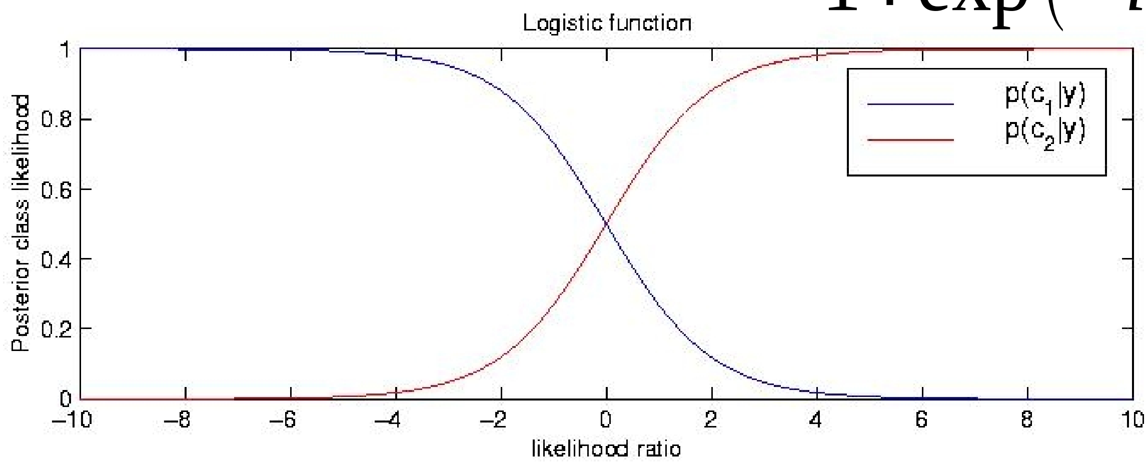
For a large class of distributions, $p(y|c)$, called exponential family, the Likelihood ratio simplifies under certain assumptions to

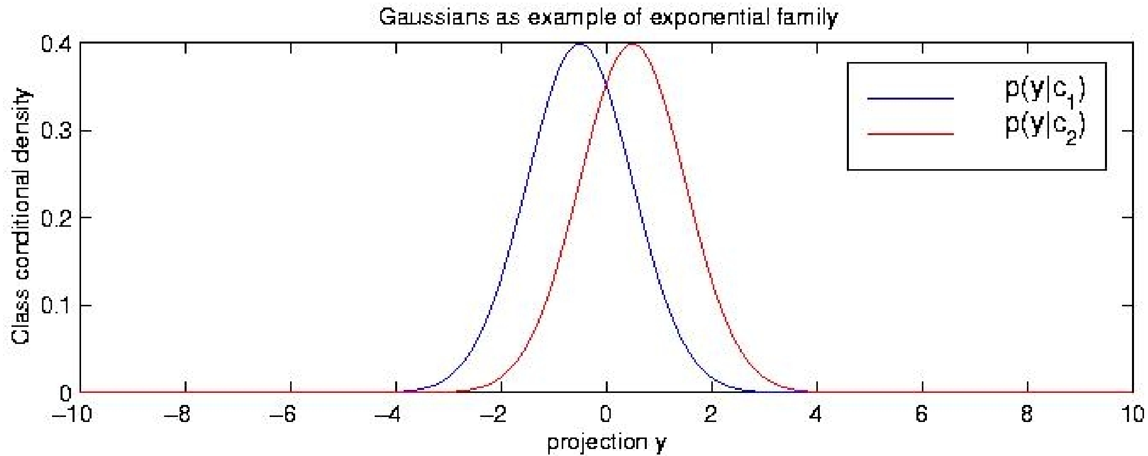
$$l=y+\ln\frac{p(c_1)}{p(c_2)}$$

Linear Discrimination - Logistic Regression

For example Gaussian projections, $p(y|c) = N(y;\mu_c,\sigma)$ with different mean for each class but the same standard deviation.

Logistic function: $$p(c_1|y) = \frac{1}{1+\exp(-l)}$$





Linear Discrimination - Gaussian Data

Example: If the two classes are Gaussian distributed with different mean for each class but the *same covariance matrix* their optimal separation is linear and solution is particularly simple:

$$p(\boldsymbol{x}|c) \propto \exp\left[-(\boldsymbol{x}-\boldsymbol{\mu_c})^T \boldsymbol{\Sigma^{-1}} (\boldsymbol{x}-\boldsymbol{\mu_c})/2\right]$$

The likelihood ration is then linear

$$l = \ln \frac{p(\boldsymbol{x}|c_1)\, p(c_1)}{p(\boldsymbol{x}|c_2)\, p(c_2)} = \boldsymbol{v}^{\boldsymbol{T}} \boldsymbol{x} + v_0$$

where the separation vector and bias are given by the means and covariance:

$$\boldsymbol{v} = \Sigma^{-1}(\mu_1 - \mu_2)$$

$$v_0 = -\frac{1}{2}\mu_1^T \Sigma^{-1} \mu_1 + \frac{1}{2}\mu_2^T \Sigma^{-1} \mu_2 + \ln \frac{p(c_1)}{p(c_2)}$$

Linear Discrimination - Logistic Regression

More generally, assume that y is distributed according to some distribution of the exponential family, which includes Gaussian, Bernoulli, Poission, and others.

The goal is then to find the optimal projection vector **v** such that the projections, $y=\boldsymbol{v}^T\boldsymbol{x}$, results in a logistic likelihood for the class labels

$$p(c_1|y)=\frac{1}{1+\exp\left(-\boldsymbol{v}^T\boldsymbol{x}\right)}=f\left(\boldsymbol{v}^T\boldsymbol{x}\right)$$

Here we have absorbed ln $p(c_1)/p(c_2)$ into v_0.

Linear Discrimination - Logistic Regression

To derive the ML solution define, c=0, and, c=1, to identify the two different classes and denote in short $p(c|y)=f$. We can write then

$$p(c|y)=f^{c}(1-f)^{1-c}$$

This is called the Bernoulli density of c, and f is the mean:

$$\hat{c}=E[c]=f(\boldsymbol{v}^{T}\boldsymbol{x})$$

Note that the expected value we wanted to compute is therefore given now by the logistic function.

Linear Discrimination - Logistic Regression

Given i.i.d. samples $x[k]$, $c[k]$ the log-likelihood of the data is

$$
\begin{aligned}
L(\boldsymbol{v}) &= \ln \prod_k p(c[k], \boldsymbol{x}[k] | \boldsymbol{v}) \\
&= \sum_k \ln p(c[k] | \boldsymbol{v}^T \boldsymbol{x}[k]) \, p(\boldsymbol{x}[k]) \\
&= \sum_k c[k] \ln f(\boldsymbol{v}^T \boldsymbol{x}[k]) + (1 - c[k]) \ln(1 - f(\boldsymbol{v}^T \boldsymbol{x}[k])) \\
&\quad + const.
\end{aligned}
$$

And the optimum solution according to ML is

$$\underset{\boldsymbol{v}}{argmin} \;\; L(\boldsymbol{v})$$

There is not close form solution for this minimum.

Linear Discrimination - Logistic Regression

However, the minimum can be computed using a fast algorithm based on Iteratively Reweighted Least Squares (IRLS). It is a type of Newton-Raphson gradient descent algorithm called Fisher Scoring method (McCullagh,Nelder 1983):

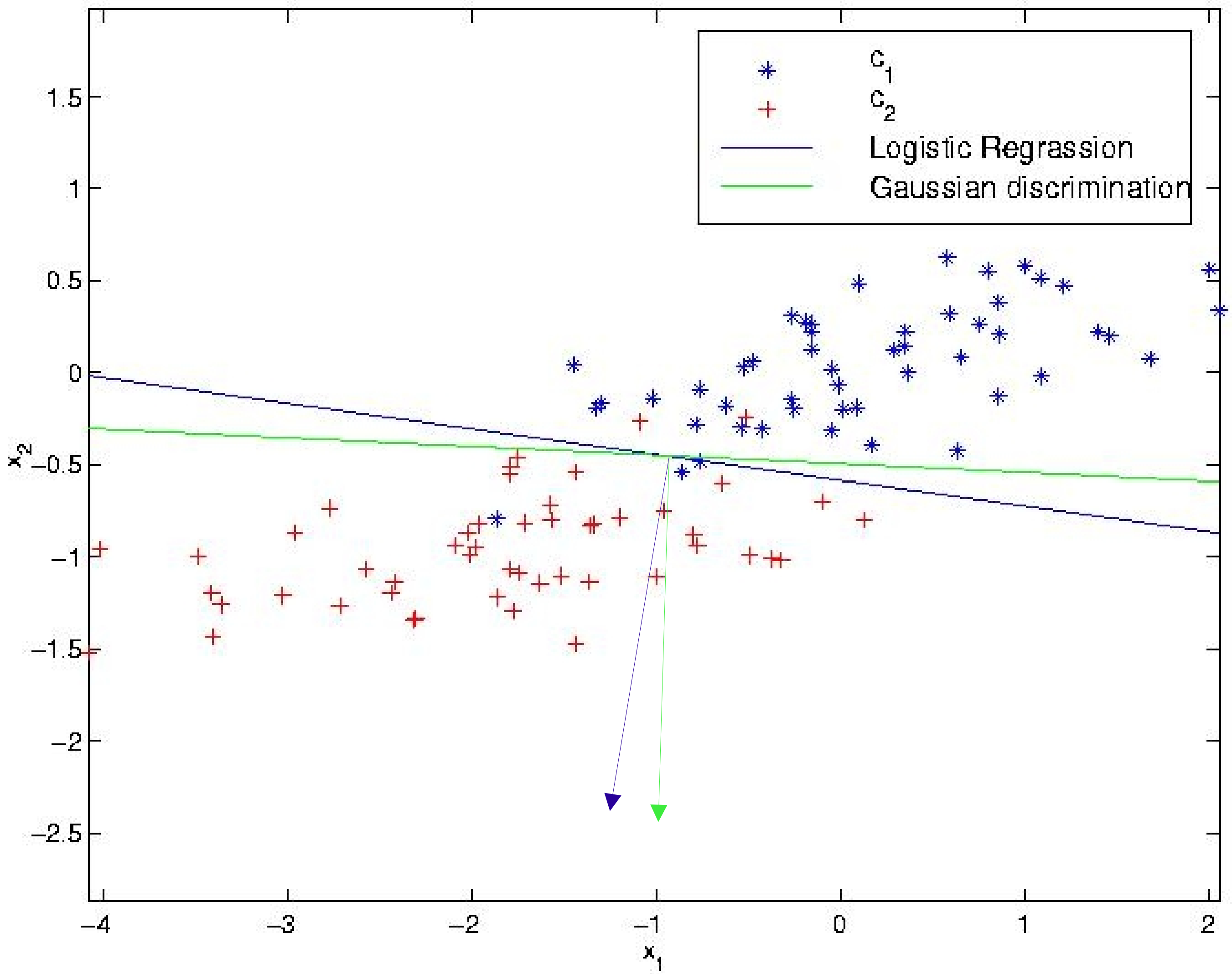
$$\boldsymbol{v}_{t+1} = \boldsymbol{v}_t - E\left[\frac{\partial L(\boldsymbol{v})}{\partial \boldsymbol{v}\, \partial \boldsymbol{v}^T}\right]^{-1} \frac{\partial L(\boldsymbol{v})}{\partial \boldsymbol{v}}$$

The expected Hessian can be computed fast and converges typically within a few iterations.

```
>> v = logist(x,c);  % not part of matlab
```

Linear Discrimination - Comparison

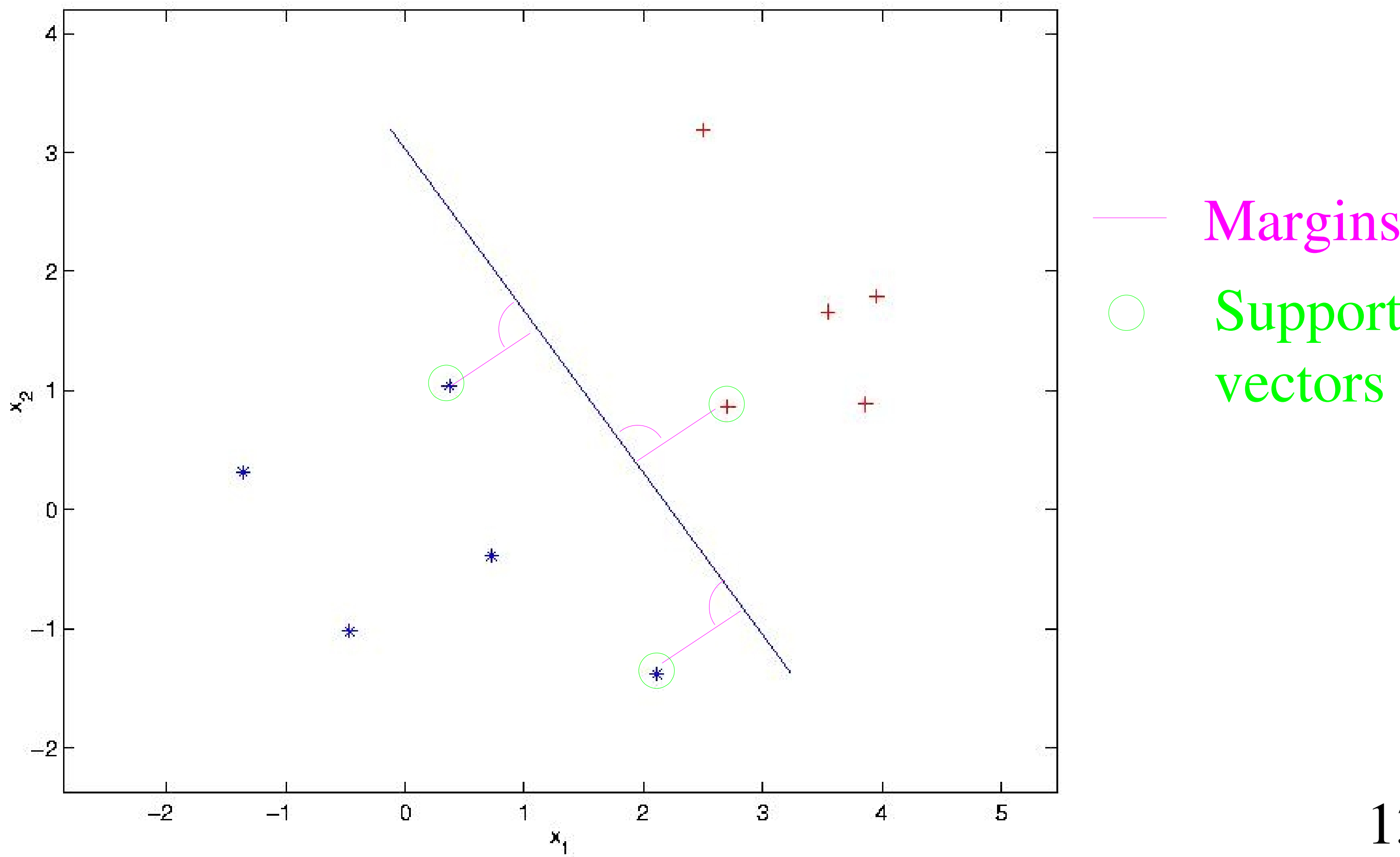
Gaussian solution and and minimum $L(\mathbf{v})$ give different results when when data not Gaussian or for small sample size.



Linear Discrimination - Support Vectors

In case of perfect separability LR is not well defined.

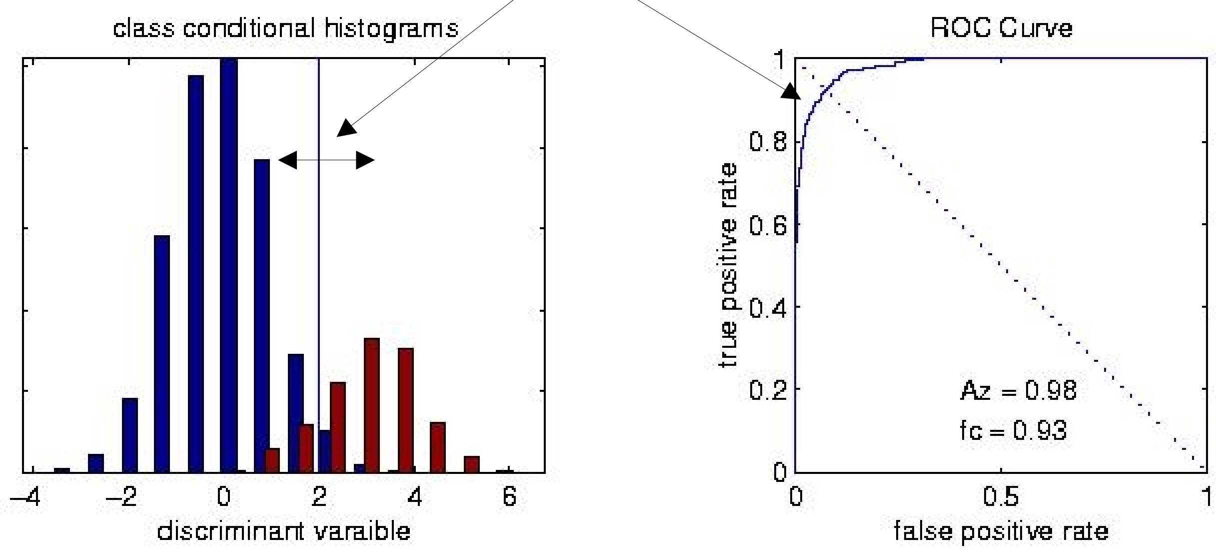
Occurs often when we have very few samples and/or high dimensions. In that case is it better to chose the separation that **maximizes the margin** to **support vectors**.



Linear Discrimination - Performance

The performance of a binary classification problem is typically evaluated with a **Receiver Operator Characteristic** (ROC) curve:

Moving threshold sweeps a curve of true positive rate versus false positive rate



Az: Area under the ROC curve measures performance independent of threshold. It is 0.5 for chance performance.
fc: Fraction correct (1-error rate) is conventionally measured where tp=1-fp. It is 0.5 for chance performance.

Linear Discrimination - Leave-One-Out

Note that the performance on the training data is biased and always better than the performance on unseen data.

Therefore ROC and Az has to always be computed on unseen test data!

If there is not sufficient data available to separate into training and test set one should use the **leave-one-out procedure**:

1. For each sample k
   Find the optimal $\boldsymbol{v}$ on all data exept $c[k],x[k]$.
   Use this optimal $\boldsymbol{v}$ to compute the leave-one-out class likelihood $p(\, c[k] \mid \boldsymbol{v}^T\mathbf{x}[k])$.

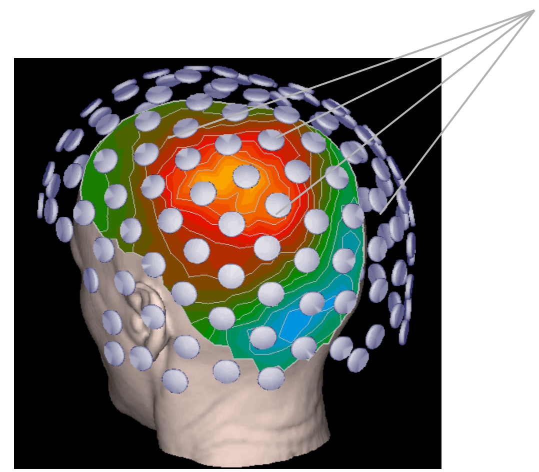
2. Compute Az using leave-one-out class likelihoods.

LD - Application to EEG and MEG

Conventional Event Related Potentials (ERP) averages over trials to increase signal to noise ratio.

The goal is to detect single trials without averaging over trials or over time. We substitute trial averaging by spatial integration.

$$y(t) = \boldsymbol{v}^T \boldsymbol{x}(t)$$

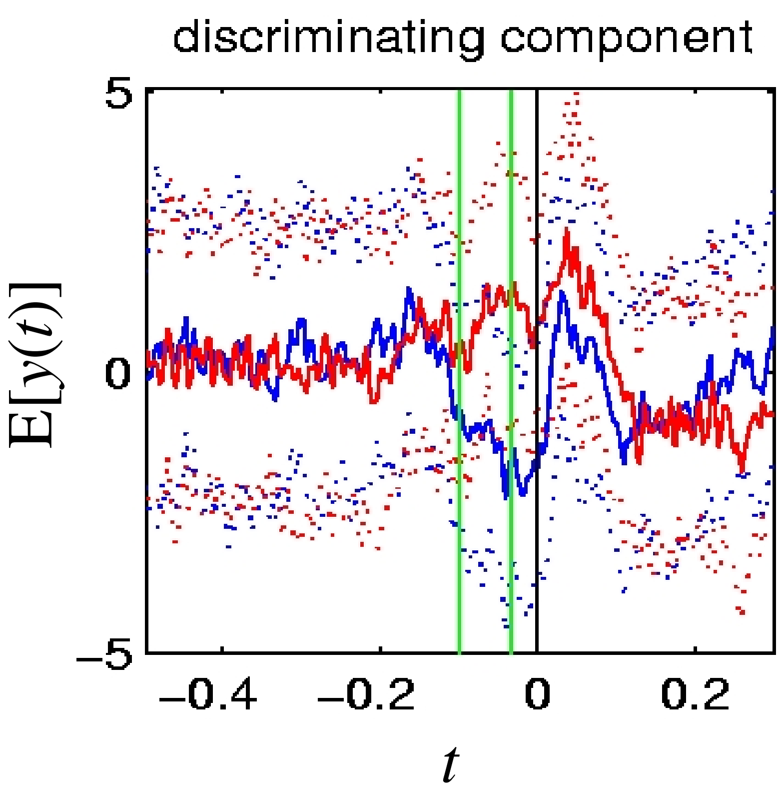


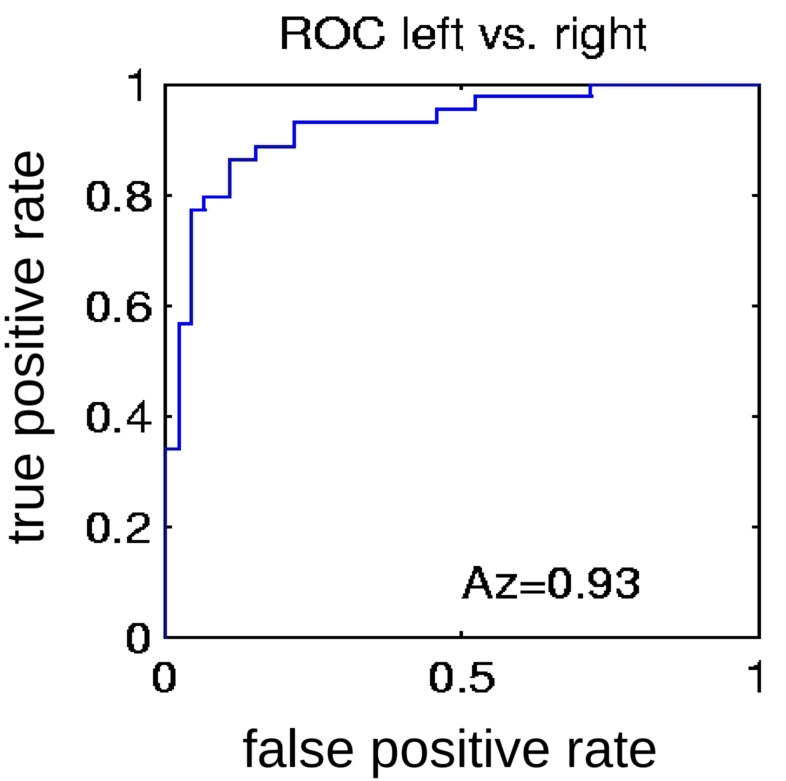
With Linear discrimination we can now compute spatial weights **v** which maximally discriminate sensor array signals $\mathbf{x}(t)$ for two different conditions observed at times t_1 and t_2.

LD - Application to EEG and MEG

Example: Find motor planing activity in MEG

Predict button press from 122 MEG sensors with linear discriminator **w** such that $y(t)$ differs the most during 100 ms to 30 ms *prior* to left (t_1) and right (t_2) button push.





LD - Application to EEG and MEG

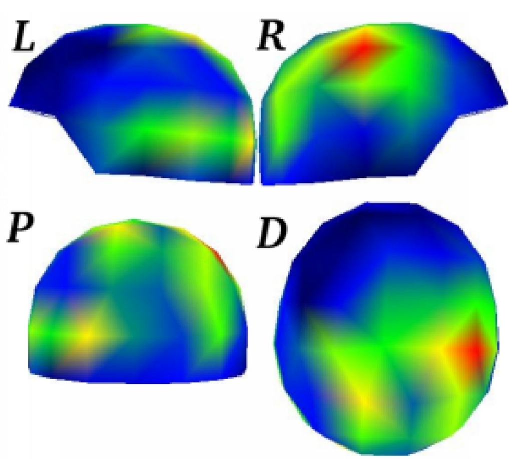
Localization of discriminating component: What is the electrical coupling $\boldsymbol{a}$ of the hypothetical source $\boldsymbol{y}$ that explains most of the activity $\boldsymbol{X}$?

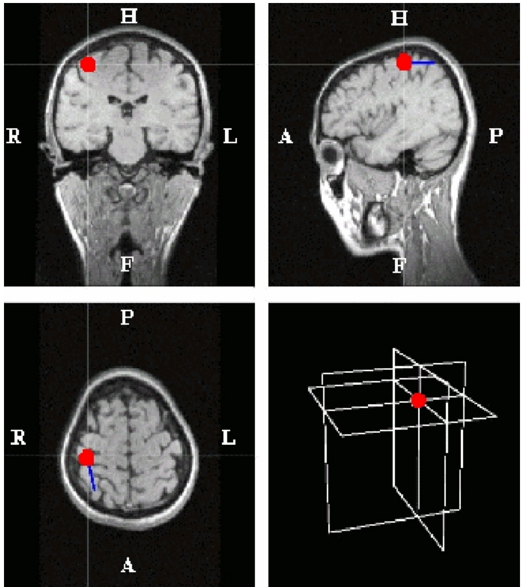
Least squares solution:

$$\boldsymbol{a} = \frac{\boldsymbol{X}\,\boldsymbol{y}}{\boldsymbol{y}^T\,\boldsymbol{y}}$$

where $\boldsymbol{X}$ has one column per sample, and $\boldsymbol{y}$ is a vector with all samples. $\boldsymbol{X}$ has to be zero mean across samples.

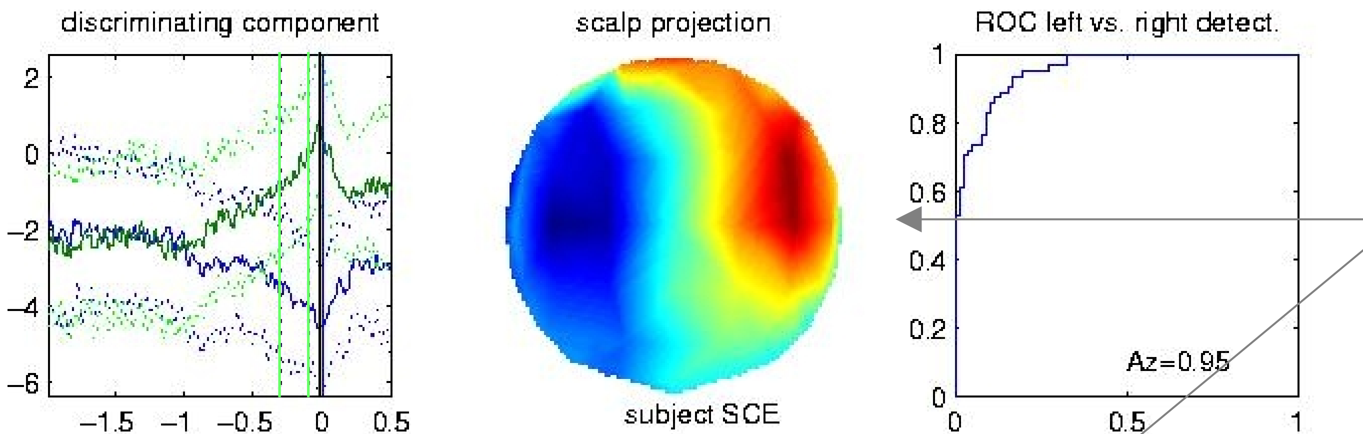
Strong coupling indicates low attenuation. Intensity on these "sensor projections" **a** indicates closeness of the source to the sensors.





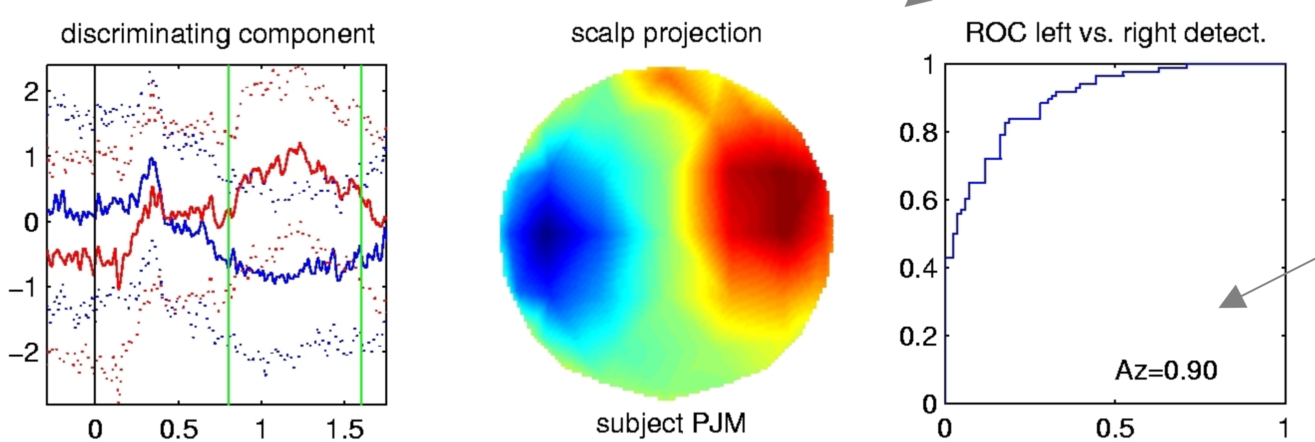
LD - Detection of Motor Planing and Imagery

Prediction of **explicit** finger tap (59 EEG sensors, 250-100ms prior)



Activity of explicit and imagined motor action have similar spatial distribution.

Detection of **imagined** finger tap (59 EEG sensors, 800 ms)

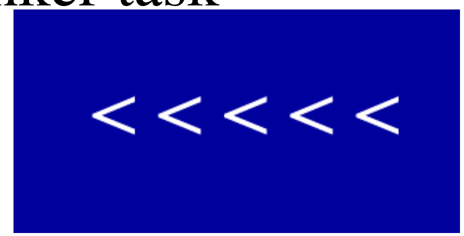


Transmission of binary information with this covert mental imagery results in communication at 12 bits/minute.

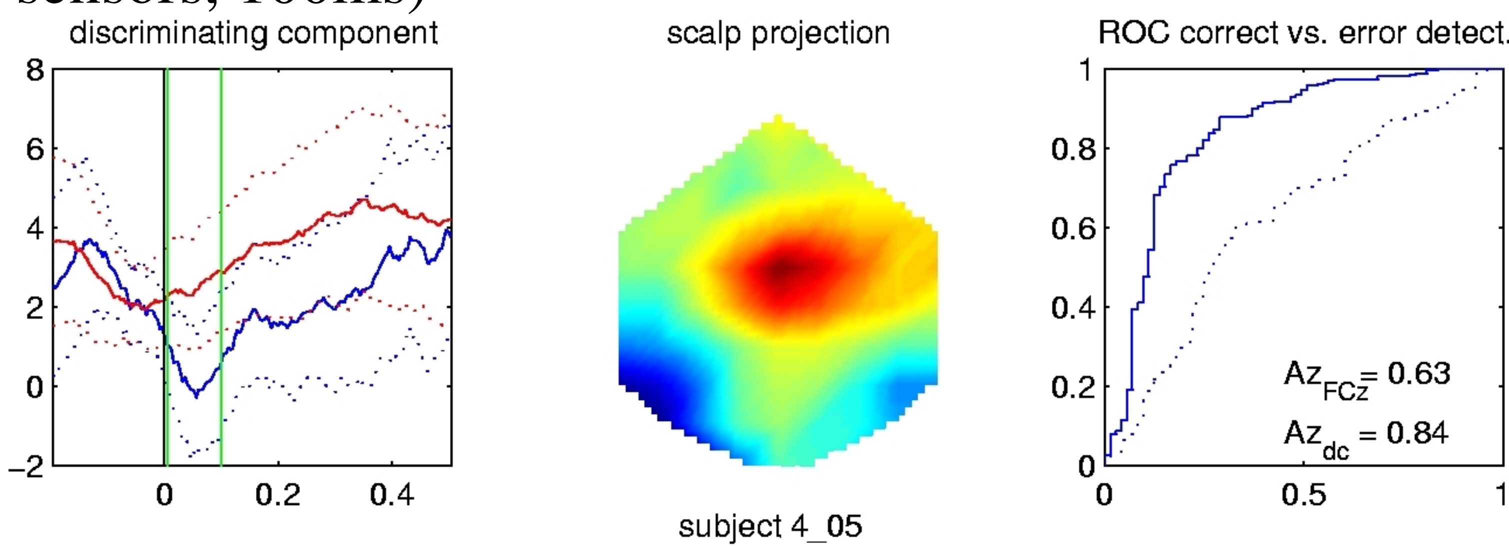
LD - Detection of Error Related Negativity

Error Related Negativity (ERN) occurs following perception of errors. It is hypothesized to originate in Anterior Cingulate and to represent response conflict or subjective loss.

Example: Erikson Flanker task

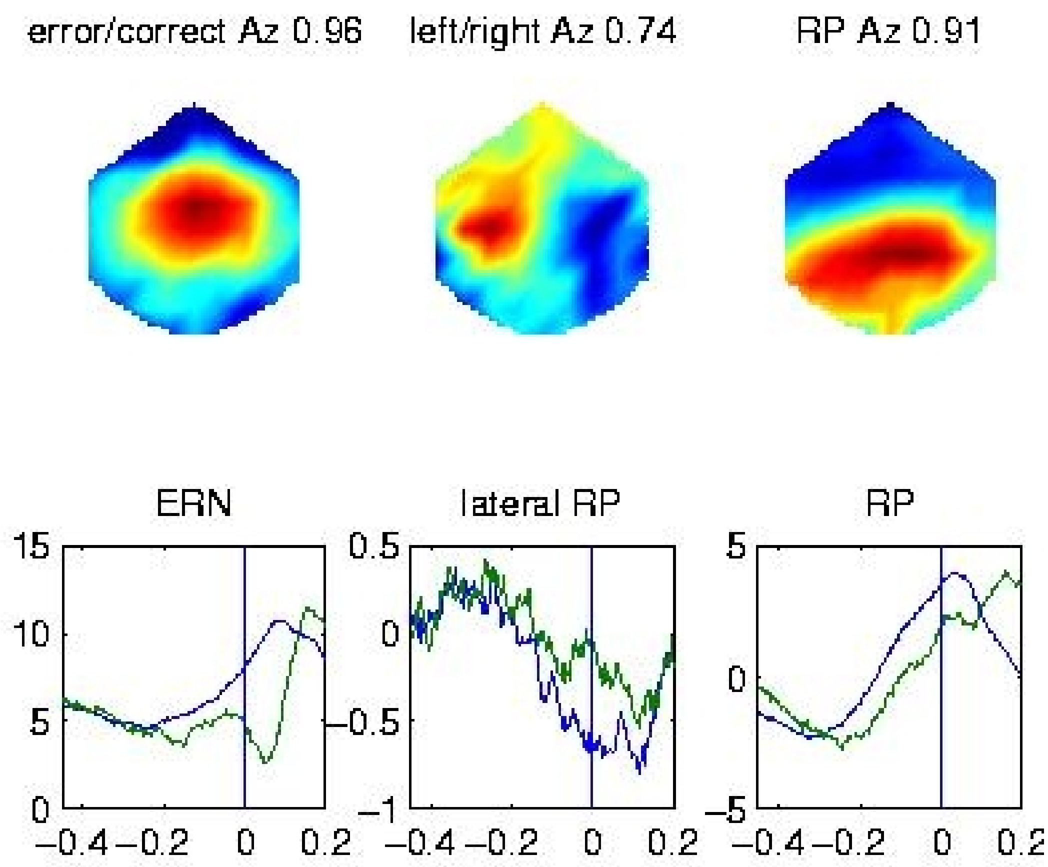


Discrimination of error versus correct response (64 EEG sensors, 100ms)



LD - Detection of Readiness Potential

In preparation of motion there is a potential buildup in the 200 ms prior to motion over motor areas. It is called readiness potential. When differentiating left/right one observed a lateralized readiness potential.



LD - Detection of Readiness Potential

Assignment 11:

Load `eeg-ern.mat` or generate overlapping random variables `x1` and `x2`.

If you use `eeg-ern.mat` note that for every trail (78 for `x1` and 300 for `x2`) there are each 25 samples. Consider them all as i.i.d. samples.

Find a linear discriminator that discriminates between `x1` and `x2` assuming Gaussian distributions.

Plot resulting **v** and show ROC and Az value on training data.

Optional:

Display scalp projection a using `scalp(coord,a)`

Show ROC curve and Az with the average *y* over 25 samples of each trial.