Linear Discriminant Analysis
LDA

Predicting Group Membership

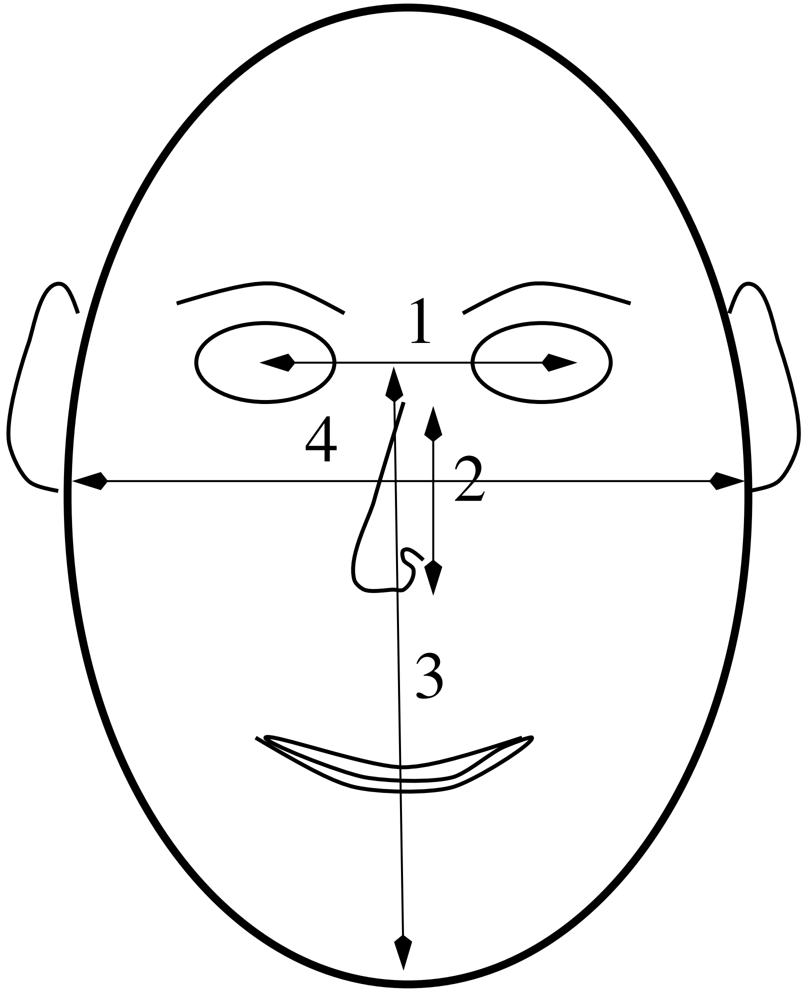
Hervé Abdi
STA 201

What is LDA for?

- Predict group membership from Variables
- X: an *I* by *J* predictor matrix
- Y: an *I* by *K* 0/1 group matrix

Boys and Girls!

Take four Measures on Each face



Predictors: 4 Variables (length in pixels)

Gender	1	2	3	4
G1	180	156	330	450
G2	168	156	360	480
G3	168	156	360	510
B1	156	144	360	480
B2	210	150	366	480
B3	162	144	342	438

C = Means	174	151	353	473

What Do We Want?

- **We want predict gender from the variables**
- So let **Y** being the "Gender" matrix (DV)

Subject	Girl	Boy
G1	1	0
G2	1	0
G3	1	0
B1	0	1
B2	0	1
B3	0	1

How to Do It?

- Best prediction = Best separation between groups
- Best prediction = largest F in ANOVA
- Recall: $F = MS_{between} / MS_{within}$
- $F = (SS_{between} / SS_{within}) * (N - K) / (K - 1)$
- $(N - K) / (K - 1)$ is fixed
- So largest $F \rightarrow$ max $(SS_{between} / SS_{within})$

IDEA: COMBINE THE PREDICTORS

- Create a new variable **f** with largest *F*
- In fact largest ratio SS_B / SS_W
- **f** = **Xq**   **f** is a linear combination of columns of **X**

In case We Have Forgotten

- SS_B = sum of squared deviations Between Groups
- SS_W = sum of squared deviations Within Groups

Here: For Example. X is Total Deviation

Gender	1	2	3	4
G1	6	5	-23	-23
G2	-6	5	7	7
G3	-6	5	7	37
B1	-18	-7	7	7
B2	36	-1	13	7
B3	-12	-7	-11	-35

Means	0	0	0	0

Basic Relation of ANOVA

- Total deviation = between group + Within groups

How to Get X_B the Between Group deviation?

First Step: Compute the Group Means

Gender	1	2	3	4
G1	6	5	-23	-23
G2	-6	5	7	7
G3	-6	5	7	37
B1	-18	-7	7	7
B2	36	-1	13	7
B3	-12	-7	-11	-35

Group	1	2	3	4
G	-2	5	-3	7
B	2	-5	3	-7

Second Step: Create Between X_B

Gender	1	2	3	4
G1	-2	5	-3	7
G2	-2	5	-3	7
G3	-2	5	-3	7
B1	2	-5	3	-7
B2	2	-5	3	-7
B3	2	-5	3	-7

Group	1	2	3	4
G	-2	5	-3	7
B	2	-5	3	-7

How to Get X_W the Within Group deviation?

FIRST STEP: COMPUTE THE GROUP MEANS

Gender	1	2	3	4
G1	6	5	-23	-23
G2	-6	5	7	7
G3	-6	5	7	37
B1	-18	-7	7	7
B2	36	-1	13	7
B3	-12	-7	-11	-35

Group	1	2	3	4
G	-2	5	-3	7
B	2	-5	3	-7

X_W

Second Step: Subtract the Means to get Within X_W: $X - M_G$

Gender	1	2	3	4
G1	8	0	-20	-30
G2	-4	0	10	0
G3	-4	0	10	30
B1	-20	-2	4	14
B2	34	4	10	14
B3	14	-2	-14	-28

Group	1	2	3	4
G	-2	5	-3	7
B	2	-5	3	-7

Fundamental Relations

- Total deviation is Between + Within. $X = X_B + X_W$
- Total Sum of Squares SS_T is: $X^T X$
- Between & Within Orthogonal: $X_B^T X_W = 0$
- And So: $SS_{Total} = SS_{Between} + SS_{Within}$
- So: $SS_T = X^T X = (X_B + X_W)^T (X_B + X_W)$

$$= X_B^T X_B + X_W^T X_W$$

Recap: $X = X_B + X_W$

1	2	3	4
6	5	-23	-23
-6	5	7	7
-6	5	7	37
-18	-7	7	7
36	-1	13	7
-12	-7	-11	-35

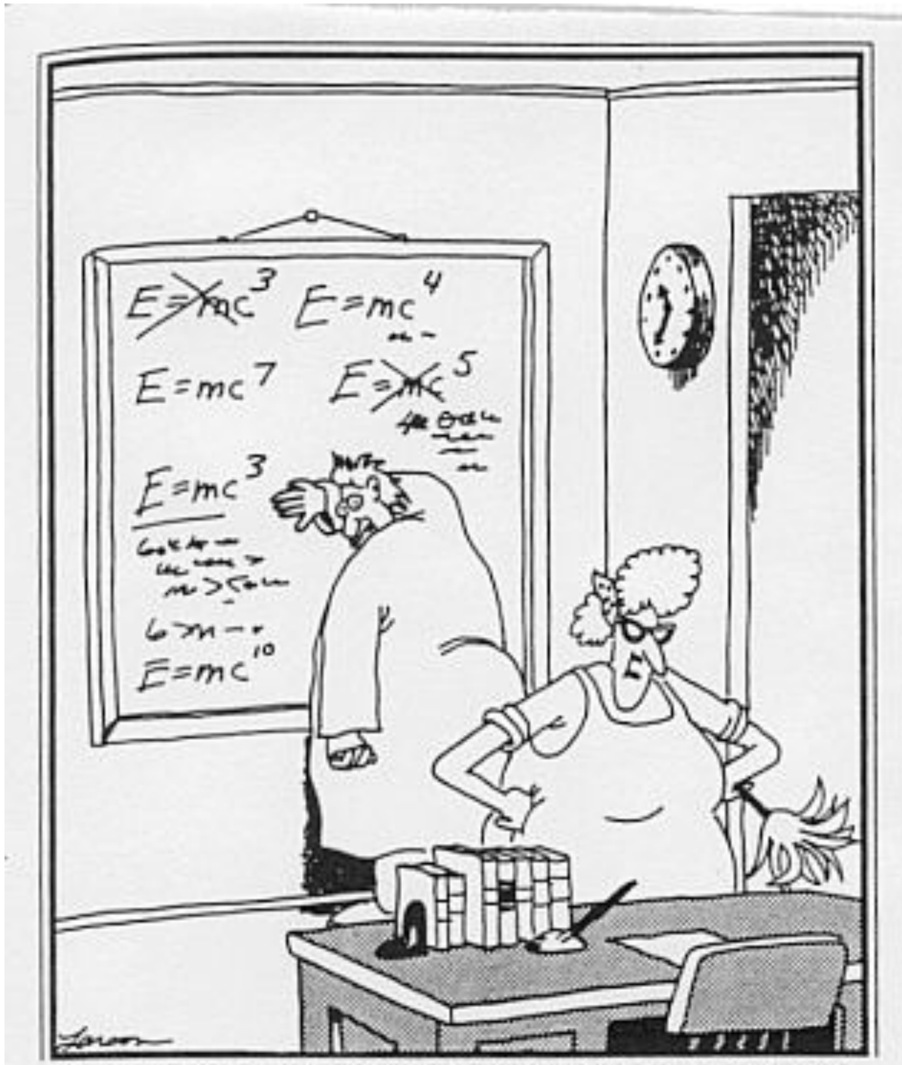
=

1	2	3	4
-2	5	-3	7
-2	5	-3	7
-2	5	-3	7
2	-5	3	-7
2	-5	3	-7
2	-5	3	-7

+

1	2	3	4
8	0	-20	-30
-4	0	10	0
-4	0	10	30
-20	-2	4	14
34	4	10	14
14	-2	-14	-28

Let us Square All That …



"Now that desk looks better. Everything's squared away, yessir, squaaaaaared away."

RECAP: $X = X_B + X_W$

1	2	3	4
6	5	-23	-23
-6	5	7	7
-6	5	7	37
-18	-7	7	7
36	-1	13	7
-12	-7	-11	-35

=

1	2	3	4
-2	5	-3	7
-2	5	-3	7
-2	5	-3	7
2	-5	3	-7
2	-5	3	-7
2	-5	3	-7

+

1	2	3	4
8	0	-20	-30
-4	0	10	0
-4	0	10	30
-20	-2	4	14
34	4	10	14
14	-2	-14	-28

TOTAL = B + W

SQUARED AWAY: $X \quad X_B \quad X_W$

$$\begin{bmatrix} 36 & 25 & 529 & 529 \\ 36 & 25 & 49 & 49 \\ 36 & 25 & 49 & 1369 \\ 324 & 49 & 49 & 49 \\ 1296 & 1 & 169 & 49 \\ 144 & 49 & 121 & 1225 \end{bmatrix} \begin{bmatrix} 4 & 25 & 9 & 49 \\ 4 & 25 & 9 & 49 \\ 4 & 25 & 9 & 49 \\ 4 & 25 & 9 & 49 \\ 4 & 25 & 9 & 49 \\ 4 & 25 & 9 & 49 \end{bmatrix} \qquad \begin{bmatrix} 64 & 0 & 400 & 900 \\ 16 & 0 & 100 & 0 \\ 16 & 0 & 100 & 900 \\ 400 & 4 & 16 & 196 \\ 1156 & 16 & 100 & 196 \\ 196 & 4 & 196 & 784 \end{bmatrix}$$

36+36+36 = (4+4+4) + (64+16+16) = 12 + 96 = 108

MAGIC OF THE SQUARES

$$SS_{\text{Total}} = SS_{\text{Between}} + SS_{\text{Within}}$$

$$\begin{bmatrix}1872 & 174 & 966 & 3270\end{bmatrix} = \begin{bmatrix}24 & 150 & 54 & 294\end{bmatrix} + \begin{bmatrix}1848 & 24 & 912 & 2976\end{bmatrix}$$

The "small F's"

$$F = \frac{SS_{\text{Between}}}{SS_{\text{Within}}} \times \frac{N-K}{K-1}$$

$$\frac{K-1}{N-K} F = \frac{SS_{\text{Between}}}{SS_{\text{Within}}}$$

$$4 \times \begin{bmatrix} 0.0130 & 6.2500 & 0.0592 & 0.0988 \end{bmatrix}$$

Idea: Combine the Predictors

- Create a new variable **f** with largest *F*

- In fact largest ratio SS_B / SS_W

- **f** = **Xq.** So **f** is a linear combination of columns of **X**

EIGENMAGIC

$$\left(\mathbf{X_W}^\top \mathbf{X_W}\right)^{-1} \left(\mathbf{X_B}^\top \mathbf{X_B}\right) =$$

$$\begin{bmatrix} 1284.00 & -3210.00 & 1926.00 & -4494.00 \\ -12167.56 & 30418.90 & -18251.34 & 42586.46 \\ 524.56 & -1311.40 & 786.84 & -1835.96 \\ -17.24 & 43.10 & -25.86 & 60.34 \end{bmatrix}$$

We need some (Eigen) magic

$$\mathbf{q} = \begin{bmatrix} 0.1048 \\ -0.9936 \\ 0.0428 \\ -0.0014 \end{bmatrix} \quad \text{and } \lambda \approx 32550$$

Discriminant scores for the means

$$\mathbf{F_G} = \mathbf{G} \times \mathbf{q}$$

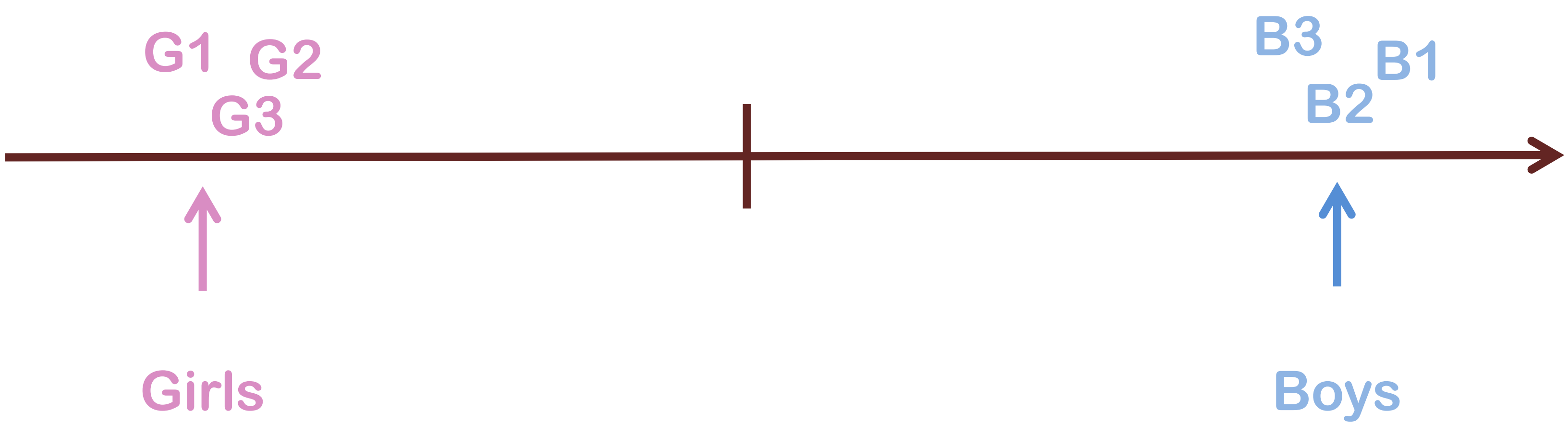
$$= \begin{bmatrix} -2 & 5 & -3 & 7 \\ 2 & -5 & 3 & -7 \end{bmatrix} \begin{bmatrix} .105 \\ -.994 \\ .043 \\ -.001 \end{bmatrix}$$

$$= \begin{bmatrix} -5.32 \\ 5.32 \end{bmatrix}$$

The Observations

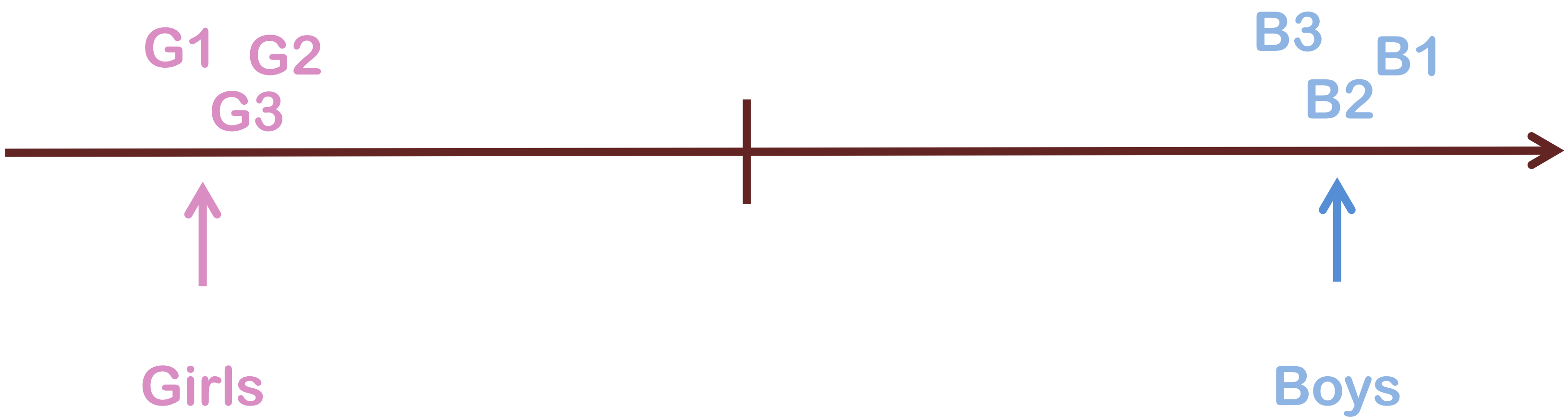
$$\mathbf{F}\mathbf{x} = \mathbf{X} \times \mathbf{q} = \begin{bmatrix} -5.29 \\ -5.31 \\ -5.35 \\ 5.36 \\ 5.32 \\ 5.27 \end{bmatrix}$$

Plot Means & Observations



A Picture

Plot Means & Observations



How to Classify? Nearest Mean

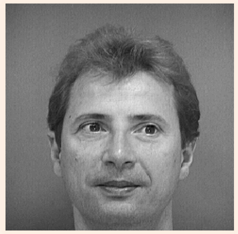
	Girls	Boys
Classified as Girls	3	0
Classified as Boys	0	3

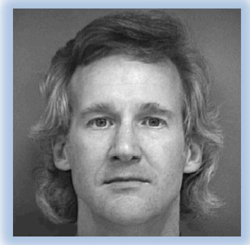
6 out of 6: Perfect …. But

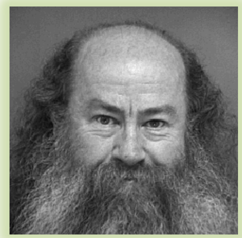
Not Surpring

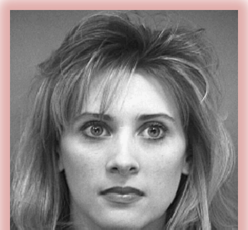
Four Variables & Six Observations

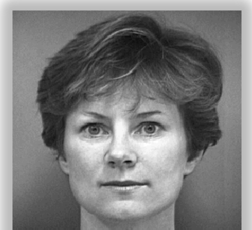
Learning & Testing Sets











PREDICTORS: 4 VARIABLES (LENGTH IN MM)

Gender	1	2	3	4
B1	180	126	318	366
B2	162	162	342	276
B3	150	150	330	384
G1	120	120	330	360
G2	168	96	300	354
G3	168	96	300	354

Old Face Means	174	151	353	473

Important: Center with the old means

$$\mathbf{X}_{\text{sup}} = \begin{bmatrix} 6 & -25 & -35 & -107 \\ -12 & 11 & -11 & -197 \\ -24 & -1 & -23 & -89 \\ -54 & -31 & -23 & -113 \\ -6 & -55 & -53 & -119 \\ -6 & -55 & -53 & -119 \end{bmatrix}$$

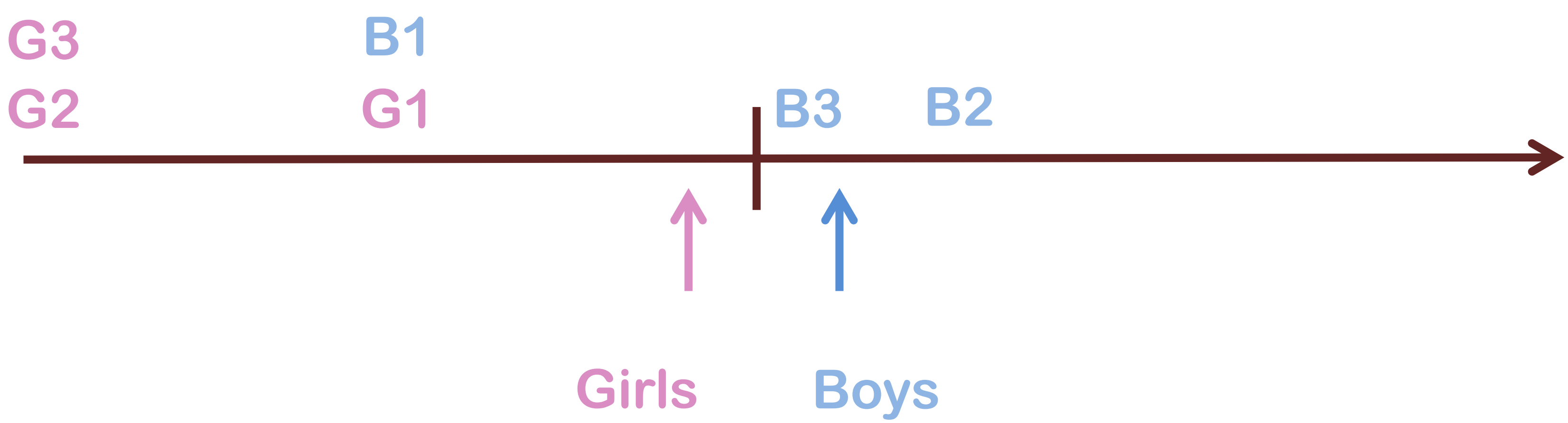
F_{SUP}

$$\mathbf{F}_{\text{sup}} = \mathbf{X}_{\text{sup}}\mathbf{q} = \begin{bmatrix} -24.1196 \\ 12.3812 \\ 2.3827 \\ -24.3126 \\ -51.9143 \\ -51.9143 \end{bmatrix}$$

A Picture

Plot Means & Observations

B1 B2 B3 G1 G2 G3



How to Classify? Nearest Mean

	Girls	Boys
Classified as Girls	3	1
Classified as Boys	0	2

5 out of 6. Not Perfect! …. (But better than I thought, though)

When $K > 2$

No Problem:

Get $K - 1$ discriminant functions

Get 2D (or More) Discriminant Maps

Mean Confidence Intervals?

No Problem: Use Bootstrap

Null Hypothesis Testing

No Problem: Use Permutation Tests

Alternatively: Use Bootstrap

Time to Wrap it Up