Discriminant Analysis
Edps/Soc 584 and Psych 594
Applied Multivariate Statistics

Carolyn J. Anderson

Department of Educational Psychology

ILLINOIS
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

Outline

Discriminant Analysis & Classification

- Introduction
- Two populations
  - Assumptions
  - Criterion and solution
  - Uses
  - Example
- Logistic Regression
- Prediction & Classification
- Two or More populations

Introduction/Overview

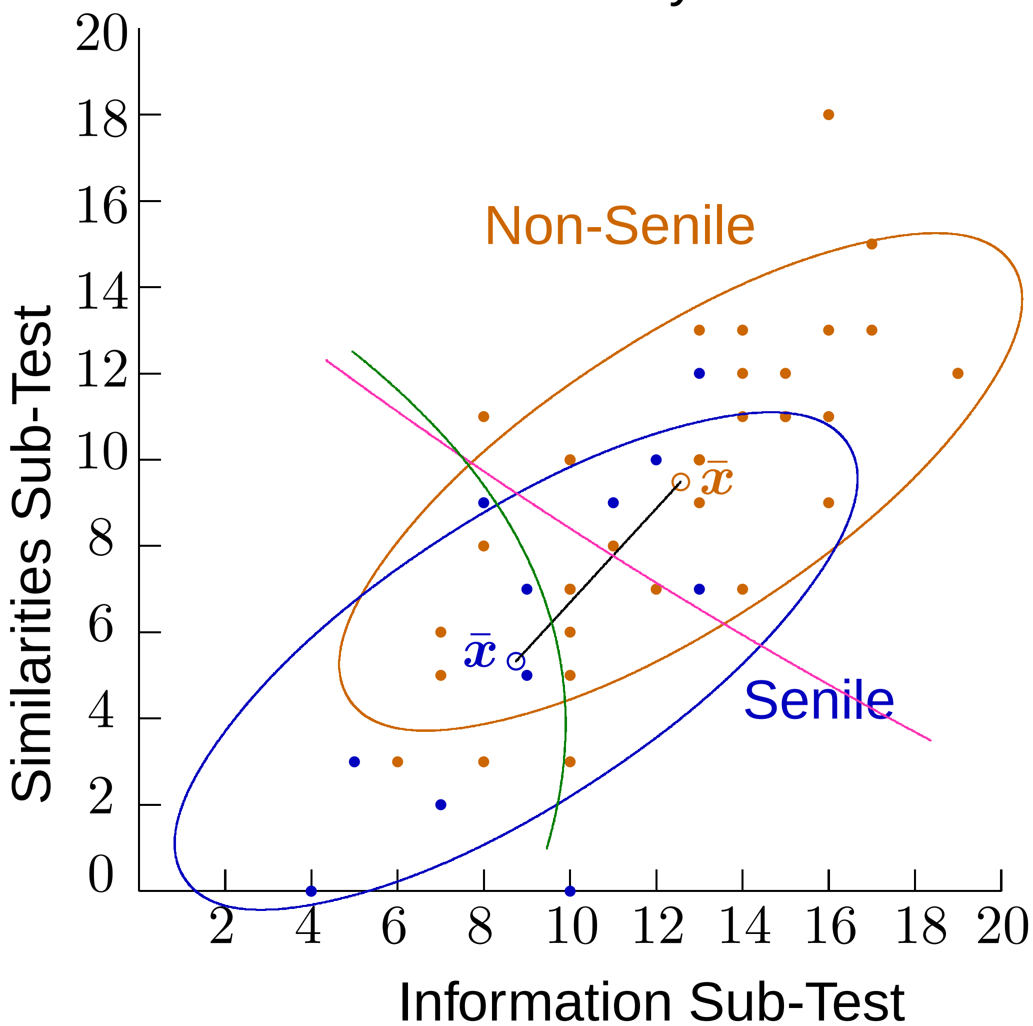
Two Goals, which often overlap (i.e., distinctions between them can become blurred):

1. Discrimination or Separation
   - Describe (graphically and/or algebraically) differences, distinguishing features of observations from several known populations/groups.
   - “Discriminants” have numerical values as different as possible for different groups.
   - Tends to be exploratory usage, 1 time (e.g., after a significant MANOVA).
2. Classification or Allocation
   - Derive a rule that can be used to optimally assign new objects to labeled classes or groups.
   - Prediction

WAIS: Two Groups with $p = 2$ Variables

Questions:

- How do we “divide” the two groups?
- How do we allocate or classify new observations?



Two Populations

R.A. Fisher pioneered discriminant analysis for discrimination purposes, and this is the method & purpose that we'll focus on.

Two Populations: π_1 and π_2, p variables, and $\boldsymbol{X}' = (X_1, X_2, \ldots, X_p)$.

Basic Idea: Transform $\boldsymbol{X}$ into a single variable Y in such a way that Y allows us to separate π_1 and π_2 as much as possible.

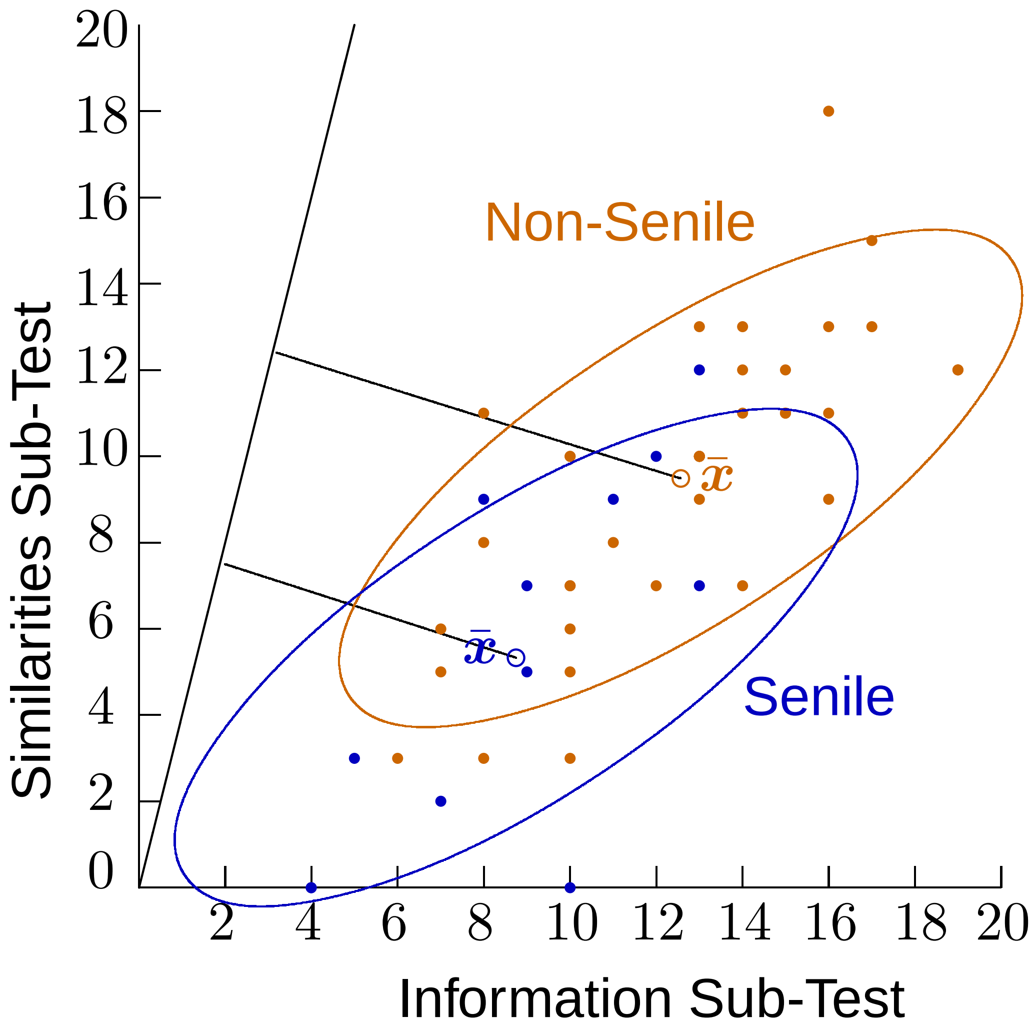
Y is a linear combination of the p X's,

$$Y = \boldsymbol{a}'\boldsymbol{X} = a_1X_1 + a_2X_2 + \cdots + a_pX_p$$

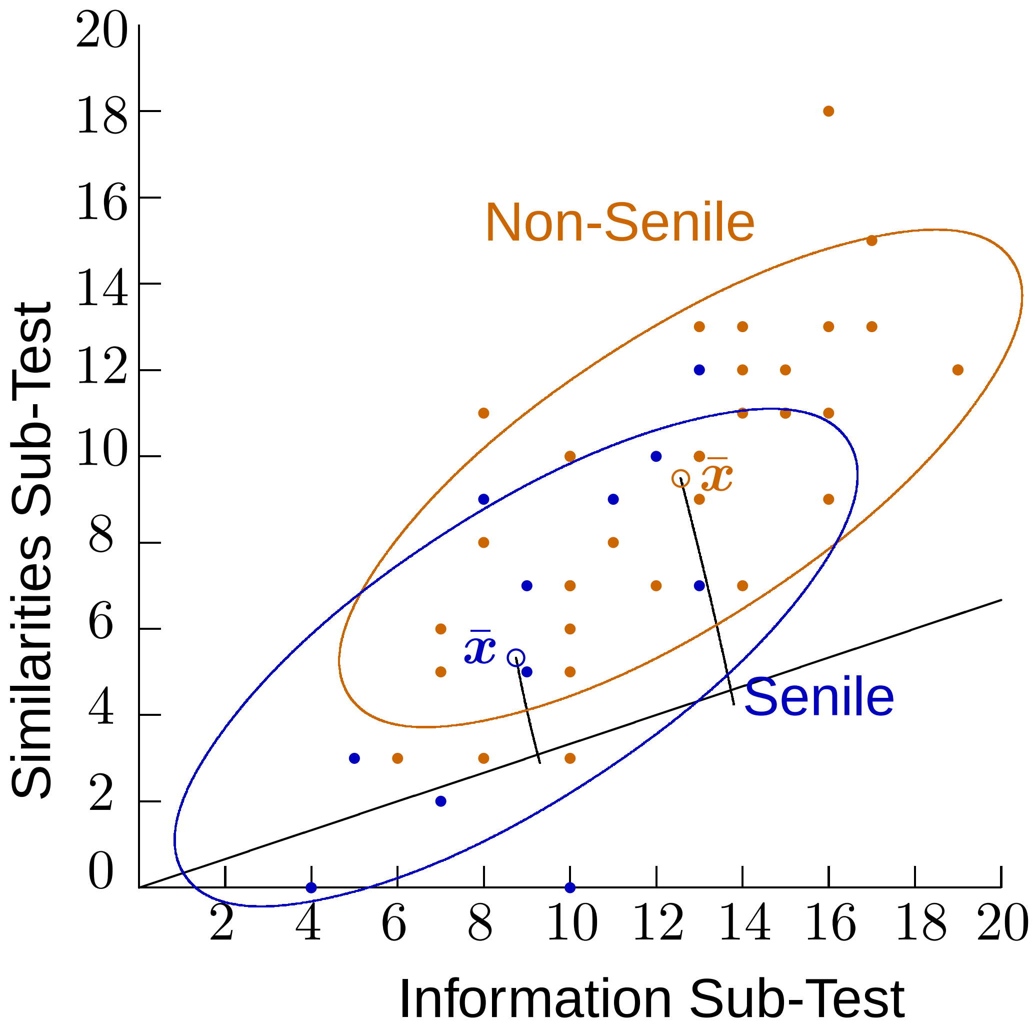
We want

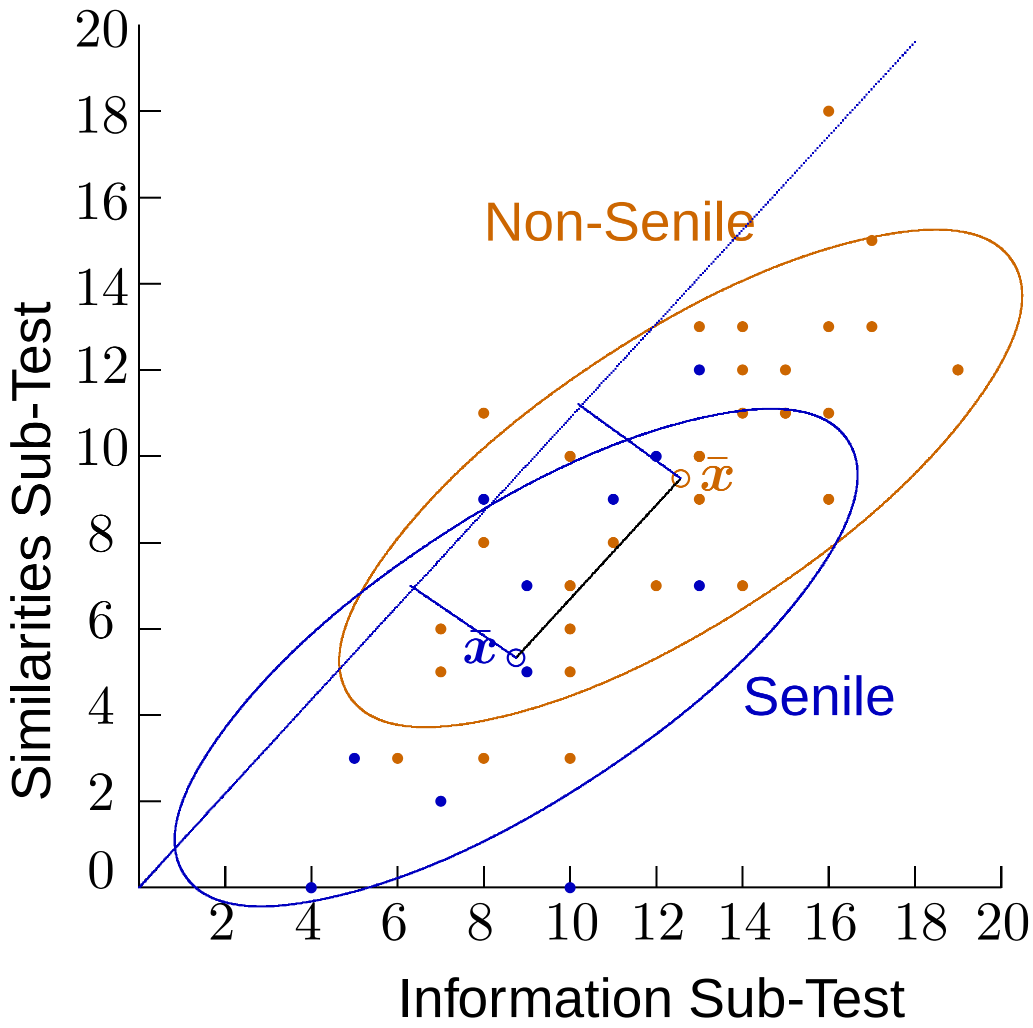
- Y to go through the origin.
- All (X_1, X_2) points are projected ("mapped") onto Y axis.
- Y that allows for the "maximum differentiation or separation between the two groups.

Graphical Illustration: Good or Bad



Graphical Illustration: Any better?




Non-Senile
Senile
$\bar{x}$
$\bar{x}$
Similarities Sub-Test
Information Sub-Test
0
2
4
6
8
10
12
14
16
18
20

Assumptions

1. Hypothesis: Two populations exist and

$$\mu_{1y} = \text{mean } Y \text{ in } 1^{st} \text{ population} = E(Y|\pi_1)$$
$$\mu_{2y} = \text{mean } Y \text{ in } 2^{nd} \text{ population} = E(Y|\pi_2)$$

   where Y is a linear combination of the X's,

$$Y = \boldsymbol{a}_1 \boldsymbol{X} \quad \text{where } \boldsymbol{a} \text{ is a } (p \times 1) \text{ vector}$$

2. $\boldsymbol{\mu}_1 = E(\boldsymbol{X}|\pi_1) =$ mean of $\boldsymbol{X}$ in 1^{st} population.
   $\boldsymbol{\mu}_2 = E(\boldsymbol{X}|\pi_2) =$ mean of $\boldsymbol{X}$ in 2^{nd} population.

   Implication: Random vector $\boldsymbol{X}$ had either mean μ_1 or μ_2, depending on which population it comes from.

3. Equal covariance matrices: $\text{cov}(\boldsymbol{X}|\pi_1) = \text{cov}(\boldsymbol{X}|\pi_2) = \boldsymbol{\Sigma}$

   Equal covariance matrices is an important assumption (& one easily violated).

Criterion to be Maximized

Since $Y = \boldsymbol{a}'\boldsymbol{X}$,

$$\begin{aligned} \mu_{1y} &= E(\boldsymbol{a}'\boldsymbol{X}|\pi_1) = \boldsymbol{a}'\boldsymbol{\mu}_1 \\ \mu_{2y} &= E(\boldsymbol{a}'\boldsymbol{X}|\pi_2) = \boldsymbol{a}'\boldsymbol{\mu}_2 \\ \text{var}(Y) &= \sigma_y^2 = \boldsymbol{a}'\boldsymbol{\Sigma}\boldsymbol{a} \end{aligned}$$

Criterion: The "best" linear combination is the maximal discriminator and is based on the ratio

$$\begin{aligned} & \frac{(\text{Squared distance between } \mu_{1y} \text{ and } \mu_{2y})}{(\text{Variance of } Y)} \\ &= (\mu_{1y} - \mu_{2y})^2/\sigma_y^2 \\ &= \frac{(\boldsymbol{a}'\boldsymbol{\mu}_1 - \boldsymbol{a}'\boldsymbol{\mu}_2)^2}{\boldsymbol{a}'\boldsymbol{\Sigma}\boldsymbol{a}} \\ &= \frac{(\boldsymbol{a}'\boldsymbol{\delta})^2}{\boldsymbol{a}'\boldsymbol{\Sigma}\boldsymbol{a}'} \qquad \text{where } \boldsymbol{\delta} = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \\ &= \frac{\boldsymbol{a}'(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{a}}{\boldsymbol{a}'\boldsymbol{\Sigma}\boldsymbol{a}'} \end{aligned}$$

Fisher's Problem

Fisher sought the $\boldsymbol{a}$ that maximizes the ratio

$$\max_{\boldsymbol{a}} \left(\frac{(\boldsymbol{a}'\boldsymbol{\delta})^2}{\boldsymbol{a}'\boldsymbol{\Sigma}\boldsymbol{a}} \right)$$

This $\boldsymbol{a}$ maximizes the standardized square distance.

The ratio

$$\frac{\boldsymbol{a}'(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{a}}{\boldsymbol{a}'\boldsymbol{\Sigma}\boldsymbol{a}}$$

is important in multivariate analysis.
This ratio is related to the F-statistic for testing

$$H_o : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \cdots = \boldsymbol{\mu}_g$$

Specifically,

$$\frac{\boldsymbol{a}'(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{a}}{\boldsymbol{a}'\boldsymbol{\Sigma}\boldsymbol{a}} = (F\text{statistic}) \left(\frac{n-2}{n-1} \right)$$

For now, we'll just consider $g = 2$.

Solution to Maximization Problem

The ratio $(\boldsymbol{a}'\boldsymbol{\delta})^2/\boldsymbol{a}'\boldsymbol{\Sigma}\boldsymbol{a}$ is maximized by the choice of

$$\boldsymbol{a} = c\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta} = c\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$$

where c is an arbitrary constant such that $c \neq 0$

For convenience and simplicity, we usually choose $c = 1$, so

$$\boldsymbol{a} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}$$

and the linear combination, our linear discriminant function, is

$$\begin{aligned} Y = \boldsymbol{a}'\boldsymbol{X} &= (\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta})'\boldsymbol{X} \\ &= \boldsymbol{\delta}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X} \\ &= (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\boldsymbol{\Sigma}^{-1}\boldsymbol{X} \end{aligned}$$

With this choice for $\boldsymbol{a}$, the maximum of the ratio equals

$$\frac{(\boldsymbol{a}'\boldsymbol{\delta})^2}{\boldsymbol{a}'\boldsymbol{\Sigma}\boldsymbol{a}} = \frac{(\boldsymbol{\delta}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta})^2}{(\boldsymbol{\delta}'\boldsymbol{\Sigma}^{-1})\boldsymbol{\Sigma}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta})} = \boldsymbol{\delta}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta} = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$$

A squared statistical distance.

Using Y for Classification

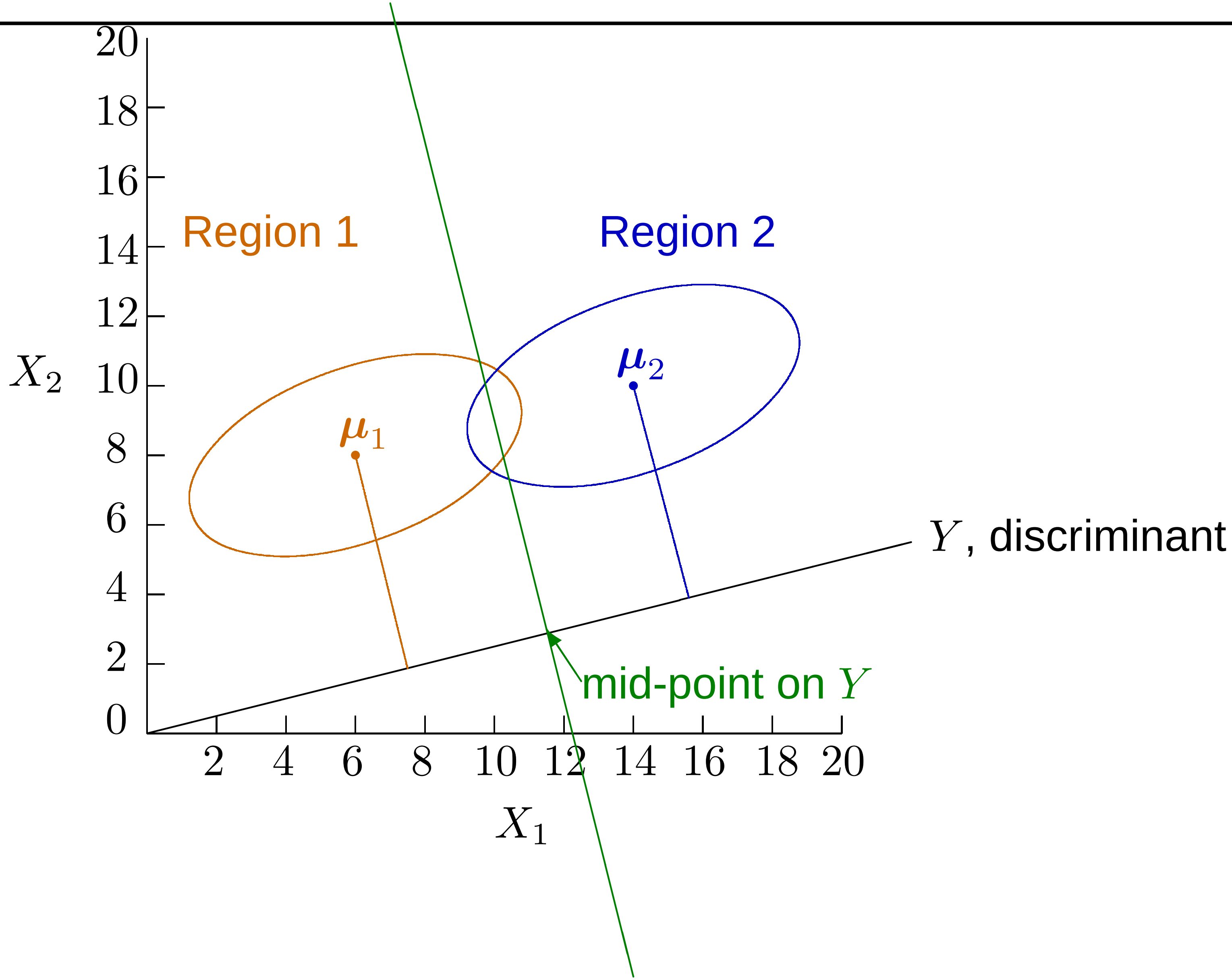
(for discrimination, a bit later)... A very simplistic method for using Y for classification or allocation of a new observation $\boldsymbol{x}_o$ to one or the other group.

1. Calculate
$$Y_o = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}\boldsymbol{x}_o$$
2. Find the mid-point between $\boldsymbol{\mu}_{1y}$ and $\boldsymbol{\mu}_{2y}$, which equals
$$\begin{aligned}\text{mid-point} = m &= \frac{1}{2}(\boldsymbol{\mu}_{1y} + \boldsymbol{\mu}_{2y}) \\ &= \frac{1}{2}(\boldsymbol{a}'\boldsymbol{\mu}_1 + \boldsymbol{a}'\boldsymbol{\mu}_2) \\ &= \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)\end{aligned}$$
3. Allocation Rule:

   Allocate y_o to π_1 if $y_o = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}\boldsymbol{x}_o \geq m$
   Allocate y_o to π_2 if $y_o = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}\boldsymbol{x}_o < m$

   Or whether $y_o - m = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\boldsymbol{\Sigma}^{-1}\boldsymbol{x}_o - m$ is $\geq$ or < 0.

Graphically What's going on



Estimation

$\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ and $\boldsymbol{\Sigma}$ are generally unknown and must be estimated from data.

Data and Estimates

Population 1: $\quad \boldsymbol{X}_{1,(n_1\times p)}$

$$\hat{\boldsymbol{\mu}}_1 = \bar{\boldsymbol{X}}_1 = \frac{1}{n_1}\sum_{j=1}^{n_1} X_{1j}$$

$$\hat{\boldsymbol{\Sigma}} = \boldsymbol{S}_1 = \frac{1}{n_1-1}\sum_{j=1}^{n_1}(X_{1j}-\bar{\boldsymbol{X}}_1)(X_{1j}-\bar{\boldsymbol{X}}_1)'$$

Population 2: $\quad \boldsymbol{X}_{2,(n_2\times p)}$

$$\hat{\boldsymbol{\mu}}_2 = \bar{\boldsymbol{X}}_2 = \frac{1}{n_2}\sum_{j=1}^{n_2} X_{2j}$$

$$\hat{\boldsymbol{\Sigma}} = \boldsymbol{S}_2 = \frac{1}{n_2-1}\sum_{j=1}^{n_2}(X_{2j}-\bar{\boldsymbol{X}}_21)(X_{2j}-\bar{\boldsymbol{X}}_2)'$$

$$\hat{\boldsymbol{\Sigma}} = \boldsymbol{S}_{pool} = \frac{(n_1-1)\boldsymbol{S}_1+(n_2-1)\boldsymbol{S}_2}{n_1+n_2-2}$$

Estimate of Fisher's linear discriminant function is

$$\hat{Y} = \hat{\boldsymbol{a}}'\boldsymbol{X} = (\bar{\boldsymbol{X}}_1-\bar{\boldsymbol{X}}_2)'\boldsymbol{S}_{pool}^{-1}\boldsymbol{X}$$

Estimation continued

The mid-point between the two sample means $\hat{Y}_1$ and $\hat{Y}_2$ is

$$\begin{aligned} \hat{m} &= \frac{1}{2}\left(\hat{\bar{Y}}_1 + \hat{\bar{Y}}_2\right) \\ &= \frac{1}{2}(\bar{\boldsymbol{X}}_1 \underbrace{-}_{minus} \bar{\boldsymbol{X}}_2)'\boldsymbol{S}_{pool}^{-1}(\bar{\boldsymbol{X}}_1 \underbrace{+}_{plus} \bar{\boldsymbol{X}}_2) \end{aligned}$$

Simplistic Allocation rule is

Allocate $\boldsymbol{x}_o$ to π_1 if $y_o = (\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2)'\boldsymbol{S}_{pool}^{-1}\boldsymbol{x}_o \geq m$; otherwise, allocate $\boldsymbol{x}_o$ to π_2.

Using Y for Discrimination

Study the elements of

$$\hat{\boldsymbol{a}}' = (\bar{x}_1 - \bar{x}_2)'\boldsymbol{S}_{pool}^{-1}$$

to see which variable(s) are the best "discriminators".
Note that

$$\hat{y} = \hat{a}_1 x_1 + \hat{a}_2 x_2 + \cdots + \hat{a}_p x_p$$

If variables are on vastly different scales, they could first be standardized and then compute the discriminant function; that is, let

$$\begin{aligned} \hat{y} &= \hat{a}_1^* z_1 + \hat{a}_2^* z_2 + \cdots + \hat{a}_p^* z_p \\ &= \hat{a}_1^* \frac{(x_1 - \bar{x}_1)}{\sqrt{s_{11}}} + \hat{a}_2^* \frac{(x_2 - \bar{x}_2)}{\sqrt{s_{22}}} + \cdots + \hat{a}_p^* \frac{(x_p - \bar{x}_p)}{\sqrt{s_{pp}}}, \end{aligned}$$

which implies that $\hat{a}_i^* = a_i/\sqrt{s_{ii}}$, or in matrix terms,

$$\boldsymbol{a}^* = \boldsymbol{D}^{-1/2}\boldsymbol{a}$$

where $\boldsymbol{D}^{-1/2} = \text{diag}(1/\sqrt{s_{ii}})$ and s_{ii} are elements of $\boldsymbol{S}_{pool}$.

Maximum of Ratio is...

Fisher's $\hat{a}$ maximizes the ratio

$$\frac{(\bar{Y}_1 - \bar{Y}_2)^2}{s_y^2} = \frac{\left(\boldsymbol{a}'\bar{X}_1 - \boldsymbol{a}'\bar{X}_2\right)^2}{\boldsymbol{a}'\boldsymbol{S}_{pool}\boldsymbol{a}}$$

Define $\boldsymbol{d} = (\bar{x}_1 - \bar{x}_2)$, and the maximum of the ratio is obtained when

$$\hat{\boldsymbol{a}}' = (\bar{x}_1 - \bar{x}_2)'\boldsymbol{S}_{pool}^{-1}$$

So we have

$$\begin{aligned}
\max_{\boldsymbol{a}} \left[\frac{(\boldsymbol{a}'\boldsymbol{d})^2}{\boldsymbol{a}'\boldsymbol{S}_{pool}\boldsymbol{a}}\right] &= \frac{(\boldsymbol{d}'\boldsymbol{S}_{pool}^{-1}\boldsymbol{d})^2}{(\boldsymbol{d}\boldsymbol{S}_{pool}^{-1})\boldsymbol{S}_{pool}(\boldsymbol{S}_{pool}^{-1}\boldsymbol{d})} \\
&= \boldsymbol{d}'\boldsymbol{S}_{pool}^{-1}\boldsymbol{d} \\
&= (\bar{x}_1 - \bar{x}_2)'\boldsymbol{S}_{pool}^{-1}(\bar{x}_1 - \bar{x}_2) \\
&= D^2 = \text{“Mahalanobis } D^2\text{”}
\end{aligned}$$

Mahalanobis D^2

- D^2 is a generalized sample squared distance between means.
- D^2 is the maximum relative separation obtained by considering linear combinations.
- In some situations D^2 is can be used to test $H_o : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$.

In particular, if we assume

$$\boldsymbol{X}_{1j} \sim \mathcal{N}_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}) \text{ and } \boldsymbol{X}_{2j} \sim \mathcal{N}_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$$

for $j = 1, \ldots, n_l$. Then the test statistic for $H_o : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ versus $H_a : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$ is

$$\begin{aligned} \frac{(n_1+n_2-p-1)}{(n_1+n_2-2)p} T^2 &= \frac{(n_1+n_2-p-1)}{(n_1+n_2-2)p}\left(\frac{n_1 n_2}{n_1+n_2}\right)(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)' \boldsymbol{S}_{pool}^{-1}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2) \\ &= \frac{(n_1+n_2-p-1)}{(n_1+n_2-2)p}\left(\frac{n_1 n_2}{n_1+n_2}\right) D^2 \end{aligned}$$

WAIS Example: Description

This example is from Morrison (2005).

Fourty nine elderly men participating in an interdisciplinary study of human aging were classified into the diagnostic categories “senile factor present” and “no senile factor” on the basis of an intensive psychiatric examination. The Wechsler Adult Intelligence Scale (WAIS) was administered to all subjects by an independent investigator.

Below are mean scores by group on some of the WAIS subtests.

	Not Senile $(n = 37)$		Senile $(n = 12)$	
Sub-Test	$\bar{x}$	std dev	$\bar{x}$	std dev
Information	12.566	3.387	8.750	3.251
Similarities	9.486	3.380	5.333	4.271
Arithmetic	11.514	3.363	8.500	3.631
Picture	7.973	1.922	4.750	3.571

Note: My results differ slightly from Morrison’s. There is either a typo in the text or in the data set. (no way to find out which).

WAIS continued

$$\boldsymbol{S}_{sen} = \begin{pmatrix} 11.47 & 8.55 & 6.39 & 2.07 \\ 8.55 & 11.42 & 5.49 & 0.29 \\ 6.39 & 5.49 & 11.31 & 1.82 \\ 2.07 & 0.29 & 1.82 & 3.69 \end{pmatrix} \boldsymbol{S}_{non} = \begin{pmatrix} 10.57 & 10.45 & 9.68 & 7.66 \\ 10.45 & 18.24 & 12.09 & 8.91 \\ 9.68 & 12.09 & 13.18 & 5.32 \\ 7.66 & 8.91 & 5.32 & 12.75 \end{pmatrix}$$

Test $H_o : \boldsymbol{\Sigma}_{non} = \boldsymbol{\Sigma}_{sen}$: $X^2 = 17.99$, $df = 10$, $p = .06$.

$$\boldsymbol{S}_{pool} = \begin{pmatrix} 11.26 & 9.00 & 7.163.38 \\ 9.00 & 13.02 & 7.042.31 \\ 7.16 & 7.01 & 11.752.64 \\ 3.38 & 2.38 & 2.645.81 \end{pmatrix}$$

Test $H_o : \boldsymbol{\mu}_{non} = \boldsymbol{\mu}_{sen}$: $T^2 = 22.41$, $F = 88.93$, $p < .01$ (note: $\mathcal{F}_{4,44}(.95) = 11.04$).

WAIS continued

Since there are differences between the groups, we need to investigate how they differ. . .

Linear Discriminant Analysis:

$$(\bar{\boldsymbol{x}}_{non} - \bar{\boldsymbol{x}}_{sen})' = (\quad 3.82, \quad 4.15, \quad 3.01, \quad 3.23)$$

$$\boldsymbol{a}' = (\bar{\boldsymbol{x}}_{non} - \bar{\boldsymbol{x}}_{sen})'\boldsymbol{S}^{-1}_{pool} = (\underbrace{0.02}_{information}, \underbrace{0.22}_{similarities}, \underbrace{0.01}_{artihmetic}, \underbrace{0.45}_{picture})$$

For interpretation, things to look at

- Do all items have about the same weight?
- Do some have relatively larger weights (i.e., does one or a sub-set dominate)?
- Is there a substantive grouping of variables?

For Description of Differences.

Logistic regression is appropriate for a dichotomous response variable and nominal, ordinal and/or numerical predictor or explanatory variables.

However,
IF $\boldsymbol{X}_{ij} \sim \mathcal{N}_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$ i.i.d. and $i = 1, 2$

THEN a logistic regression for dichotomous variables necessarily fits the data where the group is the "response" variable and the X's are explanatory variables.

Suppose

$$Y = \begin{cases} 1 & \text{for population } 1 \text{ (non-senile)} \\ 0 & \text{for population } 2 \text{ (senile)} \end{cases}$$

For the WAIS data, the logistic regression model is. . .

Logistic Regression for WAIS

$$P(Y=1) = \frac{\exp[\alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4]}{(1 + \exp[\alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4])}$$

$$P(Y=0) = \frac{1}{(1 + \exp[\alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4])}$$

where $X_1 =$ information, $X_2 =$ similarities, $X_3 =$arithmetic, and $X_4 =$ picture.

If the regression parameter (i.e., β_i) for an X variable is significant, then the senile and non-senile groups differ with respect to that variable.

Note on interpretation: The odds of a person being in the non-senile group versus the senile group when $X = x$ are $\exp(\beta)$ times the odds when $X = (x - 1)$.

In terms of probabilities:

$$\begin{aligned} P(Y=1) &= \frac{\exp[\alpha+\beta_1X_1+\beta_2X_2+\beta_3X_3+\beta_4X_4]}{(1+\exp[\alpha+\beta_1X_1+\beta_2X_2+\beta_3X_3+\beta_4X_4])} \\ P(Y=0) &= \frac{1}{(1+\exp[\alpha+\beta_1X_1+\beta_2X_2+\beta_3X_3+\beta_4X_4])} \end{aligned}$$

The odds of a subject being not senile given the explanatory variables

$$\frac{P(Y=1)}{P(Y=)} = \exp[\alpha+\beta_1X_1+\beta_2X_2+\beta_3X_3+\beta_4X_4]$$

Interpretation: The odds of not senile (versus senile) when x_i is 1 unit larger is $\exp(\beta_i)$ time the odds of not senile.

$\exp(\beta)$ is an odds ratio.

Parameter Estimates

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-3.6176	1.7559	4.2447	0.0394
information	1	0.0953	0.2079	0.2099	0.6468
simialrities	1	0.1808	0.1768	1.0447	0.3067
arithmetic	1	-0.0102	0.1574	0.0042	0.9482
picture	1	0.3682	0.1803	4.1695	0.0412

Estimated Odds Ratios

Modeling odds of not senile versus senile:

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
information	1.100	0.732	1.653
simiarlities	1.198	0.847	1.694
arithmetic	0.990	0.727	1.348
picture	1.445	1.015	2.058

- An odds ratio of 1 means no association.
- An odds ratio > 1 means non-senile is more likely.
- An odds ratio < 1 means senile is more likely.

WAIS: Prediction & Classification

We can compute the means on the linear discriminant functions for each individual, as well as means for each group:

$$\bar{y}_{non} \quad = \quad \boldsymbol{a}'\bar{\boldsymbol{x}}_{non} = (0.02, 0.22, 0.01, 0.45)\begin{pmatrix} 12.57 \\ 9.49 \\ 11.51 \\ 7.97 \end{pmatrix} = 6.07$$

$$\bar{y}_{sen} \quad = \quad \boldsymbol{a}'\bar{\boldsymbol{x}}_{sen} = (0.02, 0.22, 0.01, 0.45)\begin{pmatrix} 8.75 \\ 5.33 \\ 8.50 \\ 4.75 \end{pmatrix} = 3.59$$

Midpoint = $(1/2)(6.07 + 3.59) = 4.83$.

Classification Rule 1: Random Assignment

For demonstration, we'll allocate the 49 subjects to one of the two groups using random assignment.

Specifically, we'll use the "base rate" $= 12/49 = .24$, which equals the proportion of the sample who are in the senile group. So we'll assign 24% of the subject to the senile group and 76% to the not senile group. This gives the following:

	Actual Diagnosis		
Classification	Not Senile	Senile	
not senile	28	9	37
senile	9	3	12
	37	12	49

Under random assignment, we would expect to make $9 + 9 = 18$ errors.

Note: For random assignment, we make sure that the margins are identical and then compute $n_{i+}n_{+j}/n_{++}$.

Classification Rule 2: Midpoint

For each individual, we compute

$$y_j = 0.02x_{1j} + 0.22x_{2j} + 0.01x_{3j} + 0.45x_{4j}$$

Classification Rule 2: Assign the j^{th} individual to senile group if $y_i \leq 4.83 =$ mid-point between $\bar{y}_{senile}$ and $\bar{y}_{not}$.

	Actual Diagnosis		
Classification	Not Senile	Senile	
not senile	30	4	34
senile	7	8	15
	37	12	49

Using a simple rule with discriminant analysis we make $7 + 4 = 11$ errors. . . better than random assignment (i.e., 18 errors). . .

Other Classification Rules

Knowing the column margin proportions (24% senile and 76%), this information can be built into a prediction rule for classification.

This is artificial here because the margin doesn't reflect the population proportion; however, if you have or know this information about the population, it can be used.

General Rule: We compute the probability of a new observation, say x_o belonging to each of the groups, and then place or allocate that person to the group with highest probability.

The Probabilities can be computed in a number of different ways, but all of them depend on a distance measure.

Without going into the theory, we'll look at how a couple of them perform on our data.

Classification Rule 3: Place $\boldsymbol{x}_o$ into the group that leads to the largest probability where probabilities are computed using

$$\text{Prob}(\boldsymbol{x}_o \text{ belongs to } k) = \frac{\exp\left(\frac{1}{2}D^2_{ko}\right)}{\sum_h \exp\left(\frac{1}{2}D^2_{ho}\right)}$$

where D^2_{ko} equals

$$D^2_{ko} = (\boldsymbol{x}_o - \bar{\boldsymbol{x}}_k)'\boldsymbol{S}_k^{-1}(\boldsymbol{x}_o - \bar{x}_k) + \ln(|\boldsymbol{S}_k|) - 2\ln(\text{proportion}_k)$$

	Actual Diagnosis		
Classification	Not Senile	Senile	
not senile	34	5	39
senile	3	7	19
	37	12	49

which leads to $3 + 5 = 8$ errors.

Classification Rule 4: Non-parametric

For this I used a "kernel density".

Classification Rule 4: Place $\boldsymbol{x}_o$ into the group that leads to the largest probability where probabilities are computed using

$$\text{Prob}(\boldsymbol{x}_o \text{ belongs to } k) = \frac{m_k(\boldsymbol{x}_o)(\text{proportion})_k}{\sum_h m_h(\boldsymbol{x}_o)(\text{proportion})_h}$$

where $m_k(\boldsymbol{x}_0)$ is a function of $\boldsymbol{x}_o$ that equals the proportion of those in group k within a radius of r from $\boldsymbol{x}_o$.

When I set the radius r to

- $2 \longrightarrow 7$ errors.
- $1 \longrightarrow 1$ errors.
- $0.5 \longrightarrow 0$ errors.

Moral: If you're really interested in classification (prediction), you can do very well by using a more sophisticated rule. However, for describing differences between groups, Fisher's linear discriminant functions are fine.

DA for more than Two Groups

Uses:

- Examine group separation in a (1 or) 2 dimensional plot. When there are more than 2 groups, this requires (may require) more than 1 discriminant function.

  We can project points from p dimensional space (from each group) onto a 2-dimensional sub-space such that we obtain the best possible view of how the groups are separated.
- Find a sub-set of the original variables that separates the groups almost as well as the original set.
- Rank variables (study them) with respect to their relative contribution to group separation. For this, we use standardized discriminant function coefficients for more valid comparisons.
- Interpret the new dimensions represented by the discriminant functions.
- As a follow-up to fixed effects MANOVA, especially when there is a significant interaction.

Discriminant Analysis with $g > 2$

Suppose that we have $g =$ number of groups (> 2).

We'd like to choose just a few discriminant functions $\boldsymbol{a}_1\boldsymbol{X}$, $\boldsymbol{a}_2\boldsymbol{X}$ and $\boldsymbol{a}_3\boldsymbol{X}$ to separate the g populations.

We want the number of these functions to be much smaller than p and g so that we greatly reduce the dimensionality of the problem.

When we had $g = 2$, at most we could have 1 discriminant.

We'll extend Fisher's methodology for $g > 2$ (actually, this is Fisher's extension—basically), because

- It provides a convenient representation of the g populations that allows us to reduce the dimensionality of the original variable space from p to just a few linear discriminant functions, which are linear combinations of the $\boldsymbol{X}$'s.
- Plots of the group's mean values on the discriminants can be much easier to interpret, especially if we only need a few.
- Plots of values of each case on the first two discriminant functions help to identify outliers or weird cases in the data.

Fisher's Extension

Fisher develop this method for discrimination (description of differences between populations), but it can also be used to classify observations (prediction, allocation).

Assumptions:

- We have g populations (not necessarily normal).
- $\boldsymbol{\mu}_{i,(p\times 1)} =$ mean of population i (i.e., π_i).
- $\boldsymbol{\Sigma} = (p \times p)$ covariance matrix of the i^{th} populations
  - They are equal: $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \cdots = \boldsymbol{\Sigma}_g$
  - Full rank: If $\boldsymbol{\Sigma}$ is not full rank, then reduce the problem and re-define variables to get full rank (e.g.., use the first eigenvalues $\neq 0$, principal components).
- $n_i =$ the number of observations from population i.
- $\hat{\boldsymbol{\mu}} = (1/g)\sum_{i=1}^{g} \boldsymbol{\mu}_i$, overall population mean vector.
- $\boldsymbol{B}_o = \sum_{i=1}^{g} n_i(\boldsymbol{\mu}_i - \bar{\boldsymbol{\mu}})(\boldsymbol{\mu}_i - \bar{\boldsymbol{\mu}})'$ the Between Groups SSCP.

This definition of $\boldsymbol{B}_o$ is a bit different from what's in the text, which does not weight by n_i. We'll use n_i here because it's more consistent with what others do (e.g., SAS & MANOVA).

Linear Combination

Consider the linear combination

$$Y = \boldsymbol{a}'\boldsymbol{X}$$

with

$$\mu_{iy} = E(Y|\pi_i) = E(\boldsymbol{a}'\boldsymbol{X}|\pi_i) = \boldsymbol{a}'E(\boldsymbol{X}|\pi_i) = \boldsymbol{a}'\boldsymbol{\mu}_i$$

$$\sigma_y^2 = \text{var}(Y|\pi_i) = \text{var}(\boldsymbol{a}'\boldsymbol{X}|\pi_i) = \boldsymbol{a}'\underbrace{\boldsymbol{\Sigma}}\boldsymbol{a}$$

the equal variance assumption $\uparrow$

So we have a set of g means on Y:

$$\{\mu_{1y}, \mu_{2y}, \ldots, \mu_{gy}\}$$

and a common variance, $\sigma_y^2 = \boldsymbol{a}'\boldsymbol{\Sigma}\boldsymbol{a}$. The overall mean of the Y's is

$$\bar{\mu}_y = \frac{1}{g}\sum_{i=1}^{g}\mu_{iy} = \frac{1}{g}\sum_{i=1}^{g}\boldsymbol{a}_i'\boldsymbol{\mu}_i = \boldsymbol{a}'\left(\frac{1}{g}\sum_{i=1}^{g}\boldsymbol{a}_i'\boldsymbol{\mu}_i\right) = \boldsymbol{a}'\bar{\boldsymbol{\mu}}$$

The Criterion

Consider the ratio:

$$\frac{\left(\begin{array}{l}\text{Weighted sum of squared}\\ \text{distances from population}\\ \text{means to overall mean of } Y\end{array}\right)}{(\text{variance of } Y)} = \frac{\sum_{i=1}^{g} \overbrace{n_i}^{\text{weights}} \overbrace{(\mu_{iy} - \bar{\mu}_y)^2}^{\text{squared distance}}}{\sigma_y^2}$$

$$= \frac{\sum_{i=1}^{g} n_i(\boldsymbol{a}'\boldsymbol{\mu}_i - \boldsymbol{a}'\bar{\boldsymbol{\mu}})^2}{\boldsymbol{a}'\boldsymbol{\Sigma}\boldsymbol{a}}$$

$$= \frac{\boldsymbol{a}'\left[\sum_{i=1}^{g} n_i(\boldsymbol{\mu}_i - \bar{\boldsymbol{\mu}})(\boldsymbol{\mu}_i - \bar{\boldsymbol{\mu}})'\right]\boldsymbol{a}}{\boldsymbol{a}'\boldsymbol{\Sigma}\boldsymbol{a}}$$

$$= \frac{\boldsymbol{a}'\boldsymbol{B}_o\boldsymbol{a}}{\boldsymbol{a}'\boldsymbol{\Sigma}\boldsymbol{a}}$$

- This is the multiple group extension of the ratio $(\boldsymbol{a}'(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2))^2/(\boldsymbol{a}'\boldsymbol{\Sigma}\boldsymbol{a}) = (\boldsymbol{a}'\boldsymbol{\delta})^2/(\boldsymbol{a}'\boldsymbol{\Sigma}\boldsymbol{a})$.
- This compares the between group variability relative to the within group variability.

Maximizing the Criterion

The maximization problem

$$\max_{\boldsymbol{a}} = \frac{\boldsymbol{a}'\boldsymbol{B}_o\boldsymbol{a}}{\boldsymbol{a}'\boldsymbol{\Sigma}\boldsymbol{a}}$$

(It's like finding the $\boldsymbol{a}$ to maximized an F-statistic).

Since

- $\boldsymbol{B}_o = \boldsymbol{H}$, the hypothesis SSCP for testing $H_o : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \cdots = \boldsymbol{\mu}_g$,
- $\boldsymbol{\Sigma} \propto \boldsymbol{W} = \boldsymbol{E}$, the within groups SSCP

The Ratio to be maximized is proportional to

$$\frac{\boldsymbol{a}'\boldsymbol{B}_o\boldsymbol{a}}{\boldsymbol{a}'\boldsymbol{\Sigma}\boldsymbol{a}} \propto \frac{\boldsymbol{a}'\boldsymbol{H}\boldsymbol{a}}{\boldsymbol{a}'\boldsymbol{E}\boldsymbol{a}}$$

So our problem is equivalent to finding the $\boldsymbol{a}$ that maximizes this alternative ratio,

$$\max_{\boldsymbol{a}} = \frac{\boldsymbol{a}'\boldsymbol{H}\boldsymbol{a}}{\boldsymbol{a}'\boldsymbol{E}\boldsymbol{a}}$$

Result

Consider $\boldsymbol{\Sigma}^{-1}\boldsymbol{B}_o$ and let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_s > 0$ be the s ($s \leq \min(g-1, p)$) non-zero eigenvalues of $\boldsymbol{\Sigma}^{-1}\boldsymbol{B}_o$ and $\boldsymbol{e}_1$, $\boldsymbol{e}_2$, ..., $\boldsymbol{e}_s$ are the corresponding eigenvectors scaled so that $\boldsymbol{e}_i'\boldsymbol{\Sigma}\boldsymbol{e}_i = 1$.

The vector of coefficients $\boldsymbol{a}$ that maximizes $(\boldsymbol{a}'\boldsymbol{B}_o\boldsymbol{a})/(\boldsymbol{a}'\boldsymbol{\Sigma}\boldsymbol{a})$ is the 1^{st} eigenvector of $\boldsymbol{\Sigma}^{-1}\boldsymbol{B}_o$; that is,

$$\boldsymbol{a}_1 = \boldsymbol{e}_1$$

where $\boldsymbol{e}_1'\boldsymbol{\Sigma}\boldsymbol{e}_1 = 1$ (scaled).

The linear combination $\boldsymbol{a}_1\boldsymbol{X}$ is the first discriminant function.

The maximum of the ratio equals λ_1; that is,

$$\lambda_1 = \frac{\boldsymbol{a}_1\boldsymbol{B}_o\boldsymbol{a}_1}{\boldsymbol{a}_1\boldsymbol{\Sigma}\boldsymbol{a}_1}.$$

Eigenvalues of square non-symmetric matrix?

Eigenvalues and vectors of $\Sigma^{-1}B_o$

Why the eigenvalues and eigenvectors give solution:

Starting with our ratio, gives

$$
\begin{aligned}
\boldsymbol{a}'\boldsymbol{B}_o\boldsymbol{a} &= \lambda(\boldsymbol{a}'\boldsymbol{\Sigma}\boldsymbol{a}) \\
\boldsymbol{a}'(\boldsymbol{B}_o - \lambda\boldsymbol{\Sigma})\boldsymbol{a} &= 0 \\
(\boldsymbol{B}_o - \lambda\boldsymbol{\Sigma})\boldsymbol{a} &= 0 \qquad \text{because } \boldsymbol{a} \neq \boldsymbol{0} \\
&\text{or} \\
(\boldsymbol{\Sigma}^{-1}\boldsymbol{B}_o)\boldsymbol{a} &= \lambda\boldsymbol{a}
\end{aligned}
$$

The eigenvalues and eigenvectors of a non-symmetric matrix can be found by taking the eigenvalues and eigenvectors of $\boldsymbol{\Sigma}^{-1/2}\boldsymbol{B}_o\boldsymbol{\Sigma}^{-1/2}$ (or $\boldsymbol{E}^{-1/2}\boldsymbol{H}\boldsymbol{E}^{-1/2}$) where $\boldsymbol{\Sigma}^{-1/2}$ is the inverse of the square root matrix.

Recall: $\boldsymbol{\Sigma}^{-1/2}$ is the square root matrix that can be found by

1. Find $\boldsymbol{\Sigma} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}'$
2. $\boldsymbol{\Sigma}^{-1/2} = \boldsymbol{U}\text{diag}(1/\sqrt{\lambda_i})\boldsymbol{U}'$

Computing Eigenvalues and vectors of $\Sigma^{-1}B_o$

$$\begin{aligned}
\left(\boldsymbol{\Sigma}^{-1/2}\boldsymbol{B}_o\boldsymbol{\Sigma}^{-1/2}\right)\boldsymbol{e}^* &= \lambda\boldsymbol{e}^* \\
\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{B}_o\boldsymbol{\Sigma}^{-1/2}\boldsymbol{e}^* &= \lambda\boldsymbol{\Sigma}^{-1/2}\boldsymbol{e}^* \\
\boldsymbol{\Sigma}^{-1}\boldsymbol{B}_o\underbrace{\boldsymbol{\Sigma}^{-1/2}\boldsymbol{e}^*}_{\boldsymbol{e}} &= \lambda\underbrace{\boldsymbol{\Sigma}^{-1/2}\boldsymbol{e}^*}_{\boldsymbol{e}}
\end{aligned}$$

$\boxed{\boldsymbol{a} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{e}^* \text{ where } \boldsymbol{e}^* \text{ is eigenvector of } \boldsymbol{\Sigma}^{-1/2}\boldsymbol{B}_o\boldsymbol{\Sigma}^{-1/2}.}$

Note that

$$\begin{aligned}
\text{var}(Y_i) = \boldsymbol{a}_i'\boldsymbol{\Sigma}\boldsymbol{a}_i &= \left(\boldsymbol{e}_i^{*\prime}\underbrace{\boldsymbol{\Sigma}^{-1/2}\right)\boldsymbol{\Sigma}\left(\boldsymbol{\Sigma}^{-1/2}}_{\boldsymbol{I}}\boldsymbol{e}_i^*\right) \\
&= \boldsymbol{e}_i^{*\prime}\boldsymbol{e}_i^* \\
&= 1
\end{aligned}$$

The Next Ones

The next eigenvector $\boldsymbol{e}_2$ minimizes the ratio

$$\max_{\boldsymbol{a}} = \frac{\boldsymbol{a}'\boldsymbol{B}_o\boldsymbol{a}}{\boldsymbol{a}'\boldsymbol{\Sigma}\boldsymbol{a}}$$

subject to the restriction that the covariance between $Y_2 = \boldsymbol{a}_2'\boldsymbol{X}$ and $Y_1 = \boldsymbol{a}_1'\boldsymbol{X} = \boldsymbol{e}_1'\boldsymbol{X}$ equals 0.

$$\begin{aligned} Y_2 &= \boldsymbol{a}_2'\boldsymbol{X} = 2^{nd}\text{discriminant function} \\ \vdots & \quad \vdots \\ Y_k &= \boldsymbol{a}_k'\boldsymbol{X} = k^{th}\text{discriminant function} \\ \vdots & \quad \vdots \\ Y_s &= \boldsymbol{a}_s'\boldsymbol{X} = s^{th}\text{discriminant function} \end{aligned}$$

Note that var$(Y_i) = \boldsymbol{a}_i'\boldsymbol{\Sigma}\boldsymbol{a}_i = \boldsymbol{e}_i^{*\prime}\boldsymbol{e}_i^* = 1$, for all $i = 1, \ldots, s$ where $s \leq \min(g-1, p)$. And for $i \neq k$,

$$\text{cov}(Y_i, Y_k) = \boldsymbol{a}_i'\boldsymbol{\Sigma}\boldsymbol{a}_k = (\boldsymbol{e}_i^{*\prime}\boldsymbol{\Sigma}^{-1/2})\boldsymbol{\Sigma}(\boldsymbol{\Sigma}^{-1/2}\boldsymbol{e}_k^*) = \boldsymbol{e}_i^{*\prime}\boldsymbol{e}_k^* = 0$$

Estimation

Since $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_g$ and $\boldsymbol{\Sigma}$ are generally unknown, we need to estimate them. Suppose that we have a set of observations classified from which we can calculate sample statistics (in a prediction situation, this would be named the "training set").

$\boldsymbol{X}_i$ is the $(n_i \times p)$ data matrix for $i = 1, \ldots, g$.

$\bar{\boldsymbol{X}}_i = (1/n_i) \sum_{j=1}^{n_i} \boldsymbol{X}_{ij}$, which estimates $\boldsymbol{\mu}_i$.

$\boldsymbol{S}_i = (1/(n_i - 1)) \sum_{j=1}^{n_i} (\boldsymbol{X}_{ij} - \bar{\boldsymbol{X}}_i)(\boldsymbol{X}_{ij} - \bar{\boldsymbol{X}}_i)'$.

$\bar{\boldsymbol{X}} = (1/\sum_i n_i) \sum_{i=1}^{g} n_i \bar{\boldsymbol{X}}_i = (1/\sum_i n_i) \sum_{i=1}^{g} \sum_{j=1}^{n_i} \boldsymbol{X}_{ij}$, which estimates $\bar{\boldsymbol{\mu}}$.

The estimation of $\boldsymbol{B}_o = \boldsymbol{H}$, the sample between groups SSCP matrix

$$\hat{\boldsymbol{B}}_o = \sum_{i=1}^{g} n_i (\bar{\boldsymbol{X}}_i - \bar{\boldsymbol{X}})(\bar{\boldsymbol{X}}_i - \bar{\boldsymbol{X}})'$$

Estimation continued

and estimation of $\boldsymbol{W} = \boldsymbol{H}$, the within groups SSCP matrix

$$\hat{\boldsymbol{W}} = \sum_{i=1}^{g}\sum_{j=1}^{n_i}(\boldsymbol{X}_{ij} - \bar{\boldsymbol{X}}_i)(\boldsymbol{X}_{ij} - \bar{\boldsymbol{X}}_i)'$$

For our estimate of $\boldsymbol{\Sigma}$,

$$\boldsymbol{S}_{pool} = \frac{1}{\sum_{i=1}^{g} n_i}\boldsymbol{W}$$

If $\hat{a}_1$ maximizes

$$\frac{\boldsymbol{a}'\hat{\boldsymbol{B}}_o\boldsymbol{a}}{\boldsymbol{a}'\boldsymbol{S}_{pool}\boldsymbol{a}}$$

Then it also maximizes

$$\frac{\boldsymbol{a}'\hat{\boldsymbol{B}}_o\boldsymbol{a}}{\boldsymbol{a}'\boldsymbol{W}\boldsymbol{a}} = \frac{\boldsymbol{a}'\boldsymbol{H}\boldsymbol{a}}{\boldsymbol{a}'\boldsymbol{E}\boldsymbol{a}}$$

Fisher's Sample Discriminant Functions

Let $\hat{\lambda}_1, \cdots, \hat{\lambda}_s$ denote the $s \leq \min(g-1, p)$ non-zero eigenvalues (roots) of $\boldsymbol{W}^{-1}\hat{\boldsymbol{B}}_o$ and $\hat{\boldsymbol{e}}_1, \cdots, \hat{\boldsymbol{e}}_s$ the corresponding eigenvectors (scaled so that $\hat{\boldsymbol{e}}_i \boldsymbol{S}_{pool} \hat{\boldsymbol{e}}_i = 1$), then the vector of coefficients $\boldsymbol{a}$ that maximizes the ratio

$$\frac{\hat{\boldsymbol{a}}'\hat{\boldsymbol{B}}_o\hat{\boldsymbol{a}}}{\hat{\boldsymbol{a}}'\boldsymbol{W}\hat{\boldsymbol{a}}} = \frac{\hat{\boldsymbol{a}}'\left[\sum_{i=1}^{g} n_i(\bar{\boldsymbol{X}}_i - \bar{\boldsymbol{X}})(\bar{\boldsymbol{X}}_i - \bar{\boldsymbol{X}})'\right]\hat{\boldsymbol{a}}}{\hat{\boldsymbol{a}}'\left[\sum_{i=1}^{g}\sum_{j=1}^{n_i}(\boldsymbol{X}_{ij} - \bar{\boldsymbol{X}}_i)(\boldsymbol{X}_{ij} - \bar{\boldsymbol{X}}_i)'\right]\hat{\boldsymbol{a}}}$$

is given the

$$\hat{\boldsymbol{a}}_1 = \hat{\boldsymbol{e}}_1 = \boldsymbol{W}^{-1/2}\boldsymbol{e}_1^* \quad \text{and} \quad \hat{\boldsymbol{a}}_1\boldsymbol{X} \quad \text{is the } 1^{st} \text{ discriminant function}$$

$$\vdots \qquad \vdots$$

$$\hat{\boldsymbol{a}}_k = \hat{\boldsymbol{e}}_k = \boldsymbol{W}^{-1/2}\boldsymbol{e}_k^* \quad \text{and} \quad \hat{\boldsymbol{a}}_k\boldsymbol{X} \quad \text{is the } k^{th} \text{ discriminant function}$$

Note that $\boldsymbol{e}_i^*$ is the i^{th} eigenvector of $\boldsymbol{W}^{-1/2}\hat{\boldsymbol{B}}_o\boldsymbol{W}^{-1/2}$ and $\hat{\boldsymbol{e}}_i = \boldsymbol{W}^{-1/2}\boldsymbol{e}_i^*$.

Some programs use different scalings and even different procedures within the same program may use different scalings.

Suppose that your computed $\boldsymbol{a}$ such that $\boldsymbol{l}^{*\prime}\boldsymbol{W}\boldsymbol{l}^{*} = 1$ but you want $\boldsymbol{l}'\boldsymbol{S}_{pool}\boldsymbol{l} = 1$.

1. Compute $c = \sqrt{\boldsymbol{l}^{*\prime}\boldsymbol{S}_{pool}\boldsymbol{l}^{*}}$
2. $\boldsymbol{l} = (1/c)\boldsymbol{l}^{*}$.

Using the Discriminant for Allocation

Let $Y_k = \boldsymbol{a}_k'\boldsymbol{X} =$ the k^{th} discriminant function and

$$\boldsymbol{Y}' = (Y_1, \ldots, Y_s)$$

has mean vector for population i

$$\mu_{iy}' = (\boldsymbol{a}_1'\boldsymbol{\mu}, \cdots, \boldsymbol{a}_s'\boldsymbol{\mu})$$

where $\boldsymbol{\mu}$ is the mean of the X's.

For all populations, the covariance matrix for the Y's is

$$\text{cov}(\boldsymbol{Y}) = I.$$

The squared distance from $\boldsymbol{y}_o$ to $\boldsymbol{\mu}_{iy}$ is

$$(\boldsymbol{y}_o - \boldsymbol{\mu}_{iy})'(\boldsymbol{y}_o - \boldsymbol{\mu}_{iy}) = \sum_{l=1}^{s}(y_{ol} - \mu_{iy_l})^2$$

The (simple) Rule: Assign $\boldsymbol{y}_o$ to population k if the squared distance from $\boldsymbol{y}_o$ to μ_{ky} is the smallest squared distance between $\boldsymbol{y}_o$ to all other $\boldsymbol{\mu}_{iy}$'s ($i \neq k$).

Number of Discriminant Functions

There are limits on how many discriminant functions that you can compute. Some examples:

Number of variables (p)	Number of populations (g)	Maximum number of discriminant functions $s \leq \min(g-1, p)$
any p	$g = 2$	1
any p	$g = 3$	2
$p = 2$	any g	2

You use only say r of the discriminant functions (for description) where $r < s$. The proportion of between group differences represented by the r functions equals

$$\frac{\sum_{i=1}^{r} \lambda_i}{\sum_{k=1}^{s} \lambda_k}$$

1–Way MANOVA: Data from Tatsuoka (1988), *Multivariate Analysis: Techniques for Educational and Psychological Research*, pp 273–279.

An experiment was conducted for comparing 2 methods (A & B) of teaching shorthand to 60 female seniors in a vocational high school (a dated example). Also of interest were the effects of distributed versus massed practice

C_1: 2 hours of instruction/day for 6 weeks
C_2: 3 hours of instruction/day for 4 weeks
C_3: 4 hours of instruction/day for 3 weeks

So each subject received a total of 12 hours of instruction.
Note: $n_l = 10$ per cell of the design

Two variables (dependent measures):

$X_1 =$ speed
$X_2 =$ accuracy

Recall

We had a significant interaction, so we'll use discriminant analysis to describe the differences.

I did things a bit different and created a single variable "grp" as follows:

if conditin='C1' and method='A' then grp='A1';
else if conditin='C1' and method='B' then grp='B1';
else if conditin='C2' and method='A' then grp='A2';
else if conditin='C2' and method='B' then grp='B2';
else if conditin='C3' and method='A' then grp='A3';
else if conditin='C3' and method='B' then grp='B3';

and testing $\mu_{A1} = \mu_{B1} = \cdots = \boldsymbol{\mu}_{B3}$

$$\Lambda^* = 0.110, \quad F = 21.26, \quad \nu_1 = 10, \quad \nu_2 = 106, \quad p < .0001$$

Discriminant Functions for PROC GLM

Characteristic Roots and Vectors of: E Inverse * H, where
H = Type III SSCP Matrix for grp
E = Error SSCP Matrix

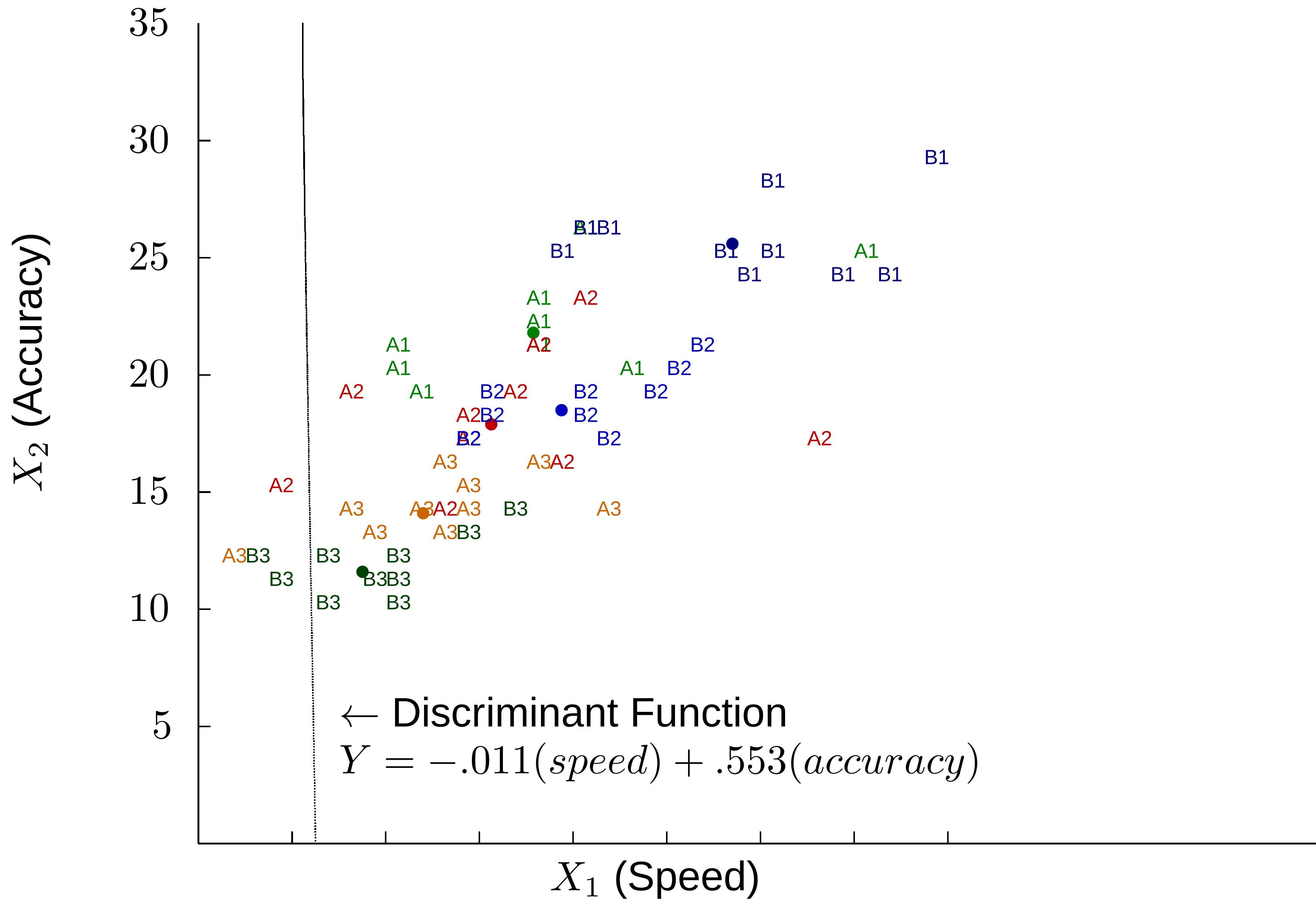
Characteristic Root	Percent	Characteristic speed	Vector V'EV=1 accuracy
6.96079625	98.10	−0.00150305	0.07520907
0.13501094	1.90	−0.03022961	0.03091280

So

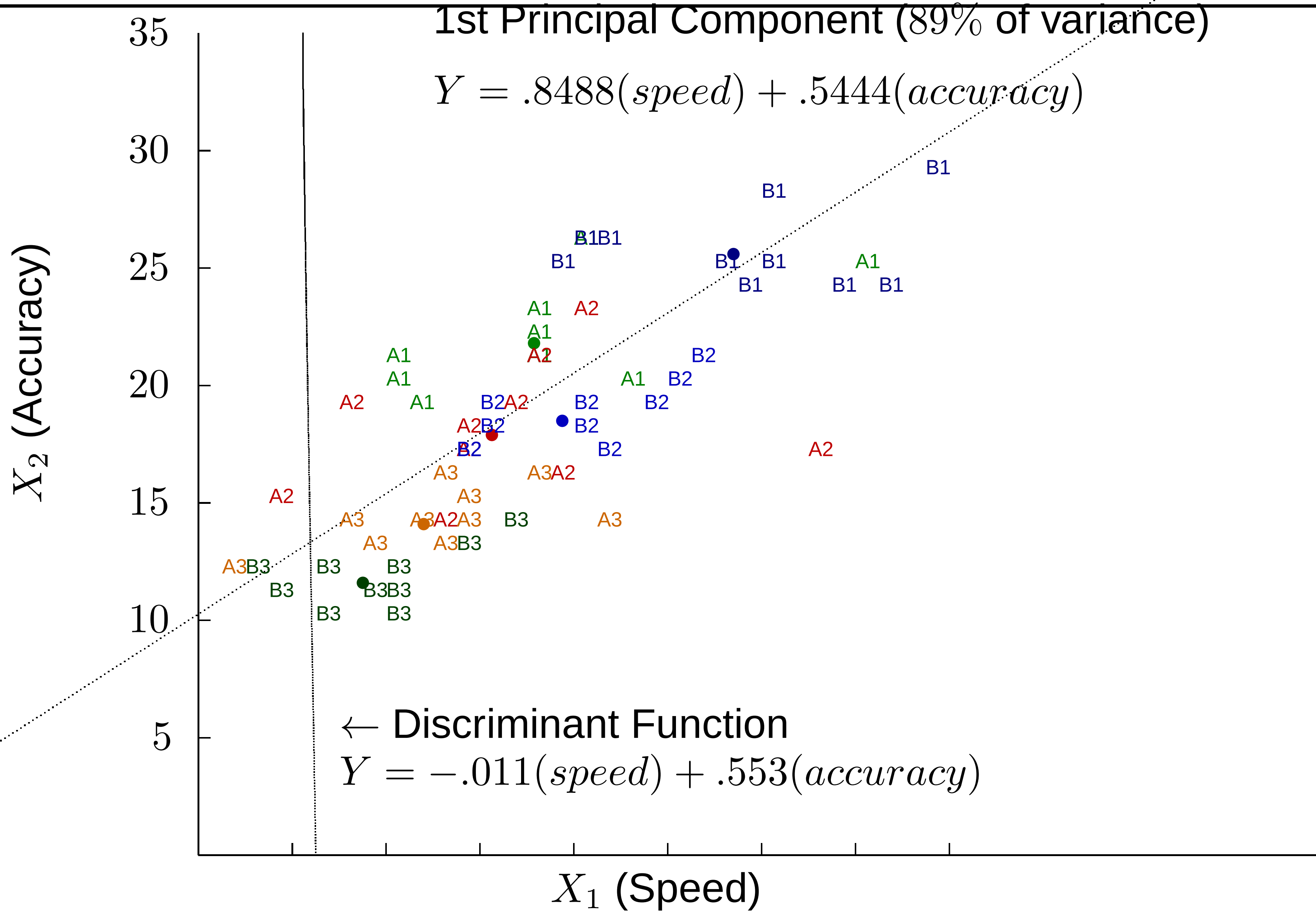
$$Y_1 = -0.00150305(speed) + 0.07520907(accuracy)$$

and this accounts for 98% of between group differences.

Data, Cell Means, and Discriminant Function



Data and Cell Means



Values of Groups on Y

(From PROC CANDISC, re-orderd from largest to smallest):

```
Class Means on Canonical Variables

grp                   Can1

B1             3.960704421
A1             1.954436320
B2             0.117366215
A2            -0.181101297
A3            -2.249222122
B3            -3.602183536
```

In PROC CANDSIC, "Canonical Variables" are linear discriminant functions.

Pairwise Distances Between Groups


```
The DISCRIM Procedure

Pairwise Generalized Squared Distances Between Groups

 2                _    _           -1  _    _
D (i|j) = (X  -  X  )' COV    (X  -  X  )
            i    j               i    j
```

Generalized Squared Distances

	A1	A2	A3	B1	B2	B3
A1	0	4.79	18.16	5.07	4.40	31.35
A2	4.79	0	4.32	17.44	0.37	11.74
A3	18.16	4.32	0	38.66	5.69	1.83
B1	5.07	17.44	38.66	0	14.77	57.30
B2	4.40	0.37	5.69	14.77	0	13.93
B3	31.35	11.74	1.83	57.30	13.93	0

(PROC CANDISCM also gives these, and more. . .)

Significant Pairwise Distances Between Groups

From PROC CANDISCRM, which computes F statistics for each pairwise distance and the corresponding p values, which are given below.

Prob $>$ Mahalanobis Distance for Squared Distance

grp	A1	A2	A3	B1	B2	B3
A1	1.00	$< .01$	$< .01$	$< .01$	$.01$	$< .01$
A2	$< .01$	1.00	$.01$	$< .01$	$.41$	$< .01$
A3	$< .01$	$.01$	1.00	$< .01$	$< .01$	$.02$
B1	$< .01$	$< .01$	$< .01$	1.00	$< .01$	$< .01$
B2	$.01$	$.41$	$< .01$	$< .01$	1.00	$< .01$
B3	$< .01$	$< .01$	$.02$	$< .01$	$< .01$	1.00

Following is SAS input

- PROC GLM
- PROC IML;
- PROC CANDISC
- PROC DISCRIM

We'll look at output in lecture.

The program and the input are on the web-site.

```
* Fisher’s Linear Discrimiant analysis using
  (a) proc glm
  (b) proc iml
  (c) proc candisc
  (d) proc discrim;

options nocenter;
data short;
 input speed accuracy method $ conditin $;
 if conditin=’C1’ and method=’A’
         then grp=’A1’;
 else if conditin=’C1’ and method=’B’
         then grp=’B1’;
 else if conditin=’C2’ and method=’A’
         then grp=’A2’;
 else if conditin=’C2’ and method=’B’
         then grp=’B2’;
 else if conditin=’C3’ and method=’A’
         then grp=’A3’;
 else if conditin=’C3’ and method=’B’
         then grp=’B3’;
* Design matrix for main effect;
 b0=1; A=0; B=0; c1=0; c2=0; c3=0;
 if conditin=’C1’             then c1=1;
    else if conditin=’C2’  then c2=1;
    else if conditin=’C3’  then c3=1;
 if method=’A’                then A=1;
```

```
    else if method='B'     then B=1;
* Design matrix for interaction;
 Ac1= A*c1;   Ac2 = A*c2;  Ac3= A*c3;
 Bc1= B*c1;   Bc2 = B*c2;  Bc3= B*c3;
datalines;

------------------------------------------------

proc glm data=short;
 class grp;
 model speed accuracy = grp;
 manova h=grp/ printh printe;
 title 'Are groups significantly different
        on the p=2 variables';
run;
------------------------------------------------

proc iml; /************* Handy module ***********
samplestats(X,Xbar,W,S,R,n);
  n = nrow(X);
  one = J(n,1);
  Xbar = X`*one/n;
  W =  (X - one*Xbar`)` * (X - one*Xbar`);
  S = W/(n-1);
  Dsqrt = sqrt(diag(S));
  R = inv(Dsqrt)*S*inv(Dsqrt);
Finish samplestats; /*****************************
```

```
use short;
read all var{speed accuracy} into X;

* Discriminant analysis using IML;
use short;
read all var{speed accuracy} into X;
read all var{speed accuracy} into Xa1
                                 where(grp='A1');
read all var{speed accuracy} into Xa2
                                  where(grp='A2');
read all var{speed accuracy} into Xa3
                                  where(grp='A3');
read all var{speed accuracy} into Xb1
                                  where(grp='B1');
read all var{speed accuracy} into Xb2
                                  where(grp='B2');
read all var{speed accuracy} into Xb3
                                  where(grp='B3');
call samplestats(X,M,W,S,R,n);
call samplestats(Xa1,Ma1,Wa1,Sa1,Ra1,na1);
call samplestats(Xa2,Ma2,Wa2,Sa2,Ra2,na2);
call samplestats(Xa3,Ma3,Wa3,Sa3,Ra3,na3);
call samplestats(Xb1,Mb1,Wb1,Sb1,Rb1,nb1);
call samplestats(Xb2,Mb2,Wb2,Sb2,Rb2,nb2);
call samplestats(Xb3,Mb3,Wb3,Sb3,Rb3,nb3);
Bo = Na1*(Ma1-M)*(Ma1-M)`+Na2*(Ma2-M)*(Ma2-M)`
   +Na3*(Ma3-M)*(Ma3-M)`+ Nb1*(Mb1-M)*(Mb1-M)`
```

```
   +Nb2*(Mb2-M)*(Mb2-M)‘+Nb3*(Mb3-M)*(Mb3-M)‘;

 W = Wa1+Wa2+Wa3+Wb1+Wb2+Wb3;


 iW = inv(W);

 *  Method 1: To get the eigenvectors
                         L of inv(W)*Bo;

 call eigen(lam,U,iw);
 iWsqrt = U*diag(sqrt(lam))*U‘;
 M = iWsqrt*Bo*iWsqrt;
 call eigen(lambda,Estar,M);
 L = iWsqrt*Estar;

* Method 2 for getting eigenvectors
                      Lstar of inv(W)*Bo;

F = root(iW);            <--- "Cholosky root"
EHE = F*Bo*F‘;
call eigen(lam2,U2,EHE);
 a = F‘*U2;
print ’Both method of computing eigenvectors
       of inv(W)Bo yield same result’, L a;
```

```
* To scale eigenvectors so L‘*E*L=1;

tmp_glm = L‘*W*L;
C_glm   = sqrt(diag(tmp_glm));
iC_glm  = inv(C_glm); L_glm = L*ic_glm;


* To scale eigenvectors so L‘*Spool*L=1;

 tmp = L‘*W*L/(N-6);
 C = sqrt(diag(tmp));
 iC = inv(C);
 L = L*iC;


 print ’Linear Discriminant Analysis (IML output
 ’Between groups SSCP   ’  Bo,
 ’Pooled (within groups) SSCP   ’ W,
 ’Linear Discriminant Functions
         (GLM scaling: L‘*W*L=1)  ’ L_glm,
 ’Linear Discriminant Functions
     (CANDISC scaling so L‘*Spool*L=1)  ’ L,
 ’Characteristic Roots   ’ lambda;

---------------------------------------------
---------------------------------------------
```

```
proc candisc distance anova data=short;
 class grp;
 var speed accuracy;
 title 'Designed for Description:
         DA via PROC CANDISC';
run;



------------------------------------------------
proc candisc data=short all ; <--"all" prints
                                  lots of stuff
 class grp;
 var speed accuracy;
 title 'Designed for Description:
         DA via PROC CANDISC';
run;
------------------------------------------------

proc discrim simple pool=yes Wcov Pcov list;
 class grp;
 var speed accuracy;
 priors equal;
 title 'Designed for Classification/Allocation:
        DA via PROC DISCRIM';
run;
```