

Chapter 5: Discriminant Analysis and Classification Analysis

Jingyuan Liu

Department of Statistics, School of Economics
Wang Yanan Institute for Studies in Economics
Xiamen University

Outline

- 1 Introduction to Discriminant and Classification Analysis
 - Intuition Example
 - Definitions and Applications
- 2 Discriminant Analysis: Description of Group Separation
 - Fisher's Linear Discriminant Analysis for Two Populations
 - Fisher's LDA for Several Populations
- 3 Classification: Allocation of Observations to Groups
 - Classification for Two Populations
 - Classification Based on Fisher's Linear Discriminant Analysis
 - Classification Based on Bayes Rule
 - Classification for Several Populations

Outline


- 1 Introduction to Discriminant and Classification Analysis
 - Intuition Example
 - Definitions and Applications
 - Classification Based on Fisher's Linear Discriminant Analysis
 - Classification Based on Bayes Rule

Intuition Example

A loan officer at a bank wishes to decide whether to approve an applicant's automobile's loan.



The personal information of the applicant - age, income, marital status, outstanding debt, and home ownership are collected.



Essentially, the loan officer need to classify the applicant into one of the two groups: The people who tend to repay loans successfully and those who default. It consists of two tasks:

- (1) Find a rule to “best” separate the two groups using the differential features (age, income, etc.) of people from the two groups based on the historical data.
- (2) Allocate this current applicant to one of the two groups using the rule.

The first task is more referred to as **discriminant analysis**, while the second as **classification analysis**.

Outline

- 1 Introduction to Discriminant and Classification Analysis
 - Intuition Example
 - Definitions and Applications
 - Classification Based on Fisher's Linear Discriminant Analysis
 - Classification Based on Bayes Rule

Introduction

- **Discriminant analysis** often refers to the **description** of group separation: It develops functions of variables (discriminant functions) to describe or elucidate the differences between two or more groups.
- **Classification analysis** more refers to the **prediction** or **allocation** of a new observation to groups: The measured values in the new observation vector are evaluated by some rules (classification functions) to find the group to which the observation most likely belongs.
- The goals of discrimination and classification frequently overlap, especially, the discriminant analysis are often used in connection with the allocation objective. So the distinction is blurred.

Application Examples

Several examples of the discrimination and classification:

- A university admissions committee wants to classify the applicants as “likely to succeed” or “likely to fail”. The variables available are the high school grades in various subject areas, standardized test scores, rating of high school, number of advanced placement courses, etc.
- A psychiatrist gives a battery of diagnostic tests in order to assign a patient to the appropriate mental illness category.
- African, or “killer” bees cannot be distinguished visually from ordinary domestic honey bees. Ten variables based on chromatograph peaks can be used to identify them.

Outline

- 2 Discriminant Analysis: Description of Group Separation
 - Fisher's Linear Discriminant Analysis for Two Populations
 - Fisher's LDA for Several Populations
 - Classification Based on Fisher's Linear Discriminant Analysis
 - Classification Based on Bayes Rule

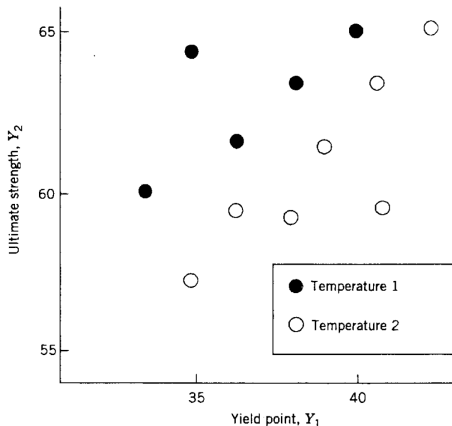
Fisher's Linear Discriminant Analysis

For the discriminant analysis, we only introduce the **Fisher's linear discriminant analysis (LDA)**, which was mentioned as a follow-up procedure of two-sample T^2 tests in Chapter 4. We here discuss it in detail as an independent technique and as the preparation of classification.

Fisher's LDA: Example

Example: Samples of steel produced at two different rolling temperatures are compared by two variables: Y_1 = yield point and Y_2 = ultimate strength. The data are as follows.

Temperature 1		Temperature 2	
y_1	y_2	y_1	y_2
33	60	35	57
36	61	36	59
35	64	38	59
38	63	39	61
40	65	41	63
		43	65
		41	59



Clearly the two groups are well separated, but not in either Y_1 or Y_2 direction. How to find a way to separate them?

Fisher's LDA: Basic Settings

■ Assumption:

The two populations to be separated have the same covariance matrix Σ but distinct mean vectors μ_1 and μ_2 . (No distribution assumptions needed.)

■ Samples and statistics:

The p -variate sample $\mathbf{y}_{11}, \dots, \mathbf{y}_{1n_1}$ has mean vector $\bar{\mathbf{y}}_1$;
The p -variate sample $\mathbf{y}_{21}, \dots, \mathbf{y}_{2n_2}$ has mean vector $\bar{\mathbf{y}}_2$.

■ Goal:

Find a linear combination of the p variables such that the statistical distance between the two transformed groups is maximized.

Fisher's Discriminant Function

The linear combinations of \mathbf{y} 's can be realized by $Z = \mathbf{a}'\mathbf{y}$:

$$z_{1i} = \mathbf{a}'\mathbf{y}_{1i} = a_1 y_{1i1} + a_2 y_{1i2} + \cdots + a_p y_{1ip}, \quad i = 1, 2, \dots, n_1$$

$$z_{2i} = \mathbf{a}'\mathbf{y}_{2i} = a_1 y_{2i1} + a_2 y_{2i2} + \cdots + a_p y_{2ip}, \quad i = 1, 2, \dots, n_2.$$

The goal is to find \mathbf{a} to maximize

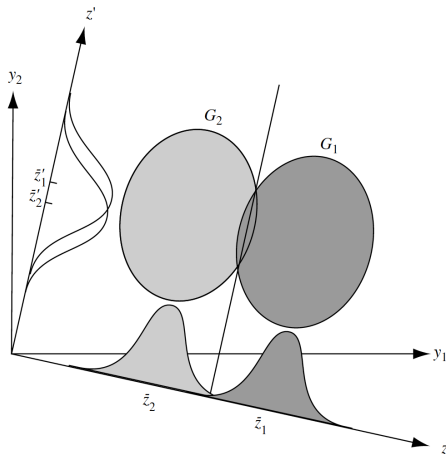
$$\frac{(\bar{z}_1 - \bar{z}_2)^2}{s_z^2} = \frac{[\mathbf{a}'(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)]^2}{\mathbf{a}'\mathbf{S}_{pl}\mathbf{a}}.$$

Then \mathbf{a} is the discriminant function coefficient (Ch 4):

$$\mathbf{a} = \mathbf{S}_{pl}^{-1}(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2) \text{ or its multiple,}$$

and $\mathbf{a}'\mathbf{y}$ is the discriminant function. The corresponding maximum is $(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)'\mathbf{S}_{pl}^{-1}(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)$,

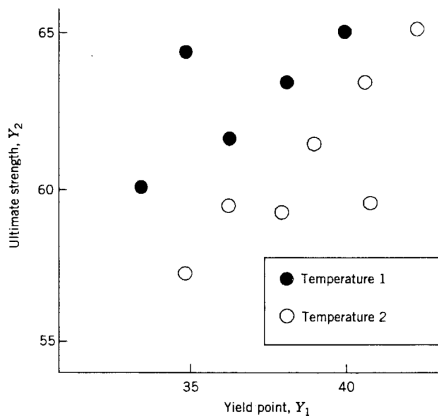
The following figure illustrates the separation of two bivariate normal populations with the same covariance matrix:



Discriminant Function: Example

Back to the previous steel example, with Y_1 = yield point and Y_2 = ultimate strength.

Temperature 1		Temperature 2	
y_1	y_2	y_1	y_2
33	60	35	57
36	61	36	59
35	64	38	59
38	63	39	61
40	65	41	63
		43	65
		41	59



If the points were projected on either y_1 or y_2 axis, there would be considerable overlap. In fact, both univariate t tests are insignificant at significance level 0.05, since the critical value $t_{0.025}(10) = 2.23$:

$$\bar{\mathbf{y}}_1 = \begin{pmatrix} 36.4 \\ 62.6 \end{pmatrix}, \quad \bar{\mathbf{y}}_2 = \begin{pmatrix} 39.0 \\ 60.4 \end{pmatrix}, \quad \mathbf{S}_{\text{pl}} = \begin{pmatrix} 7.92 & 5.68 \\ 5.68 & 6.29 \end{pmatrix}.$$

$$t_1 = \frac{\bar{y}_{11} - \bar{y}_{21}}{\sqrt{s_{11}(1/n_1 + 1/n_2)}} = -1.58,$$

$$t_2 = \frac{\bar{y}_{12} - \bar{y}_{22}}{\sqrt{s_{22}(1/n_1 + 1/n_2)}} = 1.48.$$

However, it is clear that the two groups can be separated.

The discriminant function coefficient

$$\mathbf{a} = \mathbf{S}_{pl}^{-1}(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2) = (-1.633, 1.820)'.$$

The discriminant function $Z = \mathbf{a}'\mathbf{y} = -1.633Y_1 + 1.820Y_2$, thus the values of the projected points are:

Temperature 1	Temperature 2
55.29	46.56
52.20	48.57
59.30	45.30
52.58	47.30
52.95	47.68
	48.05
	40.40

The separation between these two groups is evident.

Fisher's LDA: Remarks

Remarks:

- The discriminant function coefficient tells the direction onto which the original data should be projected in order to achieve the maximum distance, not the direction that separates the two groups.
- To interpret the discriminant function, we are more commonly interested in assessing the contribution of the variables. Refer to the previous example.
- If the p variables in \mathbf{y} are not commensurate, i.e., measured on the same scale and with comparable variances, we often first standardize them before doing the discriminant analysis.

Outline

- 2 Discriminant Analysis: Description of Group Separation
 - Fisher's Linear Discriminant Analysis for Two Populations
 - Fisher's LDA for Several Populations
 - Classification Based on Fisher's Linear Discriminant Analysis
 - Classification Based on Bayes Rule

Several-population Case: Intuition

To generalize the Fisher's LDA to the several-population case (also with equal covariance matrix $\mathbf{\Sigma}$), reconsider the rationale of the two-population case: Find \mathbf{a} to maximize

$$\frac{(\bar{z}_1 - \bar{z}_2)^2}{s_z^2} = \frac{[\mathbf{a}'(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)]^2}{\mathbf{a}'\mathbf{S}_p\mathbf{a}} = \frac{\mathbf{a}'(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)'\mathbf{a}}{\mathbf{a}'\mathbf{S}_p\mathbf{a}}.$$

We can think of the above as the ratio of the “between-group variation” and the “within-group variation”.

Several-population Case

Suppose now g groups are under consideration.

■ Notations:

For the k th group G_k , $k = 1, \dots, g$, the sample is denoted $\mathbf{y}_{k1}, \dots, \mathbf{y}_{kn_k}$, with sample size n_k and sample mean $\bar{\mathbf{y}}_k$. The overall mean $\bar{\mathbf{y}} = 1/g \sum_{k=1}^g \bar{\mathbf{y}}_k$.

■ Objective function:

The ratio of the “between-group variation” and the “within-group variation” now becomes

$$\frac{\mathbf{a}'\mathbf{B}\mathbf{a}}{\mathbf{a}'\mathbf{W}\mathbf{a}} = \frac{\mathbf{a}' [\sum_{k=1}^g (\bar{\mathbf{y}}_k - \bar{\mathbf{y}})(\bar{\mathbf{y}}_k - \bar{\mathbf{y}})'] \mathbf{a}}{\mathbf{a}' [\sum_{k=1}^g \sum_{i=1}^{n_k} (\mathbf{y}_{ki} - \bar{\mathbf{y}}_k)(\mathbf{y}_{ki} - \bar{\mathbf{y}}_k)'] \mathbf{a}} \quad (1)$$

Thus \mathbf{a} are to be maximize (1).

Fisher's Linear Discriminants: Theorem

Theorem (Fisher's linear discriminants)

Let $\lambda_1, \dots, \lambda_s > 0$ denote the $s \leq \min(g-1, p)$ nonzero eigenvalues of $\mathbf{W}^{-1}\mathbf{B}$ and $\mathbf{e}_1, \dots, \mathbf{e}_s$ are the corresponding eigen vectors. Then the coefficient vector \mathbf{a} that maximizes the ratio (1) is given by $\mathbf{a} = \mathbf{e}_1$, and the maximum is λ_1 . The linear combination $\mathbf{e}_1'\mathbf{y}$ is called the sample first discriminant, and the choice $\mathbf{e}_2'\mathbf{y}$ is called the sample second discriminant, and so forth.

Several-population Case: Example

Example: Some researchers collected crude-oil samples from sandstone of a petroleum reserve in California. These crude oils can be assigned to one of the three stratigraphic units

G_1 : Wilhelm sandstone

G_2 : Sub-Mulinia sandstone

G_3 : Upper sandstone

on the basis of five variables of their chemistry:

Y_1 = vanadium (in percent ash)

Y_2 = $\sqrt{\text{iron (in percent ash)}}$

Y_3 = $\sqrt{\text{beryllium (in percent ash)}}$

Y_4 = $1/[\text{saturated hydrocarbons (in percent area)}]$

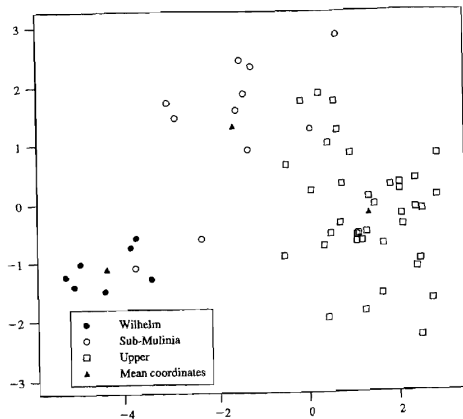
Y_5 = aromatic hydrocarbons (in percent area)

	x_1	x_2	x_3	x_4	x_5
G1	3.9	51.0	0.20	7.06	12.19
	2.7	49.0	0.07	7.14	12.23
	2.8	36.0	0.30	7.00	11.30
	3.1	45.0	0.08	7.20	13.01
	3.5	46.0	0.10	7.81	12.63
	3.9	43.0	0.07	6.25	10.42
	2.7	35.0	0.00	5.11	9.00
G2	5.0	47.0	0.07	7.06	6.10
	3.4	32.0	0.20	5.82	4.69
	1.2	12.0	0.00	5.54	3.15
	8.4	17.0	0.07	6.31	4.55
	4.2	36.0	0.50	9.25	4.95
	4.2	35.0	0.50	5.69	2.22
	3.9	41.0	0.10	5.63	2.94
	3.9	36.0	0.07	6.19	2.27
	7.3	32.0	0.30	8.02	12.92
	4.4	46.0	0.07	7.54	5.76
G3	3.0	30.0	0.00	5.12	10.77
	6.3	13.0	0.50	4.24	8.27
	1.7	5.6	1.00	5.69	4.64
	7.3	24.0	0.00	4.34	2.99
	7.8	18.0	0.50	3.92	6.09
	7.8	25.0	0.70	5.39	6.20
	7.8	26.0	1.00	5.02	2.50
	9.5	17.0	0.05	3.52	5.71
	7.7	14.0	0.30	4.65	8.63
	11.0	20.0	0.50	4.27	8.40

There are at most $s = \min(g - 1, p) = 2$ positive eigenvalues of $\mathbf{W}^{-1}\mathbf{B}$, and they are 4.354 and 0.559. The centered Fisher's linear discriminants are

$$\begin{aligned} Z_1 &= 0.312(Y_1 - 6.180) - 0.710(Y_2 - 5.081) + 2.764(Y_3 - 0.511) \\ &\quad + 11.809(Y_4 - 0.201) - 0.235(Y_5 - 6.434) \\ Z_2 &= 0.169(Y_1 - 6.180) - 0.245(Y_2 - 5.081) - 2.046(Y_3 - 0.511) \\ &\quad - 24.453(Y_4 - 0.201) - 0.378(Y_5 - 6.434) \end{aligned}$$

The separation of the three group means is fully explained in the two-dimensional discriminant coordinate system.



The separation is quite good according to the Fisher's discriminant plot.

Several-population Case: Remarks

Remarks:

- For several-population case, the “within-group” variation $\mathbf{W}/(n_1 + \dots + n_g - g) = \mathbf{S}_{pl}$ is the estimate of $\mathbf{\Sigma}$.
- In terms of graphical display, the Fisher’s LDA for several groups could reduce the dimension from a very large number of characteristics to a relatively few linear combinations. This helps display relationships and possible groupings of the population.
- The Fisher’s LDA (for both of the two-population and several-population case) does not require normality, but it assumes equal covariance matrix for all the populations.

Outline

- 3 Classification: Allocation of Observations to Groups
 - Classification for Two Populations
 - Classification Based on Fisher's Linear Discriminant Analysis
 - Classification Based on Bayes Rule
 - Classification for Several Populations

Classification: Intuition

- The discriminant analysis is to separate the groups, while the classification is to allocate the new observations to a proper group - it is the predictive aspect of the analysis.
- Recall the loan example, where the loan officer needs to decide whether to approve the applicant's loan.
 - The essential goal is to determine whether this person belongs to the “trustable” group or the “default” group.
 - The discriminant analysis aims to find the optimal direction to project the multivariate personal information (age, income, etc.) to a single dimension such that the projected values between the existing applicants from the two groups differ the most. This should be done before classifying the new applicants to one of the groups.

Outline

- 3 Classification: Allocation of Observations to Groups
 - Classification for Two Populations
 - Classification Based on Fisher's Linear Discriminant Analysis
 - Classification Based on Bayes Rule
 - Classification for Several Populations

Two-Population Classification Based on Fisher's Linear Discriminant Analysis

■ Assumption:

The two populations G_1 and G_2 have the same covariance matrix $\mathbf{\Sigma}$, and $n_1 + n_2 - 2 \geq p$.

■ Classification idea:

Based on the discriminant function

$$Z = \mathbf{a}'\mathbf{y} = (\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)' \mathbf{S}_{pl}^{-1} \mathbf{y},$$

we just need to see if $z_0 = \mathbf{a}'\mathbf{y}_0$ is closer to the transformed mean $\bar{z}_1 = \mathbf{a}'\bar{\mathbf{y}}_1$ or to $\bar{z}_2 = \mathbf{a}'\bar{\mathbf{y}}_2$. It is easy to show that z_0 is closer to \bar{z}_1 if

$$z_0 > \frac{1}{2}(\bar{z}_1 + \bar{z}_2) = \frac{1}{2}(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)' \mathbf{S}_{pl}^{-1} (\bar{\mathbf{y}}_1 + \bar{\mathbf{y}}_2).$$

An Allocation Rule Based on Fisher's Discriminant Function

Theorem (Fisher's allocation rule for two groups)

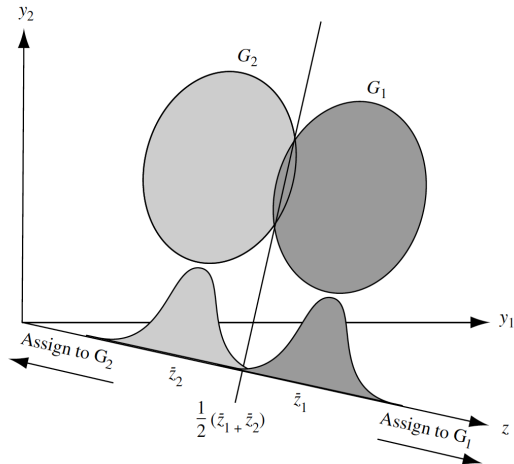
Allocate the new observation \mathbf{y}_0 to G_1 if

$$(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)' \mathbf{S}_{pl}^{-1} \mathbf{y}_0 \geq \frac{1}{2} (\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)' \mathbf{S}_{pl}^{-1} (\bar{\mathbf{y}}_1 + \bar{\mathbf{y}}_2).$$

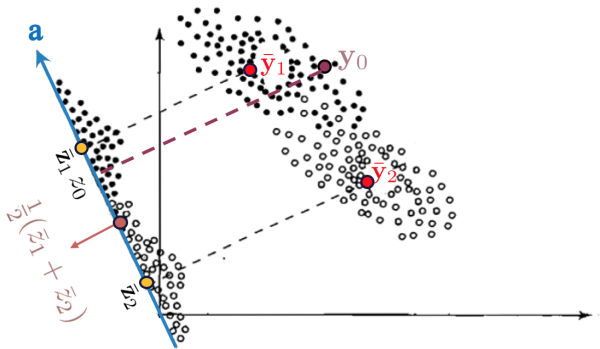
Allocate \mathbf{y}_0 to G_2 if

$$(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)' \mathbf{S}_{pl}^{-1} \mathbf{y}_0 < \frac{1}{2} (\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)' \mathbf{S}_{pl}^{-1} (\bar{\mathbf{y}}_1 + \bar{\mathbf{y}}_2).$$

For instance, if $p = 2$, i.e., $\mathbf{y} = (Y_1, Y_2)'$,



A more intuitive illustration is



Misclassification Rate

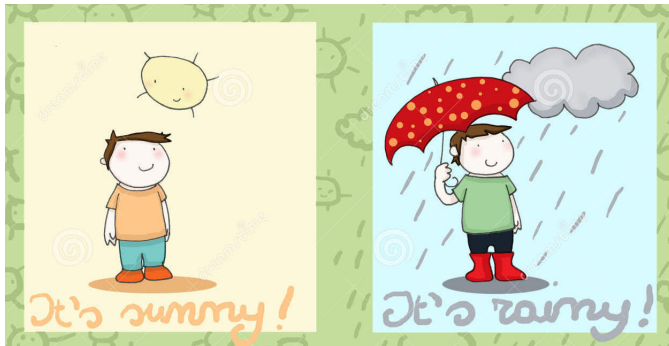
The classification rules cannot usually provide an error-free method of assignment. Then the **total probability of misclassification (TPM)** can be computed from the following count table:

		Classify as:	
		G_1	G_2
True population:	G_1	n_{11}	n_{12}
	G_2	n_{21}	n_{22}

$$\begin{aligned}\text{TPM} &= P(\text{observation is classified as } G_1 \text{ but actually from } G_2) \\ &\quad + P(\text{observation is classified as } G_2 \text{ but actually from } G_1) \\ &= \frac{n_{12} + n_{21}}{n_{11} + n_{12} + n_{21} + n_{22}}\end{aligned}$$

Fisher's Allocation Rule: Example

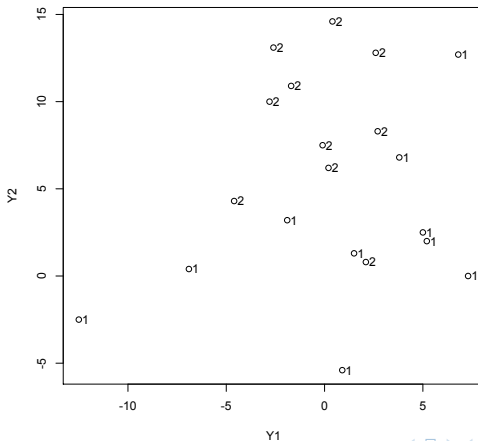
Example: The humidity difference (Y_1) and temperature difference (Y_2) between “today” and “yesterday” are two significant factors to forecast if it is rainy “tomorrow”.



The historical data are as follows.

Rainy (A)			Sunny (B)		
Group	\bar{y}_1	\bar{y}_2	Group	\bar{y}_1	\bar{y}_2
1	-1.9	3.2	2	0.2	6.2
1	-6.9	0.4	2	-0.1	7.5
1	5.2	2.0	2	0.4	14.6
1	5.0	2.5	2	2.7	8.3
1	7.3	0.0	2	2.1	0.8
1	6.8	12.7	2	-4.6	4.3
1	0.9	-5.4	2	-1.7	10.9
1	-12.5	-2.5	2	-2.6	13.1
1	1.5	1.3	2	2.6	12.8
1	3.8	6.8	2	-2.8	10.0

```
> rain=read.table("/Users/jingyuan/Documents/Teaching/Multivariate  
Analysis/R code/Chap5/rain.csv",sep="," , header=T) # read data  
> attach(rain)  
> plot(Y1,Y2)  
> text(Y1,Y2,Group,adj=-0.5)      # mark the groups
```



How to use Fisher's LDA to separate the two groups?

```
> library(MASS)
> (ld=lda(Group~Y1+Y2))           # build the LDA model
Call:
lda(Group ~ Y1 + Y2)

Prior probabilities of groups:
  1  2
0.5 0.5

Group means:
      Y1  Y2
1  0.92 2.10
2 -0.38 8.85

Coefficients of linear discriminants:
      LD1
Y1 -0.1035305
Y2  0.2247957
```

Thus, the discriminant function is $Z = -0.104Y_1 + 0.225Y_2$.

We could “re-predict” the current sample:

```
> Z<-predict(ld)                # predict the current sample
> pred.G<-Z$class
> result<-cbind(Group,pred.G,Z$x)
> colnames(result)<-c("TrueGroup", "Predicted", "TransformedData")
> result
```

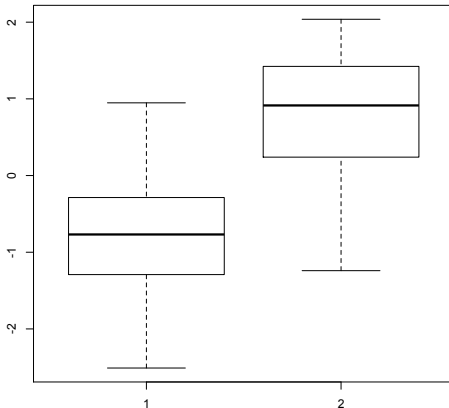
	TrueGroup	Predicted	TransformedData
1	1	1	-0.28674901
2	1	1	-0.39852439
3	1	1	-1.29157053
4	1	1	-1.15846657
5	1	1	-1.95857603
6	1	2	<u>0.94809469</u>
7	1	1	-2.50987753
8	1	1	-0.47066104
9	1	1	-1.06586461
10	1	1	-0.06760842
11	2	2	0.17022402
12	2	2	0.49351760
13	2	2	2.03780185
14	2	2	0.38346871
15	2	1	<u>-1.24038077</u>
16	2	2	0.24005867
17	2	2	1.42347182
18	2	2	2.01119984
19	2	2	1.40540244
20	2	2	1.33503926

The total probability of misclassification TPM can be also computed:

```
> (tab<-table(Group,pred.G))           # count table for classification
      pred.G
Group 1 2
  1  9 1
  2  1 9
> (TPM<-1-sum(diag(prop.table(tab))))  # total probability of misclassification
[1] 0.1
```

The boxplots of the transformed data for the two original groups show the good separation by Fisher's LDA:

```
> boxplot(Z$x~rain$G) # boxplot of the two separated transformed groups
```



If we obtain the data for today is $\mathbf{y}_0 = (8.1, 2.0)'$, how should we forecast tomorrow's weather?

```
> new<-data.frame(cbind(Y1=8.1,Y2=2)) # new observation as data frame
> predict(ld,new)
$class
[1] 1
Levels: 1 2

$posterior
      1      2
1 0.9327428 0.06725717

$x
      LD1
1 -1.591809
```

Thus, tomorrow should be rainy according to Fisher's logic...

Fisher's Allocation Rule: Remark

Remarks:

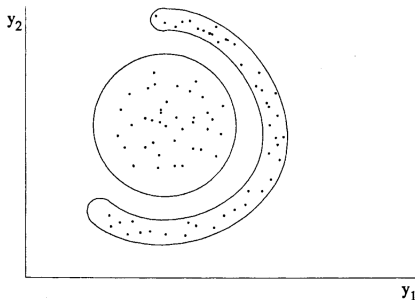
- It is easy to check mathematically that Fisher's rule is essentially comparing the Mahalanobis distance between the new observation \mathbf{y}_0 and $\bar{\mathbf{y}}_1$ and that between \mathbf{y}_0 and $\bar{\mathbf{y}}_2$. That is, allocate \mathbf{y}_0 to G_1 if \mathbf{y}_0 is “closer” to $\bar{\mathbf{y}}_1$ than to $\bar{\mathbf{y}}_2$:

$$(\mathbf{y}_0 - \bar{\mathbf{y}}_1)' \mathbf{S}_{pl}^{-1} (\mathbf{y}_0 - \bar{\mathbf{y}}_1) \leq (\mathbf{y}_0 - \bar{\mathbf{y}}_2)' \mathbf{S}_{pl}^{-1} (\mathbf{y}_0 - \bar{\mathbf{y}}_2),$$

and vice versa.

Remarks: (Cont'd)

- Fisher's rule is nonparametric as no distributional assumptions were made, but it works better for normally distributed populations or the other populations with linear trend. The following example illustrates two populations with nonlinear separation:



Outline

- 3 Classification: Allocation of Observations to Groups
 - Classification for Two Populations
 - Classification Based on Fisher's Linear Discriminant Analysis
 - Classification Based on Bayes Rule
 - Classification for Several Populations

Intuition of Classification with Bayes Rule

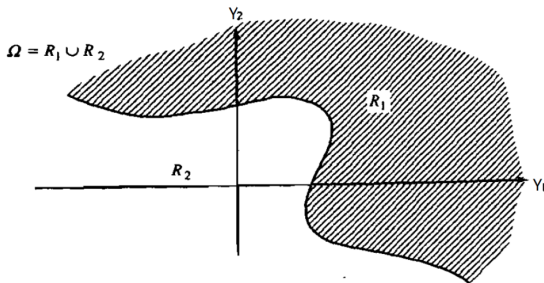
The Fisher's LDA provides a simple way to classify the new observation, but it does not consider (1) the prior probability of occurrence and (2) the cost for misclassification. E.g.

- If we really believe that the (prior) probability of a financially distressed and ultimately bankrupted firm is very small, then one should classify a randomly selected firm as nonbankrupt unless the data overwhelmingly favors bankruptcy. So (1) should be considered.
- Failing to diagnose a potentially fatal illness is substantially more “costly” than concluding that the disease is present when, in fact, it is not. So (2) should be considered if possible.

Notations to be Involved in Bayes Rule

- The p -variate random vector $\mathbf{y} = (Y_1, \dots, Y_p)'$.
- G_1 and G_2 denote the two populations, with the prior probability of p_1 and p_2 , respectively. ($p_1 + p_2 = 1$)
- $f_1(\mathbf{y})$ and $f_2(\mathbf{y})$ are the probability density functions of \mathbf{y} in populations G_1 and G_2 , respectively.
- R_1 and R_2 denote the two groups separated by the classification rule. I.e. if a new observation falls in R_k , then we claim it is from population G_k , $k = 1, 2$. R_1 and R_2 are mutually exclusive and collectively exhaustive.

E.g. The following figure depicts certain classification regions for two populations when $p = 2$.



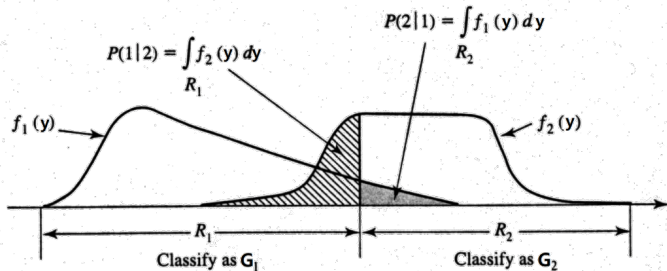
- $P(1|2)$ is the conditional probability of classifying an object \mathbf{y} as G_1 ($\mathbf{y} \in R_1$) when, in fact, it is from G_2 :


$$P(1|2) = P(\mathbf{y} \in R_1 | G_2) = \int_{R_1} f_2(\mathbf{y}) d\mathbf{y}.$$

Similarly, $P(2|1)$ is that of classifying \mathbf{y} as G_2 ($\mathbf{y} \in R_2$) when it is from G_1 :

$$P(2|1) = P(\mathbf{y} \in R_2 | G_1) = \int_{R_2 = \Omega - R_1} f_1(\mathbf{y}) d\mathbf{x}$$

E.g. The following figure depicts the misclassification probabilities when $p = 1$:





Thus the overall probabilities of incorrectly classifying objects can be derived:

$$\begin{aligned} & P(\text{observation is misclassified as } G_1) \\ = & P(\text{observation comes from } G_2 \text{ and is misclassified as } G_1) \\ = & P(\mathbf{y} \in R_1 | G_2)P(G_2) = P(1|2)p_2, \end{aligned}$$


and similarly,

$$\begin{aligned} & P(\text{observation is misclassified as } G_2) \\ = & P(\text{observation comes from } G_1 \text{ and is misclassified as } G_2) \\ = & P(\mathbf{y} \in R_2 | G_1)P(G_1) = P(2|1)p_1 \end{aligned}$$

New notations and quantities: (Cont'd)

- $c(1|2)$ is the cost of misclassifying an observation \mathbf{y} from G_2 to G_1 , i.e. $\mathbf{y} \in R_1$ but it is actually from G_2 ; $c(2|1)$ is defined reversely.

Cost	Classify as:	
	G_1	G_2
G_1	0	$c(2 1)$
G_2	$c(1 2)$	0



Based on the aforementioned quantities, can you think of any optimization objectives to obtain the classification rule?

Bayes Classification Rule

The **Bayes classification rule** aims to minimize the *expected cost of misclassification* (ECM):

$$\text{ECM} = c(2|1)P(2|1)p_1 + c(1|2)P(1|2)p_2$$

Theorem (Bayes classification rule)

The regions R_1 and R_2 that minimize the ECM are defined by the values \mathbf{x} for which the following inequalities hold:

$$R_1 : \frac{f_1(\mathbf{y})}{f_2(\mathbf{y})} \geq \left(\frac{c(1|2)}{c(2|1)} \right) \left(\frac{p_2}{p_1} \right)$$
$$R_2 : \frac{f_2(\mathbf{y})}{f_1(\mathbf{y})} \geq \left(\frac{c(2|1)}{c(1|2)} \right) \left(\frac{p_1}{p_2} \right)$$

Special Cases of Bayes Classification Rule

(a) $p_2/p_1 = 1$ (equal prior probabilities)

$$R_1 : \frac{f_1(\mathbf{y})}{f_2(\mathbf{y})} \geq \frac{c(1|2)}{c(2|1)}; \quad R_2 : \frac{f_1(\mathbf{y})}{f_2(\mathbf{y})} < \frac{c(1|2)}{c(2|1)}$$

(b) $c(1|2)/c(2|1) = 1$ (equal misclassification costs)

$$R_1 : \frac{f_1(\mathbf{y})}{f_2(\mathbf{y})} \geq \frac{p_2}{p_1}; \quad R_2 : \frac{f_1(\mathbf{y})}{f_2(\mathbf{y})} < \frac{p_2}{p_1}$$

(c) $p_2/p_1 = c(1|2)/c(2|1) = 1$

(equal prior probabilities and equal misclassification costs)

$$R_1 : \frac{f_1(\mathbf{y})}{f_2(\mathbf{y})} \geq 1; \quad R_2 : \frac{f_1(\mathbf{y})}{f_2(\mathbf{y})} < 1$$

Bayes Classification Rule: Remarks

Remarks:

- The implementation of minimum ECM rule requires (1) the density function ratio at a new observation \mathbf{y}_0 , (2) the cost ratio, and (3) the prior probability ratio. Often, it is easier to specify the ratios than their component parts.
- When the prior probabilities are unknown, they are often taken to be equal, and the rule reduces to (a); If the cost ratio is indeterminate, it is usually taken to be unity - the rule reduces to (b); If both not available, (c) is used.

Remarks: (Cont'd)

- Another criterion to derive an “optimal” classification procedure could be choosing R_1 and R_2 to minimize the *total probability of misclassification* (TPM):

$$\begin{aligned}\text{TPM} &= P(\text{misclassify a } G_1 \text{ observation or a } G_2 \text{ observation}) \\ &= p_1 \int_{R_2} f_1(\mathbf{y}) d\mathbf{y} + p_2 \int_{R_1} f_2(\mathbf{y}) d\mathbf{y}\end{aligned}$$

Mathematically, this is equivalent to (b): minimizing the ECM when the costs are equal.

Remarks: (Cont'd)

- We could also allocate the new observation \mathbf{y}_0 to the population with the largest “posterior” probability $P(G_k|\mathbf{y}_0)$, $k = 1, 2$:

$$\begin{aligned}P(G_1|\mathbf{y}_0) &= \frac{P(G_1 \text{ occurs and observe } \mathbf{y}_0)}{P(\text{observe } \mathbf{y}_0)} \\&= \frac{P(\text{observe } \mathbf{y}_0|G_1)P(G_1)}{P(\text{observe } \mathbf{y}_0|G_1)P(G_1) + P(\text{observe } \mathbf{y}_0|G_2)P(G_2)} \\&= \frac{p_1 f_1(\mathbf{y}_0)}{p_1 f_1(\mathbf{y}_0) + p_2 f_2(\mathbf{y}_0)} \\P(G_2|\mathbf{y}_0) &= 1 - P(G_1|\mathbf{y}_0) = \frac{p_2 f_2(\mathbf{y}_0)}{p_1 f_1(\mathbf{y}_0) + p_2 f_2(\mathbf{y}_0)}\end{aligned}$$

This is also equivalent to using (b).

Connecting Fisher's LDA to Bayes Rule

Suppose now G_1 and G_2 are from $N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ and $N_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$, respectively, with the common covariance matrix $\boldsymbol{\Sigma}$. Then we can simplify the estimated Bayes rule to a linear function:

Theorem (Estimated Minimum ECM Rule under Normality)

Allocate \mathbf{y}_0 to G_1 if

$$(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)' \mathbf{S}_{pl}^{-1} \mathbf{y}_0 - \frac{1}{2} (\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)' \mathbf{S}_{pl}^{-1} (\bar{\mathbf{y}}_1 + \bar{\mathbf{y}}_2) \geq \ln \left[\left(\frac{c(1|2)}{c(2|1)} \right) \left(\frac{p_2}{p_1} \right) \right]$$

Allocate \mathbf{y}_0 to G_2 otherwise.

Thus, the Fisher's LDA is actually a particular case of the Bayes rule with equal prior probabilities and equal costs of misclassification.

Fisher's LDA vs. Bayes Classification Rule

- Fisher's LDA does not need distributional assumption, while Bayes rule does need to specify $f_1(\mathbf{y})$ and $f_2(\mathbf{y})$.
- Bayes rule does not require equal covariance matrices, while Fisher's LDA does.
- Bayes can take account of the prior probability and the misclassification costs while Fisher's LDA cannot.
- As to the implementation, Bayes rule does not have a similar ready-to-use package as Fisher's LDA, since it requires the specific density function forms.
- When the normality is assumed with equal covariance matrix, Fisher's LDA corresponds to a special case of Bayes rule. But when the covariance matrices are not equal, the Fisher's LDA uses the pooled matrix, while Bayes rule yields a quadratic classification rule.

Bayes Classification Rule: Example

Example: Back to the “rainy” and “sunny” example, if we believe in this city, the chance to be rainy is twice of that to be sunny (which need not to be true - just for illustration purpose). The misclassification costs are reasonably assumed to be equal. Then we could modify the previous Fisher’s LDA:

```
> (lda2=lda(Group~Y1+Y2, prior=c(2/3,1/3))) # build
the LDA model with prior probability
Call:
lda(Group ~ Y1 + Y2, prior = c(2/3, 1/3))

Prior probabilities of groups:
      1      2
0.6666667 0.3333333

Group means:
      Y1  Y2
1  0.92 2.10
2 -0.38 8.85

Coefficients of linear discriminants:
      LD1
Y1 -0.1035305
Y2  0.2247957
```


The fitting is not as good as the Fisher's LDA, where equal prior probabilities are used:

```
> Z2<-predict(ld2) # predict the current sample
> pred.G2<-Z2$class
> result2<-cbind(Group,pred.G2,Z2$x)
> colnames(result2)<-c("TrueGroup", "Predicted", "TransformedData")
> result2
```

	TrueGroup	Predicted	TransformedData
1	1	1	-0.01142223
2	1	1	-0.12319761
3	1	1	-1.01624375
4	1	1	-0.88313979
5	1	1	-1.68324925
6	1	2	1.22342147
7	1	1	-2.23455074
8	1	1	-0.19533426
9	1	1	-0.79053783
10	1	1	0.20771836
11	2	1	0.44555080
12	2	2	0.76884438
13	2	2	2.31312863
14	2	1	0.65879549
15	2	1	-0.96505399
16	2	1	0.51538545
17	2	2	1.69879860
18	2	2	2.28652662
19	2	2	1.68072922
20	2	2	1.61036604

And the TPM is much higher than the Fisher's LDA:

```
> (tab2<-table(Group,pred.G2))
      pred.G2
Group 1 2
  1  9 1
  2  4 6
> (TPM2<-1-sum(diag(prop.table(tab2))))
[1] 0.25
> new<-data.frame(cbind(Y1=8.1,Y2=2))
> predict(ld2,new)
$class
[1] 1
Levels: 1 2

$posterior
      1      2
1 0.9652012 0.03479882

$x
      LD1
1 -1.316482
```

Thus caution needs when assuming the prior probabilities.

Outline

- 3 Classification: Allocation of Observations to Groups
 - Classification for Two Populations
 - Classification Based on Fisher's Linear Discriminant Analysis
 - Classification Based on Bayes Rule
 - Classification for Several Populations

Fisher's Rule for Several Populations

When $g > 2$ populations are involved, Fisher's allocation rule can be modified as follows:

Theorem (Fisher's allocation rule for g groups)

Allocate the new observation \mathbf{y}_0 to G_k using r discriminant functions, $r = 1, \dots, s$, $s \leq \min\{g - 1, p\}$ if

$$\sum_{j=1}^r \{\mathbf{e}_j'(\mathbf{y}_0 - \bar{\mathbf{y}}_k)\}^2 \leq \sum_{j=1}^r \{\mathbf{e}_j'(\mathbf{y}_0 - \bar{\mathbf{y}}_m)\}^2, \text{ for all } m \neq k,$$

where \mathbf{e}_j is the j th eigenvector of $\mathbf{W}^{-1}\mathbf{B}$, as defined in the Fisher's linear discriminants theorem.

Fisher's Rule for g Populations: Remarks

- As in the two-population case, when the covariance matrices are equal, the Fisher's allocation rule is equivalent to comparing the Mahalanobis distance between the new observation \mathbf{y}_0 and each sample mean $\bar{\mathbf{y}}_k$, $k = 1, \dots, g$, and assign it to the group with the smallest distance.
- The R implementation follows the same routine as that in the two-population case.

Bayes Rule for Several Populations

Incorporating misclassification costs for $g > 2$ groups is complicated because there are $g(g - 1)$ costs to consider. Hence we assume equal misclassification costs. Thus the Bayes rule extends to

Assign \mathbf{y}_0 to the group for which $p_k f_k(\mathbf{y}_0)$ is maximum,

where p_k and $f_k(\cdot)$ are the prior probability and density function defined parallel to the two-population case.

Remark: Still, when the populations are normal with equal covariance matrices and prior probabilities, Fisher's rule and Bayes rule coincide.

Classification for Several Populations: Example

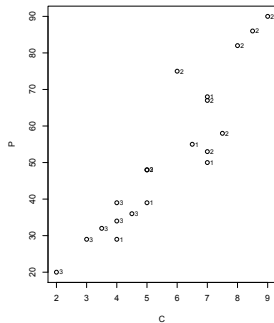
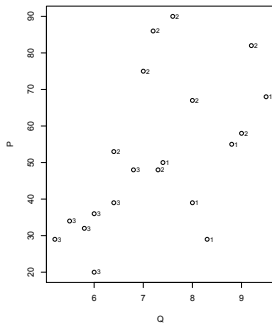
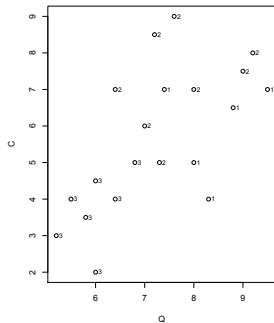
Example: The 20 brands of TV sets were investigated, and their sales status group (G): 1 for good market, 2 for average market, 3 for poor market; quality evaluation score (Q), condition evaluation score (C) and price in hundred RMB (P) are recorded. If a new brand is now promoted, with $Q = 8.0$, $C = 7.5$ and $P = 65$, how is its selling perspective?

The data are outputted as below:

```
> sale=read.table("/Users/jingyuan/Documents/Teaching/Multivariate
Analysis/R code/Chap5/sale.csv",sep=",", header=T) # read data
> attach(sale)
> sale
  G  Q  C  P
1 1 8.3 4.0 29
2 1 9.5 7.0 68
3 1 8.0 5.0 39
4 1 7.4 7.0 50
5 1 8.8 6.5 55
6 2 9.0 7.5 58
7 2 7.0 6.0 75
8 2 9.2 8.0 82
9 2 8.0 7.0 67
10 2 7.6 9.0 90
11 2 7.2 8.5 86
12 2 6.4 7.0 53
13 2 7.3 5.0 48
14 3 6.0 2.0 20
15 3 6.4 4.0 39
16 3 6.8 5.0 48
17 3 5.2 3.0 29
18 3 5.8 3.5 32
19 3 5.5 4.0 34
20 3 6.0 4.5 36
```


The pairwise scatterplots by sale status groups are obtained to illustrate the distribution of the original data:

```
> par(mfrow=c(1,3))  
> plot(Q,C);text(Q,C,G,adj=-0.8,cex=0.75)  
> plot(Q,P);text(Q,P,G,adj=-0.8,cex=0.75)  
> plot(C,P);text(C,P,G,adj=-0.8,cex=0.75)
```



The Fisher's LDA yields the two discriminant functions:

```
> library(MASS)
> (ld=lda(G~Q+C+P))
```

Call:

```
lda(G ~ Q + C + P)
```

Prior probabilities of groups:

	1	2	3
	0.25	0.40	0.35

Group means:

	Q	C	P
1	8.400000	5.900000	48.200
2	7.712500	7.250000	69.875
3	5.957143	3.714286	34.000

Coefficients of linear discriminants:

	LD1	LD2
Q	-0.81173396	0.88406311
C	-0.63090549	0.20134565
P	0.01579385	-0.08775636

Proportion of trace:

	LD1	LD2
	0.7403	0.2597

We could re-predict the current sample:

```
> Z<-predict(ld)
> pred.G<-Z$class
> cbind(G,pred.G,Z$x)
```

	G	pred.G	LD1	LD2
1	1	1	-0.1409984	2.582951755
2	1	1	-2.3918356	0.825366275
3	1	1	-0.3704452	1.641514840
4	1	1	-0.9714835	0.548448277
5	1	1	-1.7134891	1.246681993
6	2	1	-2.4593598	1.361571174
7	2	2	0.3789617	-2.200431689
8	2	2	-2.5581070	-0.467096091
9	2	2	-1.1900285	-0.412972027
10	2	2	-1.7638874	-2.382302324
11	2	2	-1.1869165	-2.485574940
12	2	2	-0.1123680	-0.598883922
13	2	3	0.3399132	0.232863397
14	3	3	2.8456561	0.936722573
15	3	3	1.5592346	0.025668216
16	3	3	0.7457802	-0.209168159
17	3	3	3.0062824	-0.358989534
18	3	3	2.2511708	0.008852067
19	3	3	2.2108260	-0.331206768
20	3	3	1.5210939	0.035984885

The total probability of misclassification can be computed:

```
> (tab<-table(G,pred.G))
      pred.G
G    1 2 3
  1 5 0 0
  2 1 6 1
  3 0 0 7
> diag(prop.table(tab,1))
      1      2      3
1.00 0.75 1.00
> (TPM<-1-sum(diag(prop.table(tab))))
[1] 0.1
```

And the new brand is allocated to $G = 2$ (average market):

```
> new<-data.frame(cbind(Q=8.0,C=7.5,P=65))
> predict(ld,new)
$class
[1] 2
Levels: 1 2 3

$posterior
      1      2      3
1 0.2114514 0.786773 0.001775594

$x
      LD1      LD2
1 -1.537069 -0.1367865
```

If Bayes rule is adopted, assuming the populations are normal and the misclassification costs are identical, we could assign the prior probability to each population.

```
> ld1=lda(G~Q+C+P,prior=c(1,1,1)/3) # Bayes rule with equal prior probability
> ld2=lda(G~Q+C+P,prior=c(5,8,7)/20) # Bayes rule with unequal prior probability
> Z1=predict(ld1) # re-predict the current sample
> Z2=predict(ld2)
> table(G,Z1$class)
```

```
G   1 2 3
  1 5 0 0
  2 1 6 1
  3 0 0 7
> table(G,Z2$class)
```

```
G   1 2 3
  1 5 0 0
  2 1 6 1
  3 0 0 7
> new<-data.frame(cbind(Q=8.0,C=7.5,P=65))
> predict(ld1,new)$class
[1] 2
Levels: 1 2 3
> predict(ld2,new)$class
[1] 2
Levels: 1 2 3
```

Summary and Take-home Messages

- What is discriminant analysis and what is classification?
- How to construct Fisher's discriminant function for two-population case? And what if more than two populations are involved?
- What is the Fisher's classification function when allocating a new observation to one of the two groups?
- What is the Bayes classification rule? What is the difference between Bayes and Fisher's rule?
- If more than two populations are considered, how to classify new observations?