Alternative Approach

In this segment, an alternative approach to discrimination will be considered.

Instead of using Fisher's discriminant function, one can design a procedure which minimizes the probability of misclassification.

It will turn out that under certain conditions, the method described herein is equivalent to Fisher's discriminant function.

To simplify calculations, we will consider only the case of two groups, say Π_1 and Π_2.

Focus on the 2-Group Case

When there are only two groups, say "+" and "–", the confusion matrix takes the following form:

	Classification	
Real Group	+	–
+	True + (TP)	False – (FN)
–	False + (FP)	True – (TN)

One can think of + and – representing the fact that an individual has COVID-19 or not. The result of a test can be positive or negative.

Terminology

Sensitivity

Probability that an individual in the $+$ group is classified as a $+$:

$$\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}.$$

Specificity

Probability that an individual in the $-$ group is classified as a $-$:

$$\frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}}.$$

The challenge is to try and have high values for both of these ratios.

Cost Considerations

A cost can sometimes be associated to a bad decision in either direction.

The following notation will be used to describe this cost:

	Classification		
Ground Truth	Π_1	Π_2	Prior Probability
Π_1	0	$C(2 \mid 1)$	q_1
Π_2	$C(1 \mid 2)$	0	q_2

For $i \in \{1, 2\}$,

$$q_i = \Pr(\mathbf{X} \in \Pi_i)$$

and

$$C(3 - i \mid i) = \text{cost associated to classifying an individual with characteristics } \mathbf{X} \text{ in the group } \Pi_{3-i} \text{ when } \mathbf{X} \in \Pi_i.$$

Problem Formulation

The group to which an individual is assigned is based on a decision rule. Ideally, one would like to find a set $R_1 \subset \mathbb{R}^p$ such that

$$\boxed{\mathbf{X} \in R_1 \quad \Leftrightarrow \quad \mathbf{X} \in \Pi_1.}$$

Alternatively, one could also define the set $R_2 = \mathbb{R}^p \setminus R_1$.

As this ideal cannot be reached, we determine the sets R_1 and R_2 by looking at the probabilities and the costs associated to a bad decision.

Probability Calculation (1–2)

For $i \in \{1, 2\}$, set

$$p_i(\mathbf{x}) = \text{density of characteristic } \mathbf{X} \text{ knowing that the individual comes from Population } \Pi_i.$$

For $i \in \{1, 2\}$, the probability of classifying someone from Population Π_i in Population Π_{3-i} is then given by

$$\begin{aligned} \Pr(3 - i \mid i) &\equiv \Pr(\text{classified in } \Pi_{3-i} \mid \text{comes from } \Pi_i) \\ &= \Pr(\mathbf{X} \in \mathbf{R}_{3-i} \mid \text{comes from } \Pi_i) \\ &= \int_{\mathbf{R}_{3-i}} p_i(\mathbf{x})\, d\mathbf{x}. \end{aligned}$$

Probability Calculation (2–2)

The probability of error depends on

- ✓ the probabilities of misclassification (both ways);
- ✓ the prior probability of observing someone from Population 1 or 2.

The probability of the event

"the individual comes from Population Π_i and
is classified in Population Π_{3-i}"

is given by

$$q_i \times \Pr(3 - i \mid i).$$

Minimization of the Expected Cost

The overall expected cost is given by

$$C(2 \mid 1)q_1 + \int_{R_1} \{C(1 \mid 2)q_2 p_2(\mathbf{x}) - C(2 \mid 1)q_1 p_1(\mathbf{x})\} d\mathbf{x}.$$

To minimize this quantity, it suffices to take

$$R_1 = \left\{ \mathbf{x} \in \mathbb{R}^p : \frac{p_1(\mathbf{x})}{p_2(\mathbf{x})} \geq \frac{C(1 \mid 2)q_2}{C(2 \mid 1)q_1} \equiv \Omega \right\}.$$

If you took MATH 324 or MATH 357, you will recognize that the proposed decision rule, viz.

$$\mathbf{x} \in R_1 \quad \Leftrightarrow \quad p_1(\mathbf{x})/p_2(\mathbf{x}) \geq \Omega,$$

is that which is suggested by the Neyman–Pearson Lemma.

Choice of Constants

By choosing q_1, q_2, $C(1 \mid 2)$ and $C(2 \mid 1)$ carefully, one can take into account prior knowledge about the problem at hand.

In the absence of prior information, it is generally assumed that

$$C(1 \mid 2) = C(2 \mid 1) \quad \text{and} \quad q_1 = q_2.$$

In this special case, the classification rule reduces to

$$\mathbf{x} \in \Pi_1 \quad \Leftrightarrow \quad \frac{p_1(\mathbf{x})}{p_2(\mathbf{x})} > 1 \quad \Leftrightarrow \quad \ln\left\{\frac{p_1(\mathbf{x})}{p_2(\mathbf{x})}\right\} > 0$$

Assumptions

Assume that no prior information is available.

Further assume that the set $\mathbf{X}$ of variables is multivariate Gaussian.

One can then minimize the expected cost under these conditions, viz.

$$\mathbf{p}_1 \sim \mathcal{N}_p\left(\mu_1, \mathbf{\Sigma}_1\right) \quad \text{and} \quad \mathbf{p}_2 \sim \mathcal{N}_p\left(\mu_2, \mathbf{\Sigma}_2\right).$$

Linear discriminant analysis assumes that $\mathbf{\Sigma}_2 = \mathbf{\Sigma}_1 = \mathbf{\Sigma}$. In contrast, quadratic discriminant analysis does not impose this constraint.

In what follows, we will concentrate on linear discriminant analysis.

Multivariate Gaussian Density

Under the previous assumptions, one has, for $i \in \{1, 2\}$,

$$p_i(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\mathbf{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x} - \mu_i)^\top \mathbf{\Sigma}^{-1}(\mathbf{x} - \mu_i)\right\}.$$

Therefore, $\ln\{p_1(\mathbf{x})/p_2(\mathbf{x})\}$ becomes

$$-\frac{1}{2}\left\{(\mathbf{x} - \mu_1)^\top \mathbf{\Sigma}^{-1}(\mathbf{x} - \mu_0) - (\mathbf{x} - \mu_2)^\top \mathbf{\Sigma}^{-1}(\mathbf{x} - \mu_2)\right\}.$$

Decision Rule

An individual is classified in Population Π_1 whenever

$$-\frac{1}{2}\left\{\mu_1^\top \mathbf{\Sigma}^{-1}\mu_1 - \mu_2^\top \mathbf{\Sigma}^{-1}\mu_2 - 2\mathbf{x}^\top \mathbf{\Sigma}^{-1}(\mu_1 - \mu_2)\right\} \geq 0.$$

Upon factorizing the above expression, one finds that an equivalent classification criterion is

$$\left(\mathbf{x} - \frac{\mu_1 + \mu_2}{2}\right)^\top \mathbf{\Sigma}^{-1}(\mu_1 - \mu_2) \geq 0.$$

Note that the last expression is linear in $\mathbf{x}$, whence the term linear discriminant analysis.

Equivalent Decision Rule

Equivalently, one sets $\mathbf{x} \in \Pi_1$ if and only if

$$(\mathbf{x} - \mu_1)\mathbf{\Sigma}^{-1}(\mathbf{x} - \mu_1) < (\mathbf{x} - \mu_2)\mathbf{\Sigma}^{-1}(\mathbf{x} - \mu_2),$$

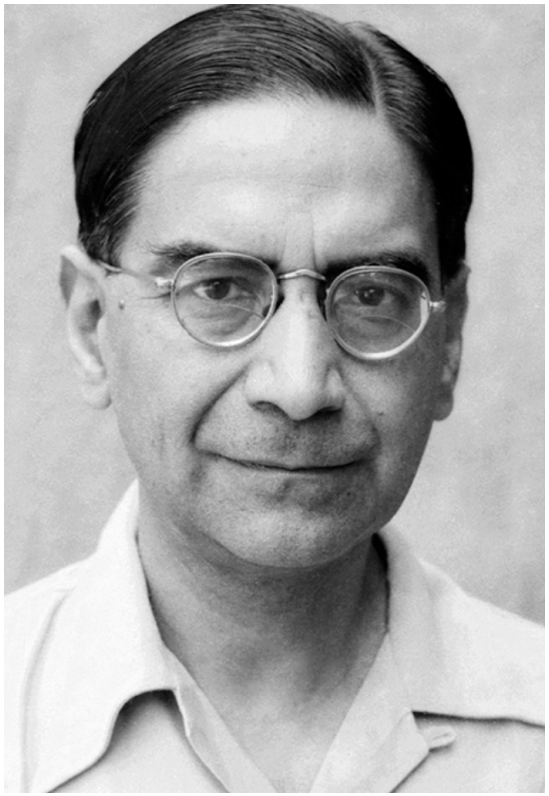
that is, whenever

$$D_1^2(\mathbf{x}, \mu_1) \leq D_2^2(\mathbf{x}, \mu_2),$$

where D_k^2 denotes the Mahalanobis distance associated to the kth population.

When $\mathbf{\Sigma}_1 \neq \mathbf{\Sigma}_2$, the above expression does not simplify and remains quadratic in $\mathbf{x}$, whence the term quadratic discriminant analysis.

P.C. Mahalanobis



প্রশান্তচন্দ্র মহলানবিশ

Prasanta Chandra Mahalanobis (1893–1972) is a famous Indian statistician.

He founded the renowned *Indian Statistical Institute* in 1931, as well as the prime Indian peer-reviewed statistics journal *Sankhyā*.

In 2006, Indian Prime Minister Manmohan Singh announced that Mahalanobis's birthdate, June 29, would be National Statistical Day in India.

"Fun fact": Mahalanobis was born June 29 and died June 28 (aged 78).

Discriminant Function

One can declare $\mathbf{x} \in \Pi_1$ if and only if

$$\mu_1^\top \mathbf{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}\mu_1^\top \mathbf{\Sigma}^{-1}\mu_1 \geq \mu_2^\top \mathbf{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}\mu_2^\top \mathbf{\Sigma}^{-1}\mu_2.$$

Because $\mu_1^\top \mathbf{\Sigma}^{-1}\mu_2 = \mu_2^\top \mathbf{\Sigma}^{-1}\mu_1$, the above expression is equivalent to

$$(\mu_1 - \mu_2)^\top \mathbf{\Sigma}^{-1}\mathbf{x} > (\mu_1 - \mu_2)^\top \mathbf{\Sigma}^{-1}\left(\frac{\mu_1 + \mu_2}{2}\right).$$

If μ_1 is estimated by $\tilde{\mathbf{x}}_1$, μ_2 is estimated by $\tilde{\mathbf{x}}_2$, and $\mathbf{\Sigma}$ is estimated by $\mathbf{S}$, one recovers Fisher's discriminant function.

Summary

Linear discriminant analysis is equivalent to Fisher's method under the following conditions:

- ✓ the vector $\mathbf{X}$ is multivariate Gaussian in both populations;
- ✓ the covariance matrix of $\mathbf{X}$ is the same in both populations;
- ✓ the prior probabilities of coming from populations 1 and 2 are equal.

Note that under these conditions, Fisher's discriminant function is not the only one which is optimal among linear functions.

However, it does minimize the global probability of misclassification.

Misclassification Probabilities

Under the same assumptions, one can compute the misclassification probabilities.

To this end, set

$$Y = \left(\mathbf{X} - \frac{\mu_1 + \mu_2}{2} \right)^\top \mathbf{\Sigma}^{-1}(\mu_1 - \mu_2).$$

Given that $\mathbf{X} \sim \mathcal{N}_p(\mu_i, \mathbf{\Sigma})$ for $i \in \{1, 2\}$, one has

$$Y \sim \mathcal{N}\left[(-1)^{i-1} \frac{\zeta^2}{2}, \zeta^2 \right],$$

where

$$\zeta^2 = (\mu_1 - \mu_2)^\top \mathbf{\Sigma}^{-1}(\mu_1 - \mu_2).$$

Proof

In the case $i = 1$, one has

$$Y = (\mu_1 - \mu_2)^\top \mathbf{\Sigma}^{-1} \left(\mathbf{X} - \frac{\mu_1 + \mu_2}{2} \right).$$

Therefore,

$$\mathrm{E}(Y) = (\mu_1 - \mu_2)^\top \mathbf{\Sigma}^{-1} \left(\mu_1 - \frac{\mu_1 + \mu_2}{2} \right) = \frac{\zeta^2}{2}.$$

Similarly, one has

$$\mathrm{var}(Y) = (\mu_1 - \mu_2)^\top \mathbf{\Sigma}^{-1} \mathbf{\Sigma} \mathbf{\Sigma}^{-1} (\mu_1 - \mu_2).$$

It follows that $\mathrm{var}(Y) = \zeta^2$. The case $i = 2$ can be handled analogously.

Special Case

In the special case where

$$q_1 = q_2 \quad \text{and} \quad C(1 \mid 2) = C(2 \mid 1),$$

one assigns $\mathbf{X}$ to Population Π_2 if and only if $Y < 0$.

Therefore,

$$P(2 \mid 1) = \Pr(Y < 0 \mid \mathbf{X} \in \Pi_1) = \Phi\left(\frac{-\zeta^2/2}{\zeta}\right) = \Phi\left(-\frac{\zeta}{2}\right)$$

and

$$P(1 \mid 2) = \Pr(Y > 0 \mid \mathbf{X} \in \Pi_2) = 1 - \Phi\left(\frac{\zeta^2/2}{\zeta}\right) = \Phi\left(-\frac{\zeta}{2}\right).$$

Generalization to the Case of q Groups

To determine to which group $\mathbf{X}$ belongs, one computes

$$D_k^2 = (\mathbf{X} - \mu_k)^\top \boldsymbol{\Sigma}^{-1} (\mathbf{X} - \mu_k),$$

namely the Mahalanobis distance between $\mathbf{X}$ and the mean μ_k of the kth population.

Observation $\mathbf{X}$ is assigned to Population Π_{k_0} whenever

$$D_{k_0}^2 = \min \left(D_1^2, \ldots, D_q^2 \right),$$

which is equivalent to maximize

$$\Pr(\mathbf{X} \in \Pi_{k_0}) = \frac{e^{-D_{k_0}^2/2}}{e^{-D_1^2/2} + \cdots + e^{-D_q^2/2}}.$$

Unbiased Estimation of μ_i and Σ

As usual, we estimate μ_i by the mean of $\mathbf{X}$ in Population Π_i. One could also estimate $\mathbf{\Sigma}$ by

$$\hat{\Sigma} = \mathbf{S}/n,$$

but this estimation is biased because the observations do not all come from the same population. An unbiased estimation of $\mathbf{\Sigma}$ is given by

$$\hat{\Sigma} = \mathbf{S}_{\text{pool}} = \frac{1}{n-q}\mathbf{W}$$

where

$$\mathbf{W} = \sum_{k=1}^{q} \sum_{i \in I_k} \left(\mathbf{X}_i - \bar{\mathbf{X}}_k\right) \left(\mathbf{X}_i - \bar{\mathbf{X}}_k\right).^{\top}$$

Exercise

You observe 34 individuals from Population Π_1 and 66 individuals from Population Π_2. Suppose

$$\bar{\mathbf{X}}_1^\top = (8, 45), \quad \bar{\mathbf{X}}_2^\top = (6, 20), \quad \hat{\Sigma} = \begin{pmatrix} 1 & 3 \\ 3 & 19 \end{pmatrix}.$$

Q1 Find Fisher's discriminant function.

Q2 In which group would one classify an observation whose $\mathbf{X}$ vector is $(7, 30)$?

Q3 What is the probability of classifying an observation in the wrong group?

Q4 What is the Mahalanobis distance between observation $(7, 30)$ and each of the two groups?