Discriminant Analysis

In the previous segment, we started investing Fisher's approach to discrimination.

The starting point was an $n \times p$ data matrix $\mathbf{X} = (X_{ij})$, where for each $i \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, p\}$,

$$X_{ij} = \text{value of the } j\text{th variable for the } i\text{th individual.}$$

For each group $k \in \{1, \ldots, q\}$, we also have at our disposal

- ✓ I_k, the set of individuals in group k,
- ✓ $n_k = |I_k|$, the size (or cardinality) of I_k,

so that $n_1 + \cdots + n_q = n$.

Fisher's Approach (1–2)

The purpose is to find a function $f : \mathbb{R}^p \to \mathbb{R}$ which can be used to compute a score

$$f(X_1, \ldots, X_p) \in \mathbb{R},$$

for each observation $(X_1, \ldots, X_p)$. This score can then be used to determine the groups via a partition of $\mathbb{R}$.

Fisher's discriminant function is defined, for each $\mathbf{X} \in \mathbb{R}^p$, by

$$f(\mathbf{X}) = \mathbf{a}^\top (\mathbf{X} - \bar{\mathbf{X}}),$$

where $\bar{\mathbf{X}} = (\bar{X}_1, \ldots, \bar{X}_q)$ is the vector of variable means and $\mathbf{a}$ is a normed eigenvector corresponding to the largest eigenvalue of $\mathbf{S}^{-1}\mathbf{B}$.

Fisher's Approach (2–2)

Here, $\mathbf{S} = (s_{jj'})$ is a $p \times p$ matrix with entries

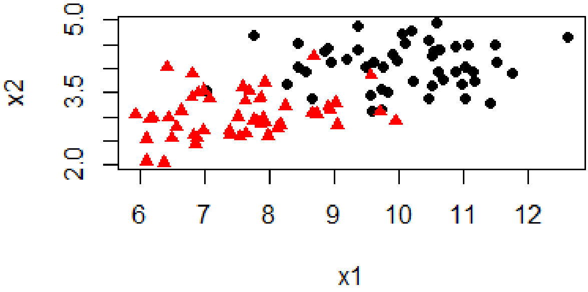
$$s_{jj'} = \sum_{i=1}^{n} \left(X_{ij} - \bar{X}_j\right) \left(X_{ij'} - \bar{X}_{j'}\right),$$

which gives the total sums of squares, and $\mathbf{B} = (b_{jj'})$ has entries

$$b_{jj'} = \sum_{k=1}^{q} n_k (\bar{X}_{kj} - \bar{X}_j)(\bar{X}_{kj'} - \bar{X}_{j'}),$$

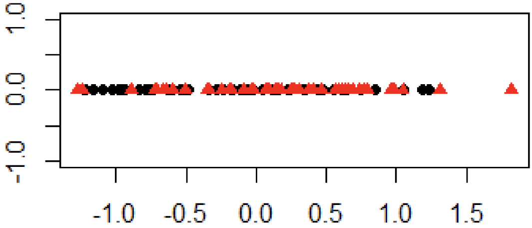
i.e., the sums of squares between groups. The scores $Y_i = \mathbf{a}^\top (\mathbf{X}_i - \bar{\mathbf{X}})$ then maximize the ratio "between variance / within variance."

Small Example (1–3)



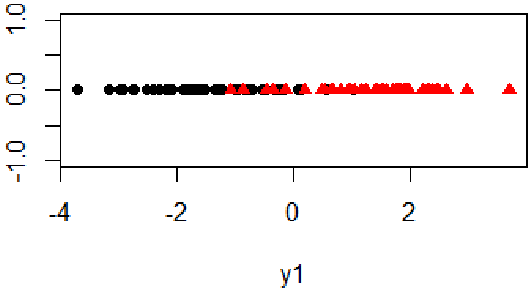
Original Data

Small Example (2–3)



Poor linear combination of X_1 and X_2

Small Example (3–3)



Optimal linear combination of X_1 and X_2

Discriminating Power (1–2)

The matrix $\mathbf{S}^{-1/2}\mathbf{B}\mathbf{S}^{-1/2}$ is symmetric and positive definite.

Therefore, its eigenvalues are all real and positive.

Moreover, one has $\mathbf{S}^{-1}\mathbf{B}\mathbf{a} = \lambda\mathbf{a}$ by definition of $\mathbf{a}$.

Consequently,

$$\mathbf{B}\mathbf{a} = \lambda\mathbf{S}\mathbf{a} \quad \Rightarrow \quad \mathbf{a}^\top\mathbf{B}\mathbf{a} = \lambda\mathbf{a}^\top\mathbf{S}\mathbf{a} \quad \Rightarrow \quad \lambda = \frac{\mathbf{a}^\top\mathbf{B}\mathbf{a}}{\mathbf{a}^\top\mathbf{S}\mathbf{a}},$$

and hence $\lambda \in [0, 1]$.

Discriminating Power (2–2)

The eigenvalue λ measures the discriminating power of f.

Limiting case 1:

$$\lambda = 1 \quad \Rightarrow \quad \mathbf{a}^\top \mathbf{B}\mathbf{a} = \mathbf{a}^\top \mathbf{S}\mathbf{a},$$

i.e., 100% of the variability is between the groups, 0% within the groups.

Limiting case 2:

$$\lambda = 0 \quad \Rightarrow \quad \mathbf{a}^\top \mathbf{B}\mathbf{a} = 0,$$

i.e., 100% of the variability is within the groups, 0% between the groups.

Discriminating Function and Classification

Once the discriminating function f has been defined, one can compute the average score of each group, given by

$$m_k = \mathbf{a}^\top \left(\bar{X}_{k1}, \ldots, \bar{X}_{kp}\right)^\top,$$

where $\bar{X}_{kj}$ denotes the mean of the jth variable taken over the individuals belonging to the kth group.

To classify a new observation $\mathbf{X}_0 \in \mathbb{R}^p$, one then proceeds as follows:

1. Compute the score $f(\mathbf{X}_0) = \mathbf{a}^\top \mathbf{X}_0$.
2. Assign $\mathbf{X}_0$ to the group k_0 such that

$$|\mathbf{a}^\top \mathbf{X}_0 - m_{k_0}| = \min_{k \in \{1,\ldots,q\}} |\mathbf{a}^\top \mathbf{X}_0 - m_k|.$$

Classification Errors

In applying the rule to the sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$ from which it was constructed, one can estimate the misclassification error rate with the *confusion matrix*:

	Classification			
Real Group	Group 1	Group 2	$\cdots$	Group q
Group 1	p_{11}	p_{12}	$\cdots$	p_{1q}
Group 2	p_{21}	p_{22}	$\cdots$	p_{2q}
$\vdots$	$\vdots$	$\vdots$	$\ldots$	$\vdots$
Group q	p_{q1}	p_{q2}	$\cdots$	p_{qq}

Case of Two Groups

When there are only two groups, the eigenvector **a** defining the discriminant function is given by

$$\mathbf{a} = \mathbf{S}^{-1}\mathbf{C} = \sqrt{n_1 n_2/n}\, \mathbf{S}^{-1}(\tilde{\mathbf{x}}_1 - \tilde{\mathbf{x}}_2),$$

where $\tilde{\mathbf{x}}_1$ and $\tilde{\mathbf{x}}_2$ denoting the $p \times 1$ means of the two groups.

To prove this claim, first use the fact that $\mathbf{B} = \mathbf{C}\mathbf{C}^\top$ (as shown at the end of this set) and check that $\mathbf{S}^{-1}\mathbf{B}\mathbf{a} = \xi\mathbf{a}$ for some $\xi > 0$. Indeed, upon substitution, one finds

$$\begin{aligned} \mathbf{S}^{-1}\mathbf{B}\mathbf{a} = \mathbf{S}^{-1}\mathbf{C}\mathbf{C}^\top\mathbf{a} = (\mathbf{S}^{-1}\mathbf{C}\mathbf{C}^\top)\mathbf{S}^{-1}\mathbf{C} \\ = \mathbf{S}^{-1}\mathbf{C}(\mathbf{C}^\top\mathbf{S}^{-1}\mathbf{C}) = \xi\mathbf{S}^{-1}\mathbf{C} = \xi\mathbf{a}, \end{aligned}$$

with $\xi = \mathbf{C}^\top\mathbf{S}^{-1}\mathbf{C}$. Again, see the end of this set of slides for more detail.

Decision Rule

Suppose that

$$m_1 = \mathbf{a}^\top \tilde{\mathbf{x}}_1 > \mathbf{a}^\top \tilde{\mathbf{x}}_2 = m_2.$$

An observation is then classified in the first group if

$$\mathbf{a}^\top \mathbf{x} > \bar{m} = (m_1 + m_2)/2 = \mathbf{a}^\top (\tilde{\mathbf{x}}_1 + \tilde{\mathbf{x}}_2)/2.$$

This happens if and only if

$$(\tilde{\mathbf{x}}_1 - \tilde{\mathbf{x}}_2)^\top \mathbf{S}^{-1} \mathbf{x} > (\tilde{\mathbf{x}}_1 - \tilde{\mathbf{x}}_2)^\top \mathbf{S}^{-1} (\tilde{\mathbf{x}}_1 + \tilde{\mathbf{x}}_2)/2.$$

Note: As the factor $\sqrt{n_1 n_2 / n}$ appears on both sides of the inequality, it need not be mentioned.

Example (1–5)

In-depth psychiatric exams were carried out on 49 elderly men. Based on the results, each one of them was classified as

- ✓ in good mental health (Group I) or
- ✓ senile (Group II).

The same subjects took four simple tests that are cheap and quick:

Test	Group I ($n_1 = 37$)	Group II ($n_2 = 12$)
Arithmetic	11.49	8.50
Drawings	7.97	4.75
Information	12.57	8.75
Similitudes	9.57	5.33

Estimation of Σ

In this study, it was found that

$$\frac{\mathbf{S}}{n} = \begin{pmatrix} 11.2553 & 9.4042 & 7.1489 & 3.3830 \\ & 13.5318 & 7.3830 & 2.5532 \\ & & 11.5744 & 2.6170 \\ & & & 5.8085 \end{pmatrix}.$$

Estimation of Σ^{-1}

Using the R command `solve()`, one gets

$$n\mathbf{S}^{-1} = \begin{pmatrix} .25907 & -0.13577 & -0.05878 & -0.064730 \\ & 0.18645 & -0.03833 & 0.01438 \\ & & 0.15098 & -0.01694 \\ & & & 0.21117 \end{pmatrix}.$$

Example (3–5)

A simple calculation yields

$$\mathbf{C}^* = \tilde{\mathbf{x}}_1 - \tilde{\mathbf{x}}_2 = (3.82, 4.24, 2.99, 3.22)^\top.$$

Accordingly,

$$\mathbf{C} = \sqrt{\frac{37 \times 12}{49}}\, \mathbf{C}^*.$$

By definition, $\mathbf{a} = \mathbf{S}^{-1}\mathbf{C}$, but one can also use

$$\mathbf{a} = n\,\mathbf{S}^{-1}\mathbf{C}^*,$$

given that the scores are always used for comparisons only. For this reason, the discriminant rule is only defined up to a linear transformation.

Computation of m_1 and m_2

In view of the previous computations, one has

$$m_1 = \mathbf{a}^\top \tilde{\mathbf{x}}_1 = 5.97 \quad \text{and} \quad m_2 = \mathbf{a}^\top \tilde{\mathbf{x}}_2 = 3.54.$$

Therefore, an individual can be declared senile on the basis of the four cheap tests whenever

$$\mathbf{a}^\top \mathbf{x} < \mathbf{a}^\top \left(\frac{m_1 + m_2}{2} \right) = 4.755.$$

Example (5–5)

Summary

	Clinical Diagnosis "OK"	Clinical Diagnosis "Senile"	Total
Classified as "OK"	29	**4**	33
Classified "Senile"	**8**	8	16
Total	37	12	49

Error rates

Global Rate	$12/49 \approx 24.5\%$
Rate Among the "OK"	$8/37 \approx 21.6\%$
Rate Among the "Seniles"	$4/12 \approx 33.3\%$

Additional Mathematical Details

We saw that the overall variability can be decomposed as follows:

$$\begin{aligned} s_{jj'} &= \sum_{k=1}^{q} \sum_{i \in I_k} (X_{ij} - \bar{X}_{kj})(X_{ij'} - \bar{X}_{kj'}) + \sum_{k=1}^{q} n_k (\bar{X}_{kj} - \bar{X}_j)(\bar{X}_{kj'} - \bar{X}_{j'}) \\ &\equiv w_{jj'} + b_{jj'}. \end{aligned}$$

When there are only two groups, one has simply

$$b_{jj'} = n_1(\bar{X}_{1j} - \bar{X}_j)(\bar{X}_{1j'} - \bar{X}_{j'}) + n_2(\bar{X}_{2j} - \bar{X}_j)(\bar{X}_{2j'} - \bar{X}_{j'}).$$

Moreover, for each $j \in \{1, \ldots, q\}$, one can compute $\bar{X}_j$ as follows:

$$\bar{X}_j = n_1 \bar{X}_{1j}/n + n_2 \bar{X}_{2j}/n.$$

Reduced Form of the Matrix $\mathbf{B}$

One then gets

$$b_{jj'} = \frac{n_1 n_2}{n} (\bar{X}_{1j} - \bar{X}_{2j})(\bar{X}_{1j'} - \bar{X}_{2j'}).$$

It follows that $\mathbf{B} = \mathbf{C}\mathbf{C}^\top$, where

$$\begin{aligned} \mathbf{C} &= \sqrt{\frac{n_1 n_2}{n}} (\bar{X}_{11} - \bar{X}_{21}, \ldots, \bar{X}_{1p} - \bar{X}_{2p})^\top \\ &\equiv \sqrt{\frac{n_1 n_2}{n}} (\tilde{\mathbf{x}}_1 - \tilde{\mathbf{x}}_2). \end{aligned}$$

Rank of the Matrix $\mathbf{B}$

Given that $\mathbf{C}$ is a $p \times 1$ vector, one has the following relationship:

$$\text{rank}(\mathbf{B}) = \text{rank}(\mathbf{S}^{-1}\mathbf{B}) = 1.$$

Indeed, the rank of a matrix remains unchanged if it is multiplied by an invertible matrix (which is thus of full rank). The constant ξ is given by

$$\begin{aligned}\xi = \text{tr}(\mathbf{S}^{-1}\mathbf{B}) = \text{tr}(\mathbf{S}^{-1}\mathbf{C}\mathbf{C}^\top) \\ = \text{tr}(\mathbf{C}^\top\mathbf{S}^{-1}\mathbf{C}) = \mathbf{C}^\top\mathbf{S}^{-1}\mathbf{C} \\ = \frac{n_1 n_2}{n}(\tilde{\mathbf{x}}_1 - \tilde{\mathbf{x}}_2)^\top\mathbf{S}^{-1}(\tilde{\mathbf{x}}_1 - \tilde{\mathbf{x}}_2).\end{aligned}$$

Insight

When there are only two classification groups, discriminant analysis is really just multiple regression, with a few tweaks.

The dependent variable is a dichotomous, categorical variable (i.e., a categorical variable that can take on only two values).

The dependent variable is expressed as a dummy variable (having values of 0 or 1).

Observations are assigned to groups, based on whether the predicted score is closer to 0 or to 1.

The regression equation is called the discriminant function.