Photo: Vladimir Batagelj, ***UNI-LJ***

Discriminant Analysis

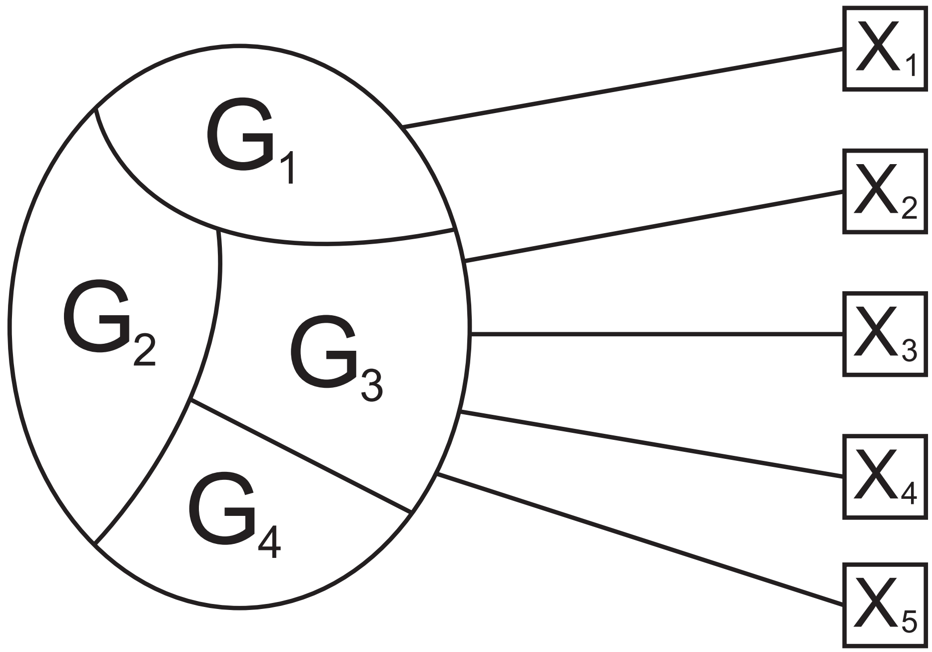
Anuška Ferligoj

UL, Ljubljana, Slovenia

NRU HSE, Moscow, Russia

Content

Introduction

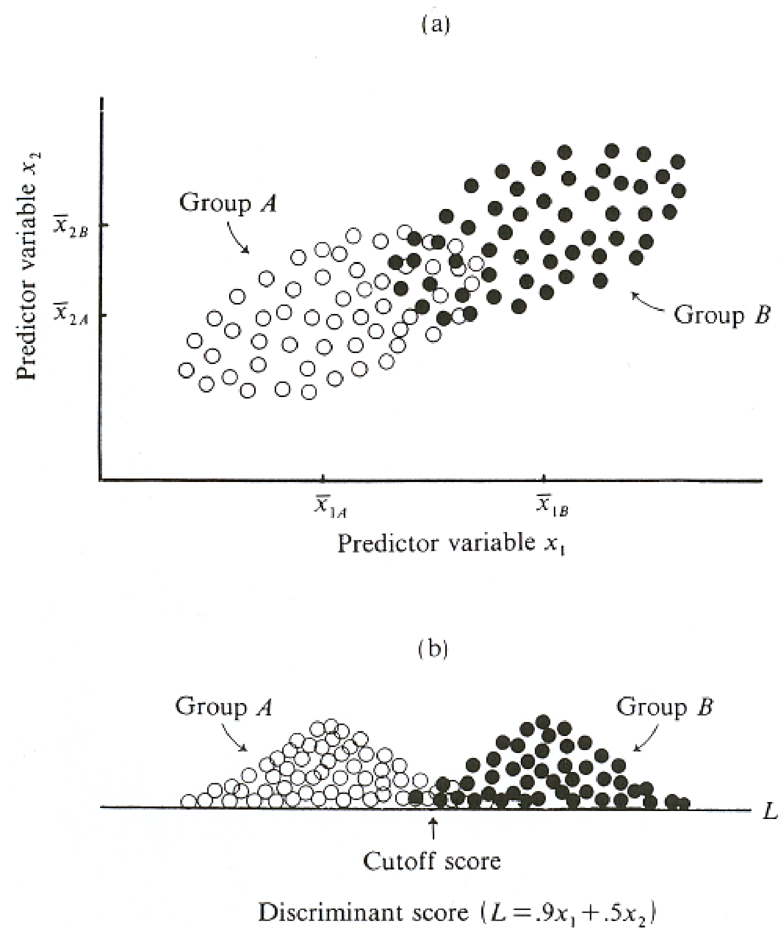


Discriminant analysis involves deriving linear combination of the measured variables that discriminate between the **a priori** defined groups in such a way that the misclassification error rates are minimized.

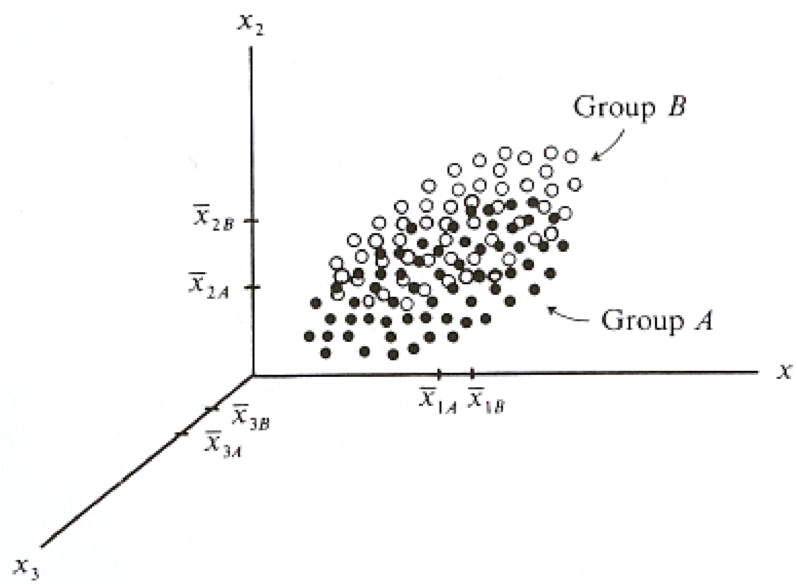
Assumptions

1. $k \geq 2$.
2. At least 2 units in each group.
3. $p < n - 2$; p is the number of variables and n the number of all units.
4. No variable is a linear combination of the other variables (multicolinearity).
5. The variables must have a multivariate normal distribution in each group when using the statistical tests.
6. The $p \times p$ variance-covariance matrix of the measured variables in each of the two groups must be the same.

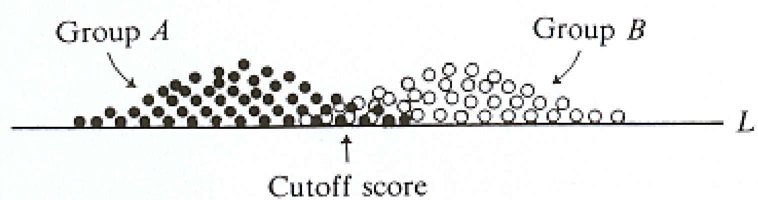
The two-group discriminant problem



(a)



(b)



Discriminant score ($L = .4x_1 + .6x_2 - .5x_3$)

groups	means	variance-covariance m.
G_1	μ_1	Σ_1
G_2	μ_2	Σ_2

The assumption: $\Sigma_1 = \Sigma_2 = \Sigma$

Fisher (1936) suggested finding a linear combination of p variables X_i

$$Y = b_0 + b_1 X_1 + b_2 X_2 + ... + b_p X_p = Xb$$

so that the ratio of the difference in the means of the linear combinations in G_1 and G_2 to its within-group variance is maximized.

The means of the linear combinations in G_1 and G_2 are:

$$\bar{Y}_1 = b'\mu_1$$

$$\bar{Y}_2 = b'\mu_2$$

The variance is

$$varY_1 = varY_2 = b'\Sigma b$$

The ratio to be maximized is

$$\frac{\bar{Y}_1 - \bar{Y}_2}{varY_1} = \frac{b'\mu_1 - b'\mu_2}{b'\Sigma b} = max$$

From this optimization criterion the discriminat loadings b can be derived. They are proportional to

$$\Sigma^{-1}(\mu_1 - \mu_2)$$

Sample-based estimates

Usually the parameters are estimated from a samples from each population G_i.

μ_i can be estimated by:

$$\bar{x}_i' = (\bar{x}_{i1}, \bar{x}_{i2}, ..., \bar{x}_{ip})$$

and Σ by pooled sample variance-covariance matrix

$$S = \frac{1}{n_1 + n_2 - 2}(X_1'X_1 + X_2'X_2)$$

where n_1 is the number of units in the sample from G_1 and n_2 is the number of units in the sample from G_2.

The estimated discriminat loadings are

$$\hat{b} = S^{-1}(\bar{x}_1 - \bar{x}_2)$$

Group centroid

The mean value of the discriminant function for the units of a group is commonly referred to as the **group centroid**.

The centroid of the group i is

$$\bar{Y}_i = b'\bar{x}_i$$

Classification rules

With the obtained (linear) discriminant variable $Y = Xb$ each (new) unit can be assigned to one of the two groups. The unit i is assigned to group G_1 if

$$y_i - \bar{Y}_1 \leq y_i - \bar{Y}_2$$

or to G_2 if

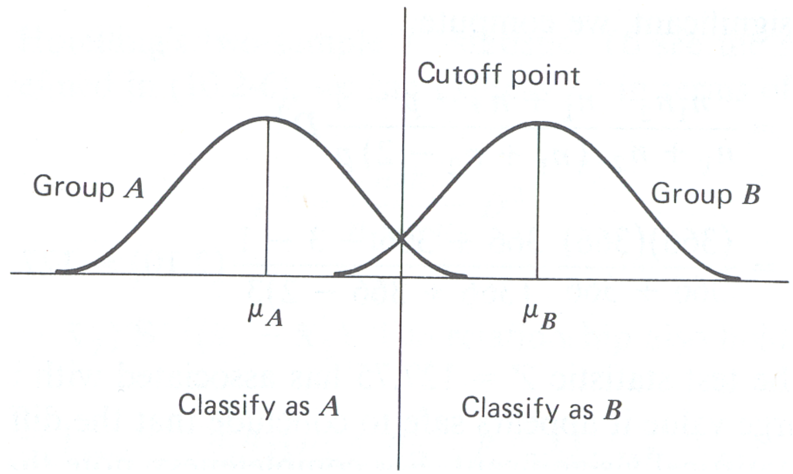
$$y_i - \bar{Y}_1 > y_i - \bar{Y}_2$$

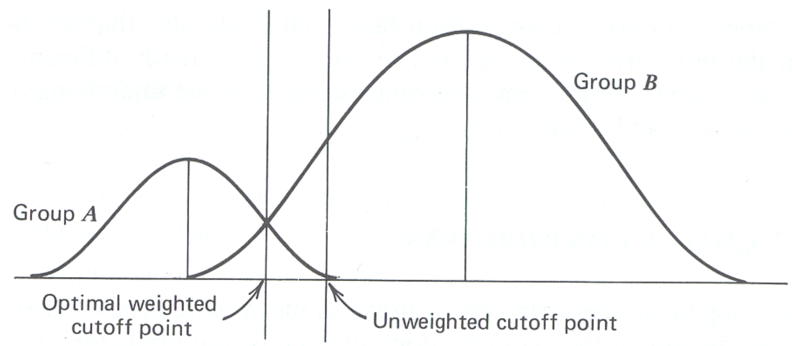
An equivalent classification rule uses **the midpoint of separation** (cutoff point). For equal sample sizes ($n_1 = n_2$) it is

$$Y_c = \frac{\bar{Y}_1 + \bar{Y}_2}{2}$$

For unequal sample sizes the point of separation is

$$Y_c = \frac{n_2 \bar{Y}_1 + n_1 \bar{Y}_2}{n_1 + n_2}$$


Cutoff point
Group A
Group B
μ_A
μ_B
Classify as A
Classify as B


Group B
Group A
Optimal weighted cutoff point
Unweighted cutoff point

Classification table

The performance of a discriminant function can be evaluated by calculating the misclassification rate. Let us apply the obtained discriminant function to the data from which it was derived. Each unit is assigned to one of the groups according to the classification rule. The following table can be produced:

	Number	Predicted Group	Membership
Actual Group	of Cases	G_1	G_2
G_1	n_1	a	b
G_2	n_2	c	d

The rate of correct classifications is

$$\frac{a+d}{n_1+n_2}$$

With equal sample size and two groups, the expacted chance accuaracy of a rule is 50%.

The estimated nonerror rates (correct classifications) are optimistically biased, since we utilize the same set of data to construct the rule and to evaluate the performance.

Example: Small companies

Let us consider the data of small companies in Slovenia. The groups are defined as follows:

- G_1 – service companies ($n_1 = 70$)
- G_2 – manufactoring companies ($n_2 = 75$)

The variables are 12 factors of business success.

Discriminant loadings

	loadings
PROD-MET	**-.54**
MARK-MET	**.40**
PRODUCT	-.00
RELATION	.01
SKIL-EMP	.22
SKIL-MAN	**.51**
FAMILY	-.33
ECON-ASO	-.18
POL-CON	**.48**
LOC-AUT	-.28
STATE	.16
COMPANY	.06

Centroids

group	centroid
service	.54
manufactoring	-.50

Classification table

Actual Group	Number of Cases	Predicted Group service	Membership manufact.
service	70	70%	30%
manufact.	75	30.7%	69.3%

The percetage of correct classifications is 70%.

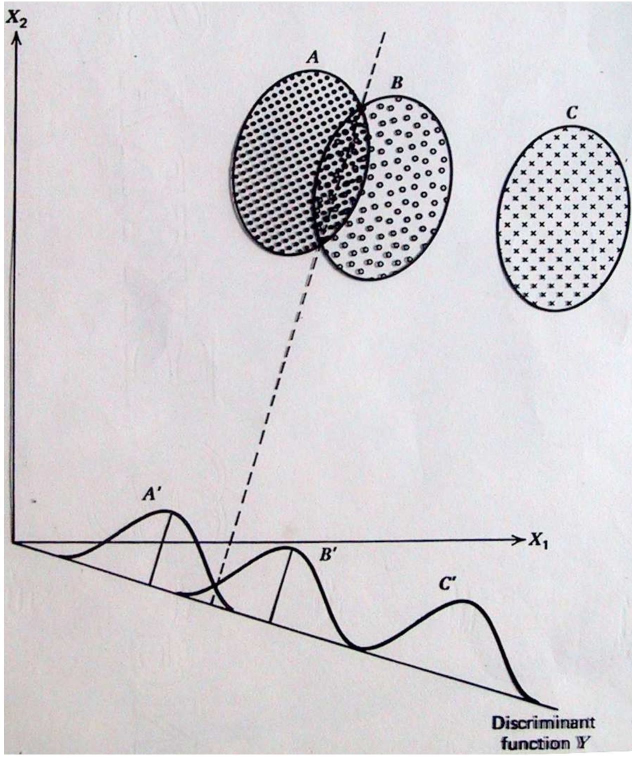
Discussion

The owners of the service sector companies and the owners of the crafts companies are the most distingushed by the following factors for the business success:

- improvement of products,
- skilled managers,
- political connections, and
- improvement of marketing methods.

The service companies owners believe more than crafts companies owners that improvement of products is less important, but more important are skilled managers, good political connections, and improvement of marketing methods.

The k-group discriminant problem



In the case of more than two groups more than one discriminant variable may be needed to characterize effectively the differences between some of the groups. The maximum number of the discriminant variables is $min(k-1, p)$, where k is the number of groups and p the number of variables.

The approach

Let us assume that we have k groups and in each n_1, n_2, ... n_k, units.

Let us denote by T the matrix of the total mean corrected sums-of-squares and cross-products for all variables on all units $n = \sum_{i=1}^{k} n_i$.

The matrix of sums-of-squares and cross-products for the ith group let be denoted by W_i.

The within-groups sums of squares and cross-products are given by

$$W = W_1 + W_2 + ... + W_k$$

The matrix of between-groups sum-of-squares and cross-products can thus be found by the difference

$$B = T - W$$

The criterion that has to be maximazed is analogous to Fisher‘s criterion

$$\frac{\text{variability between-groups}}{\text{variability within-groups groups}} = \max$$

The variance of a discriminat variable $Y = Xb$ is

$$varY = b'\Sigma b$$

The variability between-groups is then

$$varY = b'Bb$$

and the variability within-groups is then

$$varY = b'Wb$$

Therefore, the discriminant criterion that has to be maximazed is

$$\frac{b'Bb}{b'Wb} = \lambda = max$$

The best solution is obtained by calculating the eigenvalues and eigenvectors of the matrix $W^{-1}B$. The eigenvalues are λ_i . There are $r = min(k-1, p)$ obtained solutions. The largest λ and the corresponding eigenvector, whose elements are the discriminant loadings, define the first discriminant variable. The relative value of the eigenvalue λ_i gives an index of the importance of each discriminant variable:

$$\frac{\lambda_i}{\sum_{j=1}^{r} \lambda_j}$$

Relationship between discriminat analysis and canonical correlation analysis

In the case of the discriminant analysis we have k groups and p variables. The eigenvalues and eigenvectors of the matrix $W^{-1}B$ are calculated to estimate the discriminant variables. Let us denote the obtained eigenvalues by λ_j^{da}.

From the nominal variable that defines the k groups let us form $k-1$ dummy variables. With this we obtained the first set of $k-1$ variables. On the other hand we have p measured variables. We can perform canonical correlation analysis. $k-1$ eigenvalues of the matrix $\Sigma_{XX}^{-1}\Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX}$ can be denoted by λ_j^{kka}. Than it holds

$$\lambda_j^{da} = \frac{\lambda_j^{kka}}{1-\lambda_j^{kka}}$$