Discriminating between two groups using eigenvectors

Anton Buhagiar
Department of Mathematics.

Introduction

Consider g populations or groups, $g \geq 2$. The object of discriminant analysis is to allocate an individual to one of these g groups on the basis of his/her measurements on the p variables $x_1, x_2, \ldots, x_p$. It is desirable to make as few 'mistakes' as possible in classifying these individuals to the various groups.

For example, the populations might consist of different diseases and the p variables $x_1, x_2, \ldots, x_p$ might measure the symptoms of a patient, eg. blood pressure, body temperature, etc. Thus one is trying to diagnose a patient's disease on the basis of his/her symptoms. As another example, one can consider samples from three species of iris. The object is then to allocate a new iris to one of these species on the basis of its measurements eg. sepal length, sepal width, etc.

In the case of two groups, $g = 2$, in the univariate case, when $p = 1$ and x_1 is the only variable measured, it is quite easy to see when the two groups are well separated from each other. For this purpose, one can perform a t-test on $x_1$to see whether the two groups have significantly different means. Equivalently, one can define the ratio:

$$\frac{\text{the difference } \textit{between} \text{ the means of the two samples}}{\text{deviations } \textit{within} \text{ the samples}}.$$

A large value for this ratio, which is proportional to the $t-$statistic, would indicate that the means of the samples are well separated from each other;

conversely, a small value for this ratio would imply that within sample variations are relatively large, and that readings from the two samples would tend to overlap. This would in turn lead to poor discrimination between the two groups in terms of x_1, and to a non-significant difference between the sample means for x_1.

In the case when $g \geq 2$, that is for two or more groups, and when $p = 1$, one-way analysis of variance, the F-test, can be performed to examine whether the mean of x_1 differs significantly over the groups. Equivalently, one can define the ratio:

$$\frac{\text{Variation } \textit{between} \text{ the means of the samples}}{\text{Variation } \textit{within} \text{ the samples}}. \qquad (1.0.0.1)$$

Again in this case, a large value for this ratio, which is closely related to the F-statistic, signifies good separation between the groups and a significant difference for $x_1$between the groups. In fact, in the case of two groups ($g = 2$), the F-test and the t-test are equivalent to each other, with $F = t^2$ for a given problem.

In the case when the number of variables is larger than one, p >1, one can perform separate univariate tests on each of the p variables $x_1, x_2, \ldots, x_p$. For purposes of discrimination, however, it is often preferable to define a linear combination y of the x_k's, namely $y = \sum_{k=1}^{p} a_k x_k$, with the object of maximising the ratio defined in equation (1). Finding the best linear combination which maximizes this ratio is equivalent to maximizing the statistical distance between the groups. This in turn would guarantee greater success in discriminating between the different groups. As shown below, the problem of finding the optimum choice of the coefficients a_i can be reduced to a suitable eigenvalue problem.

Partitioning the total variation of y

We will now discuss briefly a very important identity in the context of discrimination and analysis of variance. We will assume that there are g different groups in all, and that there are n_i cases in the i'th group, where $i = 1, 2, \ldots, g$. For each case, the p variables $x_1, x_2, \ldots, x_p$ are measured. We then denote by x_{ijk} the value of the k^{th} variable (x_k) for the j'th case in the

i'th group. Note here that the suffixes have the following bounds:

$$x_{ijk}: \quad 1 \leqslant k \leqslant p, 1 \leqslant j \leqslant n_i, 1 \leqslant i \leqslant g.$$

value of variable x_k for case j in sample i.

It is then easy to write down the mean of the k'th variable over the i'th sample, and the grand mean of the k'th variable over all groups:

$$x_{i.k} = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ijk} \quad ; \quad x_{..k} = \frac{1}{\sum\limits_{i=1}^{g} n_i}\sum_{i=1}^{g}\sum_{j=1}^{n_i} x_{ijk} \quad .$$

In an analogous fashion, the linear combination y for the i'th case in the j'th sample can be written as $y_{ij} = \sum\limits_{k=1}^{p} a_k x_{ijk}$. Its mean over the i'th sample and its overall (grand) mean are then given as:

$$y_{i.} = \frac{1}{n_i}\sum_{j=1}^{n_i}\sum_{k=1}^{p} a_k x_{ijk} = \sum_{k=1}^{p} a_k x_{i.k} \quad ; \quad y_{..} = \frac{1}{\sum\limits_{i=1}^{g} n_i}\sum_{i=1}^{g}\sum_{j=1}^{n_i}\sum_{k=1}^{p} a_k x_{ijk} = \sum_{k=1}^{p} a_k x_{..k} \quad .$$

The sum of the square of the deviations of the values of y_{ij} for each case from their overall (grand) mean $y_{..}$ is then given by

$$\sum_{i=1}^{g}\sum_{j=1}^{n_i}(y_{ij} - y_{..})^2 \text{ or } \sum_{i=1}^{g}\sum_{j=1}^{n_i}\left[\sum_{k=1}^{p} a_k(x_{ijk} - x_{..k})\right]^2 .$$

This quantity is often referred to as the *total variation* of y, or equivalently as the *total sum of squares*, often abbreviated as *SST*. Algebraic manipulation of the SST will result in a very important partitioning of this variation into two separate parts as follows:

SST $\equiv$ Total sum of squares

$$= \sum_{i=1}^{g}\sum_{j=1}^{n_i}\left[\sum_{k=1}^{p} a_k(x_{ijk} - x_{..k})\right]^2$$

interchange order of summation:

$$= \sum_{1\leqslant l,m\leqslant p}^{a_l a_m} \sum_{i=1}^{g}\sum_{j=1}^{n_i}(x_{ijl} - x_{..l})(x_{ijm} - x_{..m})$$

add and subtract mean of sample from which reading is taken, leaving sum unchanged:

$$= \sum_{1\leqslant l,m\leqslant p}^{a_l a_m} \sum_{i=1}^{g} \sum_{j=1}^{n_i} (x_{ijl} - x_{i.l} + x_{i.l} - x_{..l})(x_{ijm} - x_{i.m} + x_{i.m} - x_{..m})$$

multiply out the terms in pairs:

$$= \sum_{1\leqslant l,m\leqslant p}^{a_l a_m} \sum_{i=1}^{g} \sum_{j=1}^{n_i} (x_{ijl} - x_{i.l})(x_{ijm} - x_{i.m}) + \sum_{1\leqslant l,m\leqslant p}^{a_l a_m} \sum_{i=1}^{g} \sum_{j=1}^{n_i} (x_{i.l} - x_{..l})(x_{i.m} - x_{..m})$$

+ the other two cross terms which each equal zero using the definition of the sample means

simplify second term since brackets are independent of suffix j:

$$= \sum_{1\leqslant l,m\leqslant p}^{a_l a_m} \sum_{i=1}^{g} \sum_{j=1}^{n_i} (x_{ijl} - x_{i.l})(x_{ijm} - x_{i.m}) + \sum_{1\leqslant l,m\leqslant p}^{a_l a_m} \sum_{i=1}^{g} n_i (x_{i.l} - x_{..l})(x_{i.m} - x_{..m})$$

$$\equiv SSW \ + SSB$$

The first term in the penultimate line, often abbreviated as *SSW*, estimates the size of deviations of the readings from *their own* sample mean, and is often called the *within-variation* or *within sum of squares.* The second term, often abbreviated as *SSB*, estimates the size of the deviations of the sample means from the overall mean and is referred to as the *between-variation*, or the *between sums of squares.* The above identity can be therefore written as

$$\text{SST} \ = \text{SSW} \ + \ \text{SSB} \qquad (1.0.0.2)$$

or *total variation = variation within samples + variation between samples*

This important identity is often referred to as *partitioning the sums of squares.* It is important to note that the terms *SSB* and *SSW* are, respectively, the numerator and denominator in the ratio defined by equation (1). The groups are more easily separated if the ratio in equation (1), $\frac{SSB}{SSW}$, is large or equivalently $\frac{SST}{SSB}$ is small. Statistical tests have been devised using these ratios to determine whether the sample means are significantly different from each other.

Matrix formulation

The sums of squares, SST, SSW and SSB are all quadratic forms in the coefficients a_k

and can be elegantly represented in matrix form. Rearranging the p coefficients a_k as the$p \times 1$ column vector a, one can rewrite the partitioning identity (2) as

$$a^tTa \quad = \quad a^tWa \quad + \quad a^tBa$$

$$\text{SST} = \text{SSW} + \text{SSB}$$

where T, W and B are symmetric $p \times p$ matrices, the l, m'th entry of which are given by the terms multiplying $a_l a_m$ in the corresponding sum of squares. Thus,

the l, m'th entry of T is $\sum\limits_{i=1}^{g}\sum\limits_{j=1}^{n_i}(x_{ijl} - x_{..l})(x_{ijm} - x_{..m})$;

the l, m'th entry of W is $\sum\limits_{i=1}^{g}\sum\limits_{j=1}^{n_i}(x_{ijl} - x_{i.l})(x_{ijm} - x_{i.m})$; (3)

the l, m'th entry of B is $\sum\limits_{i=1}^{g} n_i(x_{i.l} - x_{..l})(x_{i.m} - x_{..m})$.

The matrices T, W and B are called *sums of squares and cross-product matrices.* Since the partitioning holds for any arbitrary vector a, these three matrices satisfy the identity

$$T = W + B \tag{1.0.0.4}$$

In fact, B is usually calculated from $B = T - W$ in practise.

Maximising the ratio of between to within variation

For optimum separation of the groups, we would therefore seek to maximise the ratio $\frac{SSB}{SSW}$. In matrix form, we would like to find a suitable column vector a with entries $a_1, a_2, \ldots, a_p$, such that $\frac{a^tBa}{a^tWa}$ is a maximum. Equivalently, since multiplying a by a scalar would not change the ratio, we can maximise the numerator, subject to the constraint that the denominator is one. Using

Lagrangian multipliers, we maximise the function $\varphi(a)$defined by

$$\varphi(a) = \quad a^t Ba + \lambda(1 - a^t Wa).$$

This function $\varphi(a)$ can then be differentiated with respect to each of the a_k's, $k = 1, 2, \ldots, p$, and the derivatives $\frac{\partial\varphi(a)}{\partial a_k}$ are each set to zero. When the resulting set of p equations are rearranged in matrix form, one obtains the homogeneous linear system

$$\frac{\partial\varphi(a)}{\partial a} \quad = \quad 2Ba - 2\lambda Wa \quad = \quad 0,$$

where 0 is the $p \times 1$ column vector of zeros. Dividing by 2 and factorising, we then obtain the condition:

$$(B - \lambda W)a = 0, \text{ or equivalently} \qquad (1.0.0.5)$$

$$(W^{-1}B - \lambda I)a = 0. \qquad (1.0.0.6)$$

Therefore a is an eigenvector of $W^{-1}B$ and λ is its corresponding eigenvalue. Further, pre-multiplying equation (5) by a^t, we get

$$\begin{aligned} a^t(B - \lambda W)a &= 0, \text{ that is} \\ a^t Ba &= \lambda a^t Wa \text{ or} \\ \lambda &= \frac{a^t Ba}{a^t Wa}. \end{aligned} \qquad (1.0.0.7)$$

From equations 6 and 7, one can therefore conclude that the maximum possible value of the ratio $\frac{a^t Ba}{a^t Wa}$ $(\equiv \frac{SSB}{SSW})$ is the largest eigenvalue λ of $W^{-1}B$ and the optimum choice of a is the eigenvector of λ. The linear combination $y = \sum_{k=1}^{p} a_k x_k$ can be written in matrix form as $a^t x$. For this particular choice of the vector a, this linear combination is the one which best separates the groups. It is called *Fisher's discriminant function* (Fisher, 1936) after its inventor.

An example on discrimination between two groups

To illustrate the above, we now give an example of discrimination between two groups ($g = 2$) on the basis of two variables ($p = 2$). The following 'botanical' example is inspired by Fisher's classic paper on discrimination (Fisher, 1936), which is described in Mardia *et al.* (1979), whilst the numerical data are derived from Tacq (1997).

The datafile in our example contains measurements on two types of iris. The variables Y, X_1 and X_2 are defined as follows:

$$Y = \text{ type of iris } = \begin{cases} 0, \text{ if iris is of the } \textit{setosa} \text{ type (group 1);} \\ 1, \text{ if iris is of the } \textit{versicolor} \text{ type (group 2).} \end{cases}$$

$$X_1 = \text{ sepal length and } X_2 = \text{ sepal width.}$$

X_1 and X_2 are assumed to be normally distributed with similar covariance structure in the two groups (Tacq, 1997).

The data-file contains 15 cases in all, 6 in the first group (*setosa*), and 9 in the second group (*versicolor*). For each individual case (flower), we give its group membership (Y), its sepal length and sepal width (X_1 and X_2). The data-file is listed in **Table I**.

Table 1: The datafile and its statistical description.

Calculation of the matrices W, T and B, using equations (3) and equation (4).

	Y	X_1	X_2
Setosa	0	1	1
	0	1	4
	0	2	1
	0	4	5
	0	5	5
	0	5	9
Versicolor	1	4	2
	1	4	4
	1	5	6
	1	6	3
	1	6	6
	1	7	6
	1	8	7
	1	9	7
	1	9	8

Group 1: *Setosa*

Cases in sample: $n_1 = 6$

Mean: $\overline{X_1} = 3.000$ $\overline{X_2} = 4.166$

Variation: $\sum(X_1 - \overline{X_1})^2 = 18.000$ $\sum(X_2 - \overline{X_2})^2 = 44.833$

Covariation: $\sum(X_1 - \overline{X_1})(X_2 - \overline{X_2}) = 22$

$\therefore W_1 = \begin{pmatrix} 18 & 22 \\ 22 & 44.833 \end{pmatrix}$; *within variation in group* 1.

Group 2: *Versicolor*

Cases in sample: $n_2 = 9$

Mean: $\overline{X_1} = 6.444$ $\overline{X_2} = 5.444$

Variation: $\sum(X_1 - \overline{X_1})^2 = 30.222$ $\sum(X_2 - \overline{X_2})^2 = 32.222$

Covariation: $\sum(X_1 - \overline{X_1})(X_2 - \overline{X_2}) = 25.222$

$\therefore W_2 = \begin{pmatrix} 30.222 & 25.222 \\ 25.222 & 32.222 \end{pmatrix}$; *within variation in group* 2.

$\therefore W = W_1 + W_2 = \begin{pmatrix} 48.222 & 47.222 \\ 47.222 & 77.056 \end{pmatrix}$. $\therefore W^{-1} = \begin{pmatrix} 0.052 & -0.032 \\ -0.032 & 0.032 \end{pmatrix}$.

variation within samples; *inverse of W.*

Groups 1 and 2 together:

Total number of cases: $n = n_1 + n_2 = 15$

Overall mean: $\overline{X_1} = 5.067$ $\overline{X_2} = 4.933$

Total Variation: $\sum(X_1 - \overline{X_1})^2 = 90.933$ $\sum(X_2 - \overline{X_2})^2 = 82.933$

Total Covariation: $\sum(X_1 - \overline{X_1})(X_2 - \overline{X_2}) = 63.067$

$\therefore T = \begin{pmatrix} 90.933 & 63.067 \\ 63.067 & 82.933 \end{pmatrix}$. $\therefore B = T - W = \begin{pmatrix} 42.711 & 15.844 \\ 15.844 & 5.858 \end{pmatrix}$.

total variation *variation between samples*

In **Table I**, we also give the statistical description of each group separately, and of both groups pooled together. In particular, we give the means of the two variables X_1 and X_2, namely $\overline{X}_1$ and $\overline{X}_2$, for each sample separately, and the within sums of squares for each sample, W_1 and W_2, from which the within sum of squares matrix W for both groups could be simply calculated using $W = W_1 + W_2$. The groups are then pooled together, to obtain the grand means of X_1 and X_2, and hence the total sum of squares matrix T. The between sum of squares matrix B is then calculated as $B = T - W$. The reader is referred to **Table I** for the calculation of the 2x2 matrices W, B, T and W^{-1}.

One can then calculate $W^{-1}B$ as follows:

$$W^{-1}B = \begin{pmatrix} 0.052 & -0.032 \\ -0.032 & 0.032 \end{pmatrix} \begin{pmatrix} 42.711 & 15.844 \\ 15.844 & 5.858 \end{pmatrix} = \begin{pmatrix} 1.711 & 0.635 \\ -0.843 & -0.313 \end{pmatrix}.$$

This matrix has non-zero eigenvalue λ =1.399, with unit eigenvector $a = \begin{pmatrix} 0.897 \\ -0.442 \end{pmatrix}$.

Fisher's discriminant function is therefore given by $a^t x = 0.897X_1 - 0.442X_2$. This is the linear combination which gives the largest value (=λ) of the ratio $\frac{SSB}{SSW}$ in equation (2), namely, the ratio of the variation *between* samples to the variation *within* samples.

Test of significance on the eigenvalue.

One normally performs Hotelling's T^2 test to see whether the mean of the discriminant function $a^t x$ differs significantly between the two groups.

The T^2 statistic is defined as

$$T^2 = (n-2)\lambda,$$

where $n = n_1 + n_2$ is the total number of cases in the two samples.

T^2 should be 'large' if the means of the two groups are well separated.

Conversely, T^2 is 'small' if there is no significant difference between the two sample means. In this case, Hotelling showed that the quantity $\frac{(n-p-1)}{p(n-2)}T^2$ should be distributed according to the F-distribution with p, $n-p-1$ degrees of freedom, where p is the number of variables featuring in the discriminant function and n is the total number of cases in the two groups.

In this application, $p = 2$, $n = n_1 + n_2 = 6 + 9 = 15$,

$$T^2 = (n-2)\lambda = (15-2)(1.399) = 18.182,$$

$$F = \frac{(n-p-1)}{p(n-2)}T^2 = \frac{(15-2-1)}{2(15-2)}(18.182) = 8.392.$$

Degrees of freedom for F-test $= p,\ n-p-1 = 2,\ 15-2-1$
$= 2,\ 12$.

In our case therefore, if there is no significant difference between the groups, the F-statistic should be distributed according to the F-distribution with 2, 12 degrees of freedom.

From the tables, the critical F-value for 2, 12 degrees of freedom with α= 0.05 is 3.89. Since $8.392 > 3.89$, we can conclude that the means of the two groups are significantly different. For this reason, discriminant analysis could be done profitably on this dataset.

A typical statistical package would also include the following items in the output of a discriminant analysis:

1. a classification rule to determine the group to which a given case is assigned;

2. application of this classification rule to the cases whose group membership is known *a priori*, so as to obtain an estimate of the misclassification rate;

3. application of this classification to classify cases of unknown type.

We now describe briefly the classification rule and its application.

The Classification Rule

The discriminant function is often used to establish a classification rule whereby group membership of a given case can be determined. This could be done both for cases whose group membership is known *a priori*, and also for cases with unknown group membership.

One classification rule can be set up in the following way.

The value of the discriminant function $a^t x$ is first calculated at the centroid of each group:

Group 1: 0.897(3.000)-0.442(4.167) = 0.849,
Group 2: 0.897(6.444)-0.442(5.444) = 3.374.
The cut-off is then taken to be the average of these two values:

$$t_c = \frac{0.849 + 3.374}{2} = 2.112.$$

Then any case $(X_1, X_2)$is assigned to Group 1 if $0.897X_1 - 0.442X_2 < 2.112$, and to Group 2 otherwise.

Using this rule, one can classify the original cases to find how good the discriminant analysis is. Prior group membership could be compared to the posterior grouping predicted by the classification rule. This comparison is summarized in a classification table. One can also use this rule to classify new cases for which group membership is not known. The use of the classification rule is illustrated in **Table II**.

Table II: Use of the classification rule:

i) to classify original cases and hence
ii) to obtain a prior versus post classification table; and
iii) to classify new cases with unknown group membership.

i) Classification of original cases:

	Y	X_1	X_2	$0.897X_1 - 0.442X_2$	Posterior Classification
Setosa	0	1	1	0.455	0
	0	1	4	-0.871	0
	0	2	1	1.352	0
	0	4	5	1.378	0
	0	5	5	2.275	1
	0	5	9	0.507	0
Versi-	1	4	2	2.704	1
Color	1	4	4	1.820	0
	1	5	6	1.833	0
	1	6	3	4.056	1
	1	6	6	2.730	1
	1	7	6	3.627	1
	1	8	7	4.082	1
	1	9	7	4.979	1
	1	9	8	4.537	1

ii) Classification Table:

		Posterior Classification:	
		Group 1	Group 2
Prior Classification :	Group 1	5	1
	Group 2	2	7

80% of the cases are classified correctly.

iii) Classification of new cases with unknown group membership:

Y	X_1	X_2	$0.897X_1 - 0.442X_2$	Posterior Classification
?	6	5	3.172	1
?	5	6	1.833	0
?	3	7	-0.403	0
?	4	3	2.262	1
?	6	4	3.614	1

Conclusion and suggestions for further reading

Discriminant analysis is a very popular multivariate technique. Like many other techniques in multivariate statistics, the method is based on the algebraic eigenvalue problem. In this respect it is very similar to principal component analysis, factor analysis, correspondence analysis and multivariate analysis of variance (Manova), in all of which one has to find the eigenvalues and eigenvectors of a suitable matrix (Lebart, Morineau and Warwick, 1984). The eigenvalue problem defined by equation (5) is also important in the solution of vibrational problems of mechanics (Lunn, 1990 and Segerlind, 1983) and in the buckling of structures (Dawe, 1983).

Discriminant analysis is also related to linear regression and logistic regression, where group membership, y, is regressed on the measured variables x_i, (Flury and Riedwyl, 1993).

Most books on multivariate statistics have a chapter on discriminant analysis. The books by Tacq (1997), Manly (1986), and Flury and Riedwyl (1993) are very readable and should be reasonably easy to an undergraduate in mathematics or statistics.

For students who wish to read further on discriminant analysis, one can suggest more mathematical texts such as Morrison (1990), Everett and Dunn (1991), Johnson and Wichern (1992) and Mardia, Kent and Bibby (1979). In addition to the statistical theory, these books also give many practical examples of this technique.

References

- Dawe D.J., *Matrix and Finite Element Displacement Analysis of Structures*, Clarendon Press, Oxford, 1984.

- Everett B.S. and Dunn G., *Applied Multivariate Data Analysis*, Edward Arnold, London, 1991.

- Fisher R.A., *The use of multiple measurements in taxonomic problems*, Annals of Eugenics, 7, 179-188, 1936.

- Flury B. and Riedwyl H., *Multivariate Statistics: A Practical Approach*, Chapman and Hall, London, 1988.