# Predictive Discriminant Analysis

**Hervé Abdi***

Predictive discriminant analysis (DA, also called linear discriminant analysis, LDA) predicts group membership of observations that are described by several quantitative variables and when the group membership of (at least some) the observations is known *a priori*. The variables describing the observations are also called *predictors* or independent variables. DA is closely related to analysis of variance (ANOVA) and multivariate analysis of variance (MANOVA whose defining equations are formally equivalent to DA). Specifically, ANOVA and MANOVA use a qualitative independent variable (i.e., group membership) to predict one or more quantitative variables, whereas DA uses one or more quantitative variables (i.e., the predictors) to predict a qualitative dependent variable (group membership).

## THE MAIN IDEA

With $J$ *a priori* groups, DA linearly combines the predictors to create a set of $(J-1)$ new orthogonal variables called *discriminant variables* that maximally separate the groups. These discriminant variables best separate the *a priori* groups because an ANOVA performed with the first of these variables will give the largest possible $F$, whereas the second discriminant variable will give the second largest $F$ (while being orthogonal to the first variable) and so on, up to the last (i.e., $J-1$) discriminant variable.

## NOTATIONS

Essential notations are briefly defined here (but see the entry *matrix algebra* for details, examples, and explanations). Scalars are denoted by italic letters, with lower case letter referring to an item and with upper case letters being used to denote the cardinal of a set (e.g., $I$ is the number of observations and $i$ refers to a specific observation). Vectors are denoted by lower case bold letter (e.g., $\mathbf{a}$) and are by default column vectors. Matrices are de-

*In Bruce Frey (Ed.), *The SAGE Encyclopedia of Research Design.*
Thousand Oaks, CA: Sage. 2022.
Address correspondence to: Hervé Abdi
Program in Cognition and Neurosciences, MS: Gr.4.1,
The University of Texas at Dallas,
Richardson, TX 75080–3021 USA
***E-mail:*** herve@utdallas.edu https://personal.utdallas.edu/~herve

noted by upper case bold letter (e.g., $\mathbf{A}$) and in some cases with their number of rows and columns indicated as subscripts (e.g., $\underset{K \times J}{\mathbf{X}}$). The transpose operation is indicated by the superscript T (e.g. $\mathbf{X}^T$ is the transpose of $\mathbf{X}$). The identity matrix is denoted $\mathbf{I}$. The inverse of a matrix is indicated by the superscript $-1$ (e.g. $\mathbf{X}^{-1}$ is the inverse of $\mathbf{X}$).

### GENERALIZED SINGULAR VALUE DECOMPOSITION

The generalized singular value-decomposition (GSVD) applies to an I by J rectangular matrix of rank L denoted $\mathbf{X}$. The GSVD decomposes $\mathbf{X}$ under the constraints expressed by two symmetric positive definite matrices (called constraint matrices or sometimes metric matrices) $\mathbf{M}$ (of dimensions I by I) and $\mathbf{W}$ (of dimensions J by J) into three matrices such that

$$\mathbf{X} = \widetilde{\mathbf{U}}\widetilde{\mathbf{\Delta}}\widetilde{\mathbf{V}}^T \quad \text{with } \widetilde{\mathbf{V}}^T\mathbf{M}\widetilde{\mathbf{V}} = \widetilde{\mathbf{U}}^T\mathbf{W}\widetilde{\mathbf{U}} = \underset{L \times L}{\mathbf{I}} \tag{1}$$

with $\widetilde{\mathbf{\Delta}}$ being a diagonal matrix of positive numbers called the *singular values* of $\mathbf{X}$ and with $\widetilde{\mathbf{U}}$ and $\widetilde{\mathbf{V}}$ being called the left and right singular vectors of $\mathbf{X}$.

## *ANOVA FOR DA*

Recall that ANOVA evaluates if the means of a set of groups significantly differ from each other. To make explicit the relationship between LDA and ANOVA, we show below how ANOVA can be written with notations similar to DA. Take, for example, two groups of five participants each, with the following scores:

$$\text{Groupe 1: } \begin{bmatrix}1 & 2 & 5 & 6 & 6\end{bmatrix}^T \text{ et Groupe 2: } \begin{bmatrix}8 & 8 & 9 & 11 & 14\end{bmatrix}^T . \tag{2}$$

First, organize these scores into one vector denoted $\mathbf{y}$:

$$\mathbf{y} = \begin{bmatrix}1 & 2 & 5 & 6 & 6 & 8 & 8 & 9 & 11 & 14\end{bmatrix}^T . \tag{3}$$

The group membership of the observations is stored in a group matrix $\mathbf{X}$. A group matrix has as many rows as there are observations and as many columns as there are groups. In this matrix (containing only 0s and 1s), the element $x_{ij}$ is equal to 1 if the ith observation belongs to Group j and 0 if not. Here we have:

$$\mathbf{X} = \begin{bmatrix}1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1\end{bmatrix}^T . \tag{4}$$

The number of groups is denoted J (ici $J = 2$), the number of observations *per* group is denoted N (here $N = 5$) and the total number of observations is denoted I (here $I = 10 = 2 \times 5$).

The first step of the analysis computes the mean of all observations called the grand mean (also called the grand barycenter) as

$$G = \frac{1}{I} \underset{1 \times I}{\mathbf{1}}^{T} \mathbf{y} = 7 \,. \tag{5}$$

Next, we compute the vector of group means (also called group barycenters) denoted $\mathbf{g}$ and computed as:

$$\begin{aligned} \mathbf{g} &= \left(\mathbf{X}^{T}\mathbf{X}\right)^{-1}\mathbf{X}^{T}\mathbf{y} = \left(\frac{1}{N}\underset{J \times J}{\mathbf{I}}\right)\mathbf{X}^{T}\mathbf{y} = \frac{1}{N}\mathbf{X}^{T}\mathbf{y} \\ &= \begin{bmatrix} 5 & 0 \\ 0 & 5 \end{bmatrix}^{-1} \times \begin{bmatrix} 20 \\ 50 \end{bmatrix} = \begin{bmatrix} \frac{1}{5} & 0 \\ 0 & \frac{1}{5} \end{bmatrix} \times \begin{bmatrix} 20 \\ 50 \end{bmatrix} = \begin{bmatrix} 4 \\ 10 \end{bmatrix} \,. \end{aligned} \tag{6}$$

We then compute three I-dimensional vectors that store the distances (also called deviations) at the core of the ANOVA procedure. The first vector—called the total distance vector—stores the distance from each observation to the grand mean. Denoted $\mathbf{t}$ (for "*t*otal"), it is computed as:

$$\mathbf{t} = \mathbf{y} - \underset{I \times 1}{\mathbf{1}} \times G = \begin{bmatrix} 1 \\ 2 \\ 5 \\ 6 \\ 6 \\ 8 \\ 8 \\ 9 \\ 11 \\ 14 \end{bmatrix} - \begin{bmatrix} 7 \\ 7 \\ 7 \\ 7 \\ 7 \\ 7 \\ 7 \\ 7 \\ 7 \\ 7 \end{bmatrix} = \begin{bmatrix} -6 \\ -5 \\ -2 \\ -1 \\ -1 \\ 1 \\ 1 \\ 2 \\ 4 \\ 7 \end{bmatrix} \,. \tag{7}$$

The second vector—called the within distance vector— stores the distance from each observation to the mean of its group. Denoted $\boldsymbol{w}$ (for "*w*ithin" groups), this vector is computed as:

$$\boldsymbol{w} = \mathbf{y} - \mathbf{X}\mathbf{g} = \begin{bmatrix} 1 \\ 2 \\ 5 \\ 6 \\ 6 \\ 8 \\ 8 \\ 9 \\ 11 \\ 14 \end{bmatrix} - \begin{bmatrix} 4 \\ 4 \\ 4 \\ 4 \\ 4 \\ 10 \\ 10 \\ 10 \\ 10 \\ 10 \end{bmatrix} = \begin{bmatrix} -3 \\ -2 \\ 1 \\ 2 \\ 2 \\ -2 \\ -2 \\ -1 \\ 1 \\ 4 \end{bmatrix} \,. \tag{8}$$

The third vector—called the between distance vector— stores for each observation the distance from the mean of its group to the grand mean

Denoted $\mathbf{b}$ (for "*b*etween" groups), this vector is computed as:

$$\mathbf{b} = \mathbf{X}\mathbf{g} - \underset{I \times 1}{\mathbf{1}} \times G = \begin{bmatrix} 4 \\ 4 \\ 4 \\ 4 \\ 4 \\ 10 \\ 10 \\ 10 \\ 10 \\ 10 \end{bmatrix} - \begin{bmatrix} 7 \\ 7 \\ 7 \\ 7 \\ 7 \\ 7 \\ 7 \\ 7 \\ 7 \\ 7 \end{bmatrix} = \begin{bmatrix} -3 \\ -3 \\ -3 \\ -3 \\ -3 \\ 3 \\ 3 \\ 3 \\ 3 \\ 3 \end{bmatrix} . \tag{9}$$

These three vectors define the fundamental equation of the ANOVA which indicates that the total distance vector is the sum of the between distance vector and the within distance vector:

$$\mathbf{t} = \mathbf{b} + \boldsymbol{w} \tag{10}$$

(this equation is obtained by replacing $\mathbf{b}$ and $\boldsymbol{w}$ by their definition from Equations 9 and 8).

These three distance vectors are then used to compute the corresponding sum of squares. The distances within a group express the experimental error (i.e., what is *not* predicted by the model), when squared and summed, these distances give the sum of squares within groups (denoted $SS_{\text{within}}$) computed as

$$SS_{\text{within}} = \boldsymbol{w}^{\mathsf{T}}\boldsymbol{w} = 48 . \tag{11}$$

The distances between groups express the experimental effect (i.e., what is *predicted* by the model), when squared and summed these distances give the sum of squares between groups (denoted $SS_{\text{between}}$) computed as

$$SS_{\text{between}} = \mathbf{b}^{\mathsf{T}}\mathbf{b} = 90 . \tag{12}$$

A simple algebraic manipulation shows that the vectors $\mathbf{b}$ et $\boldsymbol{w}$ are othogonal to each other (*i.e.*, $\boldsymbol{w}^{\mathsf{T}}\mathbf{b} = \mathbf{b}^{\mathsf{T}}\boldsymbol{w} = 0$) and therefore that the total sum of squares is equal to the sum of squares within *plus* the sum of squares between:

$$\mathbf{t}^{\mathsf{T}}\mathbf{t} = \boldsymbol{w}^{\mathsf{T}}\boldsymbol{w} + \mathbf{b}^{\mathsf{T}}\mathbf{b} = 90 + 48 = 138 . \tag{13}$$

To be comparable in magnitude, these sums of squares are transformed into mean squares by dividing each sum of squares by its number of degrees of freedom. These degrees of freedom correspond to the number of data points that can be chosen taken into account the constraints in the data. So, for example, the sum of squares between is obtained from J deviations (so here two) but, because, the grand mean is known before computing the deviations. when the first $(J - 1)$ have been chosen, the value

of the Jth deviation is also automatically known. Therefore, the sum of square between has only $(J-1)$ degrees of freedom and so, here, we have $df_{\text{between}} = J-1 = 1$. A similar argument shows that the number of degrees of freedom for $SS_{\text{within}}$ is equal to $df_{\text{within}} = I - J = 8$. This way, we obtain the following mean squares

$$MS_{\text{within}} = \frac{SS_{\text{within}}}{df_{\text{within}}} = \frac{48}{8} = 6 \quad \text{et} \quad MS_{\text{between}} = \frac{SS_{\text{between}}}{df_{\text{between}}} = \frac{90}{1} = 90\,. \tag{14}$$

Finally, the ratio of the mean squares gives the standard F-ratio statistics of the ANOVA:

$$F = \frac{MS_{\text{between}}}{MS_{\text{within}}} = \frac{90}{6} = 15.00\,. \tag{15}$$

Under the usual statistical hypotheses (e.g., normality of the residuals and homoscedasticity) the F ratio will—when the null hypothesis is true—follow a Fisher distribution with $\nu_1 = J-1$ (ici $\nu_1 = 1$) et $\nu_2 = N-J$ (ici $\nu_2 = 8$) degrees of freedom. For this example, the p value associated to an $F = 15.00$ (with $\nu_1 = 1$ and $\nu_2 = 8$) is equal to $p = .0047$——a value small enough to reject the null hypothesis at the usual thresholds.

The formula for F can be rewritten in a way that makes the discriminant analysis optimization problem easier to state. Specifically, $\mathbf{F}$ can be rewritten as

$$F = \frac{MS_{\text{between}}}{MS_{\text{within}}} = \frac{SS_{\text{between}}}{SS_{\text{within}}} \times \frac{df_{\text{within}}}{df_{\text{between}}} = \frac{\boldsymbol{w}^\top \boldsymbol{w}}{\mathbf{b}^\top \mathbf{b}} \times \frac{I-J}{J-1}\,. \tag{16}$$

This shows that for a specific problem (i.e., number of groups and number of observations being fixed), the F ratio depends only upon the ratio $\dfrac{\boldsymbol{w}^\top \boldsymbol{w}}{\mathbf{b}^\top \mathbf{b}}$.

## BACK TO DA: MAXIMIZATION PROBLEM

Recall that DA optimally combines the original predictors to create the discriminant variables which would give the largest F if used to compute an ANOVA on the groups. This maximization problem—akind to principal component analysis and canonical correlation analysis—can be expressed from Equation 16 by first noting that the term $\frac{I-J}{J-1}$ plays the rôle of a constant and can therefore be ignored. So, we are looking for the coefficients of one (or more) linear transformation (which can be stored in a vector $\boldsymbol{\nu}$ or in a matrix $\mathbf{V}$ if there are several such linear transformations) of the predictors that will maximize the F ratio from Equation 16. Specifically, if we denote by $\mathbf{Z}$ the matrix of the centered predictors (which could also be normalized or transformed into Z-scores), we are looking for a linear transformation of the columns of $\mathbf{Z}$ that will give a new variable called $\mathbf{h}$ obtained as a linear combination of the columns of $\mathbf{Z}$ with coefficients stored in the vector denoted $\boldsymbol{z}$ such that $\mathbf{h} = \mathbf{Z}\boldsymbol{\nu}$ with the condition that this new variable $\mathbf{h}$ maximizes the F ratio.

There are several equivalent ways of solving the maximization problem of DA, but they all involve the same matrices. Specifically, with K predictors if we denoted (by analogy with the ANOVA) $\mathbf{D_g}$ the $K \times J$ matrix of the between group distances and $\mathbf{C}$ the $J \times J$ within group variance covariance matrix, then the DA maximization problem can be solved in terms of the generalized singular value decomposition as:

$$\mathbf{D_g} = \widetilde{\mathbf{U}}\widetilde{\mathbf{\Delta}}\widetilde{\mathbf{V}}^{\mathsf{T}} \quad \text{with } \widetilde{\mathbf{V}}^{\mathsf{T}}\left(\mathbf{C}^{-1}\right)\widetilde{\mathbf{V}} = \widetilde{\mathbf{U}}^{\mathsf{T}}\left(\frac{1}{I}\underset{I \times I}{\mathbf{I}}\right)\widetilde{\mathbf{U}} = \underset{L \times L}{\mathbf{I}} \tag{17}$$

where $L \leq \min(K-1)(J-1)$ is the rank of $\mathbf{D_g}$. The matrix $\mathbf{H_g} = \widetilde{\mathbf{U}}\widetilde{\mathbf{\Delta}}$ gives the optimum scores for separating the group means. This decomposition shows that DA finds the best possible separation (i.e., create the largest variance) between the groups means under the constraints imposed by the matrix $\mathbf{C}^{-1}$ (sometimes called the Mahalonobis distance matrix, from the eponym Indian statistician). The first generalized singular vectors $\widetilde{\mathbf{U}}$ et $\widetilde{\mathbf{V}}$—which are associated to the largest singular value—give the largest F ratios. The following pairs of singular vectors give the subsequent optimal discriminant variables (with the constraint that discriminant variables are pairwise orthogonal). To obtain the values of the discriminant variables for the original observations the mean of each variable is subtracted from this variable. Denote by $\mathbf{Z}$ this centered matrix, then the value of the discriminant variables is obtained by multiplying $\mathbf{Z}$ by the matrix $\mathbf{V} = \left(\mathbf{C}^{-1}\right)\widetilde{\mathbf{V}}$. The whole procedure is illustrated by the following example.

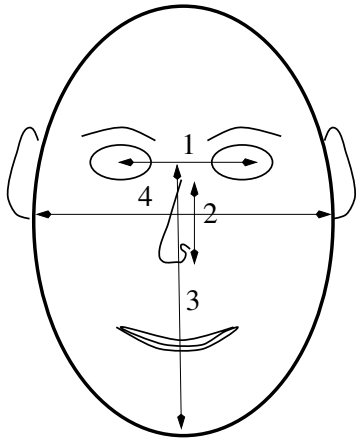


Figure 1: The four variables measured on the faces used in the example: 1) between eye distance, 2) length of the nose, 3) distance from eyes to chin, and 4) width of the face.

Table 1
Four variables measured on the six participants (from the Figure in Abdi & Valentin, 2009, p. 159). All these measurements are in millimeters and were collected on pictures of size 8cm $\times$ 5cm.

| ID Number | Between-Eyes Distance | Nose Length | Eye-to-Chin Distance | Width |
|---|---|---|---|---|
| | | Females | | |
| $F_1$ | 30 | 26 | 55 | 75 |
| $F_2$ | 28 | 26 | 60 | 80 |
| $F_3$ | 28 | 26 | 60 | 85 |
| | | Males | | |
| $G_1$ | 26 | 24 | 60 | 80 |
| $G_2$ | 35 | 25 | 61 | 80 |
| $G_3$ | 27 | 24 | 57 | 73 |

## EXAMPLE: SEX IDENTIFICATION

This small example illustrate DA with a face processing example. Here the task is to identify the sex of a person from measurements performed on a picture of their face. The pictures of six persons (three females and three males, obtained from Figure V.6 at page 159 of the 2009 book of Hervé Abdi and Dominique Valentin) were used for the exercise. Four measurements (in millimeters) were collected from these pictures: 1) between eye distance, 2) length of the nose, 3) distance from eyes to chin, and 4) width of the face. (see Figure 1 for details and Table 1 for the data).

The raw data are stored into the I (here $I = 6$) by K (here $K = 4$) data matrix $\mathbf{X}$:

$$\mathbf{X} = \begin{bmatrix} 30 & 26 & 55 & 75 \\ 28 & 26 & 60 & 80 \\ 28 & 26 & 60 & 85 \\ 26 & 24 & 60 & 80 \\ 35 & 25 & 61 & 80 \\ 27 & 24 & 57 & 73 \end{bmatrix} . \tag{18}$$

The number of groups to identify is J (here $J = 2$) as we have a balanced design, the number of observations per group is denoted N (here $N = 3$).

The next step is to create the dependent variable group matrix. This I by J matrix, denoted $\mathbf{Y}$, contains only 0s and 1s with a value of 1 indicating group membership:

$$\mathbf{Y} = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \end{bmatrix}^{\mathsf{T}} . \tag{19}$$

(Note, incidentally that, in DA, the rôles of $\mathbf{X}$ and $\mathbf{Y}$ are flipped compared to the ANOVA, *cf.* Equation 4).

The K by 1 vector of the grand means for all $K = 4$ variables is denoted $\mathbf{g}$ and computed as

$$\mathbf{g} = \mathbf{X}^{\mathsf{T}} \times \left( \frac{1}{I} \underset{I \times 1}{\mathbf{1}} \right) = \begin{bmatrix} 29.00 & 25.17 & 58.83 & 78.83 \end{bmatrix}^{\mathsf{T}} . \tag{20}$$

To eliminate size effects the matrix is centered to give matrix $\mathbf{Z}$ (note that in most applications, $\mathbf{Z}$ would also be normalized or transformed into $\mathbf{Z}$ scores):

$$\mathbf{Z} = \mathbf{X} - \left( \underset{I \times 1}{\mathbf{1}} \mathbf{g}^{\mathsf{T}} \right) = \begin{bmatrix} 1.00 & 0.83 & -3.83 & -3.83 \\ -1.00 & 0.83 & 1.17 & 1.17 \\ -1.00 & 0.83 & 1.17 & 6.17 \\ -3.00 & -1.17 & 1.17 & 1.17 \\ 6.00 & -0.17 & 2.17 & 1.17 \\ -2.00 & -1.17 & -1.83 & -5.83 \end{bmatrix} . \tag{21}$$

The means of the groups are collected in the $K \times J$ matrix denoted $\mathbf{O}$

$$\mathbf{O} = \left( \mathbf{Y}^{\mathsf{T}} \mathbf{Y} \right)^{-1} \mathbf{Y}^{\mathsf{T}} \mathbf{X} = \begin{bmatrix} 28.67 & 26.00 & 58.33 & 80.00 \\ 29.33 & 24.33 & 59.33 & 77.67 \end{bmatrix} . \tag{22}$$

In Matrix $\mathbf{O}$, the intersection of row $k$ and column $j$ stores the value of the mean of the $k$th group for the $j$th variable.

The distances of the group means to the grand means of the variables are stored in the $K \times J$ matrix denoted $\mathbf{D}_{\mathbf{g}}$ computed as:

$$\mathbf{D}_{\mathbf{g}} = \mathbf{O} - \left( \underset{K \times 1}{\mathbf{1}} \mathbf{g}^{\mathsf{T}} \right) = \begin{bmatrix} -0.33 & 0.83 & -0.50 & 1.17 \\ 0.33 & -0.83 & 0.50 & -1.17 \end{bmatrix} . \tag{23}$$

The value of the distances of the observations to their group means are collected in the I by J matrix $\mathbf{D}_{\mathbf{W}}$ computed as:

$$\mathbf{D}_{\mathbf{W}} = \mathbf{X} - (\mathbf{Y} \times \mathbf{O}) = \begin{bmatrix} 1.33 & 0.00 & -3.33 & -5.00 \\ -0.67 & 0.00 & 1.67 & 0.00 \\ -0.67 & 0.00 & 1.67 & 5.00 \\ -3.3 & -0.33 & 0.67 & 2.33 \\ 5.67 & 0.67 & 1.67 & 2.33 \\ -2.33 & -0.33 & -2.33 & -4.67 \end{bmatrix} . \tag{24}$$

Matrix $\mathbf{D}_W$ is then used to compute the within groups variance/covariance matrix denoted $\mathbf{C}$:

$$\mathbf{C} = \frac{1}{I - K} \left( \mathbf{D}_{\mathbf{W}}^{\mathsf{T}} \mathbf{D}_{\mathbf{W}} \right) = \begin{bmatrix} 12.83 & 1.42 & 1.50 & 1.58 \\ 1.42 & 0.17 & 0.42 & 0.58 \\ 1.50 & 0.42 & 6.33 & 10.33 \\ 1.58 & 0.58 & 10.33 & 20.67 \end{bmatrix} , \tag{25}$$

whose inverse is:

$$\mathbf{C}^{-1} = \begin{bmatrix} 304.00 & -2880.00 & 124.00 & -4.00 \\ -2880.00 & 27291.52 & -1175.52 & 38.08 \\ 124.00 & -1175.52 & 51.52 & -2.08 \\ -4.00 & 38.08 & -2.08 & 0.32 \end{bmatrix}. \quad (26)$$

These matrices can now be used to compute the generalized SVD $\mathbf{D_g}$ :

$$\begin{aligned} \mathbf{D_g} &= \widetilde{\mathbf{U}}\widetilde{\mathbf{\Delta}}\widetilde{\mathbf{V}}^{\mathsf{T}} \\ &= \underbrace{\begin{bmatrix} -1.73 \\ 1.73 \end{bmatrix}}_{\widetilde{\mathbf{u}} \text{ with: } \widetilde{\mathbf{u}}^{\mathsf{T}}\left(\frac{1}{I}\mathbf{I}\right)\widetilde{\mathbf{u}}=\mathbf{I}} \times \underbrace{85.05}_{\widetilde{\Delta}} \\ &\quad \times \underbrace{\begin{bmatrix} 0.0023 & -0.0057 & 0.0034 & -0.0079 \end{bmatrix}}_{\widetilde{\mathbf{V}}^{\mathsf{T}} \text{ with: } \widetilde{\mathbf{V}}^{\mathsf{T}}\left(\mathbf{C}^{-1}\right)\widetilde{\mathbf{V}}=\mathbf{I}}. \end{aligned} \quad (27)$$

From this GSVD, the values of the discriminant function for the group means are obtained as:

$$\begin{aligned} \mathbf{H_g} &= \widetilde{\mathbf{U}}\widetilde{\mathbf{\Delta}} \\ &= \mathbf{D_g}\left(\mathbf{C}^{-1}\right)\widetilde{\mathbf{V}} = \mathbf{D_g} \times \mathbf{V} \\ &= \begin{bmatrix} -147.3094 \\ 147.3094 \end{bmatrix}, \end{aligned} \quad (28)$$

(with $\mathbf{V} = \mathbf{C}^{-1}\widetilde{\mathbf{V}}$, note also that matrix $\mathbf{H_g}$ is column centered). Finally the values of the discriminant function for the original observations are stored in matrix $\mathbf{H}$ computed as (cf. Equation 28):

$$\begin{aligned} \mathbf{H} &= \mathbf{Z}\left(\mathbf{C}^{-1}\right)\widetilde{\mathbf{V}} = \mathbf{Z} \times \mathbf{V} \quad (29) \\ &= \begin{bmatrix} -146.64 & -147.06 & -148.23 & 148.47 & 147.29 & 146.17 \end{bmatrix}^{\mathsf{T}}. \end{aligned}$$

These factor scores can be used to compute the I by J Euclidean distance matrix between observations and group which is denoted $\widehat{\mathbf{D}}$ and equal to

$$\widehat{\mathbf{D}} = \begin{bmatrix} 0.67 & 0.25 & 0.92 & 295.78 & 294.60 & 293.48 \\ 293.95 & 294.37 & 295.54 & 1.16 & 0.02 & 1.14 \end{bmatrix}^{\mathsf{T}}. \quad (30)$$

Using the distance matrix, the observations are then assigned to the closest group

$$\widehat{\mathbf{Y}} = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \end{bmatrix}^{\mathsf{T}}. \quad (31)$$

Comparing $\hat{\mathbf{Y}}$ with the group matrix $\mathbf{Y}$ (from Equation 19) shows that prediction is perfect all 6 observations are assigned to their groups.

Note, incidentally that the computing a Fisher F using the scores from Equation 29 will a maximum value of (see Equation 15):

$$F = \frac{130200.3265}{1} = 130200.3265\,. \tag{32}$$

A value larger than any of the original Fs.

### Evaluating the model performance with new observations

The perfect performance of the discriminant model evaluated previously is a positively biased estimation because the data used to evaluate the model were also the data used to build the model. There are several ways to obtain a less biased (ideally an unbiased) estimate of the performance of the discriminant model. Some parametric approaches exist, but they often rely on assumptions such as multivariate normality that make these approaches undesirable. Contemporary practice therefore prefers to rely on computational cross-validation procedures. Probably the most straightforward of these procedures is to evaluate the model with *new* data (i.e. data that were, therefore, not used to compute the discriminant function). These new data can be obtained, for example, by sequestering part of the data (i.e., keeping some observations hidden) prior to the analysis and evaluating these observations only *after* the analysis has been performed. Here, for example, we sequestered 4 observations (2 females and 2 males), stored in a matrix denoted $\mathbf{X}_{\text{sup}}$:

$$\mathbf{X}_{\text{sup}} = \begin{bmatrix} F_1 & 29 & 26 & 58 & 77 \\ F_2 & 30 & 23 & 60 & 83 \\ M_1 & 28 & 25 & 60 & 80 \\ M_2 & 25 & 25 & 50 & 80 \end{bmatrix}\,. \tag{33}$$

To predict the sex of these observations, we use the parameters estimated from the DA to compute the discriminant function which will be used, in turn, to assign these observations to the groups. The first step is to pre-process the data matrix in way identical to the original data. Here the data were only centered so we will only center the new data by subtracting the grand mean vector $\mathbf{g}$ from each observation (likewise, if we had normalized the original data we would normalize the new observations using the normalization parameters used in the original analysis). This gives

$$\mathbf{Z}_{\text{sup}} = \mathbf{X}_{\text{sup}} - \left(\underset{I_{\text{sup}} \times 1}{\mathbf{1}}\,\mathbf{g}^T\right) = \begin{bmatrix} 0.00 & 0.83 & -0.83 & -1.80 \\ 1.00 & -2.17 & 1.17 & 4.17 \\ -1.00 & -0.17 & 1.17 & 1.17 \\ -4.00 & -0.17 & -8.83 & 1.17 \end{bmatrix}\,. \tag{34}$$

The matrix of discriminant scores for these new observations is obtained by replacing $\mathbf{Z}$ by $\mathbf{Z}_{\text{sup}}$ in Equation 29

$$\begin{aligned}\mathbf{H}_{\text{sup}} &= \mathbf{Z}_{\text{sup}}\left(\mathbf{C}^{-1}\right)\widetilde{\mathbf{V}} = \mathbf{Z}\times\mathbf{V} \qquad (35)\\ &= \begin{bmatrix} 143.17025 & -382.69389 & -18.13598 & 105.38093 \end{bmatrix}^{\mathsf{T}}.\end{aligned}$$

The new observations are now correctly classified only half of the time: a performance much less impressive that with the original DA prediction. When there is no available external datum, a substitute approach could be to predict each observation (or subset of observations) by a model built from all the other observations. This approach called the "leave one out (LOO)" procedure is often favored for small samples.

## CONCLUSION

In addition of being mathematically straightforward and easily implemented, DA remains a very popular method for predicting group membership observations described by multiple quantitative variables. It has, however some limitations that current research is trying to alleviate. A first limitation of DA is to necessitate quantitative predictors; this limitation is addressed by discriminant correspondence analysis that adapts correspondence analysis to handle qualitative variables.

Another problem with DA originates from its assigment rule: An observation is assigned to the closest group and this gives a 0 or a 1 probability for an observation to belong to a group. This assigment rule seems too strict, as can be illustrated by the example of the prediction of the new faces: a male face with a discriminant score of $-18.14$ was misclassified as female and a female face was misclassified as male with a discriminant score of $-143.17$: The first prediction error feels weaker than the second, but DA gives the same score to both incorrect prediction.s. *Logistic regression* generalizes DA by providing probabilities of assignment to a class and would preserve the difference between the predictions by assigning different probabilities to them. Note that logistic regression is, however, even more sensitive than DA to the inversion problem discussed below.

Another problem of DA originates from the inversion step of matrix $\mathbf{C}$, because this inversion requires well conditioned data, it *de facto* precludes using DA with large data sets for which the variables outnumber th observations (a configuration often called the $\mathsf{P} >> \mathsf{N}$ problem). Modern techniques such as ridge can minimize this problem as long as the variables do not *vastly* outnumber the observations. A lot of recent work in artificial intelligence, pattern recognition, and statistics (e.g., deep learning, neural networks, statistical learning, and support vector machines, to cite but a few) is dedicated to develop the next generation of discriminant

methods that would overcome the limits of DA while preserving its predictive power.

## SEE ALSO

Barycentric discriminant analysis, Correspondence analysis, Matrix algebra, Principal component analysis, Analysis of variance (ANOVA), Multivariate analysis of variance (MANOVA)

# Further Readings


[1] Abdi, H., & Beaton, D. (2022) *Principal Component and Correspondence Analyses Using R*. New York: Springer Verlag.

[2] Abdi, H., Williams, L.J., Beaton, D., Posamentier, M., Harris, T.S., Krishnan, A., & Devous, M.D. (2012). Analysis of regional cerebral blood flow data to discriminate among Alzheimer's disease, fronto-temporal dementia, and elderly controls: A multi-block barycentric discriminant analysis (MUBADA) methodology. *Journal of Alzheimer Disease, 31*, s189–s201.

[3] Abdi, H., & Valentin, D. (2009). *Mathématiques pour les Sciences Cognitives*. Grenoble: Presses Universaitaires de Grenoble.

[4] Abdi, H., Valentin, D. & Edelman, B. (1999). *Neural Networks*. Thousand Oaks: Sage.

[5] Beaton, D., Chin Fatt C.R., & Abdi, H. (2014). An ExPosition of multivariate analysis with the Singular Value Decomposition in R. *Computational Statistics & Data Analysis, 72*, 176–189.

[6] Deisenroth, M.P., Faisal, A.A., & Ong, C.S., (2021). *Mathematics of Machine Learning*. Cambrige: Cambridge University Press.

[7] Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics, 7*, 179–18

[8] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning with Applications in R* (2nd Ed.). New York: Springer.

[9] McLachlan, G. J. (2004). *Discriminant Analysis and Statistical Pattern Recognition*. New York: Wiley.