# Biplots in Practice

**MICHAEL GREENACRE**

Professor of Statistics at the Pompeu Fabra University

Chapter 11 Offprint

# Discriminant Analysis Biplots

First published: September 2010
ISBN: 978-84-923846-8-6

Supporting websites:
http://www.fbbva.es
http://www.multivariatestatistics.org

Chapter 11

# Discriminant Analysis Biplots

Discriminant analysis is characterized by a classification of the cases in a data set into groups, and the study of these groups in terms of the variables observed on the cases. The ultimate aim is to discover which variables are important in distinguishing the groups, or to develop a classification rule for predicting the groups. Up to now biplots have been displaying, as accurately as possible, data on individual case points. One exception, which is a type of discriminant analysis, was in Chapter 9 where the biplot was designed to show differences between demographic groups in the data rather than show individual differences. This biplot of group differences did not take into account correlations between the variables, while other approaches—such as Fisher's linear discriminant analysis—use a distance function between cases which does take into account correlations. In this chapter the geometry of these two approaches is explained and biplots are developed to display the cases and their group averages, along with the variables, in a discriminant space that shows the group differences optimally.

## Contents



### Analyzing groups

A common feature of a cases-by-variables data matrix is the classification of the cases (or the variables) into groups. For example, in the case of the *Arctic charr* fish in the "morphology" data set, each fish is classified as male or female, and also whether it was caught near the shore or in the open waters. Are these groups of fish different in terms of their morphology? In the "women" data set we displayed differences between individual respondents in Exhibit 10.3, whereas in Exhibit 9.3 group differences were displayed. In all of these methods where individual-level differences are displayed, it is also possible to display aggregate-level differences. Analyzing group differences based on multivariate data is called *discriminant analysis*, or DA. The basic idea underlying DA is the analysis of group means, or centroids, rather than the individual data. In other words, what we have been doing up to now in analyzing $N$ cases, say, can be thought of as a DA of $N$ groups, each consisting of 1 case, whereas now we consider the DA of $G$ groups, of size $N_1, \ldots, N_G$, where $N_1 + \cdots + N_G = N$. Although not so common, the same idea can be applied to the variables: instead of analyzing all the morphometric variables individually in the "morphology" data set, for example, the variables can be amalgamated by summation or averaging into predetermined groups.

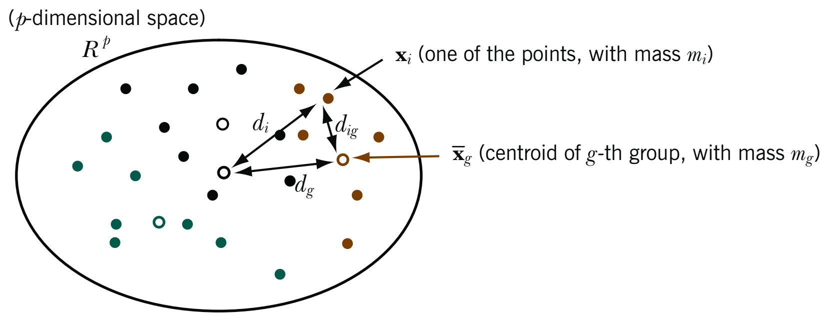
Between- and within-group variance/inertia

When cases are grouped, there is a decomposition of the total variance (or inertia in CA and LRA when points have different weights) into two parts: variance between groups, which quantifies how different the groups are, and variance within groups, which quantifies how internally heterogeneous the groups are:

$$\text{Total variance} = \text{Between-group variance} + \text{Within-group variance} \tag{11.1}$$

The greater the between-group variance, the more homogeneous the groups must be—in the extreme case where groups consist of single cases, between-group variance is the total variance and within-group variance is zero. At the other extreme where there is no group separation, between-group variance is zero and the within-group variance is the total variance. The decomposition in (11.1) is the basis of analysis of variance for one variable, whereas in our context that variance is measured in multidimensional space, using the distance measure of the particular method, be it PCA, CA/MCA or LRA.

**Exhibit 11.1:**
*The open circles represent the centroids of three groups (coloured in green, black and brown). Points have a distance $d_i$ to the overall centroid, represented by the bold open circle. The distance of a member of group $g$ to its group centroid is $d_{ig}$, and the distance from the centroid of group $g$ to the overall centroid is $d_g$. Points have masses $m_i$ and the aggregated mass in group $g$ is $m_g$, which is assigned to the respective group centroid*




Total inertia = Between-group inertia + Within-group inertia

$$\textstyle\sum_i m_i d_i^2 \;=\; \sum_g m_g d_g^2 \;+\; \sum_g (\sum_i m_i d_{ig}^2)$$

In PCA (principal component analysis) each case point typically has weight $1/N$ and the weights assigned to the group average points $\bar{\mathbf{x}}_1, \bar{\mathbf{x}}_2, \ldots, \bar{\mathbf{x}}_G$ are then $N_1/N$, $N_2/N$, …, $N_G/N$. In LRA and CA/MCA the case points can have varying weights, $r_i$, in which case the group means are weighted by the accumulated weights of their respective cases. Exhibit 11.1 illustrates the decomposition of variance/inertia in a schematic way—here the groups are shown almost perfectly separated, but in practice there is a high degree of overlap and the group averages, or centroids, separate out much less. For example, in the "morphology" data set of Chapter 7, where LRA was used to visualize the 75 *Arctic charr* fish, there were four groups of fish defined by the combinations of sex (male/female, abbreviated as m/f) and habitat (littoral/pelagic, abbreviated as L/P)—these were labelled in Exhibit 7.3 as fL, mL, fP and mP. The total inertia in the full multidimensional space of the LRA was equal to 0.001961. The decomposition of inertia (11.1) with respect to the four sex/habitat groups turns out to be:

$$0.001961 = 0.000128 + 0.001833$$

The between-group inertia (0.000128) is only 6.5% of the total. In the Computational Appendix a permutation test is performed, showing that the between-group differences, although small, are statistically significant ($p = 0.015$) and worth investigating. A small between-group variance or inertia does not mean that there is no meaningful separation of the groups—groups can be separable and still have a high percentage of within-group variance. In Exhibit 7.3, however, the objective of the biplot was to separate the individual fish optimally in the two-dimensional view, not the groups of fish—separating the groups optimally in a low-dimensional display is the job of DA.

Example: LRA-DA biplot

The log-ratio discriminant analysis (LRA-DA) biplot of the four fish groups is shown in Exhibit 11.2. This is achieved by performing a regular LRA on the $4 \times 26$ matrix of group centroids, weighted by their respective aggregated masses (remember that in LRA, as in CA, the mass of a point is proportional to its marginal sum, so that the mass $r_i$ of each fish is proportional to the total of its morphometric values). The dimensionality of the four centroids is three, so that dimension reduction to two dimensions means sacrificing only one dimension. The biplot shows that the main difference (along the horizontal axis) is that between the two littoral groups on the left and the two pelagic groups on the right. The second axis separates the females from the males, especially the female and male littorals.

To find out which ratios might be associated with these separations, the log-ratio of *Bc* relative to *Jw* is the longest horizontal link corresponding to the left-to-right littoral-pelagic contrast in Exhibit 11.2. Performing a two-group *t*-test between the

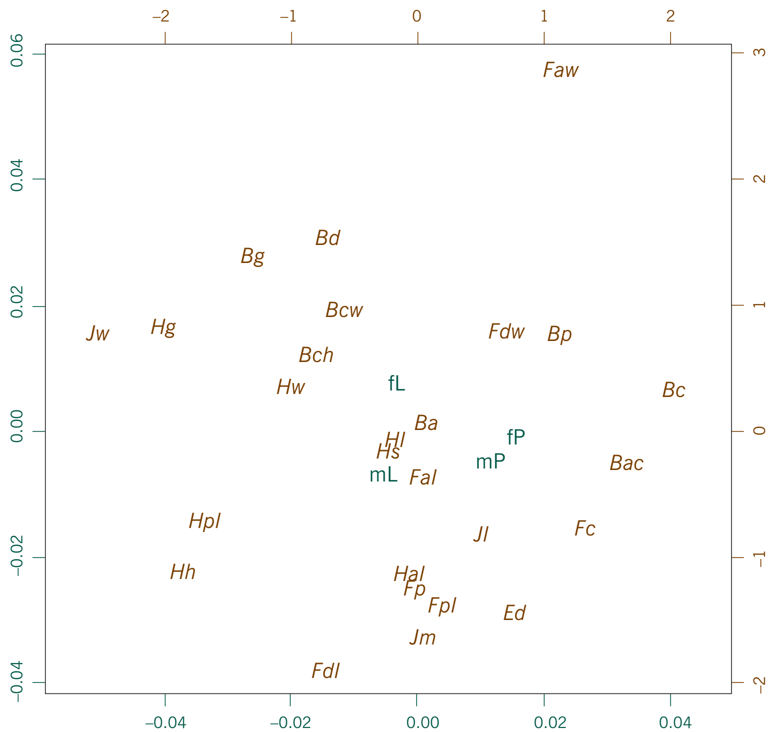



**Exhibit 11.2:** *LRA-DA biplot of the four fish groups in the "morphology" data set, with fish group centroids in principal coordinates and variables in standard coordinates. Because of the very small inertia of the centroids (0.000128), they are shown on a different scale. 79.9% of this inertia of the centroids is explained in the biplot*

pelagic and littoral groups on this log-ratio gives a difference on the log-scale of 0.083, which translates to a 8.7% increased ratio of *Bc/Jw* in pelagics compared to littorals, significant at $p = 0.006$. As for the difference between females and males on the vertical axis, the biplot suggests the log of *Faw/Fdl* to be a possible candidate. Indeed, the *t*-test shows a difference between females and males of 0.073 on the log-scale, which is a 7.6% increase in the *Faw/Fdl* ratio in females compared to males, with $p < 0.0001$.

## Example: CA-DA biplot

Exhibit 7.3 was already an example of a DA between groups in a CA context, although there we showed the differences between several groupings of the cases simultaneously. To focus on only one set of groups, let us consider visualize and quantify the differences in attitudes between the marital status groups m1 to m5: respectively, "married", "widowed", "divorced", "separated" and "single". What we call a CA-DA is just a CA of the concatenated tables with the 5 rows being the marital groups, and 39 columns the categories of the 8 variables (remember that *H4* and *H5* were combined to give one column *H4,5*). Exhibit 11.3 shows the CA contribution biplot, comparable to Exhibit 9.6. Again, *G2* is an important category for the conservative end of the scale, typified by the group "widowed". In Exhibit 9.6 we saw that disagreement (category 4) to statements *C*, *D*, *E* and *G* all correlated and were associated with younger males. In Exhibit 11.3 these categories are more spread out vertically, with *C4* an important category for "single" (disagree that family life suffers when a woman works) while *E4* and *G4* are important for "divorced" (disagree that running a household is just as satisfying as a paid job, and disagree that a man's job is to work and a woman's is the household).

Again, even though these between-group differences are meaningful and actually highly significant statistically (chi-square tests between marital status and each question all have $p < 0.0001$), the between-group inertia relative to the total is very small. This is because the total inertia of the original data in indicator matrix form is a fixed value and very high—see (10.3)—equal to 3.875 in this example. The between-group inertia could only attain this value in the extreme case that each of the five marital status groups gave identical responses within each group. In practice, the inertias of condensed tables like this one are very much smaller than the indicator matrix: in this example the total inertia of the five groups is 0.03554, which is only 0.917% of the total inertia of the indicator matrix. In the Computational Appendix we explain how to perform a permutation test to quantify the statistical significance of the between-group inertia.

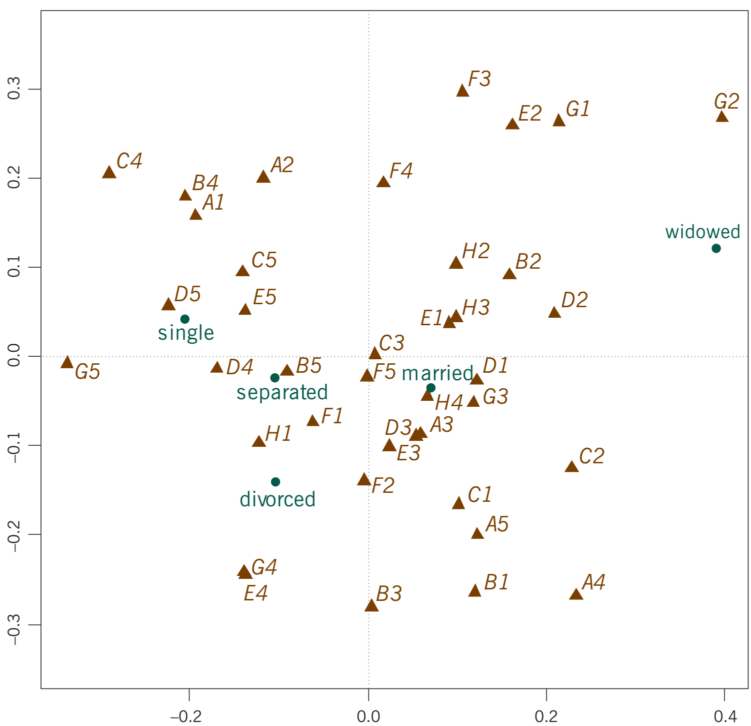



**Exhibit 11.3:**
*CA-DA of marital status groups in the "women" data set, in terms of the 8 questions on women working. 90.7% of the inertia is displayed here*

Mahalanobis distance

In the different variants of discriminant analysis described above, no mention was made at all of the effect of correlations between variables, which can distort the distances between group centroids. Suppose there are two groups of bivariate points, indicated by "x" and "o" in Exhibit 11.4, with their group means in boldface. The coloured ellipses in the left hand picture summarize the spread of each group, and show that the two variables are highly correlated within each group. The two centroids are at a certain distance apart. Now suppose that the "o" group lay in the position indicated by the dashed ellipse—its centroid is at the same distance from the "x" centroid, but the groups overlap much more. This anomaly can be removed by performing a transformation on the variables to de-correlate them, shown on the right hand side. The elliptical dispersion of the points is now spherical and, indeed, the "o" centroid is now much further away from the "x" centroid compared to the alternative in the dashed circle. The transformation involved is more than a standardization of the variables, because in the left hand picture of Exhibit 11.4 the two variables have the same variance. Rather, what is needed is a stretching out of the points in a direction more or less at right-angles to the axis of dispersion of the points—this is achieved by defining what is called the *Mahalanobis distance* between the cases, named after a famous Indian statistician.

Suppose $\mathbf{C}$ is the *within-groups covariance matrix* between the variables (this is defined in the next section). Then the Mahalanobis distance between two points $\mathbf{x}$ and $\mathbf{y}$ in multivariate space is

$$\text{Mahalanobis distance} = \sqrt{(\mathbf{x}-\mathbf{y})^{\mathsf{T}}\mathbf{C}^{-1}(\mathbf{x}-\mathbf{y})} \qquad (11.2)$$

If we omit the off-diagonal covariances in $\mathbf{C}$ so that $\mathbf{C}$ is the diagonal matrix of variances, then (11.2) is just the regular standardization of the variables. The presence of the off-diagonal covariances in $\mathbf{C}$ decorrelates the variables.

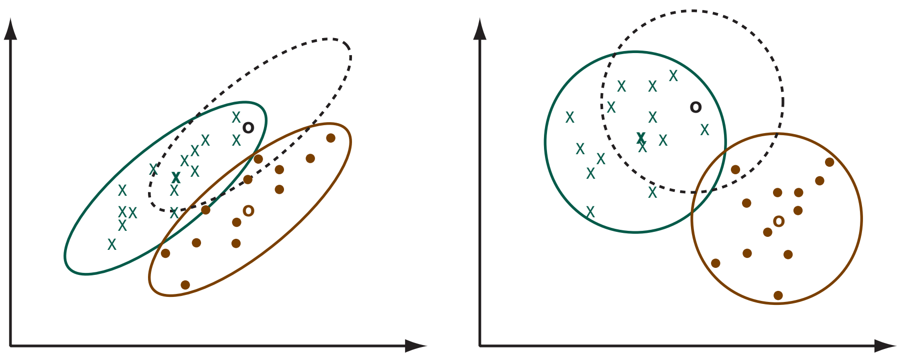


**Exhibit 11.4:** *The effect of high correlation between variables on the measure of between-group distance. On the right a transformation has been performed to remove the correlation—now the distances between points are Mahalanobis distances*

Linear discriminant analysis (LDA)

With this distance function in the space of the cases the same analysis of the centroids is performed as before—this is called *linear discriminant analysis* (LDA), attributed to the statistician R.A. Fisher, so sometimes called *Fisher discriminant analysis*. Given an cases × variables data matrix $\mathbf{X}$ ($I \times J$) where the cases are classified into $G$ groups, denote by $\mathbf{x}_{ig}$ the vector of observations for the $i$-th case in the $g$-th group, with weight (mass) $w_{ig}$. The group masses are $w_1, w_2, \ldots, w_G$ ($w_g = \sum_i w_{ig}$) and the centroids $\bar{\mathbf{x}}_1, \bar{\mathbf{x}}_2, \ldots, \bar{\mathbf{x}}_G$ ($\bar{\mathbf{x}}_g = \sum_i (w_{ig}/w_g)\mathbf{x}_{ig}$). Then the within-groups covariance matrix $\mathbf{C}$ is the weighted average of the covariance matrices computed for the groups separately:

$$\mathbf{C} = \sum_g w_g \mathbf{C}_g, \text{ where } \mathbf{C}_g = \sum_{i=1}^{N_g} (w_{ig} / w_g)(\mathbf{x}_{ig} - \bar{\mathbf{x}}_g)(\mathbf{x}_{ig} - \bar{\mathbf{x}}_g)^{\mathsf{T}} \tag{11.3}$$

The theory of generalized PCA and contribution biplots in Chapters 5 and 6 applies here:

– Centre the group means: $\bar{\mathbf{Y}} = \bar{\mathbf{X}} - \mathbf{1}\mathbf{w}^{\mathsf{T}}\bar{\mathbf{X}} = (\mathbf{I} - \mathbf{1}\mathbf{w}^{\mathsf{T}})\bar{\mathbf{X}}$, where $\bar{\mathbf{X}}$ is the matrix of centroids in the rows, $\mathbf{w}$ is the vector of group weights; since the overall centroid $\bar{\mathbf{x}}^{\mathsf{T}}$ (written as a row vector) of the individual-level data in $\mathbf{X}$ is identical to the centroid $\mathbf{w}^{\mathsf{T}}\bar{\mathbf{X}}$ of the group centroids, we could also write $\bar{\mathbf{Y}} = \bar{\mathbf{X}} - \mathbf{1}\bar{\mathbf{x}}^{\mathsf{T}}$.

– Transform to Mahalanobis distance and weight by the masses before computing the SVD (singular value decomposition):

$$\mathbf{S} = \mathbf{D}_w^{\frac{1}{2}} \bar{\mathbf{Y}} \mathbf{C}^{-\frac{1}{2}} (1/J)^{\frac{1}{2}} = \mathbf{U}\mathbf{D}_\eta \mathbf{V}^{\mathsf{T}} \tag{11.4}$$

  where $\mathbf{C}^{-\frac{1}{2}}$ is the inverse of the *symmetric square root* of $\mathbf{C}$—this is calculated using the eigenvalue decomposition[7] of $\mathbf{C}$: $\mathbf{C} = \mathbf{X}\mathbf{D}_\lambda \mathbf{X}^{\mathsf{T}}$, hence $\mathbf{C}^{-\frac{1}{2}} = \mathbf{X}\mathbf{D}_\lambda^{-\frac{1}{2}} \mathbf{X}^{\mathsf{T}}$.

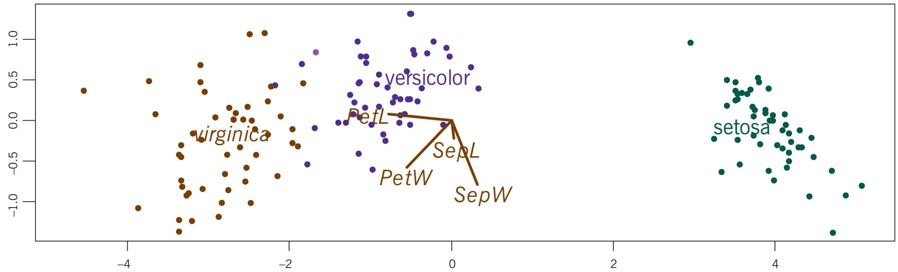
– Calculate the principal coordinates of the group centroids: $\mathbf{F} = \mathbf{D}_w^{-\frac{1}{2}} \mathbf{U}\mathbf{D}_\eta$ and the coordinates of the variables for the contribution biplot (for example): $\mathbf{\Gamma} = \mathbf{V}$.

This theory is illustrated with Fisher's famous "iris" data set, available in the R package (see the Computational Appendix). In this case, the decomposition of variance is:

$$9.119 = 8.119 + 1.000$$

7. Notice that the eigenvalue decomposition of a square symmetric matrix is the same as the singular value decomposition when the eigenvalues are all non-negative, as is the case here for the covariance matrix **C**. Thus it is possible to calculate the square roots of the eigenvalues of **C**.

**Exhibit 11.5:**
*LDA contribution biplot of Fisher "iris" data. 99.1% of the variance of the centroids is explained by the first axis, on which PetL (petal length) is the highest contributor*




Notice that the within-group variance is equal to 1, by construction. In this example the variance of the centroids accounts for 89.0% of the total variance, and the separation is excellent, as shown in Exhibit 11.5. The first principal axis of the centroids totally dominates, accounting for 99.1% of their variance. Petal length (*PetL*) and petal width (*PetW*) are seen to be the most important variables on this axis.

Notice that the individual points have been added to this LDA biplot, as supplementary points. To derive the coordinates of the individuals, notice from the algorithm above that the principal coordinates $\mathbf{F} = \mathbf{D}_w^{-½}\mathbf{U}\mathbf{D}_\eta$ of the group centroids are equivalently obtained by the transformation of the column variable points $\mathbf{V}$: $\bar{\mathbf{Y}}\mathbf{C}^{-½}(1/J)^{½}\mathbf{V}$. The coordinates $\mathbf{F}_{\text{case}}$ of individual case points are obtained in a similar way, using the centred matrix $\mathbf{Y}$ for the original individual-level data:

$$\mathbf{F}_{\text{case}} = \mathbf{Y}\mathbf{C}^{-½}(1/J)^{½}\mathbf{V} \tag{11.5}$$

In a similar fashion all the individual cases could have been added to the DA biplots of Exhibits 11.2 and 11.3, using the appropriate relationship in each analysis between the row and column points—this relationship is often referred to as the *transition formula* between rows and columns.

SUMMARY:
Discriminant Analysis Biplots

1. Discriminant analysis is the analysis of group means, or centroids, of a set of multivariate points classified into pre-specified groups.
2. The centroids have masses (weights) equal to the sum of the masses of their members.
3. There is a decomposition of total variance/inertia of the set of points into that of the centroids, the between-group variance/inertia, plus the weighted average variance/inertia within the groups, the within-group variance/inertia. (Inertia is simply the alternative term for variance when the points have different weights; or conversely, variance is the special case of inertia when the weights of all points are equal).

4. The dimension reduction of the centroids follows the same algorithm as the corresponding PCA, LRA or CA/MCA method.

5. *Linear discriminant analysis* (LDA) is also a dimension-reduction method on a set of centroids, but uses the Mahalanobis distance based on the within-groups covariance matrix to decorrelate the data.

6. In all these variations of DA the contribution biplot displays the centroids in an optimal map of their positions, along with the variables so that the most important (i.e., most discriminating) variables are quickly identified.