Biplots in Practice

MICHAEL GREENACRE

Professor of Statistics at the Pompeu Fabra University

Chapter 8 Offprint

Correspondence Analysis Biplots

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

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

CHAPTER 8

Correspondence Analysis Biplots

Correspondence analysis is the most versatile of the methods based on the SVD (singular value decomposition) for visualizing data. It applies primarily to a cross-tabulation (also called a contingency table) of two categorical variables but can be extended to frequency tables, ratio-scale data in general, binary data, preferences and fuzzy-coded continuous data. Like log-ratio analysis, correspondence analysis treats the rows and columns of a table in a symmetric fashion. There are several equivalent asymmetric ways of thinking about the analysis, however, and there are different associated biplots depending on whether the rows or columns are regarded as the "variables" of the table. In this chapter we define and illustrate the basic algorithm of correspondence analysis and list its properties and relationships to principal component analysis, log-ratio analysis, multidimensional scaling and regression biplots. In subsequent chapters various extensions of correspondence analysis will be described: the multiple form in Chapters 9 and 10, and the constrained form in Chapter 12.

Contents

Profiles, masses and chi-square distances: data set "smoking" 79
Correspondence analysis (CA) 82
Asymmetric maps 83
Connection with PCA and MDS 84
Connection with regression biplots 85
Inertia and inertia decomposition 85
Data set "benthos" 86
Contribution CA biplot 87
SUMMARY: Correspondence Analysis Biplots 88

Profiles, masses and chi-square distances: data set "smoking"

All the methods in this book are based on what the French call a *triple* (triplet) of information for a data set: the definition of (1) objects in a multidimensional space, (2) their weights, and (3) the distances between them. In MDS (multidimensional scaling) the distances form the original data set and an approximate map of the objects is produced. The objects could, however, have different

weights and the analysis would then represent distances involving points with higher weight better than those with lower weight. In PCA (principal component analysis), the original data is in the form of a rectangular data matrix and each row (or column) defines a point in multidimensional space. The points could be assigned different weights here as well, and if the distance function between the points is Euclidean, then the SVD provides the solution for the low-dimensional visualization of the points. In correspondence analysis (CA), these three concepts of points, weights and distances are called profiles, masses and chi-square distances, respectively. We review their definitions using the classic "smoking" data set (available in the **ca** package in R), given in Exhibit 8.1. The first table is the cross-tabulation of all 193 staff members of an organization according to their category in the organization and their level of smoking.

The row *profiles*, given in the second table, are the frequencies in the rows divided by their row sums (e.g., 0.364 = 4/11). The last row contains the average row profile, which is the profile of the column sums of the original table (e.g., 0.316 = 61/193). Similarly, the column profiles in the third table are the frequencies in the columns divided by the column sums and the average column profile in the last column is the profile of the row sums in the original table (e.g., 0.057 = 11/193). The row profiles are the points visualized in the row problem, and the column profiles are those visualized in the column problem.

Each profile has a weight called a *mass*, equal to the marginal sum of that row or column as the case may be, divided by the grand total of the table. For example, the first row profile has mass 11/193 = 0.057, which is identical to the first element of the average column profile. Thus the average column profile contains the row masses, and the average row profile contains the column masses. The masses are used to weight the profiles in the analysis, so that profiles based on larger counts have a stronger role in the analysis.

Distances between profiles are calculated using the chi-square distance, which has already been introduced in Chapter 4. The average row profile, for example, apart from serving to centre the row profiles, defines the distance function between row profiles, using the inverses of its values. For example, the distance between the first two row profiles is:

$$\sqrt{\frac{(0.364-0.222)^2}{0.316}+\frac{(0.182-0.167)^2}{0.233}+\frac{(0.273-0.389)^2}{0.321}+\frac{(0.182-0.222)^2}{0.130}}=0.345$$

This is a natural default standardization for frequency data, which tend to have higher variances if their means are higher. Similarly, chi-square distances can be

Exhibit 8.1:
Data set "smoking" and its row and column profiles, as well as their respective average profiles

Original cross-tabulation:

Staff group		Smoking class				
		None	*Light*	*Medium*	*Heavy*	*Sum*
Senior managers	SM	4	2	3	2	*11*
Junior managers	JM	4	3	7	4	*18*
Senior employees	SE	25	10	12	4	*51*
Junior employees	JE	18	24	33	13	*88*
Secretaries	SC	10	6	7	2	*25*
	Sum	*61*	*45*	*62*	*25*	*193*

Row profiles:

	Smoking class			
	None	*Light*	*Medium*	*Heavy*
SM	0.364	0.182	0.273	0.182
JM	0.222	0.167	0.389	0.222
SE	0.490	0.196	0.235	0.078
JE	0.205	0.273	0.375	0.148
SC	0.400	0.240	0.280	0.080
Average	*0.316*	*0.233*	*0.321*	*0.130*

Column profiles:

	Smoking class				
	None	*Light*	*Medium*	*Heavy*	*Average*
SM	0.066	0.044	0.048	0.080	*0.057*
JM	0.066	0.067	0.113	0.160	*0.093*
SE	0.410	0.222	0.194	0.160	*0.264*
JE	0.295	0.533	0.532	0.520	*0.456*
SC	0.164	0.133	0.113	0.080	*0.130*

defined between the column profiles, using the inverses of the elements of the average column profile.

Correspondence analysis (CA)

CA has many equivalent definitions and we give just one of them here. It is—at the same time—a generalized PCA of the row profiles and a generalized PCA of the column profiles, and the treatment of the rows and columns is the same, just as in LRA (log-ratio analysis) of the previous chapter. And again, both the row and column problems rely on the same matrix decomposition, as follows:

– First divide the original data table $\mathbf{N}$ by its grand total $n = \sum_i \sum_j n_{ij}$: $\mathbf{P} = (1/n)\mathbf{N}$

Denote by $\mathbf{r}$ and $\mathbf{c}$ the marginal sums of $\mathbf{P}$: $\mathbf{r} = \mathbf{P1}$, $\mathbf{c} = \mathbf{P}^\mathsf{T}\mathbf{1}$ (8.1)
(these are identical to $\mathbf{r}$ and $\mathbf{c}$ defined in (7.1)).

– Calculate the matrix of standardized residuals $\frac{p_{ij} - r_i c_j}{\sqrt{r_i c_j}}$ and its SVD:

$$\mathbf{S} = \mathbf{D}_r^{-½}(\mathbf{P} - \mathbf{rc}^\mathsf{T})\mathbf{D}_c^{-½} = \mathbf{U}\mathbf{D}_\alpha\mathbf{V}^\mathsf{T} \tag{8.2}$$

– Calculate the coordinates:

Principal coordinates of rows: $\mathbf{F} = \mathbf{D}_r^{-½}\mathbf{U}\mathbf{D}_\alpha$, of columns: $\mathbf{G} = \mathbf{D}_c^{-½}\mathbf{V}\mathbf{D}_\alpha$ (8.3)

Standard coordinates of rows: $\mathbf{\Phi} = \mathbf{D}_r^{-½}\mathbf{U}$, of columns: $\mathbf{\Gamma} = \mathbf{D}_c^{-½}\mathbf{V}$ (8.4)

Notice how similar this algorithm is to that of log-ratio analysis, formulated in (7.1)–(7.6) of Chapter 7; in fact, the two algorithms can be seen to be even more similar if the $\mathbf{S}$ matrix in (8.2) is rewritten in the equivalent form:

$$\text{[matrix for SVD in CA]} \qquad \mathbf{S} = \mathbf{D}_r^{½}(\mathbf{I} - \mathbf{1r}^\mathsf{T})\,(\mathbf{D}_r^{-1}\mathbf{P}\,\mathbf{D}_c^{-1})\,(\mathbf{I} - \mathbf{1c}^\mathsf{T})^\mathsf{T}\mathbf{D}_c^{½} \tag{8.5}$$

whereas in log-ratio analysis, from (7.2), (7.3) and (7.4):

$$\text{[matrix for SVD in LRA]} \qquad \mathbf{S} = \mathbf{D}_r^{½}(\mathbf{I} - \mathbf{1r}^\mathsf{T})\log(\mathbf{N})\,(\mathbf{I} - \mathbf{1c}^\mathsf{T})^\mathsf{T}\mathbf{D}_c^{½} \tag{8.6}$$

So the difference is that CA analyzes the contingency ratios $p_{ij}/(r_i c_j)$ in $\mathbf{D}_r^{-1}\mathbf{P}\,\mathbf{D}_c^{-1}$, whereas LRA analyses the logarithms of the data $\log(\mathbf{N})$. Since the double-centring removes any additive row or column constant, $\log(\mathbf{N})$ in (8.6) can be replaced by $\log(\mathbf{D}_r^{-1}\mathbf{P}\,\mathbf{D}_c^{-1})$ without changing the matrix for the SVD. So the only real difference between LRA and CA is the logarithmic transformation!

Asymmetric maps

As in LRA, there are two biplots that result in CA: the row-principal biplot ($\mathbf{F}, \mathbf{\Gamma}$) and the column-principal biplot ($\mathbf{G}, \mathbf{\Phi}$). In CA, however, the points in standard coordinates have an additional geometric interpretation: they are the extreme *unit profiles* or *vertices* of the profile space. Consider the row profiles of the "smoking" data, for example, and their associated biplot coordinates: $\mathbf{F}$ for the rows and $\mathbf{\Gamma}$ for the columns. In a two-dimensional display (using the first two columns of $\mathbf{F}$ and $\mathbf{\Gamma}$) the five row points are projections of the row profiles onto the best-fitting plane. The four column points, in standard coordinates, are the projections onto the same plane of the unit profiles [1 0 0 0], [0 1 0 0], [0 0 1 0] and [0 0 0 1]. Since any row profile $[p_1\ p_2\ p_3\ p_4]$ with elements adding up to 1 can be expressed as p_1 [1 0 0 0] + p_2 [0 1 0 0] + p_3 [0 0 1 0] + p_4 [0 0 0 1], it follows that the row profiles are at weighted averages of the column points, the weights being the profile elements. It is this weighted average (or centroid) property that makes CA so popular in ecological applications—if the columns follow an ecological gradient (for example, rainfall in a botanical study) then the weighted averages of the columns points for each row profile would situate the row on that gradient. Because the row and column points in this biplot lie in the same space, with the column points defining the most extreme profiles possible, the resultant display is also called a *map*, specifically an *asymmetric map*.

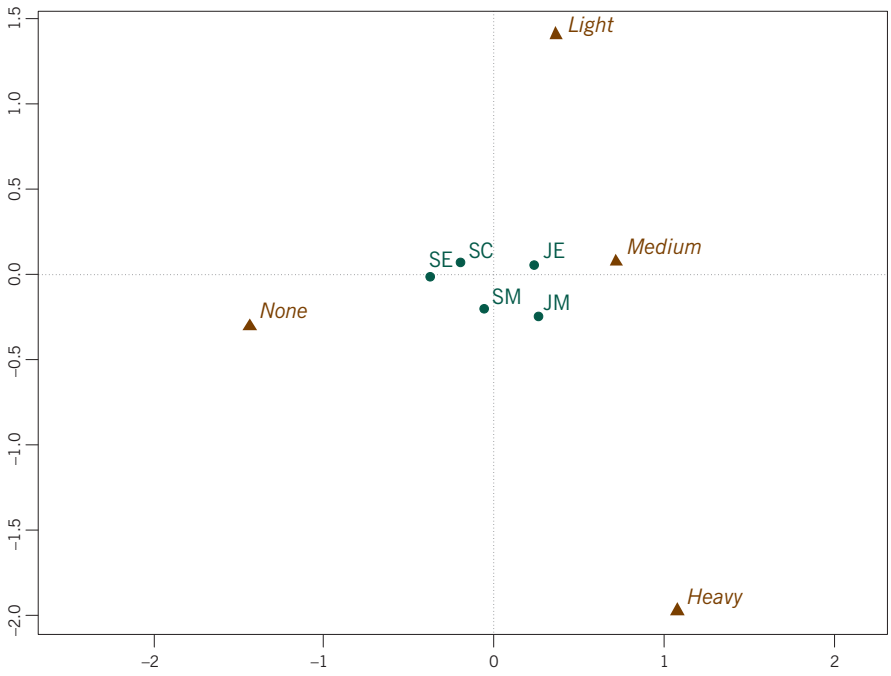


Exhibit 8.2:
Row asymmetric CA map (i.e., row principal biplot) of the "smoking" data, with rows in principal coordinates and columns in standard coordinates. This map is reproduced directly from the **ca** *package in* R *—see the Computational Appendix*

Exhibit 8.2 shows the row asymmetric map of the "smoking" data set. Because the two sets of points co-exist in the same space, the amount of variation between the row profiles can be seen in relation to the extreme vertex profiles. The five row profiles actually lie inside a tetrahedron in three-dimensional space, which has vertices defined by the four column points. As explained above, each row profile is at the weighted average of the four vertex points in this three-dimensional space, and so also in the projected map of Exhibit 8.2. Thus secretaries (SE) lie to the left because they must have higher than average proportion of non-smokers, whereas junior employees and managers (JE and JM) lie to the right because they have higher than average levels of smokers, with junior managers tending towards the high smoking group. All these deductions from the map can be confirmed in the data of Exhibit 8.1. In fact, we can be absolutely sure of these conclusions because 99.5% of the variance in Exhibit 8.1 is displayed in the map.

Connection with PCA and MDS

In (8.1)–(8.5) the algorithm is presented as a decomposition of a matrix where rows and columns have been treated symmetrically, but the same decomposition can also be thought of asymmetrically as an analysis of rows or an analysis of columns, as we in fact introduced it originally in the context of Exhibit 8.1. The matrix formulation **S** in (8.2) (equivalently in (8.5)) can be written either as:

$$\mathbf{S} = \mathbf{D}_r^{\frac{1}{2}}(\mathbf{D}_r^{-1}\mathbf{P} - \mathbf{1}\mathbf{c}^{\mathsf{T}})\mathbf{D}_c^{-\frac{1}{2}} \tag{8.7}$$

or, in transposed form:

$$\mathbf{S}^{\mathsf{T}} = \mathbf{D}_c^{\frac{1}{2}}(\mathbf{D}_c^{-1}\mathbf{P}^{\mathsf{T}} - \mathbf{1}\mathbf{r}^{\mathsf{T}})\mathbf{D}_r^{-\frac{1}{2}} \tag{8.8}$$

These two formulations show that CA can be thought of either (in (8.7)) as a weighted PCA (see Chapter 5) of the row profiles in the rows of $\mathbf{D}_r^{-1}\mathbf{P}$, weighted by the row masses in **r**, centred at their average profile $\mathbf{c}^{\mathsf{T}}$, in the chi-square metric defined by $\mathbf{D}_c^{-1}$; or (in (8.8)) as a weighted PCA of the column profiles in the rows of $\mathbf{D}_c^{-1}\mathbf{P}^{\mathsf{T}}$, weighted by the column masses in **c**, centred at their average profile $\mathbf{r}^{\mathsf{T}}$, in the chi-square metric defined by $\mathbf{D}_r^{-1}$. In the former row problem, the asymmetric map represents the row profiles in principal coordinates, with the unit profiles representing the columns in standard coordinates; in the latter asymmetric map, the columns profiles are in principal coordinates with the unit profiles representing the rows in standard coordinates.

Exactly the same principal coordinates can be obtained if CA is formulated as a pair of MDS problems. For example, chi-square distances are calculated between row profiles using the metric $\mathbf{D}_c^{-1}$ and with row points weighted by the row masses in **r**. Then by applying the classical MDS algorithm, with weights, in (5.11) and

(5.12), the principal row coordinates are recovered exactly. A symmetric result holds for the column profiles.

Connection with regression biplots

There are several ways to explain CA as a regression biplot, as described in Chapter 2—we explain it here in terms of definition (8.7), the weighted PCA of the row profiles, weighted by the row masses in the chi-square metric based on $\mathbf{D}_c^{-1}$ (in other words, the asymmetric map of the row profiles). The j-th column of the row profile matrix has elements p_{1j}/r_1, p_{2j}/r_2, ⋯, p_{Ij}/r_I. Centring is with respect to the average profile element c_j and the chi-square standardization implies dividing the centred profile by the square root of the corresponding average profile element, $c_j^{1/2}$. An appropriate regression is then when these standardized values $((p_{ij}/r_i) - c_j)/c_j^{1/2}$ $(i = 1, \ldots, I)$ constitute the response variable and the standard row coordinates on the first two dimensions, say, form the explanatory variables, then applying weighted least-squares fitting with weights equal to the row masses. The solution gives a constant equal to 0 and coefficients equal to $c_j^{1/2}$ times the principal coordinates of the j-th column (this result is illustrated in the Computational Appendix). This result implies that, if the regression is performed on the principal row coordinates rather than the standard ones, then the coefficients in the solution will be exactly the coordinates of the columns in the contribution biplot (see later in this chapter).

Inertia and inertia decomposition

The total variance in CA has a close connection with the chi-square statistic χ^2 often calculated on cross-tabulations as a measure of statistical association between rows and columns. The total variance of the **S** matrix decomposed in (8.2) (equivalently, (8.5), (8.7) or (8.8)) is:

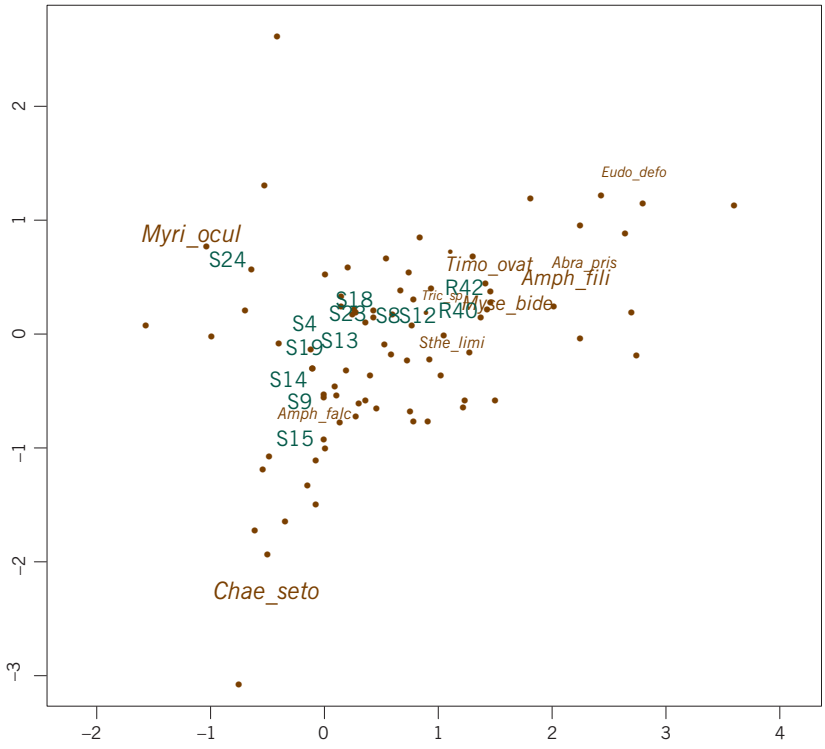
$$\sum_i \sum_j s_{ij}^2 = \sum_i \sum_j \frac{(p_{ij} - r_i c_j)^2}{r_i c_j} \tag{8.9}$$

In CA terminology this quantity is called the *total inertia* of the data matrix, or simply the *inertia*. It is easily shown that multiplying the inertia by the grand total n of the matrix gives the chi-square statistic: $\chi^2 = n \times$ inertia. As explained in general in Exhibit 6.4, there is a decomposition of total variance across points and across principal axes, leading to two ways of defining contributions for the rows as well as for the columns. First, contributions of each profile point to the inertia of each axis (column proportions in Exhibit 6.4) are used to interpret each axis—in the **ca** package in R, these are denoted by the acronym CTR and expressed as permills (see Computational Appendix). Second, contributions of the axes to the inertia of each point (row proportions in Exhibit 6.4) are squared angle cosines between the axes and the points, interpreted as squared correlations or as proportions of inertia explained at the point level rather than for all points together—these are denoted by the acronym COR in the **ca** package and also multiplied by 1000.

Data set "benthos"

This data set consists of 13 columns, the sites at which samples have been taken on the sea bed in the North Sea near an oilfield to study the effect of oil exploration on marine life. In each sample the benthic ("sea bed") species have been identified and counted, leading to an ecological abundance table where the large number of variables (the species) form the rows and the smaller number of samples the columns: in this case, 92 rows (species) by 13 columns (sites). Two of the sites, labelled R40 and R42, are reference stations far from the oilfield and regarded as an unpolluted environment. CA is regularly applied to such abundance tables to visualize the sites in relation to their species composition (the "column problem", the way the matrix is organized here) or to visualize the species' distribution across sites (the "row problem"). Exhibit 8.3 shows the column principal asymmetric map, where sites in principal coordinates are at weighted averages of species points in standard coordinates. The abbreviated species names are shown only if they contribute more than 1% to the two-dimensional map—referring to Exhibit 6.4, this percentage is calculated as $100(w_i f_{i1}^2 + w_i f_{i2}^2)/(\lambda_1 + \lambda_2)$. Of the 92 species, only 10 contribute more than 1% each, totalling 85% of the two-dimensional solution, while the remaining 82 collectively contribute only 15%.

Exhibit 8.3:
Column principal CA biplot of the "benthos" data, with columns (sites) in principal coordinates and rows (species) in standard coordinates. The 10 species with abbreviated labels each make a contribution of more than 1% to the solution, the others are indicated by dots. Total inertia is 0.783, with 57.5% explained in the biplot



Contribution CA biplot

The points that contribute highly to a CA map such as Exhibit 8.3 are generally the high-frequency points, while the low-frequency points contribute very little. The low frequency points, however, often have unusual profiles and lie on the periphery of the map, giving an impression of high importance—for example, in the "benthos" application a very rare species, occurring in just two or three sites, will have a profile at the outer reaches of the profile space. If we are interested only in the direction vectors for the species in the biplot, then this is an excellent situation to use the contribution biplot (see the end of Chapter 6). Rather than use the standard coordinates to represent the species, as in Exhibit 8.3, these coordinates are multiplied by the square roots of the corresponding species masses, causing rare species to be pulled towards the centre, as shown in Exhibit 8.4. The species vectors now show which ones are important to interpret, because their lengths now reflect the contributions of the species to the solution. Exhibit 8.4 shows which are the important species that separate out the unpolluted sites R40 and R42 to the right, while species *Chaetesona setosa* is generally found at polluted sites, particularly S15 which is close to the oilfield. There is a very high abundance of *Myriochele oculata* at site S24 which is not related to the pollution

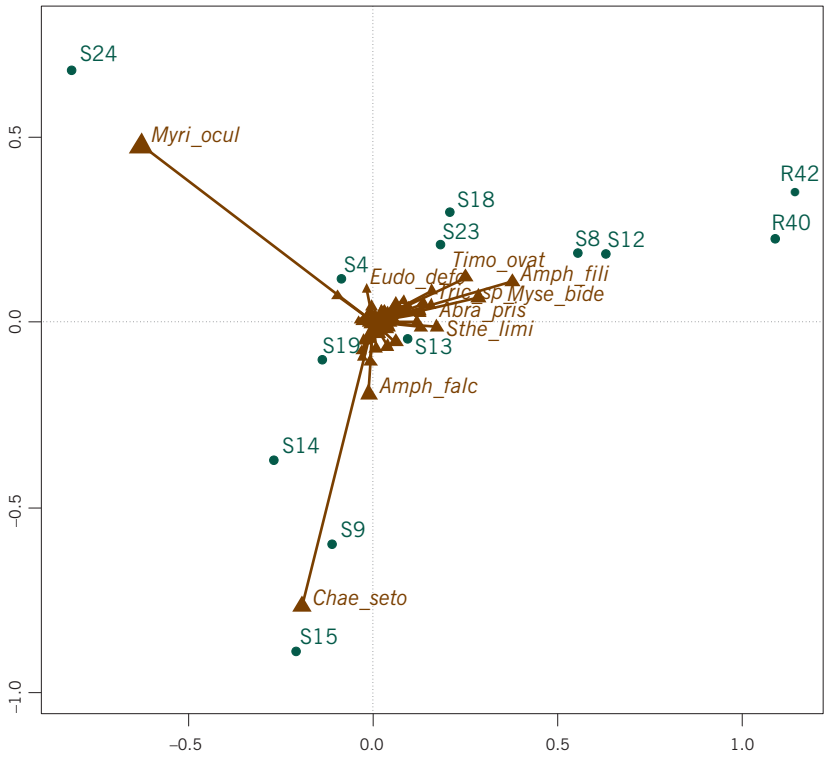


Exhibit 8.4:
Contribution CA biplot of the "benthos" data, with sites in principal coordinates and species in standard coordinates multiplied by the square roots of their masses. The position of each species on each axis is now directly related to its contribution to that axis. The 10 highly contributing species of Exhibit 8.3 (labelled) now stand out in the biplot and all the others collapse to the centre. In this graphic the size of the triangle at each species point, rather than the label, is related to the species total abundance level

gradient—this gradient emerges as a curve from the bottom (sites S15, S9 and S14) to the right (the reference stations).

As described earlier in this chapter, the coordinates of the species in this contribution biplot turn out to be the regression coefficients when the species are regressed on the two CA dimensions, where (i) the species "response" is defined by its elements in the matrix of site profiles, centred and normalized in a chi-square fashion, (ii) the "predictors" are the principal coordinates of the sites, and (iii) the regression is fitted by weighted least squares, using the site masses as weights. This is an additional interpretation of the contribution biplot in the case of CA, giving even more meaning to the positions of the species points in Exhibit 8.4.

SUMMARY: Correspondence Analysis Biplots

1. Correspondence analysis is applicable to a table of nonnegative data, the primary example being a cross-tabulation of two categorical variables, that is a contingency table.
2. The method can be thought of as an analysis of row or column *profiles* of the data matrix—these are the rows or columns expressed relative to their marginal totals.
3. Each profile receives a weight equal to the relative marginal total, called a *mass*.
4. Distances between profiles are defined by the *chi-square metric*. This is essentially a type of standardization of the profile values similar to that used in PCA, but using the average profile element as an estimate of variance rather than the variance itself.
5. The total variance, called *inertia*, in the data is numerically equal to the chi-square statistic for the table divided by the table's grand total.
6. Two types of asymmetric maps, both of which are biplots, are possible, depending on whether row or column profiles (and thus their interpoint chi-square distances) are visualized. Both are form biplots.
7. The contribution biplot can be particularly useful in CA applications, especially when there are quite different levels in rows or in columns (i.e., large differences in the masses). This biplot pulls in the points represented in standard coordinates by the square roots of their respective masses. For each such point, the squares of its rescaled coordinates are equal to the part contributions that the point makes to the respective principal axes.
8. In the contribution biplot, suppose that rows are in principal coordinates (i.e., row profiles are being visualized) and columns in "shrunken" standard coordinates. Then these latter coordinates for each column are also regression coefficients when the standardized values for that column in the row profile matrix are regressed on the principal coordinates of the rows, using weighted least squares with weights equal to the row masses.