Biplots in Practice

MICHAEL GREENACRE

Professor of Statistics at the Pompeu Fabra University

Chapter 6 Offprint

Principal Component Analysis Biplots

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

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

Chapter 6

Principal Component Analysis Biplots

Principal component analysis (PCA) is one of the most popular multivariate methods in a wide variety of research areas, ranging from physics to genomics and marketing. The origins of PCA can be traced back to early 20th century literature in biometrics (Karl Pearson) and psychometrics (Harold Hotelling). The method is inextricably linked to the singular value decomposition (SVD)—this powerful result in matrix theory provides the solution to the classical PCA problem and, conveniently for us, the solution is in a format leading directly to the biplot display. In this section we shall consider various applications of PCA and interpret the associated biplot displays. We will also introduce the notion of the contribution biplot, which is a variation of the biplot that will be especially useful when the rows and/or columns have different weights.

Contents

PCA of data set "attributes"

The last section of Chapter 5 defined a generalized form of PCA where rows and columns were weighted. If we consider the 13×6 data matrix of Exhibit 4.3, there is no need to differentially weight the rows or the columns: on the one hand, the countries should be treated equally, while on the other hand, the variables are all on the same 1 to 9 scale, and so there is no need to up- or downweight any variable with respect to the others (if variables were on different scales, the usual way to equalize out their roles in the analysis is to standardize them). So in this example all rows and all columns obtain the same weight, i.e. $\mathbf{w} = (1/13)\mathbf{1}$ and $\mathbf{q} = (1/6)\mathbf{1}$, where $\mathbf{1}$ is an appropriate vector of ones in each case. Referring to (5.9), the matrix $\mathbf{D}_q$ defines the metric between the row points (i.e., the countries in this example), so that distances between countries are the average squared differences between the six variables.

The computational steps are then the following, as laid out in the last section of Chapter 5. Here we use the general notation of a data matrix $\mathbf{X}$ with I rows and J columns, so that for the present example $I = 13$ and $J = 6$.

– Centring (cf. (5.8)): $\mathbf{Y} = [\mathbf{I} - (1/I)\mathbf{1}\mathbf{1}^\mathsf{T}]\mathbf{X}$ (6.1)

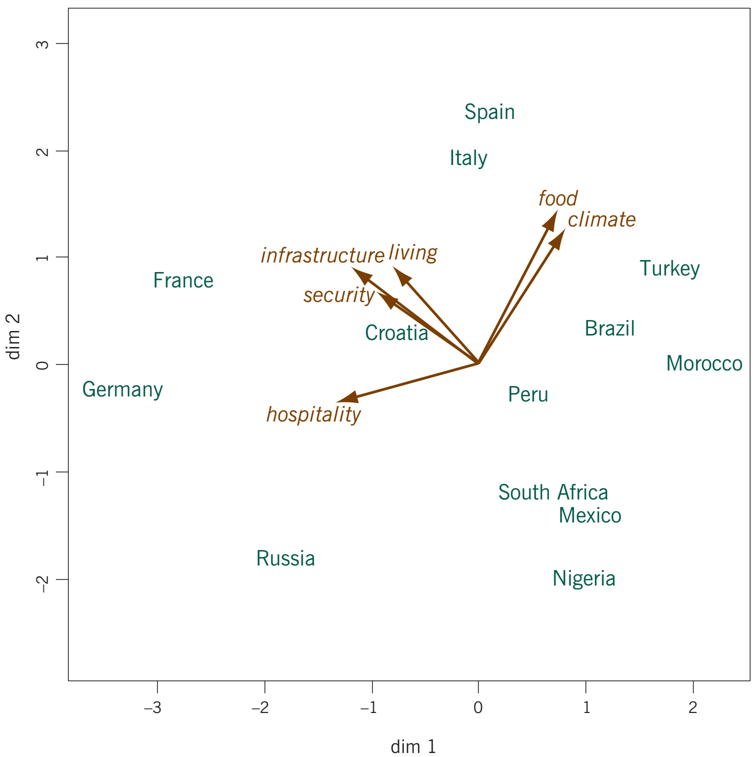
– Weighted SVD (cf. (5.4) and (5.5)):

$$\mathbf{S} = (1/I)^{½}\,\mathbf{Y}(1/J)^{½} = (1/IJ)^{½}\,\mathbf{Y} = \mathbf{U}\mathbf{D}_\beta\mathbf{V}^\mathsf{T} \qquad (6.2)$$

– Calculation of coordinates; i.e., the left and right cf. (5.6) and (5.10)):

$$\mathbf{F} = I^{½}\,\mathbf{U}\mathbf{D}_\beta \text{ and } \mathbf{\Gamma} = J^{½}\,\mathbf{V} \qquad (6.3)$$

Exhibit 6.1:
PCA biplot of the data in Exhibit 4.3, with the rows in principal coordinates, and the columns in standard coordinates, as given in (6.3). This is the row-metric-preserving biplot, or form biplot (explained on following page). Remember that the question about hospitality was worded negatively, so that the pole "friendly" is in the opposite direction to the vector "hospitality"—see Exhibit 4.3



We use the term "weighted SVD" above even though there is no differential weighting of the rows and columns: the whole of the centred matrix **Y** is simply multiplied by a constant, $(1/IJ)^{½}$. The resultant biplot is given in Exhibit 6.1.

Principal and standard coordinates

When the singular values are assigned totally to the left or to the right, the resultant coordinates are called *principal coordinates*. The other matrix, to which no part of the singular values is assigned, contains the so-called *standard coordinates*. The choice made in (6.3) is thus to plot the rows in principal coordinates and the columns in standard coordinates. From (6.3) it can be easily shown that the principal coordinates on a particular dimension have average sum of squared coordinates equal to the square of the corresponding singular value; for example, for **F** defined in (6.3):

$$\mathbf{F}^\mathsf{T}\mathbf{D}_w\mathbf{F} = (1/I)\ (I^{½}\mathbf{U}\mathbf{D}_\beta)^\mathsf{T}(I^{½}\mathbf{U}\mathbf{D}_\beta) = \mathbf{D}_\beta{}^\mathsf{T}\mathbf{U}^\mathsf{T}\mathbf{U}\mathbf{D}_\beta = \mathbf{D}_\beta^2 \tag{6.4}$$

By contrast, the standard coordinates on a particular dimension have average sum of squared coordinates equal to 1 (hence the term "standard"); for example, for $\mathbf{\Gamma}$ defined in (6.3):

$$\mathbf{\Gamma}^\mathsf{T}\mathbf{D}_q\mathbf{\Gamma} = (1/J)\,(J^{½}\mathbf{V})^\mathsf{T}(J^{½}\mathbf{V}) = \mathbf{V}^\mathsf{T}\mathbf{V} = \mathbf{I} \tag{6.5}$$

In this example the first two squared singular values are $\beta_1^2 = 2.752$ and $\beta_2^2 = 1.665$. If **F** has elements f_{ik}, then the normalization of the row principal coordinates on the first two dimensions is $(1/13)\sum_i f_{i1}^2 = 2.752$ and $(1/13)\sum_i f_{i2}^2 = 1.665$. For the column standard coordinates γ_{jk} the corresponding normalizations are $(1/6)\sum_j \gamma_{j1}^2 = 1$ and $(1/6)\sum_j \gamma_{j2}^2 = 1$.

Form biplot

There are various names in the literature for this type of biplot. It can be called the *row-metric-preserving biplot*, since the configuration of the row points approximates the interpoint distances between the rows of the data matrix. It is also called the *form biplot*, because the row configuration is an approximation of the *form matrix*, composed of all the scalar products $\mathbf{Y}\mathbf{D}_q\mathbf{Y}^\mathsf{T}$ of the rows of **Y**:

$$\mathbf{Y}\mathbf{D}_q\mathbf{Y}^\mathsf{T} = \mathbf{F}\mathbf{\Gamma}^\mathsf{T}\mathbf{D}_q\mathbf{\Gamma}\mathbf{F}^\mathsf{T} = \mathbf{F}\mathbf{F}^\mathsf{T} \tag{6.6}$$

In fact, it is the form matrix which is being optimally represented by the row points, and—by implication—the inter-row distances which depend on the scalar products.

Covariance biplot

If the singular values are assigned totally to the right singular vectors in (6.2), then we get an alternative biplot called the *covariance biplot*, because it shows the inter-variable covariance structure. It is also called the *column-metric-preserving biplot*. The left and right matrices are then defined as (cf. (6.3)):

– Coordinates in covariance biplot: $\boldsymbol{\Phi} = I^{½}\mathbf{U}$ and $\mathbf{G} = J^{½}\mathbf{V}\mathbf{D}_\beta$ (6.7)

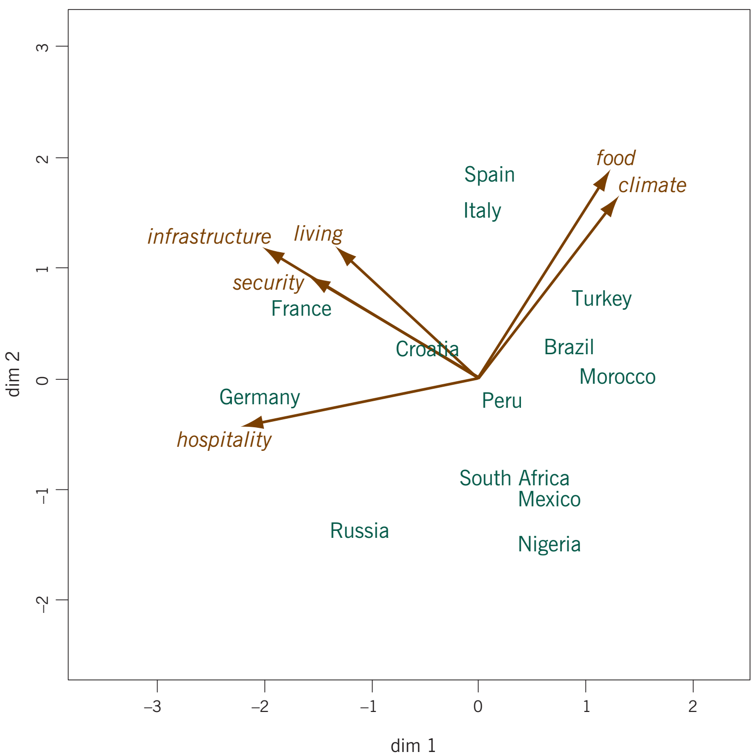
The covariance biplot is shown in Exhibit 6.2.

Apart from the changes in scale along the respective principal axes in the row and column configurations, this biplot hardly differs from the form biplot in Exhibit 6.1. In Exhibit 6.1 the countries have weighted sum of squares with respect to each axis equal to the corresponding squared singular value, while in Exhibit 6.2 it is the attributes that have weighted sum of squares equal to the squared singular values. In each biplot the other set of points has unit normalization on both principal axes.

In the covariance biplot the covariance matrix $\mathbf{Y}^\mathsf{T}\mathbf{D}_w\mathbf{Y} = (1/I)\mathbf{Y}^\mathsf{T}\mathbf{Y}$ between the variables is equal to the scalar product matrix between the column points using all the principal axes:

$$\mathbf{Y}^\mathsf{T}\mathbf{D}_w\mathbf{Y} = \mathbf{G}\boldsymbol{\Phi}^\mathsf{T}\mathbf{D}_w\boldsymbol{\Phi}\mathbf{G}^\mathsf{T} = \mathbf{G}\mathbf{G}^\mathsf{T} \tag{6.8}$$

Exhibit 6.2:
PCA biplot of the data in Exhibit 4.3, with the columns in principal coordinates, and the rows in standard coordinates, as given in (6.7). This is the column-metric-preserving biplot, or covariance biplot



and is thus approximated in a low-dimensional display using the major principal axes. Hence, the squared lengths of the vectors in Exhibit 6.2 approximate the variances of the corresponding variables—this approximation is said to be "from below", just like the approximation of distances in the classical scaling of Chapter 4. By implication it follows that the lengths of the vectors approximate the standard deviations and also that the cosines of the angles between vectors approximate the correlations between the variables. If the variables in **Y** were normalized to have unit variances, then the lengths of the variable vectors in the biplot would be less than one—a unit circle may then be drawn in the display, with vectors extending closer to the unit circle indicating variables that are better represented.

Connection with regression biplots

It can be easily shown that in both the form and covariance biplots the coordinates of the variables, usually depicted as vectors, are the regression coefficients of the variables on the dimensions of the biplot. For example, in the case of the covariance biplot, the regression coefficients of **Y** on $\mathbf{\Phi} = I^{½}\mathbf{U}$ are:

$$(\mathbf{\Phi}^{\mathsf{T}}\mathbf{\Phi})^{-1}\mathbf{\Phi}^{\mathsf{T}}\mathbf{Y} = (I\,\mathbf{U}^{\mathsf{T}}\mathbf{U})^{-1}(I^{½}\,\mathbf{U})^{\mathsf{T}}(IJ)^{½}\,\mathbf{U}\mathbf{D}_{\beta}\mathbf{V}^{\mathsf{T}} \qquad \text{(from (6.2))}$$

$$= J^{½}\mathbf{D}_{\beta}\mathbf{V}^{\mathsf{T}} \qquad \text{(because } \mathbf{U}^{\mathsf{T}}\mathbf{U} = \mathbf{I}\text{)}$$

which is the transpose of the coordinate matrix **G** in (6.7). In this case the regression coefficients are not correlations between the variables and the axes (see Chapter 2), because the variables in **Y** are not standardized—instead, the regression coefficients are the covariances between the variables and the axes. These covariances are equal to the correlations multiplied by the standard deviations of the respective variables. Notice that in the calculation of covariances and standard deviations, sums of squares should be divided by I (=13 in this example) and not by $I-1$ as in the usual computation of sample variance.

Dual biplots

The form biplot and the covariance biplot defined above are called *dual biplots*: each is the *dual* of the other. Technically, the only difference between them is the way the singular values are allocated, either to the left singular vectors in the form biplot, which thus visualizes the spatial form of the rows (cases), or to the right singular vectors in the covariance biplot, visualizing the covariance structure of the columns (variables). Substantively, there is a big difference between these two options, even though they look so similar. We shall see throughout the following chapters that dual biplots exist for all the multivariate situations treated. An alternative display, especially prevalent in correspondence analysis (Chapter 8), represents both sets of points in principal coordinates, thus displaying row and column structures simultaneously, that is both row- and column-metric-preserving. The additional benefit is that both sets of points have the same dispersion along the principal axes and avoid the large differences in scale that are sometimes observed between principal and standard coordinates. However, this choice is not a biplot and its benefits go at the expense of losing the scalar-product interpretation and the ability to project one set of points onto the other. This loss is not so great when the singular values on the two axes being displayed are close to each other—in fact, the closer they are to each other, the more the scalar-product interpretation remains valid and intact. But if there is a large difference between the singular values, then the scalar-product approximation of the data becomes degraded (see the Epilogue for further discussion of this point).

Squared singular values are eigenvalues

From (6.6) and (6.8) it can be shown that the squared singular values are eigenvalues. If (6.6) is multiplied on the right by $\mathbf{D}_w\mathbf{F}$ and (6.8) is similarly multiplied on the right by $\mathbf{D}_q\mathbf{G}$, and using the normalizations of $\mathbf{F}$ and $\mathbf{G}$, the following pair of eigenequations is obtained:

$$\mathbf{Y}\mathbf{D}_q\mathbf{Y}^\mathsf{T}\mathbf{D}_w\mathbf{F} = \mathbf{F}\mathbf{D}_\beta^2 \quad \text{and} \quad \mathbf{Y}^\mathsf{T}\mathbf{D}_w\mathbf{Y}\mathbf{D}_q\mathbf{G} = \mathbf{G}\mathbf{D}_\beta^2, \quad \text{where} \quad \mathbf{F}^\mathsf{T}\mathbf{D}_w\mathbf{F} = \mathbf{G}^\mathsf{T}\mathbf{D}_q\mathbf{G} = \mathbf{D}_\beta^2 \qquad (6.9)$$

The matrices $\mathbf{Y}\mathbf{D}_q\mathbf{Y}^\mathsf{T}\mathbf{D}_w$ and $\mathbf{Y}^\mathsf{T}\mathbf{D}_w\mathbf{Y}\mathbf{D}_q$ are square but non-symmetric. They are easily symmetrized by writing them in the equivalent form:

$$\mathbf{D}_w^{\frac{1}{2}}\mathbf{Y}\mathbf{D}_q\mathbf{Y}^\mathsf{T}\mathbf{D}_w^{\frac{1}{2}}\mathbf{D}_w^{\frac{1}{2}}\mathbf{F} = \mathbf{D}_w^{\frac{1}{2}}\mathbf{F}\mathbf{D}_\beta^2 \quad \text{and} \quad \mathbf{D}_q^{\frac{1}{2}}\mathbf{Y}^\mathsf{T}\mathbf{D}_w\mathbf{Y}\mathbf{D}_q^{\frac{1}{2}}\mathbf{D}_q^{\frac{1}{2}}\mathbf{G} = \mathbf{D}_q^{\frac{1}{2}}\mathbf{G}\mathbf{D}_\beta^2$$

which in turn can be written as:

$$\mathbf{S}\mathbf{S}^\mathsf{T}\mathbf{U} = \mathbf{D}_\beta^2 \quad \text{and} \quad \mathbf{S}^\mathsf{T}\mathbf{S}\mathbf{V} = \mathbf{V}\mathbf{D}_\beta^2, \quad \text{where} \quad \mathbf{U}^\mathsf{T}\mathbf{U} = \mathbf{V}^\mathsf{T}\mathbf{V} = \mathbf{I} \qquad (6.10)$$

These are two symmetric eigenequations with eigenvalues $\lambda_k = \beta_k^2$ for $k = 1, 2, \ldots$

Scree plot

The eigenvalues (i.e., squared singular values) are the primary numerical diagnostics for assessing how well the data matrix is represented in the biplot. They are customarily expressed relative to the total sum of squares of all the singular values—this sum quantifies the total variance in the matrix, that is the sum of squares of the matrix decomposed by the SVD (the matrix $\mathbf{S}$ in (6.2) in this example). The values of the eigenvalues and a bar chart of their percentages of the total are given in Exhibit 6.3—this is called a *scree plot*.

It is clear that the first two values explain the major part of the variance, 57.6% and 34.8% respectively, which means that the biplots in Exhibits 6.1 and 6.2 explain 92.4% of the variance. The pattern in the sequence of eigenvalues in the bar chart is typical of almost all matrix approximations in practice: there are a few

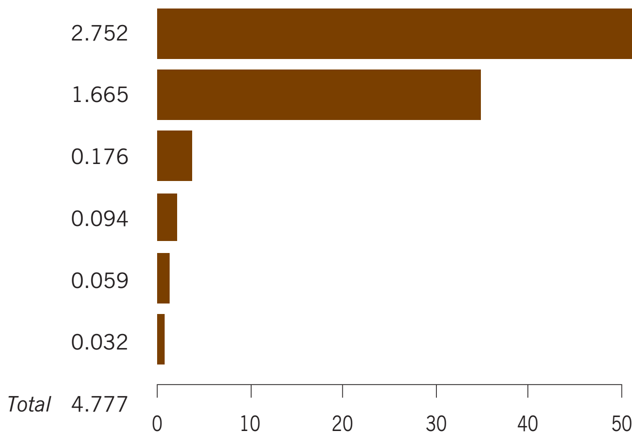


Exhibit 6.3:
Scree plot of the six squared singular values $\lambda_1, \lambda_2, \ldots, \lambda_6$, and a horizontal bar chart of their percentages relative to their total

eigenvalues that dominate and separate themselves from the remainder, with this remaining set showing a slow "dying out" pattern associated with "noise" in the data that has no structure. The point at which the separation occurs (between the second and third values in Exhibit 6.3) is often called the *elbow* of the scree plot. Other rules of thumb for deciding which axes reflect "signal" in the data, as opposed to "noise", is to calculate the average variance per axis, in this case 4.777/6 = 0.796. Axes with eigenvalues greater than the average are generally considered worth plotting.

Contributions to variance

Just like the eigenvalues quantify how much variance is accounted for by each principal axis, usually expressed as a percentage, so we can decompose the vari-

$$
\begin{array}{cc|cccc|c}
 & & & p \text{ dimensions} & & & \\
 & & 1 & 2 & \cdots & p & \\
\hline
 & 1 & w_1 f_{11}^2 & w_1 f_{12}^2 & \cdots & w_1 f_{1p}^2 & w_1 \Sigma_k f_{1k}^2 \\
 & 2 & w_2 f_{21}^2 & w_2 f_{22}^2 & \cdots & w_2 f_{2p}^2 & w_2 \Sigma_k f_{2k}^2 \\
 & \cdot & \cdot & \cdot & & \cdot & \cdot \\
n \text{ row points} & \cdot & \cdot & \cdot & & \cdot & \cdot \\
 & \cdot & \cdot & \cdot & & \cdot & \cdot \\
 & \cdot & \cdot & \cdot & & \cdot & \cdot \\
 & n & w_n f_{n1}^2 & w_n f_{n2}^2 & \cdots & w_n f_{np}^2 & w_n \Sigma_k f_{nk}^2 \\
\hline
 & & \lambda_1 & \lambda_2 & \cdots & \lambda_p & \text{Total variance}
\end{array}
$$

Exhibit 6.4:
Decomposition of total variance by dimensions and points: the row sums are the variances of the row points and the columns sums are the variances of the dimensions

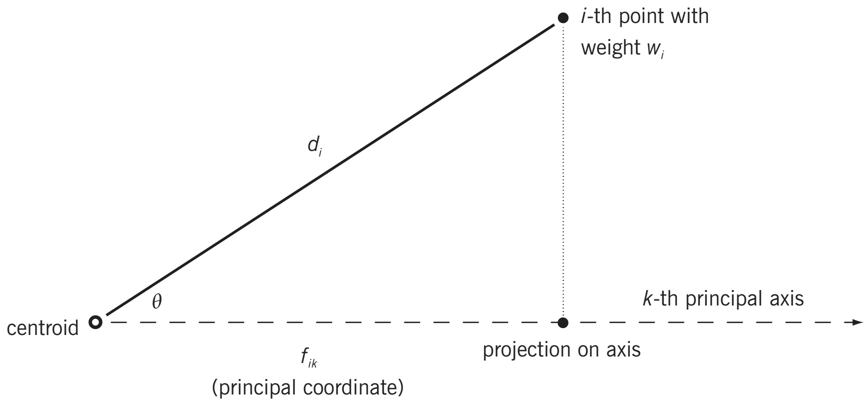


Exhibit 6.5: *Geometry of variance contributions: f_{ik} is the principal coordinate of the i-th point, with weight w_i, on the k-th principal axis. The point is at distance $d_i = \sum_k f_{ik}^2$ from the centroid of the points, which is the origin of the display, and θ is the angle between the point vector (in the full space) and the principal axis. The square cosine of θ is $\cos^2(\theta) = f_{ik}^2 / d_i^2$ (i.e., the proportion of point i's variance accounted for by axis k) and $w_i f_{i1}^2$ is the contribution of the i-th point to the variance on the k-th axis*

ance of each individual point, row or column, along principal axes. But we can also decompose the variance on each axis in terms of contributions made by each point. This leads to two sets of contributions to variance, which constitute important numerical diagnostics for the biplot. These contributions are best understood in terms of the principal coordinates of the rows and columns (e.g., **F** and **G** defined above). For example, writing the elements $w_i f_{ik}^2$ in an $n \times p$ matrix (n rows, p dimensions) as shown in Exhibit 6.4. The solution in two dimensions, for example, displays $100(\lambda_1 + \lambda_2)/(\lambda_1 + \lambda_2 + \cdots + \lambda_p)\%$ of the total variance. On the first dimension $100\, w_i f_{i1}^2/\lambda_1\%$ of this dimension's variance is accounted for by point i (similarly for the second dimension)—this involves expressing each column in Exhibit 6.4 relative to its sum. Conversely, expressing each row of the table relative to its row sum, $100\, w_i f_{i1}^2 / w_i \sum_k f_{ik}^2\%$ of point i's variance is accounted for by the first dimension and $100\, w_i f_{i2}^2 / w_i \sum_k f_{ik}^2\%$ is accounted for by the second dimension. Notice that in this latter case (row elements relative to row sums) the row weights cancel out. The ratio $w_i f_{i1}^2 / w_i \sum_k f_{ik}^2$, for example, equals $f_{i1}^2 / \sum_k f_{ik}^2$, which is equal to the squared cosine of the angle between the i-th row point and the first principal axis, as illustrated in Exhibit 6.5.

The contribution biplot

In the principal component biplots defined in (6.3) and (6.7) the points in standard coordinates are related to their contributions to the principal axes. The following result can be easily shown. Suppose that we rescale the standard coordinates by the square roots of the respective point weights, that is we recover the corresponding singular vectors:

$$\text{from (6.3): } (1/J)^{½}\boldsymbol{\Gamma} = (1/J)^{½} J^{½}\mathbf{V} = \mathbf{V},$$

$$\text{and from (6.7): } (1/I)^{½}\boldsymbol{\Phi} = (1/I)^{½} I^{½}\mathbf{U} = \mathbf{U}$$

Then these coordinates display the relative values of the contributions of the variables and cases respectively to the principal axes. In the covariance biplot, for example, the squared value u_{ik}^2 (where u_{ik} is the (i,k)-th element of the singular vector, or rescaled standard coordinate of case i on axis k) is equal to $(w_i f_{ik}^2)/\lambda_k$, where $w_i = 1/I$ here. This variant of the covariance biplot (plotting **G** and **U** together) or the form biplot (plotting **F** and **V** together) is known as the *contribution biplot*[4] and is particularly useful for displaying the variables. For example, plotting **F** and **V** jointly does not change the direction of the biplot axes, but changes the lengths of the displayed vectors so that they have a specific interpretation in terms of the contributions. In Exhibit 6.1, since the weights are all equal, the lengths of the variables along the principal axes are directly related to their contributions to the axes, since they just need to be multiplied by a constant $(1/J)^{½} = (1/6)^{½} = 0.41$. For PCA this is a trivial variation of the original definition —because the point masses are equal, it just involves an overall rescaling of the standard coordinates. But in other methods where the point masses are different, for example in log-ratio analysis and correspondence analysis, this alternative biplot will prove to be very useful—we will return to this subject in the following chapters.

SUMMARY: Principal Component Analysis Biplots

1. Principal component analysis of a cases-by-variables matrix reduces to a singular value decomposition of the centred (and optionally variable-standardized) data matrix.

2. Two types of biplot are possible, depending on the assignment of the singular values to the left or right singular values of the decomposition. In both the projections of one set of points on the other approximate the centred (and optionally standardized) data.

3. The *form biplot*, where singular values are assigned to the left vectors corresponding to the cases, displays approximate Euclidean distances between the cases.

4. The *covariance biplot*, where singular values are assigned to the right vectors corresponding to the variables, displays approximate standard deviations and correlations of the variables. If the variables had been pre-standardized to have standard deviation equal to 1, a unit circle is often drawn on the covariance biplot because the variable points all have lengths less than or equal to 1 —the closer a variable point is to the unit circle, the better it is being displayed.

4. This variation of the scaling of the biplot is also called the *standard biplot* because the projections of points onto vectors are approximations to the data with variables on standard scales, and in addition because it can be used across a wide spectrum of methods and wide range of inherent variances.

5. The *contribution biplot* is a variant of the form or covariance biplots where the points in standard coordinates are rescaled by the square roots of the weights of the respective points. These rescaled coordinates are exactly the square roots of the part contributions of the respective points to the principal axes, so this biplot gives an immediate idea of which cases or variables are most responsible for the given display.