Biplots in Practice

MICHAEL GREENACRE

Professor of Statistics at the Pompeu Fabra University

Appendix D Offprint

Epilogue

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

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

Epilogue

Up to now this book has presented known facts about the theory and practice of biplots. In this final section I give my personal opinions about biplots and their use in practice. For, as the title of the book declares, this book is mainly about the practice and indeed the usefulness of this method as a research tool. I start off with a reflection of what the term "biplot" means and then treat some specific aspects which have a greater or lesser repercussion when it comes to practical applications.

What constitutes a biplot?

In my understanding of the term, a biplot is a representation of the rows and the columns of a data matrix in a joint display, with few dimensions, usually two-dimensional but nowadays possibly three-dimensional when viewed with special software such as R's **rgl** package. Because of the orthogonality and nested property of the principal axes of a biplot, further dimensions can be studied separately, for example, by considering axes 3 and 4 of the solution space in a planar display (see Exhibit 14.11, for example, where different planar projections were displayed). The essential feature of the biplot is that it displays scalar products between row and column points of a target matrix (the data matrix, appropriately centred and normalized), according to the fundamental result in Gabriel's original paper, formulated in Chapter 1 as:

$$\textit{target matrix} = \textit{left matrix} \cdot \textit{right matrix}$$

(see also the abstract from Ruben Gabriel's original article, which is reproduced in the Bibliography). The left and right matrices of low rank (dimensionality), obtained conveniently from the singular value decomposition (SVD), provide respective row and column coordinates that are used to plot the rows and columns as points or vectors as the case may be.

Let us suppose for convenience of description that the rows have been plotted as points and the columns as vectors drawn from the origin of the display. The idea of plotting columns as vectors gives the idea that each column has been regressed on the axes of the biplot and the vector actually represents the regression plane (or hyperplane for a three-dimensional biplot). This plane is defined uniquely by the vector that indicates the direction of steepest ascent of the plane, that is the gradient vector with elements equal to the regression coefficients. Since contours

(or isolines) of a plane are at right-angles to this gradient vector, estimated values of each row for that column can be obtained by projecting them onto the biplot axis through the vector.

These are thus the basic properties of a biplot. Variations exist, for example the row points could be approximating a particular interpoint distance, or they could be standardized along principal axes. Then there are nonlinear transformations of the data, which induce biplot axes with nonlinear scales, or weighting of the points when determining the solution space, but the biplot basically results in two sets of points, one of which is optionally drawn as a set of vectors.

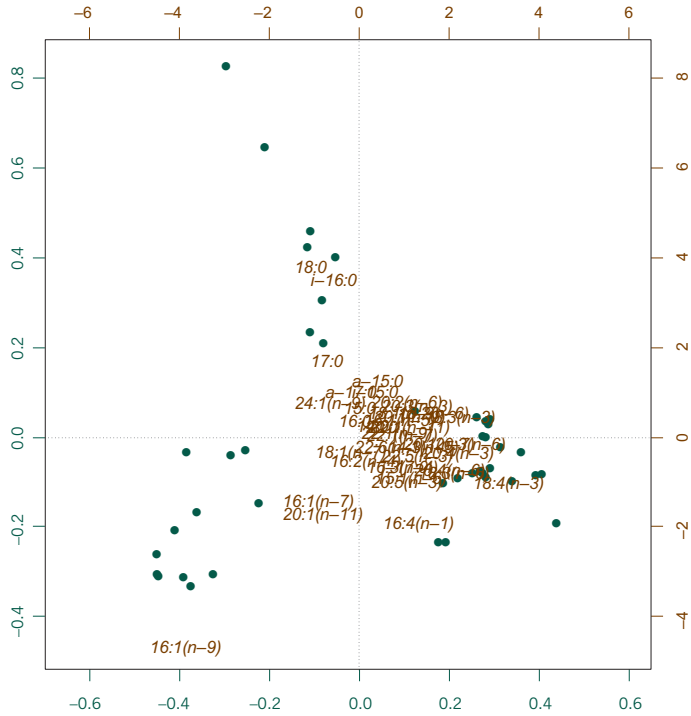
Calibration of biplot axes

The idea of calibrating biplot axes adds understanding about how the biplot works and how to interpret it. As shown in Chapters 2 and 3, adding tic marks to a biplot axis that passes through a biplot vector and then calibrating the axis according to the original scale of the corresponding variable, gives insight to what the biplot vector actually represents. But, when a user has digested this fact once and for all, I can see no practical purpose in leaving the calibrations on the final analyses reported in research findings. Firstly, it is only possible to include calibrations when biplot axes are few, which is seldom the case apart from small educational examples; secondly, they clutter up the display and detract from its simplicity; and thirdly, they do not allow one to define a meaningful length of the biplot vector, which can be a useful aspect of the biplot's interpretation (see, for example, the contribution biplot, discussed again below).

Good biplot design

The comments above about calibrated biplots lead naturally into the subject of what constitutes good graphical design for a biplot display. The objective should be to include as little as possible, but enough for the user to make a correct interpretation of what is presented. Overloading the biplot with calibrations, for example, for every biplot axis is not necessary, since it is known that the centre of the biplot represents the (weighted) average of each variable (data matrices are almost always centred) and all one needs to see is how the points line up on a biplot axis above and below their average. Knowing the scale of the biplot is relevant, but the tic marks and calibrations on the principal axes should be few and discrete; sometimes we have added principal variances (or inertias) and their percentages to axes, or simply mentioned these in the text or caption. Including what Tufte calls "chartjunk" just increases the biplot's ink-to-information ratio unnecessarily, but things like coloured labels, symbols of different sizes and textures are all useful instruments for communicating more about the plot. Omitting labels of uninteresting points is also a useful strategy. Consider the two versions of the same correspondence analysis solution in Exhibit D.1, performed on a table of fatty acid compositions for a sample of fish. The upper biplot is the asymmetric one with row points (fish, displayed by dots) in principal coordinates and column

a)



b)

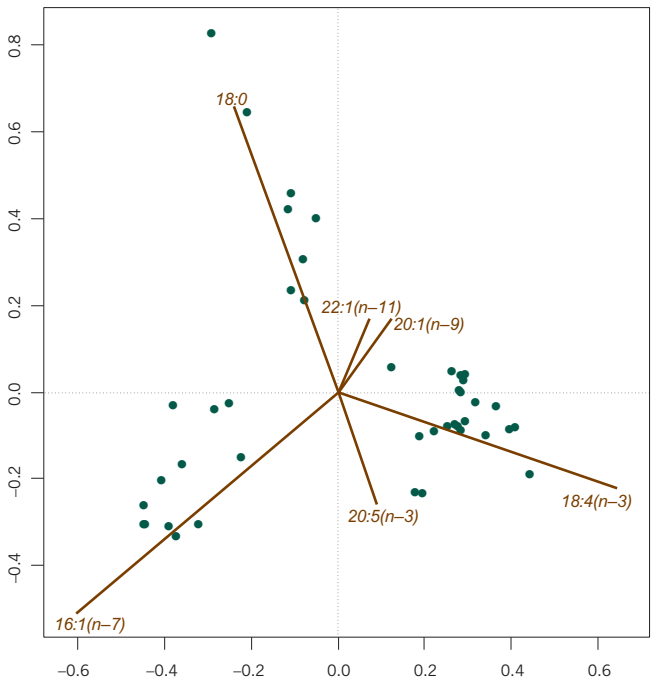


Exhibit D.1:

Two versions of the same correspondence analysis of a data set of compositions of 40 fatty acids in 42 fish: on top, the asymmetric ("rowprincipal") biplot, including all the points; at the bottom, the contribution biplot, with only the most contributing fatty acids shown

points in standard coordinates, and because of the very low inertia in these data two scales are necessary. This biplot is very cluttered with points and the fact that there are three groups of fish is partly obscured; also, it would be even worse if we added the vectors to the variable points (the fatty acids). The lower biplot is the contribution biplot, which only needs one scale and where only fatty acids are displayed that contribute more than average to the principal axes, which gives a much cleaner and easier to interpret solution. Only six out of the original 40 fatty acids remain and it is seen clearly that there are mainly three fatty acids, each associated with one of the three groups. Furthermore, the fatty acid *16:1(n–7)* which is responsible for characterizing the group of fish at bottom left is not at all clear in the upper biplot, where one might have thought that the most important one was *16:1(n–9).* Since these six fatty acids are the major contributors, the biplot would remain almost the same if the other 34 fatty acids were removed from the data set and a subset correspondence analysis performed.

Quality of a biplot

The dimension-reduction step is necessary to be able to visualize a high-dimensional data set in a few dimensions, but variance (or inertia) is lost in the process. There is a measure of variance retained and variance lost, in raw amounts or percentages, and also numerical diagnostics for how much variance of each point, row or column, is retained in the solution and how much is lost. On the other hand, the solution space is determined by different rows and columns in varying amounts. A point may be displayed accurately in a biplot, with little variance lost, but it could have played almost no role in determining the solution (the reverse is not true: points that generally determine the solution are usually quite accurately displayed). The contributions, both of the solution to the variances of individual points and of the points to the solution space, are important numerical diagnostics that support the interpretation of the biplot.

The contribution biplot

The idea of incorporating the contribution of a column (for example) into the length of its corresponding vector is, in my opinion, one of the most important variations of the biplot display. Users have difficulty in deciding which vectors are important for the interpretation of the biplot, so by rescaling individual vectors to correspond to their part contributions to the principal axes, the important vectors are immediately made more evident to the user, since their lengths along principal axes are the longest.

So why do we not always use the contribution biplot? The answer is, very simply, that in gaining this property of the interpretation, another property is inevitably lost. For example, in the correspondence analysis (CA) of the "benthos" data set, the standard coordinates of the species points indicate vertex, or extreme unit profile, positions and each sample point lies at a weighted average of the species points, using the sample's profile elements across species as weights (see Exhibit

8.3)—this is often called the *barycentric property* of CA. When the standard coordinates are multiplied by the square roots of their masses to reduce them to their contribution coordinates, this property is lost but now the main contributing species are visible. In the log-ratio analysis of the data set "morphology" in Chapter 7, the ability to detect equilibrium relationships when variables fall on straight lines (see Exhibit 7.3) would clearly be lost if each variable were rescaled into its position in terms of contribution coordinates. So, as I said in the Epilogue to *Correspondence Analysis in Practice*, you cannot "have your cake and eat it too"—all the desirable properties one might like to have in a biplot cannot be included in one display, although we can introduce additional graphical "tricks" such as omitting the labels of low contributing points and making the size of symbols related to an omitted aspect of the data, such as the point masses (see Exhibit 8.3, for example).

Computation of a biplot

The solution of a biplot is found by performing a weighted least-squares fit of the product of the left and right matrices to the target matrix, a solution that is conveniently encapsulated in the SVD. One way of computing the SVD is by a process known as *alternating least squares*. Suppose that the target matrix is $\mathbf{S}$ and the approximation $\mathbf{S} \approx \mathbf{X}\mathbf{Y}^\mathsf{T}$ is sought. Writing this approximation as an equality by including a matrix of residuals, or "errors":

$$\mathbf{S} = \mathbf{X}\mathbf{Y}^\mathsf{T} + \mathbf{E}$$

is recognizable as a regression problem if either $\mathbf{X}$ or $\mathbf{Y}^\mathsf{T}$ is fixed. For example, for a fixed $\mathbf{X}$ the least-squares solution for $\mathbf{Y}^\mathsf{T}$ is $(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{S}$. If there are weights associated with the rows of the matrix, where the weights are in the diagonal matrix $\mathbf{D}_w$, then the weighted least-squares solution for $\mathbf{Y}^\mathsf{T}$ would be $(\mathbf{X}^\mathsf{T}\mathbf{D}_w\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{D}_w\mathbf{S}$. Having estimated $\mathbf{Y}^\mathsf{T}$, it is regarded as fixed and a similar regression, with or without weights, is performed to estimate a new $\mathbf{X}$. From step to step it can be proved that the residual sum of squared errors reduces, so the process is repeated until the solution converges. At each step the estimates have to be orthonormalized: for example, the left matrix $\mathbf{X}$ would be orthonormalized so that $\mathbf{X}^\mathsf{T}\mathbf{D}_w\mathbf{X} = \mathbf{I}$, which means that the regression step is just a matrix multiplication.[10] The main point is that the left and right matrices are solutions of alternative regressions, with or without weights.

10. This is the more complicated aspect of this algorithm, which we omit here. In practice the dimensions can be computed one at a time: start with any vector $\mathbf{x}$ with the same number of elements as rows of $\mathbf{S}$ and which has been normalized as $\mathbf{x}^\mathsf{T}\mathbf{x} = 1$, then a solution for $\mathbf{y}$ simplifies as $\mathbf{S}^\mathsf{T}\mathbf{x}$; then normalize $\mathbf{y}$ so that $\mathbf{y}^\mathsf{T}\mathbf{y} = 1$ and estimate $\mathbf{x}$ as $\mathbf{S}\mathbf{y}$; normalize $\mathbf{x}$ and continue this process until convergence, which gives the first pair of singular vectors, the singular value α being the norm of the final $\mathbf{x}$ or $\mathbf{y}$ before being normalized to 1. This first dimension is then subtracted out from $\mathbf{S}$,. i.e. $\mathbf{S}$ is replaced by $\mathbf{S} - \alpha\mathbf{x}\mathbf{y}^\mathsf{T}$ (using normalized $\mathbf{x}$ and $\mathbf{y}$) and the process is repeated to find the solution for the second dimension. The only difference when weighted solutions are required is to normalize $\mathbf{x}$ and $\mathbf{y}$ at each step using the weights, and use weighted least-squares regression in each step, which leads to the solution of a weighted SVD.

The optimality of a biplot

Since there are variations of the biplot display, the question arises as to how each variation approximates the original data. The answer is quite simply that the approximation is always the same, with the definition of row and column weights depending on the scaling of the biplot coordinates. The different biplots of a CA illustrate what is meant. In Chapter 8, equation (8.2) defined correspondence analysis in terms of the regular unweighted SVD of the matrix of standardized residuals $\mathbf{D}_r^{-½}(\mathbf{P} - \mathbf{r}\mathbf{c}^\mathsf{T})\mathbf{D}_c^{-½}$:

$$\mathbf{D}_r^{-½}(\mathbf{P} - \mathbf{r}\mathbf{c}^\mathsf{T})\mathbf{D}_c^{-½} = \mathbf{U}\mathbf{D}_\alpha\mathbf{V}^\mathsf{T}, \text{ where } \mathbf{U}^\mathsf{T}\mathbf{U} = \mathbf{V}^\mathsf{T}\mathbf{V} = \mathbf{I}$$

(notice here that the identification conditions on $\mathbf{U}$ and $\mathbf{V}$ do not contain weights). If we were interested in biplotting the standardized residuals themselves, we would use coordinate matrices such as $\mathbf{U}\mathbf{D}_\alpha$ and $\mathbf{V}$, or $\mathbf{U}\mathbf{D}_\alpha^{½}$ and $\mathbf{V}\mathbf{D}_\alpha^{½}$ for example. But CA is not biplotting the standardized residuals—in the case of the asymmetric CA biplots, for example, one set of points is plotted in principal coordinates and the other set in standard coordinates, for example $\mathbf{F} = \mathbf{D}_r^{-½}\mathbf{U}\mathbf{D}_\alpha$ for rows and $\mathbf{\Gamma} = \mathbf{D}_c^{-½}\mathbf{V}$ for columns, or $\mathbf{\Phi} = \mathbf{D}_r^{-½}\mathbf{U}$ for rows and $\mathbf{G} = \mathbf{D}_c^{-½}\mathbf{V}\mathbf{D}_\alpha$ for columns—see the definitions in (8.3) and (8.4). In order to have the corresponding matrix of scalar products on the right hand side of the CA definition, the target matrix in the defining equation becomes:

$$\mathbf{D}_r^{-1}\mathbf{P}\mathbf{D}_c^{-1} - \mathbf{1}\mathbf{1}^\mathsf{T} = \mathbf{\Phi}\mathbf{D}_\alpha\mathbf{\Gamma}^\mathsf{T}, \text{ where } \mathbf{\Phi}^\mathsf{T}\mathbf{D}_r\mathbf{\Phi} = \mathbf{\Gamma}^\mathsf{T}\mathbf{D}_c\mathbf{\Gamma} = \mathbf{I}$$

i.e., in various scalar forms: $p_{ij}/(r_ic_j) - 1 = (p_{ij}/r_i - c_j)/c_j = (p_{ij}/c_j - r_i)/r_i = \sum_k \alpha_k \phi_{ik}\gamma_{jk}$ (the first form shows the ratio of the data element to its expected value, while the second and third forms show how the asymmetric map represents the profiles' deviations from their expected values relative to their expected values). This defines a *generalized* (or *weighted*) SVD where the rows and columns are weighted by the row masses $\mathbf{r}$ and column masses $\mathbf{c}$ respectively. Associating the singular values with the left or right singular vectors in this version of the definition (the standard coordinates) will give the two types of asymmetric biplot and the low-dimensional approximation of the scalar products to the target matrix is by weighted least-squares using the masses.

The form of the definition for the contribution biplot, where the column points, for example, are rescaled by the square roots of their respective masses, plots $\mathbf{F}$ and $\mathbf{V}$ jointly. This implies the generalized SVD in terms of the row profiles $\mathbf{D}_r^{-1}\mathbf{P}$:

$$(\mathbf{D}_r^{-1}\mathbf{P} - \mathbf{1}\mathbf{c}^\mathsf{T})\mathbf{D}_c^{-½} = \mathbf{\Phi}\mathbf{D}_\alpha\mathbf{V}^\mathsf{T}, \text{ where } \mathbf{\Phi}^\mathsf{T}\mathbf{D}_r\mathbf{\Phi} = \mathbf{V}^\mathsf{T}\mathbf{V} = \mathbf{I}$$

that is, with rows weighted by the row masses $\mathbf{r}$, and columns unweighted. The target matrix consists of the standardized profile elements $(p_{ij}/r_i - c_j)/c_j^{½}$, hence the

alternative name of standard biplot for the contribution biplot. The weighting makes sense since the row profiles should be weighted but not the columns, which have already been standardized.

Nonlinear biplots

Another development of the biplot is that of so-called "nonlinear" biplots, where variables are represented by curves and estimation of the values for each case on a particular variable is performed by finding the point on the curve closest to the case. While being of theoretical interest, nonlinear biplots are unlikely to find favour amongst users because of their complexity of interpretation, which detracts from the simplicity of the "linear" biplot treated in this book and its desirable properties such as decomposition of variance, nesting of dimensions and parallel contours for each biplot axis. It is extremely difficult to make any deductions about the properties of variables and their inter-relationships when they are represented by different curves. My view is that it is much better, from a practical point of view, to consider appropriate nonlinear transformations of the original variables, which users and non-specialists can understand, and then use the linear biplot, bearing in mind the nonlinear scales of the resulting biplot axes.

The future of biplots

The SVD as a mathematical result has been known for more than a century, and its property of identifying matrix approximations of any desired rank by (weighted) least squares makes it the most useful matrix result in the area of multivariate data analysis. Algorithms for computing the SVD are well-researched and provide global optima for the dimension-reducing methods that have been presented in this book. The SVD has appeared, is appearing and will appear in every area of scientific research where tabular data are collected. Wherever there is an SVD, there is a biplot. Data are often collected by a painstaking and expensive process and I have always thought it a pity that the richness of a data set is not fully exposed to the researcher who has taken so much trouble to collect it. The biplot is a tool for exploring complex data sets of all types and sizes. The future of the biplot is in its further application in many different areas of research, to make data more transparent to the researcher, and to assist in the interpretation and in the discovery of structures and patterns, both suspected and unsuspected.