Biplots in Practice

MICHAEL GREENACRE

Professor of Statistics at the Pompeu Fabra University

Chapter 4 Offprint

Multidimensional Scaling Biplots

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

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

CHAPTER 4

Multidimensional Scaling Biplots

Multidimensional scaling is the graphical representation of a set of objects based on their interpoint distances. The method originated in psychological experiments where people were asked to judge the similarities of the objects, examples of which were a set of paintings, a set of writings, a set of countries or a set of products. This approach is often called *perceptual mapping* because it results in a spatial map of the respondents' perceptions of the set of objects. These maps are multidimensional, although they are usually represented in their best two-dimensional aspects, appearing like a scatterplot of the objects. In this map the horizontal and vertical axes are known as *principal axes*, which are artificially created to provide the support on which the objects are represented. If in addition there are variables characterizing the set of objects, then these variables can be added as biplot axes to the multidimensional scaling map.

Contents

Multidimensional scaling—data set "countries"

In the first class that I give in my postgraduate course "Methods of Marketing Research", I ask the students, who usually come from a number of different countries, to evaluate the similarities and differences between these countries on a scale from 1 (most similar) to 9 (most different). Exhibit 4.1 is an example of a table given by one of the students, with initials MT, for the 13 countries represented in that particular class. A low value given to a pair of countries, for example a 2 between Italy and Spain, means that MT perceives these countries as being very similar to one another, whereas a high value, for example a 9 between Russia and Spain, means that he perceives them to be very different. The idea in *multidimensional scaling* (MDS) is to represent the countries in a spatial map such that the physical

Exhibit 4.1:
Student MT's ratings of the similarities/differences between 13 countries, on a scale 1 = most similar to 9 = most different. The column labels are the international codes for the countries used in the MDS maps

Countries	I	E	HR	BR	RU	D	TR	MA	PE	NG	F	MX	ZA
Italy	0	2	5	5	8	7	3	5	6	8	4	5	7
Spain	2	0	5	4	9	7	4	7	3	8	4	4	6
Croatia	5	5	0	7	4	3	6	7	7	8	4	6	7
Brazil	5	4	7	0	9	8	3	6	2	7	4	3	5
Russia	8	9	4	9	0	4	7	7	8	8	7	7	7
Germany	7	7	3	8	4	0	7	8	8	8	4	8	8
Turkey	3	4	6	3	7	7	0	5	4	5	6	4	5
Morocco	5	7	7	6	7	8	5	0	7	4	6	6	4
Peru	6	3	7	2	8	8	4	7	0	6	7	2	4
Nigeria	8	8	8	7	8	8	5	4	6	0	6	3	3
France	4	4	4	4	7	4	6	6	7	6	0	8	7
Mexico	5	4	6	3	7	8	4	6	2	3	8	0	4
South Africa	7	6	7	5	7	8	5	4	4	3	7	4	0

distances in the map approximate as closely as possible the values in the matrix. The way this approximation is measured and optimized distinguishes the different methods of MDS, but we do not enter into those details specifically here (see the Bibliographical Appendix for some recommended literature).

Classical scaling

The result of one approach, called classical MDS (function `cmdscale` in R—see Computational Appendix) is given in Exhibit 4.2. The countries are depicted as points and the distances between pairs of countries are approximations of the numbers in Exhibit 4.1. In this map we can see that Russia and Spain indeed turn out to be the furthest apart, while Italy and Spain appear close together, so at a first glance it seems like we have a good representation. We can approximately measure the interpoint distances in Exhibit 4.2 according to the scale shown on the sides, then the distances are always less than those in the table of ratings: for example, the distance between Italy and Spain in Exhibit 4.2 is about 1 unit whereas the given rating is 2. This is because classical scaling approximates the distances "from below"—the country points actually reside in a higher-dimensional space and have been projected onto a two-dimensional plane within this space. So all distances become shortened by this projection.

To measure how good the map is, a quality of approximation is measured in a similar way as it is done in regression analysis. In Exhibit 4.2 56.7% of the variance is accounted for. If we added a third dimension to the solution, depicting the countries in a three-dimensional map, a further 12.9% of the variance would be visualized, bringing the overall quality to 69.6%. In the Web Appendix a three-dimensional rotation of these country points is shown to illustrate the additional benefit of viewing the results in a three-dimensional space. For

our present purpose, however, we shall use the two-dimensional map. In Chapter 5 the topic of dimension reduction is explained more fully, with some technical details.

Principal axes

Exhibit 4.2 differs from the maps up to now (for example, Exhibits 2.5, 3.2 and 3.6) in one important respect: previously these maps were drawn using two observed variables, the (standardized) pollution and depth variables, whereas in MDS the axes on which the plot is constructed are so-called *principal axes*. These are not observed, but derived from the data with the objective of explaining the most variance possible: alternative names for the principal axes are *latent variables* or *factors*. As mentioned above, Exhibit 4.2 is the best view of the country points that can be achieved by projecting them onto a plane—in this plane the two axes are defined in order to be able to orientate the representation. These principal axes have the property that they are uncorrelated and the variance of the country points along each axis is equal to that part of the variance accounted for by that axis. The principal axes are also *nested*, which means that the first principal axis gives the best one-dimensional solution, explaining 33.3% of the variance in

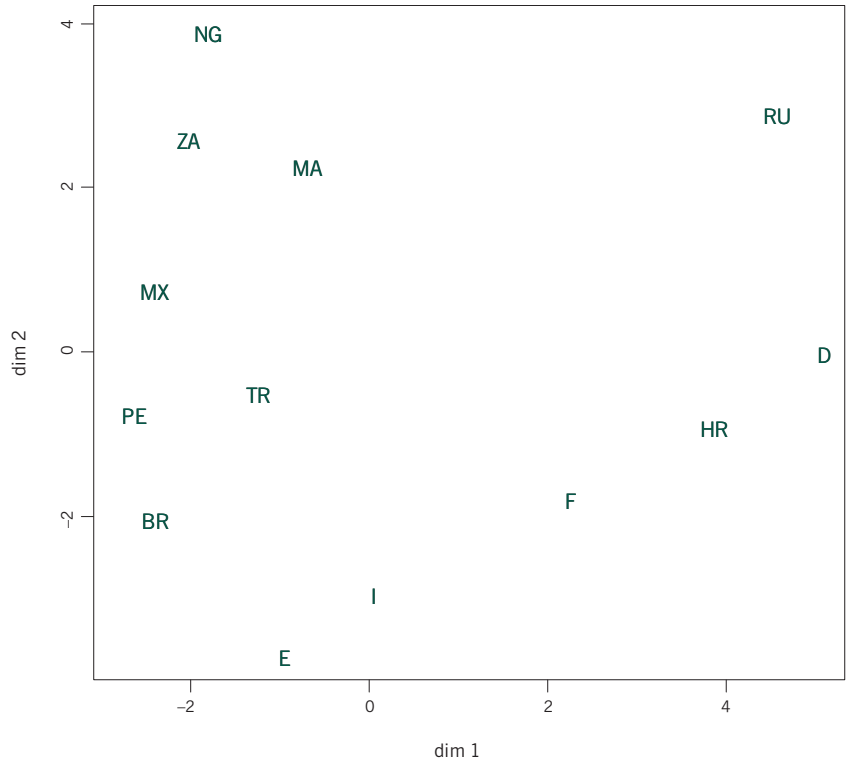

Exhibit 4.2:
MDS map of the 13 countries according to the ratings in Exhibit 4.1. The percentage of variance accounted for is 56.7%, with 33.3% on the first dimension, and 23.4% on the second

Exhibit 4.3: *Student MT's ratings of the 13 countries on six attributes:* standard of living *(1 = low,…,9 = high);* climate *(1 = terrible,…, 9 = excellent);* food *(1 = awful,…, 9 = delicious);* security *(1 = dangerous,…, 9 = safe);* hospitality *(1 = friendly,…, 9 = unfriendly);* infrastructure *(1 = poor,…,9 = excellent). On the right are the coordinates of the country points in the MDS map of Exhibit 4.2*

Countries	*living*	*climate*	*food*	*security*	*hospitality*	*infrastructure*	*dim1*	*dim2*
Italy	7	8	9	5	3	7	0.01	–2.94
Spain	7	9	9	5	2	8	–1.02	–3.68
Croatia	5	6	6	6	5	6	3.70	–0.88
Brazil	5	8	7	3	2	3	–2.56	–2.01
Russia	6	2	2	3	7	6	4.41	2.91
Germany	8	3	2	8	7	9	5.01	0.00
Turkey	5	8	9	3	1	3	–1.38	–0.48
Morocco	4	7	8	2	1	2	–0.87	2.27
Peru	5	6	6	3	4	4	–2.77	–0.74
Nigeria	2	4	4	2	3	2	–1.97	3.91
France	8	4	7	7	9	8	2.18	–1.76
Mexico	2	5	5	2	3	3	–2.58	0.77
South Africa	4	4	5	3	3	3	–2.16	2.62

this example, the space defined by the first two principal axes gives the best two-dimensional solution, explaining 33.3 + 23.4 = 56.7% of the variance, and so on. Each principal axis simply builds on the previous ones to explain additional variance, but in decreasing amounts. This is identical to the situation in stepwise regression when all the explanatory variables are uncorrelated.

Multidimensional scaling biplot—data set "attributes"

Suppose now that we had additional variables about the 13 countries, which could be economic or social indicators or even further ratings by the same student. In fact, each student had to supply, in addition to the inter-country ratings, a set of ratings on six attributes, listed in Exhibit 4.3. The idea is now to relate these ratings to the MDS map in exactly the same way as we did before, and represent each of these attributes as a biplot vector. This will give us some idea of how these attributes relate to the general perception summarized in Exhibit 4.2. Each of the variables in Exhibit 4.3 is linearly regressed on the two dimensions of Exhibit 4.2 (the country coordinates used as predictors are given in Exhibit 4.3 as well), giving the regression coefficients in Exhibit 4.4.

The regression coefficients for the two dimensions again define biplot vectors which can be overlaid on the MDS plot—see Exhibit 4.5. Since the dimensions are centred in the MDS, the constants are the means for each attribute, situated at the origin of Exhibit 4.5. Each of the biplot axes through the biplot vectors could then be calibrated by working out what one unit is on its axis, as before. A unit will be inversely proportional to the length of the biplot vector, so the tic marks for "infrastructure", one of the longest vectors, will be closer together than those for "security", a shorter vector. Thus, even though both of

	Constant	*dim1*	*dim2*	R^2
Living	5.231	0.423	–0.513	0.754
Climate	5.692	–0.395	–0.618	0.693
Food	6.077	–0.399	-0.645	0.610
Security	4.000	0.502	–0.444	0.781
Hospitality	3.846	0.660	0.010	0.569
Infrastructure	4.923	0.627	–0.591	0.818

Exhibit 4.4:
The regression coefficients for the regressions of the six attributes on the two dimensions of the MDS solution in Exhibit 4.2, as well as the measure of fit (R^2) in each case

these vectors point in exactly the same direction, there will be more variance in the projections of the countries onto "infrastructure" than onto "security". Notice that "hospitality" is worded negatively, so that the biplot vector is pointing to the "unfriendly" end of the scale: "friendly" would point to the left. It seems that the perception of the student in separating the South American countries on the left is due to their friendly hospitality, and that Brazil is not only hospitable but has a good climate and food as well.

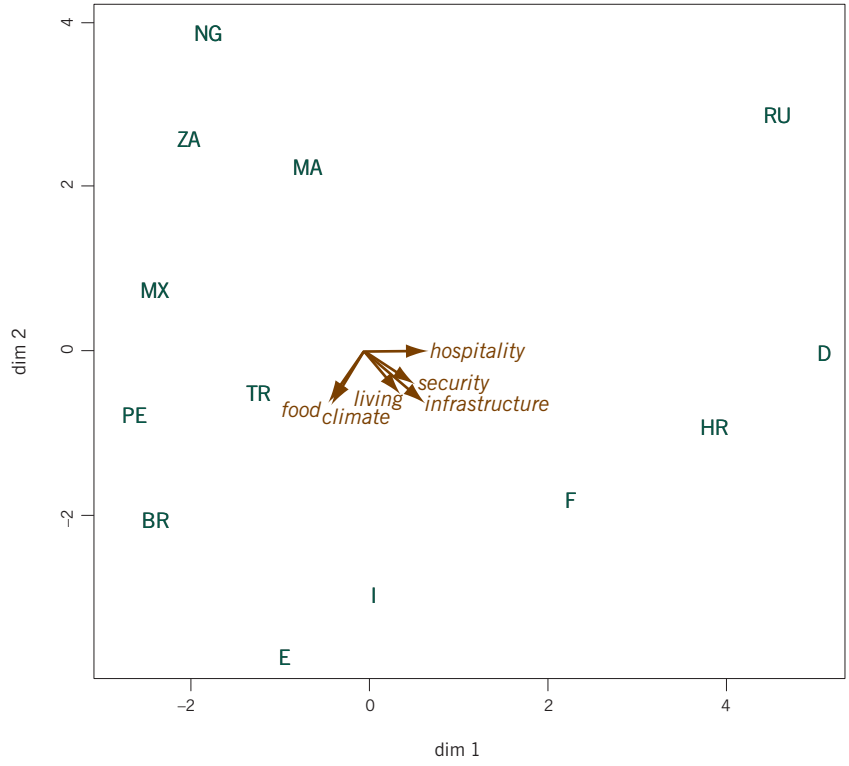

Exhibit 4.5:
MDS biplot, showing the countries according to the data of Exhibit 4.1 (i.e., the map of Exhibit 4.2), with the six attributes added as biplot vectors. Each biplot vector can be calibrated, as before, in its units from 1 to 9

Remember that the countries are positioned according to the student's overall perception of between-country differences, while the biplot vectors indicate how the ratings on specific attributes relate to this perception. The parts of variance explained (R^2) in Exhibit 4.4 show how well the attribute variables are represented. We can check some features of the map against the data in Exhibit 4.3 to get a qualitative idea of how well, or how badly, the biplot is performing. The three attributes "standard of living", "security" and "infrastructure" all point towards the European countries Germany, Croatia, France, Italy and Spain—these all fall above the averages for these three variables, but in decreasing value as one can see if one projects them onto the three corresponding biplot axes. To see if this agrees with the data, we average the three ratings for these variables, obtaining 8.3, 5.7, 7.7, 6.3 and 6.7 respectively, which are all above the average (which is equal to 4.7 in this case—Russia has an average of 5.0, which agrees with its position in the map, while all the other countries are below average and on the negative side of these three attributes). The value of 5.7 for Croatia is most out of line with the data—the general perception of Croatia, based on inter-country differences, places Croatia between Germany and France, but according to the attribute data Croatia should be lower down on the direction defined by these three correlating attributes. Hence the lack of fit, or unexplained variance, in the attributes in Exhibit 4.5 includes this "error" in their display with respect to Croatia. But, of course, Exhibit 4.5 is not designed to show the attributes optimally —these have been superimposed *a posteriori* on the map. In Chapter 6 we shall perform an analysis specifically of the attribute data, in which we will see the attributes represented with much higher R^2, and showing Croatia in a position more in accordance with them.

Chi-square distance biplot

We can anticipate the chapter on correspondence analysis (Chapter 8) by reconsidering data set "bioenv" of Exhibit 2.1. Previously we performed regressions of the five species variables on two of the concomitant variables "pollution" and "depth" and showed the results in the space of these two environmental variables. We now take the MDS biplot approach, performing an MDS of the 30 stations in terms of their species information and then show how the explanatory variables relate to the MDS map. The only decision we need to make is how to measure distance between the 30 stations—in contrast to the "countries" example above, the distances are not the original data but need to be calculated from the species data. Here the *chi-square distance* will be used, the distance function that is the basis of correspondence analysis. This distance is based on the relative frequencies of the species at each station and also adjusts the contribution of each species according to its average across the 30 stations. This will be explained more in Chapter 8, but to give one example, consider the distance between stations s1 and s2 (see Exhibit 2.1). The relative frequencies of the species at these two stations are, respectively, [0 0.074 0.333 0.519 0.074] and [0.481 0.074 0.241 0.204 0]—for

example, for station s2, a count of 26 for species *a* is 0.481 of the total of 54 individuals counted at that station (the second row total). The totals for each species (column totals) have relative frequencies [0.303 0.196 0.189 0.245 0.067], showing that species *a* is the most common and species *e* the most rare. The chi-square distance between the two stations is computed as:

$$\sqrt{\frac{(0-0.481)^2}{0.303}+\frac{(0.074-0.074)^2}{0.196}+\frac{(0.333-0.241)^2}{0.189}+\frac{(0.519-0.204)^2}{0.245}+\frac{(0.074-0)^2}{0.067}}=1.139$$

The division of each squared difference by the overall species proportion is a form of standardization of the relative frequencies. There are bigger differences between the more common species and smaller differences between rare species, and the standardization serves to compensate for this natural variability found in frequency data. Having computed the 30 × 30 chi-square distance matrix between the 30 stations, the MDS procedure leads to a visualization of these distances, shown in Exhibit 4.6. Then, by performing the regressions of the species variables

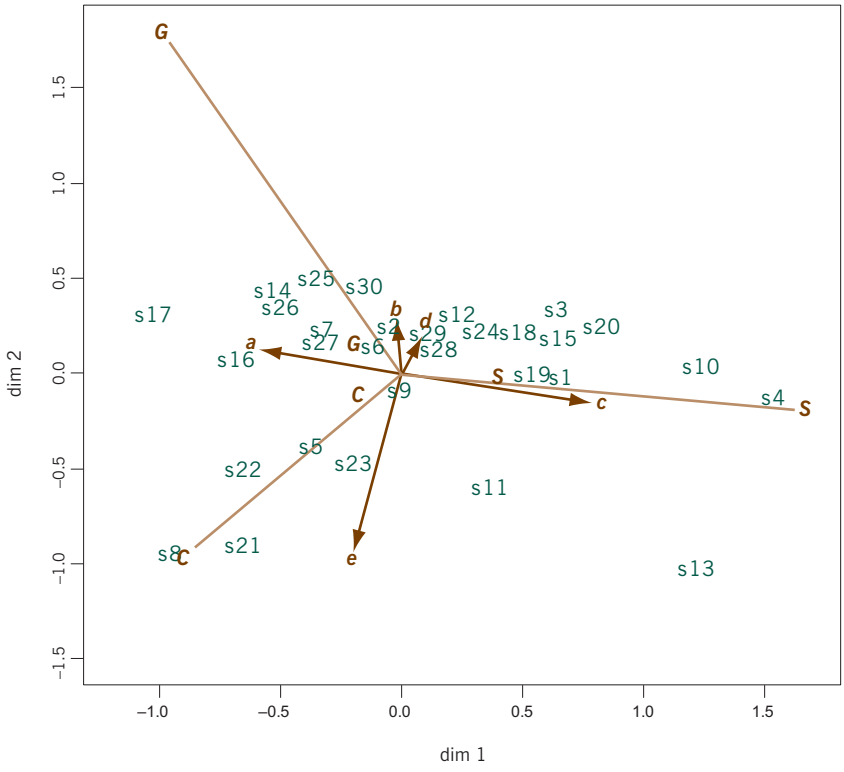

Exhibit 4.6:
MDS biplot, showing approximate chi-square distances between sites, upon which are added the biplot vectors of the five species (using linear regression), biplot vectors of the three sediment types (using logistic regression) and the averages of the stations according to the three sediment types

on the two dimensions of the map, these can be depicted on the map (here we use the relative frequencies, divided by the square roots of their respective averages, to be consistent with the way the chi-square distances were calculated). In addition, the three categories of the variable sediment can be shown, either as the averages of the site points in for each category or using the logistic regression biplot. Notice once more that there are two levels of error in this biplot: first, the chi-square distances between sites are not perfectly displayed (74.4% explained in the map—52.4% on dimension 1, 22.0% on dimension 2—i.e., 25.6% error); and second, the two dimensions only explain parts of variance of each species (*a*: 76.7%, *b*: 18.6%, *c*: 92.1%, *d*: 14.6%, *e*: 95.8%) and of each category of sediment (expressed as percentages of deviance explained in the logistic regressions, *C*: 7.3%, *G*: 11.8%, *S*: 15.8%).

SUMMARY: Multidimensional Scaling Biplots

1. Multidimensional scaling (MDS) is a method that represents a set of objects as a set of points in a map of chosen dimensionality, usually two-dimensional, based on their given interpoint distances. The objective is to maximize the agreement between the displayed interpoint distances and the given ones.
2. Any variable observed on the same set of objects can be superimposed on such a map using the regression coefficients obtained from the regression of the variable (or its standardized equivalent) on the dimensions of the MDS. The resultant joint plot is a biplot: the objects can be projected onto the biplot vectors of the variables to approximate the values of the variables. The optional standardization of the variable only changes the lengths of the biplot vectors, not their directions.
3. There are two different levels of error in the display. First, there is the error incurred in the MDS, because the distances are not perfectly displayed. Second, there are the errors in the regressions of the variables on the dimensions.