Biplots in Practice

MICHAEL GREENACRE

Professor of Statistics at the Pompeu Fabra University

Chapter 9 Offprint

Multiple Correspondence Analysis Biplots I

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

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

Chapter 9

Multiple Correspondence Analysis Biplots I

Multiple correspondence analysis is the extension of simple correspondence analysis of a cross-tabulation of two categorical variables to the case of several variables. The method is used mostly in the visualization of social survey data, where respondents reply to a series of questions on discrete scales such as "yes/no" or "agree/unsure/disagree". A type of data that is intermediate between simple and multiple correspondence analysis is a concatenated, or stacked, table—this is a block matrix composed of several two-way cross-tabulations of the same sample of respondents, where each cross-tabulation is between a demographic and a substantive variable. In this chapter we show how CA biplots of a single table can be extended to concatenated tables and then, in the following chapter, to multiple correspondence analysis where several variables are cross-tabulated with one another. The way total variance is measured and how it is decomposed into parts is a recurrent theme in this area, and it will be shown how the biplot concept can clarify this issue.

Contents

Data set "women" 89
Concatenated table 90
Symmetric CA map 92
Asymmetric map/biplot for concatenated table 92
Between- and within category inertia 93
Contribution biplot for concatenated table 94
Supplementary points 96
SUMMARY: Multiple Correspondence Analysis Biplots I 96

Data set "women"

The *International Social Survey Program* (ISSP) is an annual co-operative program between many countries where social surveys are conducted to ask people in each country the same questions on a different theme. The data we shall consider in this chapter are from the third Family and Changing Gender Roles survey conducted in 2002. Even though data are available from more than 30 countries, we shall just treat the Spanish data here (see the second case study in Chapter 14 for a more detailed analysis). Also, because we want to avoid the issue of missing values for the moment, we have deleted 364 cases with missing data, leaving 2107 of the original 2471 respondents. The questions we focus on are those related to working women and the effect on the family, specifically the following eight statements to which the respondents could either (1) strongly agree, (2) agree, (3) neither agree nor disagree, (4) disgree, or (5) strongly disagree:

A: a working mother can establish a warm relationship with her child
B: a pre-school child suffers if his or her mother works
C: when a woman works the family life suffers
D: what women really want is a home and kids
E: running a household is just as satisfying as a paid job
F: work is best for a woman's independence
G: a man's job is to work; a woman's job is the household
H: working women should get paid maternity leave

There were also several demographic variables, of which we retained the following:

g: gender (1 = male, 2 = female)
m: marital status (1 = married/living as married, 2 = widowed, 3 = divorced, 4 = separated, but married, 5 = single, never married)
e: education (0 = no formal education, 1 = lowest education, 2 = above lowest education, 3 = higher secondary completed, 4 = above higher secondary level, below full university, 5 = university degree completed)
a: age (1 = 16-25 years, 2 = 26-35, 3 = 36-45, 4 = 46-55, 5 = 56-65, 6 = 66 and older)

Abbreviations in the analyses that follow are constructed in the obvious way: for example, *C2* is an agreement to statement *C*, and e5 is category 5 of education. The only exception is for the variable *H* for which there were only two respondents who strongly disagreed—these were combined with the disagree category, leading to a new category denoted as *H4,5*. To demonstrate what is called *interactive coding* of two variables, a variable with 12 categories was constructed from the gender and age variables, with categories denoted by ma1 to ma6 (six age groups for males) and fa1 to fa6 (six age groups for females).

Concatenated table

In simple CA a single demographic variable would be cross-tabulated with a single substantive question, for example the cross-tabulation of education (6 categories) with question *A* (5 categories). The pairwise cross-tabulations of each of the demographic variables with each of the substantive questions can be assembled in a block matrix called a *concatenated table*. Exhibit 9.1 shows just a part of this 23 × 39 table (one less column because of the combining of *H4* and *H5*), with

	A1	A2	A3	A4	A5	B1	B2	B3	B4	B5		n
m1	192	486	54	381	56	80	550	138	334	67	···	1169
m2	21	68	9	63	11	13	101	16	37	5	···	172
m3	10	17	2	12	4	5	16	10	12	2	···	45
m4	12	32	5	16	4	4	30	7	24	4	···	69
m5	162	329	21	126	14	22	259	66	258	47	···	652
e0	23	97	16	90	14	20	138	26	52	4	···	240
e1	76	203	21	192	30	39	286	62	114	21	···	522
e2	99	263	28	168	22	37	264	69	182	28	···	580
e3	100	203	16	95	17	17	178	43	160	33	···	431
e4	48	81	2	32	3	5	52	13	78	18	···	166
e5	51	85	8	21	3	6	38	24	79	21	···	168
ma1	38	80	3	27	3	6	70	13	55	7	···	151
ma2	41	116	9	53	7	10	92	23	92	9	···	226
ma3	30	94	13	37	7	14	77	18	60	12	···	181
ma4	25	65	9	40	4	13	68	12	43	7	···	143
ma5	16	54	0	50	5	15	72	17	18	3	···	125
ma6	15	43	9	64	15	18	95	13	18	2	···	146
fa1	48	83	7	39	6	5	56	21	84	17	···	183
fa2	59	92	5	58	13	10	82	23	86	26	···	227
fa3	46	81	9	51	6	7	78	21	68	19	···	193
fa4	37	75	7	60	3	7	76	30	55	14	···	182
fa5	21	52	5	42	6	6	65	15	34	6	···	126
fa6	21	97	15	77	14	13	125	31	52	3	···	224

Exhibit 9.1:

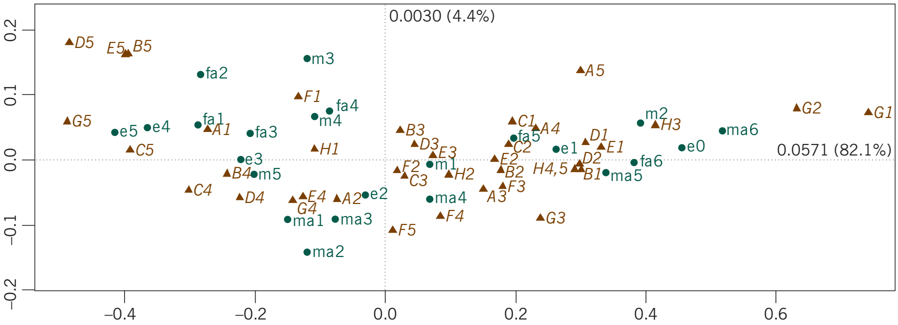
Part of the 23 × 39 concatenated table for the "women" data set, showing the first 10 columns corresponding to the response categories of questions A and B. The 40 column categories are reduced to 39 because H4 and H5 are combined. The sample size for each demographic category is given in the last column. There are 3 × 8 = 24 cross-tabulations in this concatenated table

the rows being the categories of marital status (5 categories), education (6 categories) and the gender/age combinations (12 categories). Notice that the gender and age groups themselves are not included along with their combinations, although they can be added as so-called *supplementary points* in the analysis, a subject to be discussed later in this chapter.

This type of table is also called a *block matrix*, composed of 3 blocks row-wise and 8 blocks column-wise, that is 24 subtables in total which form the blocks, or subtables, of the matrix. Each of the 24 subtables has the same grand total, which is the number of respondents, equal to 2107. Each of the 8 subtables in a row block has the same row margins and each of the 3 subtables in a column block has the same column margins: for example, the row sums of the table cross-tabulating marital status (m1 to m5) with question *A* (*A1* to *A5*) are the same as the row sums of the table cross-tabulating marital status with question *B*—these row sums are the sample sizes given in column "*n*" of Exhibit 9.1. As a consequence of this equality of marginal sums, it is easy to show the useful result that the total inertia of the concatenated table is the average of the inertias of its 24 subtables.

Exhibit 9.2:

Symmetric CA map of the concatenated table of Exhibit 9.1. This is not a biplot since both the row and column points are displayed in principal coordinates



Symmetric CA map

The most common way of showing the results of a CA is in the form of the symmetric map, shown in Exhibit 9.2. In this map both rows and columns are displayed in principal coordinates, with the result that the graphic is strictly speaking not a biplot. However, interpoint chi-square distances are approximately displayed both between rows and between columns. Since in this case the result is very one-dimensional, with 82.1% of the inertia on the first dimension, we initially interpret only the left-to-right dispersion of the points. Clearly, categories on the left hand side correspond to attitudes favourable to women working while those on the right hand side correspond to the traditional view that they should not work but look after the household and children. Correspondingly there is a lining up of the demographic categories from left to right: for example, highest education is on left and lowest on the right, and the age groups similarly vary from youngest on the left to oldest on the right.

Asymmetric map/biplot for concatenated table

The total inertia in this example is equal to 0.06957, which is a very low value in absolute terms: on average the associations between the demographic variables and the question responses are low, which is quite typical for social science data. Geometrically, this means that the row profiles, for example, are scattered close to the average row profile, with the column vertex points at the outer extremities of the profile space very far out from the set of row profile points. This is clear in the asymmetric map/biplot of Exhibit 9.3, where the column points (in standard coordinates) are so far away from the demographic row points (in principal coordinates) that the latter are too close to one another to label.

Notice that for a concatenated table a vertex point consisting of zeros with a single 1 does not have the same geometric meaning as in simple CA because it is a point impossible to observe—a sample can not be present just for one variable and non-existent elsewhere. But one can think about the row–column relationship as an average of separate CA-type relationships across the column variables as follows. Each row, for example education group e5, has a profile across each

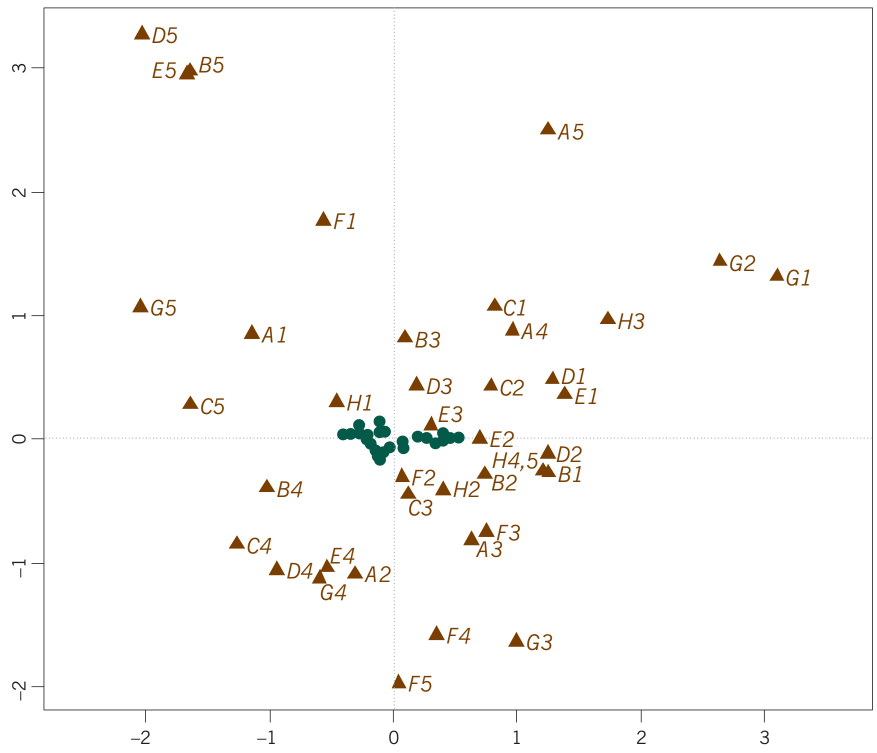


Exhibit 9.3:

Asymmetric CA map of the concatenated table of Exhibit 9.1. The positions of the row points (in green) are identical to those in Exhibit 9.2, as well as the inertias and percentages of inertia

column variable. There are thus 8 different weighted average positions that one could compute for `e5`, computed the same way as in a simple CA—`e5`'s position is then the ordinary average of these 8 positions. Exhibit 9.4 illustrates the idea by showing the eight averages around `e5` and also those around `ma4`, males in the fourth age group. There is much more variance around `e5` compared to `ma4`. Respondents in the highest education group `e5` react to statement *G* ("a man's job is to work; a woman's job is the household") with a relatively high level of disagreement, whereas with respect to statement *E* ("running a household is just as satisfying as a paid job") they are closer to the average opinion. Taking all 8 statements into account, their average position is the most extreme on the liberal side of the map. The attitudes to the individual questions by males in the 45-55 years age group, on the other hand, are much more similar, slightly to the conservative/traditional side of average.

Between- and within category inertia

This way of showing each demographic category's position as an average across the questions suggests an interesting decomposition of inertia for each demographic category, into a "within-category" component across the 8 questions and a "between-category" component. The "between" component is nothing else but

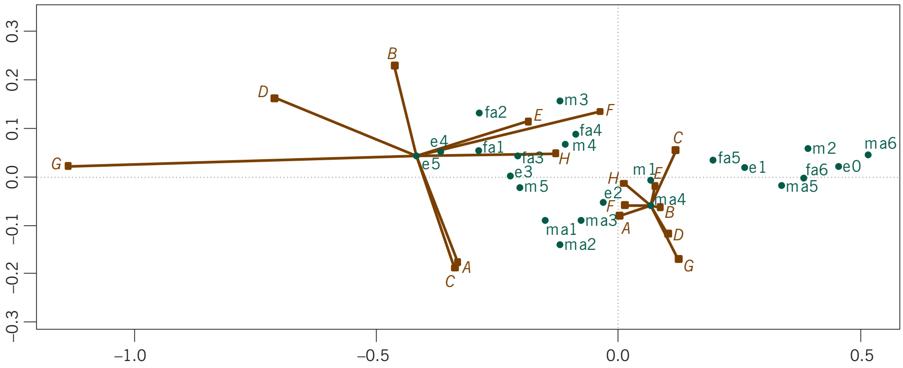


Exhibit 9.4: *Map of row points (demographic categories) of the concatenated table, illustrating two points,* e5 *and* ma4*, at the average of their positions with respect to the 8 variables (for example, "G" for point* e5 *is the weighted average position of* e5 *with respect to the categories for question G, using the profile values for* e5 *across G)*

the usual measure of inertia, which is the measure of dispersion of the demographic categories, whereas the "within" component is a measure of the dispersion across the 8 questions for each demographic category. The table in Exhibit 9.5 summarizes how much each demographic category contributes to the total "within" component, measured in permills (thousandths). This table shows that ma4 has the smallest contribution (5/1000) of all categories, and e5 an above average contribution (61/1000). The lowest two education groups and the oldest age group, both male and female, have the highest contributions—thus e0, for example, would show a much higher dispersion than e5 across the questions if its 8 individual points (of which it is the average) were drawn as in Exhibit 9.4.

Contribution biplot for concatenated table

Both the symmetric map of Exhibit 9.2 and the asymmetric biplot of Exhibit 9.3 have their particular advantages but neither tells the analyst which categories of the variables are driving the solution—this can be seen using the contribution biplot (see Chapter 8), which multiplies the standard coordinates of the column categories by the square roots of their corresponding masses. The contribution

Exhibit 9.5: *Permill contributions of each category to the dispersion across the questions, or "within-category" inertia*

Marital Status		*Education*		*Gender × Age*			
				Male		*Female*	
m1	14	e0	149	ma1	11	fa1	37
m2	81	e1	94	ma2	16	fa2	50
m3	8	e2	17	ma3	7	fa3	22
m4	5	e3	64	ma4	2	fa4	12
m5	55	e4	37	ma5	37	fa5	11
		e5	61	ma6	103	fa6	107
TOTALS	*163*		*422*		*176*		*239*

biplot is shown in Exhibit 9.6—the directions of the category points are identical to those of Exhibit 9.3 (thus, calibrations along these directions would be identical), but now the squares of their coordinates are equal to their contributions to the respective principal axes. Immediately it is clear that variable *G* ("a man's job is to work; a woman's job is the household"), especially categories *G5* opposed to *G2*, is the biggest contributor to the first (horizontal) axis—in fact, these two categories alone contribute 31% to the first axis, which is itself explaining 82.1% of the total inertia in the data. Notice that it is the "agree" category on the right which opposes the "strongly disagree" on the left—in Exhibits 9.2 and 9.3, which show this category's positional information as a scale value, the "strongly agree" category *G1* is situated further to the right than *G2*, as expected, but as far as contributing to this axis is concerned *G1* is less important, probably due to the fact that not many people give this response.

Now that we see which categories of attitude are driving the solution, there is an interesting interpretation of the vertical second dimension on the left hand side of Exhibit 9.6, even though this dimension explains only 4.4% of the total inertia. The biggest contributors are (at the top) *D5*, *E5* and *F1*, expressing the strongest support for women working, whereas at the bottom we have *G4*, *A2*, *D4* and *E4*, expressing moderate support for women working. The corresponding contrast is between the divorced marital group and the female groups up to the ages of 55 on top, and the male groups up to age 45. This contrast between males and females

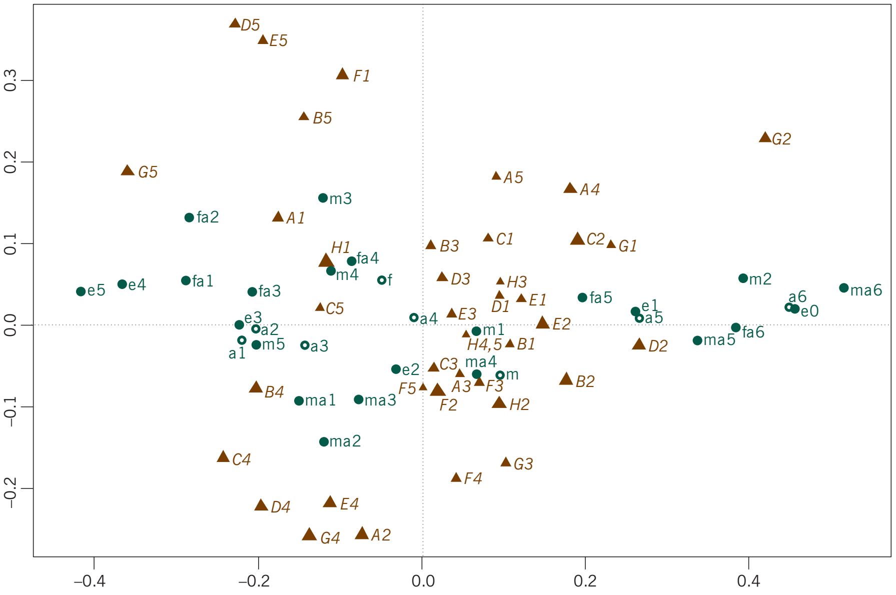


Exhibit 9.6:
Contribution biplot of the concatenated table of Exhibit 9.1, with column coordinates equal to the standard coordinates multiplied by the square roots of the respective column masses. The gender and age groups have been added as supplementary points (empty circle symbols). The positions of the row points (in green) are identical to those in Exhibits 9.2 and 9.3, as well as the inertias and percentages of inertia

does not exist in the older age groups on the right hand side of the display, where the demographic groups are closer together on the vertical axis.

Supplementary points

Exhibit 9.6 shows some additional points, for the two gender and six age groups. In the analysis these two variables had been combined interactively to form 12 groups, but the original categories can also be displayed. From the graphical viewpoint these points are just the weighted averages (centroids) of their displayed component groups: for example, the point `a1`, denoting the youngest age group, is between the female and male points for this group, `fa1` and `ma1`. It is at the weighted average of these two points, weighted by the numbers in the respective female and male subgroups. Similarly, the point `f` for females, is at the weighted average of the six female subgroups, `fa1` to `fa6`.

Analytically, supplementary points define additional profiles that are not used to establish the solution space, but are projected onto that space afterwards. The coordinates of the supplementary row points in this example are obtained by computing scalar products between the profile elements and the standard column coordinates: $\mathbf{D}_r^{-1}\mathbf{P}\mathbf{\Gamma}$ in the notation of Chapter 8, where the profile is calculated across all Q variables (i.e., summing to 1 across all the variables). Equivalently, following the way the joint display in Exhibit 9.3 was interpreted, compute for each column variable the weighted average position of the row using its profile just across that variable (i.e., summing to 1 across that variable), and then average these positions (8 of them in this case) to situate the supplementary point.

SUMMARY: Multiple Correspondence Analysis Biplots I

1. A *concatenated table* is a block matrix composed of several contingency tables cross-tabulating the same sample of cases between two sets of variables. If there are L variables in the first set and Q in the second set, then there are $L \times Q$ subtables constituting the concatenated table, and each subtable has the same grand total, equal to the sample size.
2. The CA of a concatenated table is an average picture of the pairwise relationships between the two sets of variables. Its inertia is the average of the inertia of the subtables and the graphical display is the best approximation to all the subtables. One can think of this analysis as a compromise among all possible simple CAs of the subtables, using only one set category points for each row and column variable.
3. The asymmetric biplot of a concatenated table usually shows the set of points in principal coordinates close to the origin and far from the other set in standard coordinates, in which case a separate plot of the "inner" set of points is required.
4. Each category in principal coordinates, say a row category, is the average of a mini-cloud of points, one point for each of the column variables. It is useful to

measure the dispersion within each of these mini-clouds because this gives information about the variance of the category across the column variables.

5. The contribution biplot of a concatenated table displays one set of points in principal coordinates to show their interpoint distances, and the other points in standard coordinates multiplied by the square root of the respective masses (these are the usual masses that sum to 1 across all the variables). The latter set of points then indicates how they contribute to the construction of the axes of the representation space.

6. A *supplementary point* is an additional row or column of data with a profile that is displayed afterwards by projection onto the biplot. One can think of this row or column being in the analysis from the start but with zero mass assigned to it, hence having no influence on the solution.