Biplots in Practice

MICHAEL GREENACRE

Professor of Statistics at the Pompeu Fabra University

Chapter 14 Offprint

CASE STUDY–SOCIO-ECONOMICS

Positioning the "Middle" Category in Survey Research

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

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

CHAPTER 14

Case Study 2: Positioning the "Middle" Category in Survey Research

Offering a middle alternative, for example "neither agree nor disagree", as a response on an attitudinal scale can have various consequences in survey research. In this case study, after a general investigation of a large survey data set, special attention is given to the middle categories, specifically how they associate (a) with one another, (b) with the "adjacent" categories between which they are supposed to lie and (c) with demographic characteristics. Missing responses enter our study in a natural way as additional non-substantive responses and we can also see how these are interrelated and related in turn to the substantive responses. This approach is illustrated with the ISSP data on attitudes to women working (this is an expanded version of the data set "women" used in Chapter 9).

Contents

Data set "womenALL"

The data set "women" was introduced in Chapter 9, consisting of eight questions eliciting attitudes about working women. In that chapter, to simplify the explanation, data from only one country, Spain, were considered and all respondents with some missing data were eliminated. In this case study all the data from 46 638 respondents in 32 countries are considered, and no respondents are eliminated —this data set is referred to as "womenALL". Exhibit 14.1 lists the countries surveyed and the abbreviations used in the graphics.

Exhibit 14.1:
Countries surveyed in the third Family and Changing Gender Roles survey of the ISSP in 2002 (former West and East Germany are still sampled separately for research purposes). The abbreviations are used in subsequent biplots

AU	Australia	SE	Sweden	SK	Slovakia
DW	Germany (west)	CZ	Czech Republic	CY	Cyprus
DE	Germany (east)	SI	Slovenia	PT	Portugal
GB	Great Britain	PL	Poland	RC	China
NI	Northern Ireland	BG	Bulgaria	DK	Denmark
AT	Austria	NZ	New Zealand	CH	Switzerland
US	USA	RP	Philippines	FL	Belgium (Flanders)
HU	Hungary	IL	Israel	BR	Brazil
IE	Ireland	JP	Japan	SF	Finland
NL	Netherlands	ES	Spain	TW	Taiwan
NW	Norway	LV	Latvia		

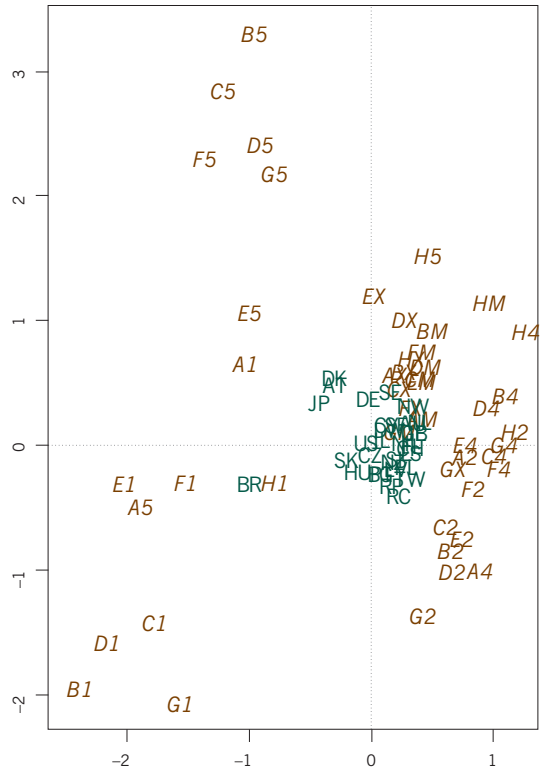
Cross-national comparison using CA biplot of concatenated tables

To get a broad overview of the differences between countries, a concatenated matrix (see Chapter 9) is assembled, cross-tabulating the countries with each of the eight questions.

Since each question has five substantive response categories plus a missing category, there are six categories per question, and the concatenated matrix has 32 rows and 48 columns. The CA asymmetric map/biplot is shown in Exhibit 14.2, with a separate amplification of the row points, which as usual are bunched up near the middle of the biplot. The missing categories of response, labelled *AX* to *HX*, are all in a group near the origin of the biplot. So too are the middle response categories (the category 3's), which we have labelled *AM* to *HM* here. Generally all the extreme response categories (1's and 5's) are to the left of centre, while the moderate response categories (2's and 4's) are to the right. The conservative-to-liberal attitude scale runs from bottom to top, with strong agreement to statements B, C, D and G at bottom left (see chapter 9 for the statement wording). Brazil is in an isolated position, showing that its respondents tend to use the extreme conservative response categories. China is also at the conservative extreme of these countries, but using more of the moderate response categories. At the top Denmark is at the most liberal position, followed by Austria and Sweden, with the Swedish using the more moderate responses.

Multiple correspondence analysis of respondent-level data

This aggregate-level picture of the countries does not reflect associations between response categories at the level of the individual respondent. Applying MCA to the original 46,638 × 8 matrix gives the biplot in Exhibit 14.3. The result is typical of social-science applications, with all the non-substantive missing response categories separating out (bottom left) and opposing the substantive responses which themselves split into the moderate and middle categories (upper left) and the extreme categories (upper right). The fact that these three types of response



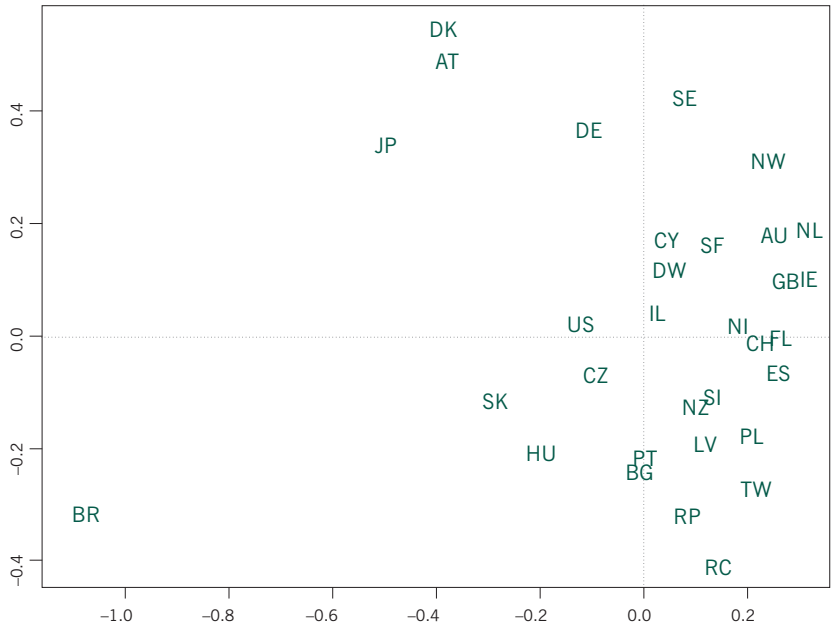
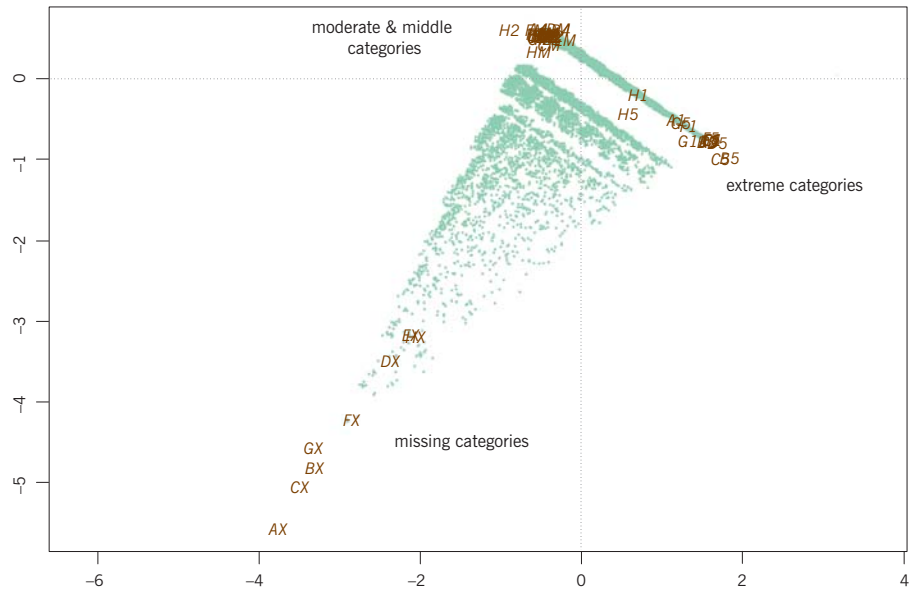


Exhibit 14.2:
CA biplot of the concatenated countries by categories matrix, and a separate plot of the countries alone

Exhibit 14.3:
MCA biplot of the respondent-level data: each dot represents one of the 46,638 respondents at the average position of his or her eight response categories



(missing, moderate and extreme) group together means that the categories within each group associate strongly: for example, someone who gives a missing response on one question is likely to give more missing responses on others, and similarly for the moderate and extreme responses. The respective groupings of the extreme and moderate responses, irrespective of the wording of the question, are stronger than the underlying attitude scale. Each respondent is displayed by a dot at the average position of his or her eight response categories. Thus the thick band of points at the top correspond to those with no missing values, followed by bands of respondents with increasing numbers of missing values as we move towards the bottom left.

Subset multiple correspondence analysis eliminating missing categories

Since we are more interested in the substantive responses, we can eliminate the strong effect of the missing categories by performing a so-called *subset analysis*. This is variant of correspondence analysis which visualizes a subset of points while maintaining the original geometry of the complete set: the centre, masses and metric of the space remain the same, what changes is the elimination of certain points from the dimension reduction process (this is not the same as declaring the missing categories as supplementary points, which would change the geometry of the active data set). Exhibit 14.4 shows the result of omitting the missing categories in the subset analysis—as in Exhibit 14.2, the first (horizontal) axis opposes extreme response categories on the left and moderate categories on the right. But now the attitude scale itself appears vertically with conservative response categories lower down and liberal responses at the top. Exceptions are

questions *F* ("work is best for a woman's independence") and *H* ("working women should get paid maternity leave"), where opinions appear unrelated to the general scale of attitude towards whether women should work or not.

In Exhibit 14.4 the middle ("*M*") categories appear grouped between the moderate ones, as one might expect. If we bring in the third dimension of the subset analysis, however, these categories are seen to separate as a group, and do not lie between the "*2*"s and the "*4*"s. Exhibit 14.5 shows the planar view of the third dimension (horizontal) and second dimension (vertical), and the "*M*"s no longer lie in their expected positions on the scale. For question responses that are consistent with an underlying ordinal attitudinal scale the MCA configuration should take the approximate form of polynomials of increasing order, as shown in Exhibit 14.6 (usually the scale appears on the first dimension, and the polynomials are with respect to the first dimension—in this example, the scale is found on the second dimension because of the strong extreme versus moderate response effect). Exhibit 14.4 fits the pattern on the left in Exhibit 14.6, while all the response categories except the middle ones in Exhibit 14.5 fit the pattern on the

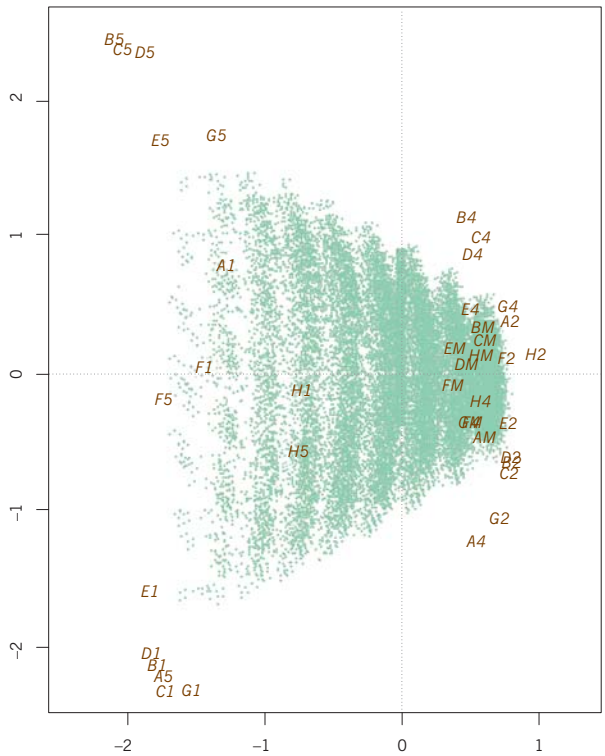


Exhibit 14.4:
Subset MCA biplot of the respondent-level data: each dot represents one of the 46,638 respondents at the average position of his or her eight response categories

Exhibit 14.5:
Subset MCA biplot of the respondent-level data, showing dimension 2 vertically, as in Exhibit 14.4, but dimension 3 horizontally. The separation of the middle categories (encircled) is now apparent

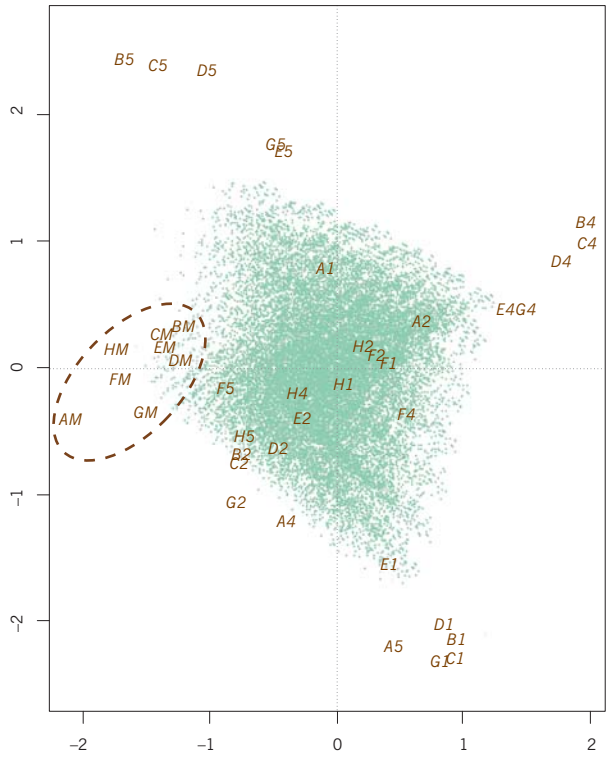
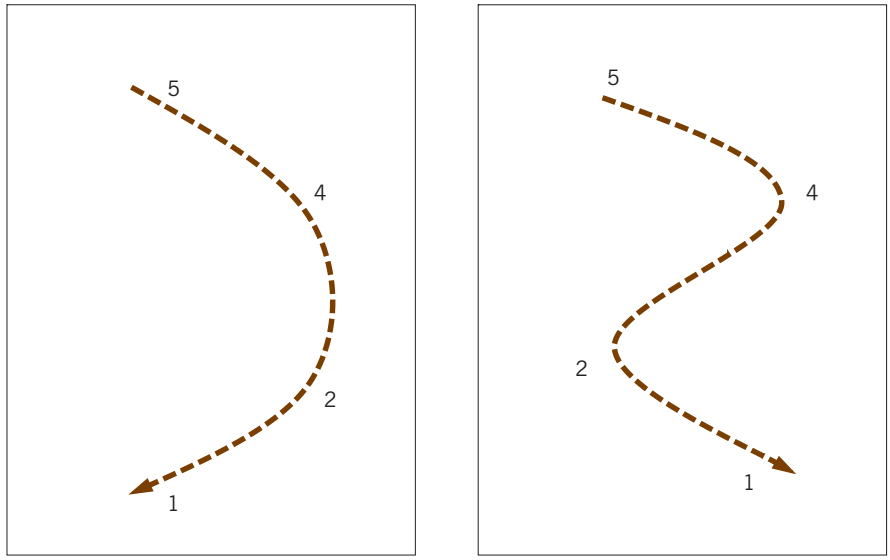


Exhibit 14.6:
General patterns in Exhibits 14.4 and 14.5 (for questions B, C, D and G, for example, all worded negatively towards working women), showing their respective quadratic and cubic patterns



right in Exhibit 14.6. What we are discovering is that the middle response categories are partially fitting into the attitudinal scale but show a distinct separation as a different response type, which might be a type of non-response or so-called "satisficing" effect, which is the respondents' way of giving an acceptable answer without having to spend time and effort forming an opinion.

The dimensions of "middleness"

If we look at the numerical diagnostics of the subset MCA, it turns out—not unexpectedly—that the middle response categories are somewhat correlated with dimension 3 (as seen in Exhibit 14.5), but these categories will have parts of their variance on many other dimensions too. If we really want to study the effect of the middle response, we need to isolate exactly where the middle categories are. Then we can see if any demographic characteristic is linked to those dimensions. There are two ways we can do this: using subset analysis again, but just on the middle categories, or using canonical correspondence analysis (CCA), where we define "middleness" as an explanatory variable. A difference between the two approaches will be that the subset analysis will focus on all the dimensions of "middleness", of which there are 8 since there are 8 questions, so that dimension reduction will be necessary, whereas the way we implement the CCA just one dimension of "middleness" will be imposed as a constraining variable on the solution, which makes it easier to relate to the demographics.

Canonical correspondence analysis to focus on middles and missings

To implement the CCA we create the explanatory variable "number of middle responses" by counting, for each respondent, how many middle responses are in his or her response set. This variable can vary from 0 to 8. Since this is the addition of the 8 columns of the 46,638 × 48 indicator matrix, the variable we are creating is actually the centroid of the 8 middle response points. It is this centroid on which the CCA will focus. In fact, we can do exactly the same for the missing values: instead of performing a subset MCA, we can add a variable "number of missing responses" and then use both of these as explanatory variables, thus giving a two-dimensional restricted space of middles and missings. This will allow a convenient investigation of possible associations with the demographics. The set-up for what amounts to a canonical MCA is shown in Exhibit 14.7.

The canonical MCA with the two constraining variables is shown in Exhibit 14.8. The missing count variable is almost exactly aligned with the horizontal first dimension, and the middle count variable slightly more than 90 degrees away in a vertical direction. The 46,638 respondents occur in only $1 + 2 + \cdots + 9 = 45$ different combinations of the two constraining variables, which are indicated by circles with an area proportional to the corresponding number of respondents. Thus the largest circle at bottom left corresponds to 0 middles and 0 missings (3107 respondents), and the next circle vertically corresponds to 1 middle and 0 missings (2354 respondents) up to the topmost circle for all 8 middles (254 respondents).

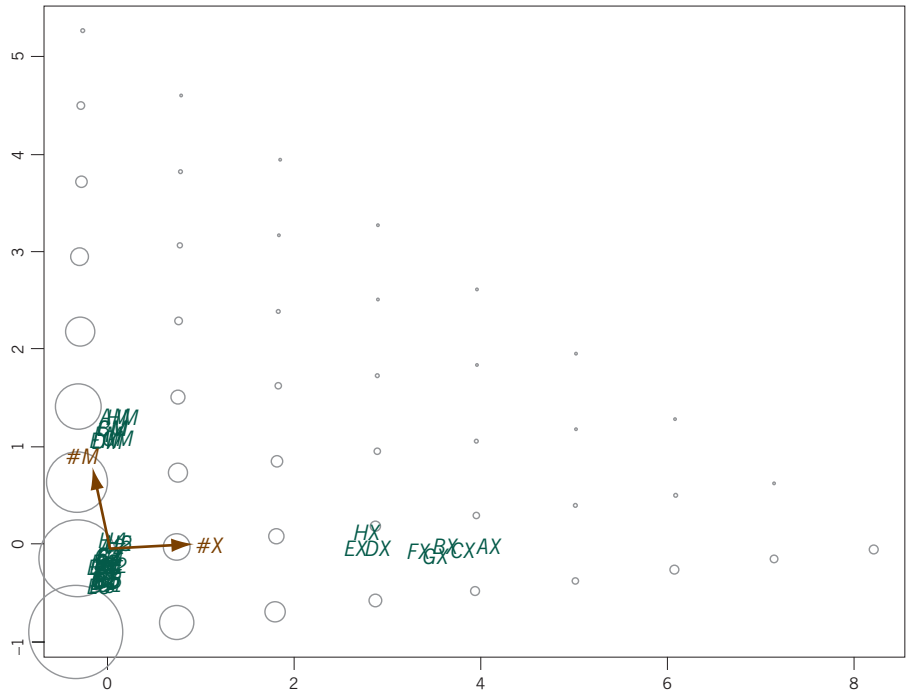
Exhibit 14.7:
Data set-up for canonical MCA biplot, showing first 10 rows of the original data on the left and recoded data on the right used for the analysis. The columns #M and #X are the sums of the M and X columns of the indicator matrix, i.e. the counts of middle and missing responses respectively

A	B	C	D	E	F	G	H	A1	A2	AM	A4	A5	AX	B1	B2	BM	B4	B5	BX	· · ·	#M	#X
4	2	2	3	3	2	3	4	0	0	0	1	0	0	0	1	0	0	0	0	· · ·	3	0
1	5	5	5	1	1	5	2	1	0	0	0	0	0	0	0	0	0	1	0	· · ·	0	0
2	3	2	3	2	2	4	2	0	1	0	0	0	0	0	0	1	0	0	0	· · ·	2	0
4	2	1	4	4	4	4	4	0	0	0	1	0	0	0	1	0	0	0	0	· · ·	0	0
3	2	2	3	2	4	3	2	0	0	1	0	0	0	0	1	0	0	0	0	· · ·	3	0
4	9	2	3	2	3	3	5	0	0	0	1	0	0	0	0	0	0	0	1	· · ·	3	1
2	3	3	3	3	2	3	3	0	1	0	0	0	0	0	0	1	0	0	0	· · ·	6	0
1	5	4	4	3	2	4	2	1	0	0	0	0	0	0	0	0	0	1	0	· · ·	1	0
4	2	2	2	4	3	3	1	0	0	0	1	0	0	0	1	0	0	0	0	· · ·	2	0
5	1	1	1	1	4	2	2	0	0	0	0	1	0	1	0	0	0	0	0	· · ·	0	0
·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	· · ·	·	·
·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	· · ·	·	·
·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	· · ·	·	·

Moving horizontally we have 1 missing, 2 missings, and so on, with a triangular matrix structure of the respondents due to the near orthogonality of the two variables.

To visualize the demographic groups, centroids are calculated of respondent points in Exhibit 14.8 for each country (Exhibit 14.9) and for each of the age and education groups (Exhibit 14.10). The biggest dispersion is seen in Exhib-

Exhibit 14.8:
Canonical MCA of the indicator matrix with constraining variables the counts of middles and missings (#M and #X). The respondents pile up at discrete positions at the centres of the circles, the areas of which indicate the frequencies



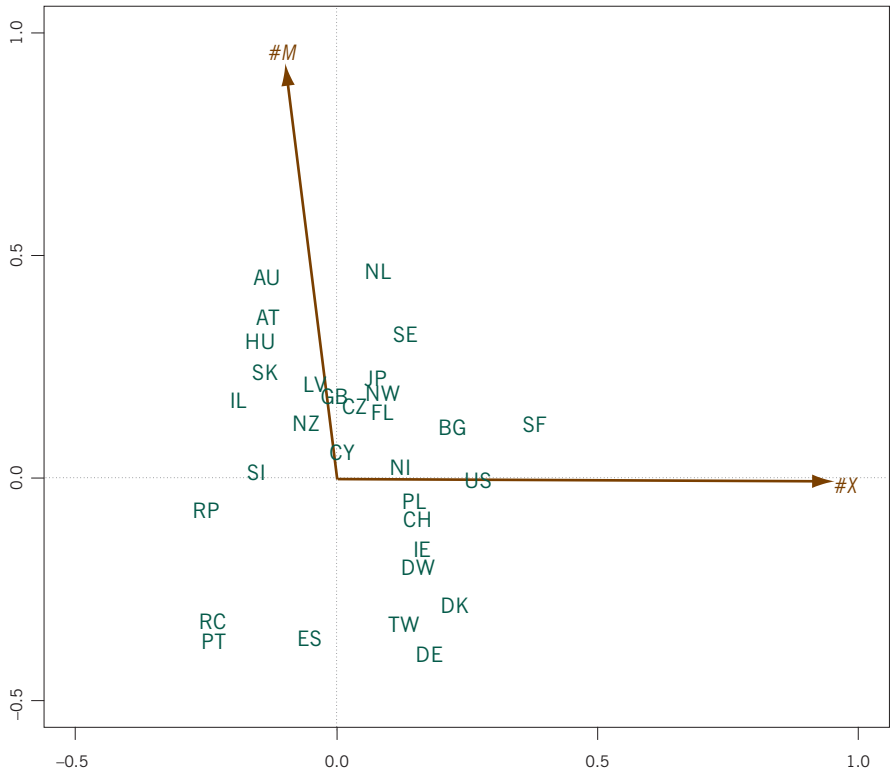


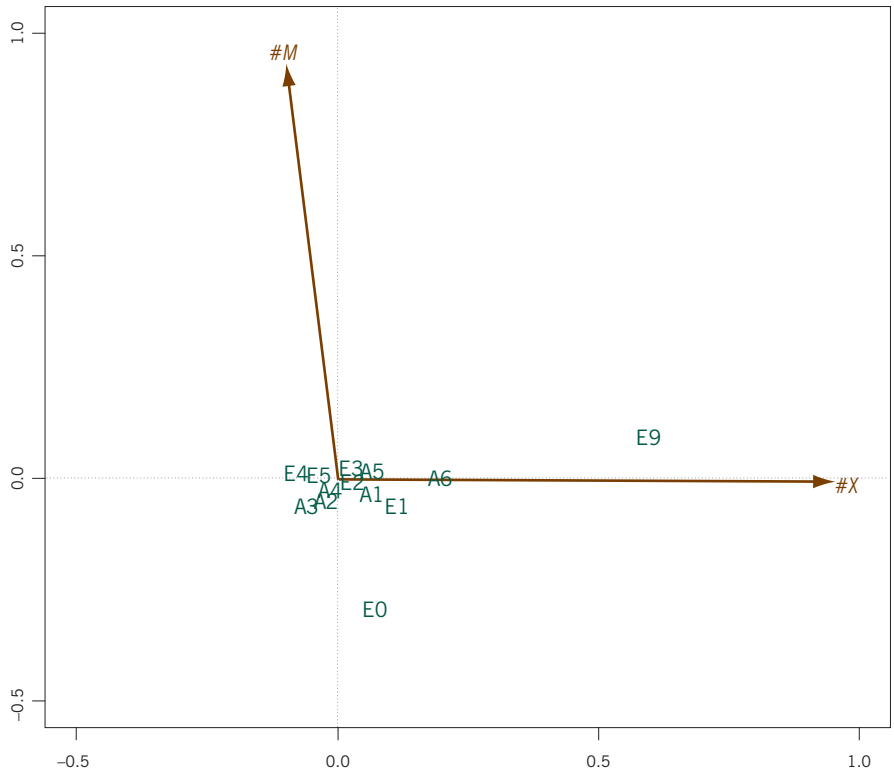
Exhibit 14.9:
Country centroids of the respondents in Exhibit 14.8

it 14.9 for the countries, with Australia and the Netherlands showing the highest use of the middle responses, and Finland the highest on missing responses. Portugal and China give missings and middles the least amongst this group of countries. In Exhibit 14.10 respondents which have their education group missing (E9, 520 cases) are far in the direction of missing responses, while the lowest education group E0 is less than average on middles; the highest education groups E4 and E5 are slightly less than average on missings. The age groups A2, A3 and A4 (26–55 years) are also slightly less than average on missings, while the youngest and oldest groups are slightly higher on average, especially A6 (66+ years). There is no age or education group with a tendency to use the middle responses.

Subset analysis of middle categories

In contrast to the canonical analysis which focuses on a single dimension of middle responses, a subset analysis focuses on the dimensions of all 8 middle categories. Exhibit 14.11 shows two views of the subset MCA of the middle categories, in standard coordinates. The first dimension (horizontal dimension in left hand map) puts all middle categories on the right, so this dimension will coincide with the biplot arrow "*#M*" in Exhibits 14.8–14.10 which simply counts the middles. The second and third dimensions, in the right hand map of Exhibit 14.11, shows

Exhibit 14.10:
Education and age group centroids of the respondents in Exhibit 14.8



that there is a clustering of the middle categories of the first three questions *AM*, *BM*, *CM* (top left), then those of the next four questions *DM*, *EM*, *FM*, *GM* (top right), and quite separately the last question's middle category *HM* at the bottom. The right hand map of Exhibit 14.11 contains information about grouping of middle responses that was not evident in the canonical MCA. The fact that the middle categories group together according to the sequences of questions might indicate a certain type of behaviour on the part of the respondents where they give middle responses to sequences of questions. We can investigate if there is any demographic variable that coincides with this phenomenon.

As in all MCA analyses, each respondent has a position in the map, so any demographic grouping can be represented by a set of centroids. Exhibit 14.12 shows the average positions of the 32 countries. Corresponding to the slightly diagonal orientation of the two clusters at the top of the right hand map of Exhibit 14.11, there are two sets of countries extending from bottom left to top right, indicated by two dashed lines. These correspond to countries that have more than average middle responses on the two clusters of questions, whereas their vertical

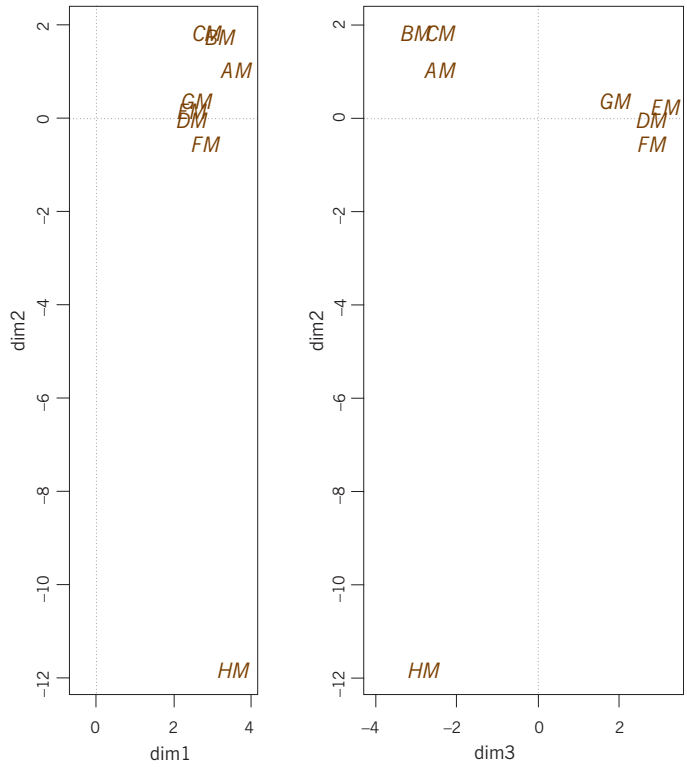


Exhibit 14.11:
Subset MCA of the 8 middle response categories, dimensions 1 by 2 (left) and dimensions 3 by 2 (right). Three clusters are evident in the right hand map

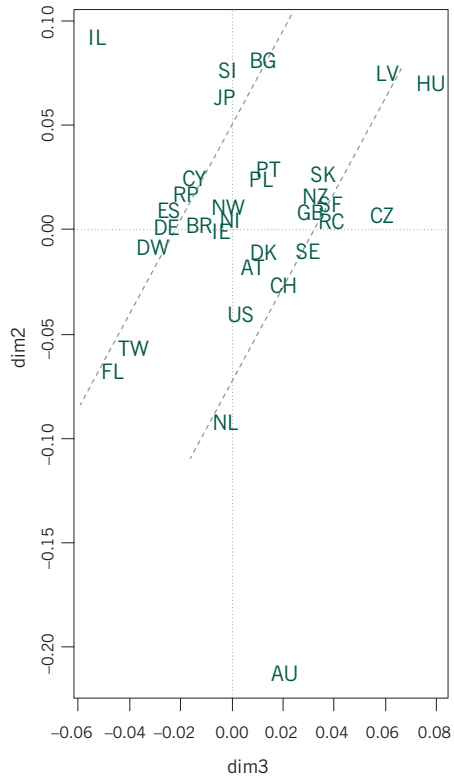
position depends on the number of middle responses on the isolated question *H*. Australia and the Netherlands, for example, which we previously saw had a high level of middle responses, both have a particularly high level on question *H*: going back to the original data, 20.5% of Australians and 13.8% Dutch responded *HM*, whereas for the other 30 countries the average response rate for this category was only 3.9%.

SUMMARY

We have shown how multiple correspondence analysis and its subset and canonical variants can allow a detailed investigation of the patterns of response in a large data set from a social survey. Based on similar studies that we have conducted on several different survey data sets of this type, from the ISSP and Eurobarometer, several results that emanate here appear to be typical:

1. In the analysis of individual responses, the missing response categories dominate. This is partially due to responses sets (all missings) by many respondents, which inflate the associations between the missing categories, but this association is strong even when the response sets are eliminated.

Exhibit 14.12:
Centroids of the countries in the right hand map of Exhibit 14.11. dashed lines indicate a set of countries with more than average middle responses on the first three questions (on left) *and on the next four questions* (on right), *with vertical spreads depending on the incidence of middle response on the last question*



2. Although the middle response categories might appear to be positioned "correctly" amongst the moderate categories in the initial best planar view of the data, inspection of further dimensions shows them to be associated among one another, and separated from their expected positions between the moderate categories. Again, we have found that this phenomenon persists even when response sets (all middles) are removed.

3. A middle response on one question is not generally associated with non-response on other questions. It seems that, if the middle response is used as a satisficing alternative to a non-response, there are some respondents that generally give just middle responses while there are others that give just non-responses.

4. Canonical MCA can be used to isolate a single dimension corresponding to the middle responses, similarly for the missing responses (or whatever responses the researcher is interested in). The positions of the respondents on these dimensions can be averaged within demographic groups to investigate their possible dispersions on these dimensions.

5. Subset MCA can be used to study the middle responses in more detail. As many dimensions as there are middle responses are analysed, so that dimension reduction is necessary to create a map. Generally, the first dimension in this analysis corresponds to the single constraining variable of "middleness" in the canonical MCA, so that the following dimensions reveal the more detailed patterns in middle response.
6. The same approach can be used to investigate patterns of any particular response category or categories. For example, the dimensions of the set of extreme response categories (1's and 5's) could be studied on their own and related to the demographic characteristics.