Three More Examples

10

To conclude these first 10 introductory chapters to the CA of a two-way table, we now give three additional examples: (i) a table which summarizes the classification of scientists from ten research areas into different categories of research funding; (ii) a table of counts of 92 marine species at a number of sampling points on the ocean floor; (iii) a linguistic example, where the letters of the alphabet have been counted in samples of texts by six English authors. In the course of these examples we shall discuss some further issues concerning two-dimensional displays, such as the interpretation of dimensions, the difference between asymmetric and symmetric maps, and the importance of the aspect ratio of the map.

Contents

Data set 5: Evaluation of scientific researchers

The data come from a scientific research and development organization which has classified 796 scientific researchers into five categories for purposes of allocating research funds (Exhibit 10.1). The researchers are cross-classified according to their scientific discipline (the 10 rows of the table) and funding category (the five columns of the table). The categories are labeled *A*, *B*, *C*, *D* and *E*, and are in order from highest to lowest categories of funding. Actually, *A* to *D* are the categories for researchers who are receiving research grants, from *A* (most funded) to *D* (least funded), while *E* is a category assigned to researchers whose applications were not successful (i.e., funding application rejected).

Exhibit 10.1: *Frequencies of funding categories for 796 researchers who applied to a research agency: A is the most funded, D is the least funded, and E is not funded.*

SCIENTIFIC	*FUNDING CATEGORIES*					
AREAS	*A*	*B*	*C*	*D*	*E*	*Sum*
Geology	3	19	39	14	10	*85*
Biochemistry	1	2	13	1	12	*29*
Chemistry	6	25	49	21	29	*130*
Zoology	3	15	41	35	26	*120*
Physics	10	22	47	9	26	*114*
Engineering	3	11	25	15	34	*88*
Microbiology	1	6	14	5	11	*37*
Botany	0	12	34	17	23	*86*
Statistics	2	5	11	4	7	*29*
Mathematics	2	11	37	8	20	*78*
Sum	*31*	*128*	*310*	*129*	*198*	*796*
Average Row Profile	*3.9%*	*16.1%*	*38.9%*	*16.2%*	*24.9%*	

Decomposition of inertia

This 10×5 table lies exactly in four-dimensional space and the decomposition of inertia along the four principal axes are as follows:

Dimension	*Principal inertia*	*Percentage of inertia*
1	0.03912	47.2%
2	0.03038	36.7%
3	0.01087	13.1%
4	0.00251	3.0%

Each axis accounts for a part of the inertia, expressed as a percentage. Thus the first two dimensions account for almost 84% of the inertia. The sum of the principal inertias is 0.082879, so the χ^2 statistic is $0.082879 \times 796 = 65.97$. If one wants to perform the statistical test using the χ^2 distribution with $9 \times 4 = 36$ degrees of freedom, this value is highly significant ($P = 0.002$).

Asymmetric map of row profiles

Exhibit 10.2 shows the asymmetric map of the row profiles and the column vertices. In this display we can see that the magnitude of the association between the disciplines and the research categories is fairly low; in other words the profiles do not deviate too much from the average (cf. Exhibit 4.2). This situation is fairly typical of social science data, so the asymmetric map is not so successful because all the profile points are bunched up in the middle of the display — in fact, they are so close to one another that we cannot write the full labels and have just put the first two letters of each discipline. Nevertheless, we can interpret the space easily looking at the positions of the vertices. The horizontal dimension lines up the four categories of funding in their inherent ordering, from *D* (least funded) to *A* (most funded), with *B* and *C* close together in the middle. The vertical dimension opposes category *E* (not funded) against the others, so the interpretation is fairly straightforward. The more a discipline is high up in this display the less its researchers are actually granted funding. The more a discipline lies to the right of this display, the more funding its funded researchers receive. Using marketing research terminology,

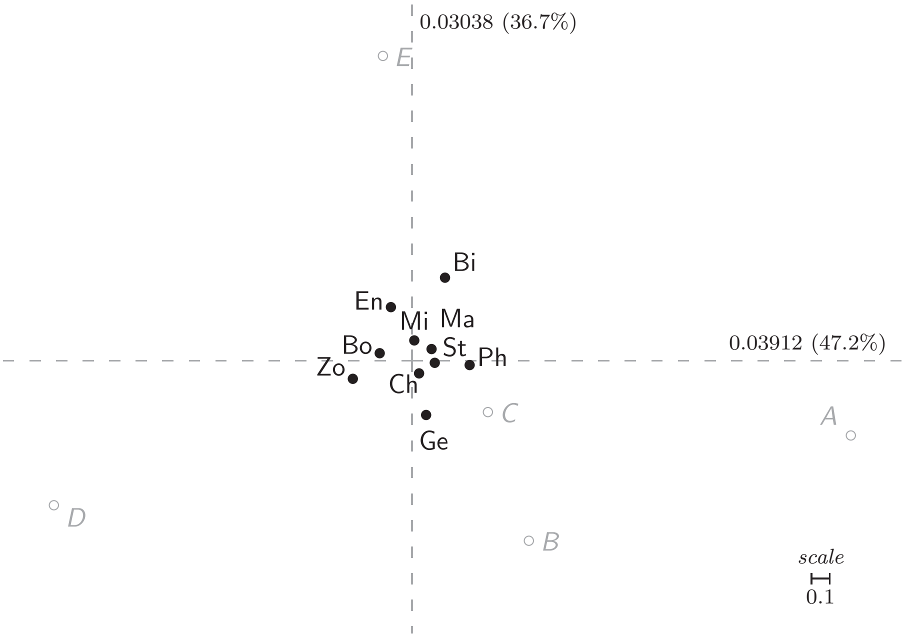


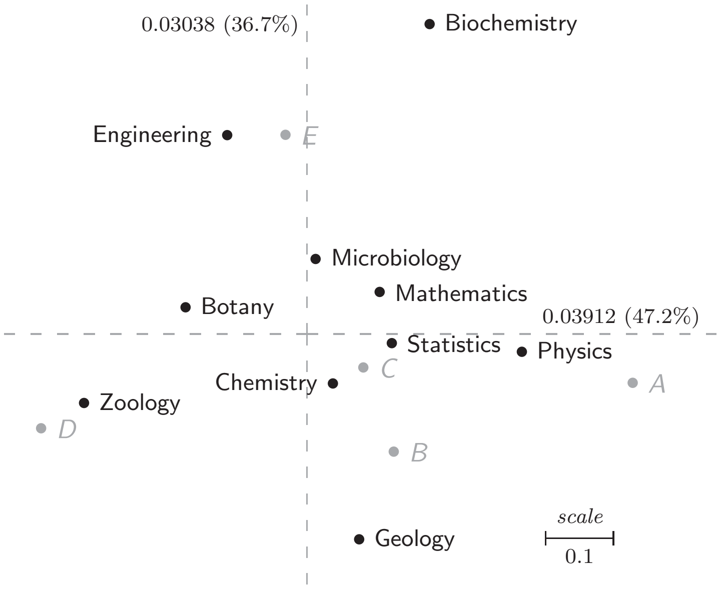
Exhibit 10.2: *Asymmetric map of the row profiles of Table 10.1 (scientific funding data).*

the "ideal point" is in the lower right of the map: more grant applications accepted (low down), and those accepted receiving good classifications (to the right). Hence, if we were doing a trend study over time, disciplines would need to move towards the bottom right-hand side to show an improvement in their funding status. At the moment there are no disciplines in this direction, although Physics is the most to the right (highest percentage — 10 out of 114, or 8.8% — of type *A* researchers), but is at the middle vertically since it has a percentage of non-funded researchers close to average (26 out of 114 not funded, or 22.8%, compared to the average of 198 out of 796, or 26.5%).

Symmetric map

Exhibit 10.3 shows the symmetric map of the same data, so that the only difference between this display and that of Exhibit 10.3 is that the column profiles are now displayed rather than the column vertices, leading to a change in scale which magnifies the display of the row profiles. This zooming in on the configuration of disciplines facilitates the interpretation of their relative positions and also gives space for fuller labels. The relative positions of the disciplines can now be seen more easily: for example, Geology, Statistics, Mathematics and Biochemistry are all at a similar position on the first axis, but widely different on the second. This means that the researchers in thse fields whose grants have been accepted have similar positions with respect to the funded categories *A* to *D* categories, but Geology has much fewer rejections (11.8% of category *E*) than Biochemistry (41.4%). In this symmetric display we cannot assess graphically the overall level of association (inertia) between the rows and the columns. This can be assessed only from the numerical

Exhibit 10.3: *Symmetric map of Table 10.1 (scientific funding data).*



value of the principal inertias along the axes, or their square roots which are the canonical correlations along each axis, namely $\sqrt{0.039117} = 0.198$ and $\sqrt{0.030381} = 0.174$, respectively. The level of row–column association can be judged graphically only in an asymmetric map such as Exhibit 10.2 (compare again the different levels of association illustrated in Exhibit 4.2).

Dimensional interpretation of maps

Whether the joint map is produced using asymmetric or symmetric scaling, the *dimensional* style of interpretation remains universally valid. This involves interpreting one axis at a time, as we did above and as is customary in factor analysis, using the relative positions of one set of points — the "variables" of the table — to give a descriptive name to the axis. For example, we used the funding category points to give a descriptive name to the axes and then interpreted the discipline points with respect to the axes. All statements in such an interpretation are relative and it is not possible to judge the absolute difference in funding profiles between the disciplines unless we refer to the original data. Putting this another way, symmetric maps similar to Exhibit 10.3 could be obtained for other data sets where there are much larger (or smaller) levels of association between the funding profiles of the disciplines.

Data set 6: Abundances of marine species in sea-bed samples

CA is used extensively to analyse ecological data, and the second example represents a typical data set in marine biology. The data, given partially in Exhibit 10.4, are the counts of 92 marine species identified in 13 samples from the sea-bed in the North Sea. Most of the samples are taken close to an oil-drilling platform where there is some pollution of the sea-bed, while two samples, regarded as reference samples and assumed unpolluted, are taken far from the drilling activities. These data, and biological data of this kind in

	STATIONS (SAMPLES)												
SPECIES	S4	S8	S9	S12	S13	S14	S15	S18	S19	S23	S24	R40	R42
Myri.ocul.	193	79	150	72	141	302	114	136	267	271	992	5	12
Chae.seto.	34	4	247	19	52	250	331	12	125	37	12	8	3
Amph.falc.	49	58	66	47	78	92	113	38	96	76	37	0	5
Myse.bide.	30	11	36	65	35	37	21	3	20	156	12	58	43
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
Eucl.sp.	0	0	0	0	1	0	0	1	1	0	0	0	0
Scal.infl.	0	1	0	0	0	1	0	0	0	0	0	0	1
Eumi.ocke.	0	0	1	0	0	1	1	0	0	0	0	0	0
Modi.modi.	0	0	0	1	1	0	0	1	0	0	0	0	0

Exhibit 10.4: *Frequencies of 92 marine species in 13 samples (the last two are reference samples); the species (rows) have been ordered in descending order of total abundance; hence four most abundant and four least abundant are shown here.*

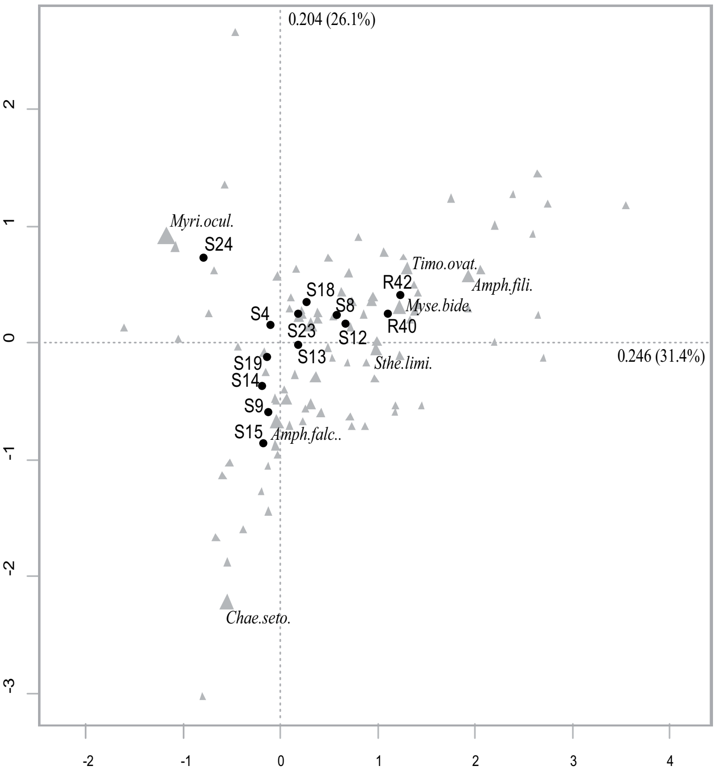


Exhibit 10.5: *Asymmetric CA map, with stations in principal coordinates and the species in standard coordinates. The species symbols have size proportional to the species abundance (mass) — some important species in the analysis are labelled with the first letter of the label being close to its corresponding triangular symbol. Inertia explained in map: 57.5%.*

general, are characterized by high variability, which can already be seen by simple inspection of the small part of the data given here. The total inertia of this table is 0.7826, much higher than in the previous examples, so we can expect the profiles to be more spread out relative to the vertices. Notice that in this example the χ^2-test is not applicable, since the data do not constitute a true contingency table — each individual count is not independent of the others, since the marine organisms often occur in groups at a sampling point.

Asymmetric CA map of species abundance data

Exhibit 10.5 shows the asymmetric map of the sample (column) profiles and species (row) vertices. Since there are 92 species points, it is impossible to label each point so we have labelled only the points which have a high contribution to the map; these are generally the most abundant ones. (The topic of how to measure this contribution is described in the Chapter 11, for the moment let us simply report that 10 out of the 92 species contribute over 85% to the construction of this map, the other 82 could effectively be removed without the map changing very much.) The stations form a curve from bottom left (actually, the most polluted stations) to top right (the least polluted), with the reference stations far from the drilling area at upper right. An exception is station 24, which separates out notably from the others, mainly because of the very high abundance of species *Myri.ocul.* (*Myriochele oculata*) which can be seen in the first row of Exhibit 10.4. The most abundant species are labelled and it is mainly these that determine the map. Notice that the asymmetric map does well in this example because the inertia is so high, which is typical of ecological data where there is high variability between the samples. The next example is the complete opposite!

Exhibit 10.6: *Letter counts in 12 samples of texts from books by six different authors, showing data for 9 out 26 letters.*

BOOKS	*a*	*b*	*c*	*d*	*e*	···	*w*	*x*	*y*	*z*	*Sum*
TD-Buck	550	116	147	374	1015	···	155	5	150	3	*7144*
EW-Buck	557	129	128	343	996	···	187	10	184	4	*7479*
Dr-Mich	515	109	172	311	827	···	156	14	137	5	*6669*
As-Mich	554	108	206	243	797	···	149	2	80	6	*6510*
LW-Clar	590	112	181	265	940	···	146	13	162	10	*7100*
PF-Clar	592	151	251	238	985	···	106	15	142	20	*7505*
FA-Hemi	589	72	129	339	866	···	225	1	155	2	*6877*
Is-Hemi	576	120	136	404	873	···	250	3	104	5	*6924*
SF7-Faul	541	109	136	228	763	···	160	11	280	1	*6885*
SF6-Faul	517	96	127	356	771	···	216	12	171	5	*6971*
Pe3-Holt	557	97	145	354	909	···	194	9	140	4	*6650*
Pe2-Holt	541	93	149	390	887	···	218	2	127	2	*6933*

Abbreviations:
TD (Three Daughters), EW (East Wind) -Buck (Pearl S. Buck)
Dr (Drifters), As (Asia) -Mich (James Michener)
LW (Lost World), PF (Profiles of Future) -Clar (Arthur C. Clarke)
FA (Farewell to Arms), Is (Islands) -Hemi (Ernest Hemingway)
SF7 and SF6 (Sound and Fury, chapters 7 and 6) -Faul (William Faulkner)
Pen3 and Pen2 (Bride of Pendorric, chapters 3 and 2) -Holt (Victoria Holt)

Data set 7: Frequencies of letters in books by six authors

This surprising example is a data set provided in the **ca** package of the R program (see Computational Appendix, pages 222–223). The data form a 12 × 26 matrix with the rows representing 12 texts which form six pairs, each pair by the same author (Exhibit 10.6 shows a part of the matrix).

The columns are the 26 letters of the alphabet, *a* to *z*. The data are the counts of these letters in a sample of text from each of the books. There are approximately 6500-7500 letter counts for each book or chapter.

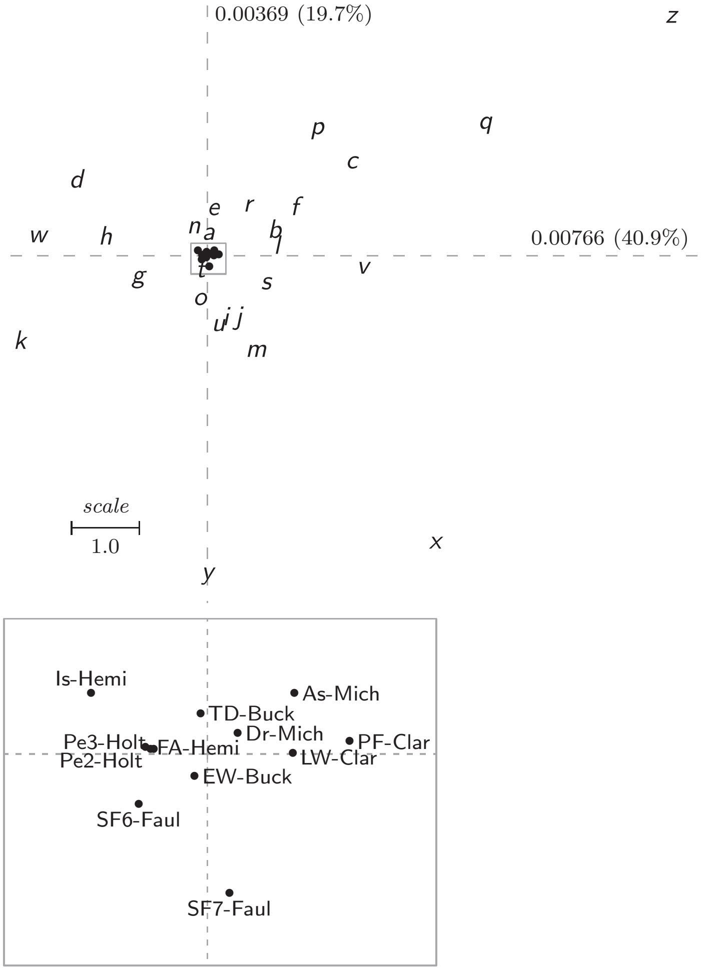


Exhibit 10.7: *Asymmetric CA map of the author data of Table 10.6, with row points (texts) in principal coordinates. The very low inertia in the table is seen in the closeness of the row profiles to the centroid. A "blow-up" of the rectangle at the centre of the map shows the relative positions of the row profiles.*

One of the lowest inertias, but with a significant structure

This data set has one of the lowest total inertias I have seen in my experience with CA: the total inertia is 0.01873, which means that the data are very close to the expected values calculated from the marginal frequencies; i.e., the profiles are almost identical. The asymmetric map of these data is shown in Exhibit 10.7, showing the letters in their vertex positions and the 12 texts as a tiny blob of points around the origin, showing how little variation there is between the texts in terms of letter distributions, which is what one would expect. If one expands the tiny blob of points, it is surprising to see how much structure there is within such tiny variation. Each pair of texts by the same author lies in the same vicinity, and the result is highly significant from a statistical viewpoint (we discuss the permutation test for testing this in Chapter 25).

Preserving a unit aspect ratio in maps

An important final remark concerns the physical plotting of two-dimensional correspondence analysis maps. Since distances in the map are of central interest, it is clear that a unit on the horizontal axis of a plot should be equal to a unit on the vertical axis. Even though this requirement seems obvious, it is commonly overlooked in many software packages and spreadsheet programs that produce scatterplots of points with different scales on the axes. For example, the points might in reality have little variation on the vertical second axis, but the map is printed in a pre-defined rectangle which then exaggerates the second axis. We say that the *aspect ratio* of the map, that is the ratio of one unit length horizontally to one unit vertically, should be equal to 1. A few options for producing good quality maps are discussed at the end of the Computational Appendix.

SUMMARY: Three More Examples

1. When applicable, it is useful to test a contingency table for significant association, using the χ^2 test. However, statistical significance is not a crucial requirement for justifying an inspection of the maps. CA should be regarded as a way of re-expressing the data in pictorial form for ease of interpretation — with this objective any table of data is worth looking at.

2. In both asymmetric and symmetric maps the dimensional style of interpretation is valid. This applies to one axis at a time and consists of using the relative positions of one set of points on a principal axis to give the dimension a conceptual name, and then separately interpreting the relative positions of the other set of points along this named dimension.

3. The asymmetric map functions well when total inertia is high, but it is problematic when total inertia is small because the profile points in principal coordinates are too close to the origin for easy labelling.

4. It is important to have plotting facilities which preserve the *aspect ratio* of the display. A unit on the horizontal axis must be as close as possible to a unit on the vertical axis of the map; otherwise distances will be distorted if the scales are different.