Chapter 9

Shot-scale data analysis

9.1 Correspondence analysis – initial example

In Section 2.2.6 the use of correspondence analysis for investigating patterns in shot-scale data was introduced (see, also, Baxter, 2012c). A more detailed exposition of the methodology, and its implementation in R is provided here, using 42 films of Alfred Hitchcock in Table 9.1 as a running example.

Title	Year	BCU	CU	MCU	MS	MLS	LS	VLS
Pleasure Garden, The	1925	13	47	66	98	104	153	18
Downhill	1927	36	25	89	93	98	154	5
Lodger, The	1927	29	94	74	78	106	119	0
Champagne	1928	37	65	78	90	116	108	5
Easy Virtue	1928	25	61	96	65	97	140	14
Farmer's Wife, The	1928	10	42	97	86	115	145	6
Manxman, The	1928	21	79	66	77	117	123	16
Ring, The	1928	41	65	90	89	98	112	6
Blackmail (Silent version)	1929	39	49	67	85	137	113	10
Blackmail (Sound version)	1929	42	41	87	74	112	128	16
Juno and the Paycock	1930	9	30	53	84	164	159	0
Murder	1931	33	69	89	74	92	138	5
Skin Game, The	1931	28	78	50	60	136	125	22
Number Seventeen	1932	42	58	51	90	103	126	29
Rich and Strange	1932	24	75	75	75	88	131	33
Waltzes from Vienna	1934	19	49	42	79	126	186	0
Man Who Knew Too Much, The	1935	35	48	90	102	103	109	13
Thirty-Nine Steps, The	1935	27	72	80	87	75	132	26
Secret Agent, The	1936	43	94	105	93	79	73	14
Sabotage	1937	63	87	104	96	64	79	8
Young and Innocent	1938	27	96	124	75	67	95	16
Jamaica Inn	1939	15	85	71	92	88	135	18
Lady Vanishes, The	1939	24	72	117	114	103	59	11
Foreign Correspondent	1940	22	126	103	79	84	77	9
Rebecca	1940	18	100	111	96	83	78	12
Saboteur	1942	29	90	74	74	90	114	28
Suspicion	1942	38	107	130	89	58	62	16
Shadow of a Doubt	1943	24	101	107	83	100	78	8
Lifeboat	1944	21	122	114	106	80	38	19
Notorious	1947	72	119	88	74	74	66	8
Paradine Case, The	1947	23	140	89	116	57	66	7
Strangers on a Train	1951	44	116	114	59	65	87	16
I Confess	1952	32	108	82	81	77	101	20
Dial M for Murder	1954	15	69	98	141	141	37	0
Rear Window	1954	17	75	113	74	72	121	29
Man Who Knew Too Much, The	1955	15	59	115	96	81	111	23
To Catch a Thief	1955	8	64	72	104	102	113	37
Trouble With Harry, The	1956	4	20	113	145	121	73	23
Wrong Man, The	1957	31	127	150	75	63	51	4
Vertigo	1958	15	113	104	84	56	92	35
North by North-West	1959	12	87	96	96	77	99	33
Birds, The	1963	59	127	111	69	57	58	19

Table 9.1: *Shot scale data (%) for films of Alfred Hitchcock.*

The data are as given in Barry Salt's database on the Cinemetrics site (i.e. scaled to sum to 500 rather than percentages as in Section 2.2.6[1]). The first nine films, to the silent version of *Blackmail*, are British silent films; the 14 from the sound version of *Blackmail* through to *The Lady Vanishes* are British sound films; the remaining 19 are American. The conventional bar chart representation of the data for the silent films is shown in Figure 9.1; bar charts for the other films are in Section 2.2.6.

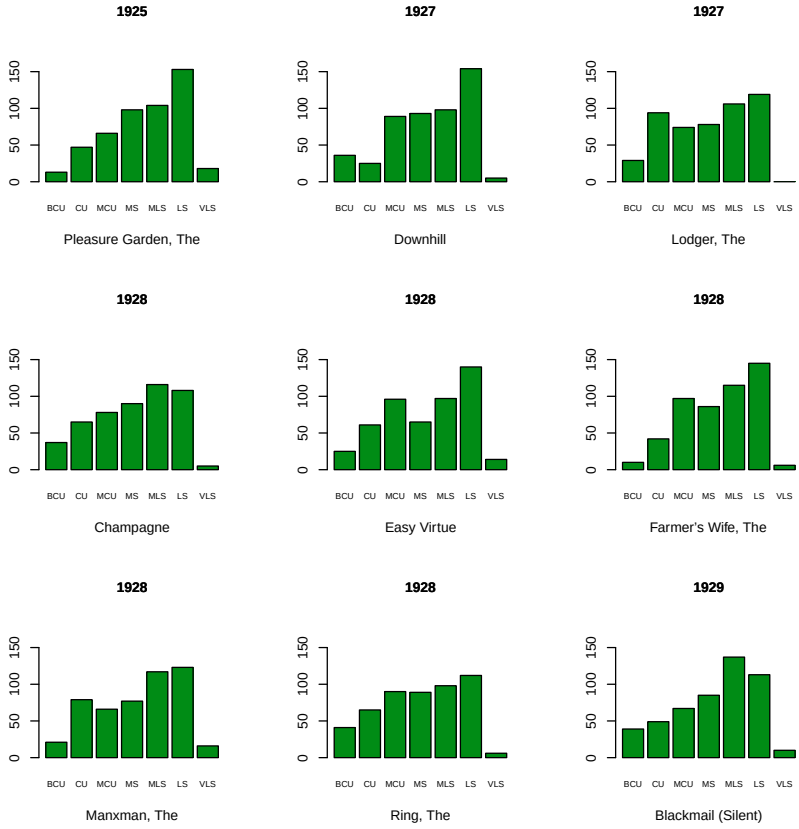

Figure 9.1: *Bar charts of shot-scale data for Hitchcock's silent films.*

Before looking at the idea behind correspondence analysis (CA) it is convenient to have an example to hand. A basic CA is shown in Figure 9.2, and was obtained using the `ca` function in the `ca` package, which needs to be imported then loaded using `library(ca)`. The data are read into `R`, named `Hitchcock` say, with shot-scales extracted as `shot <- Hitchcock[, 3:9]`. All that is then needed to get the figure is `plot(ca(shot))`.

The blue dots in the plot correspond to films and are labelled, by default, according to their order in the table. Films close to each other on the plot have similar shot-scale profiles; films at some distance apart have comparatively different profiles. The silent films are labelled 1-9, for example, and plot fairly close to each other so, as a broad statement – to which there are specific exceptions – it can be inferred that they tend to be more similar to each other than they are to most later films. Their similarity to each other is also evident in Figure 9.1.

The red triangles correspond to the different shot-scales and are labelled accordingly. Their interpretation was discussed at a little length in connection with Figure 2.9 and the salient points will be reiterated. For the moment it is sufficient to note that their positions allow useful inferences to be drawn about why points corresponding to films on the plot are similar or distant. For

[1]The scaling chosen has no effect on the analysis

example, it can be inferred that films to the left, plotting in the same general region as LSs and MLSs will feature comparatively more of these kind of shots than films to the right; conversely, those to the right are more likely to favour the various degrees of close-up.

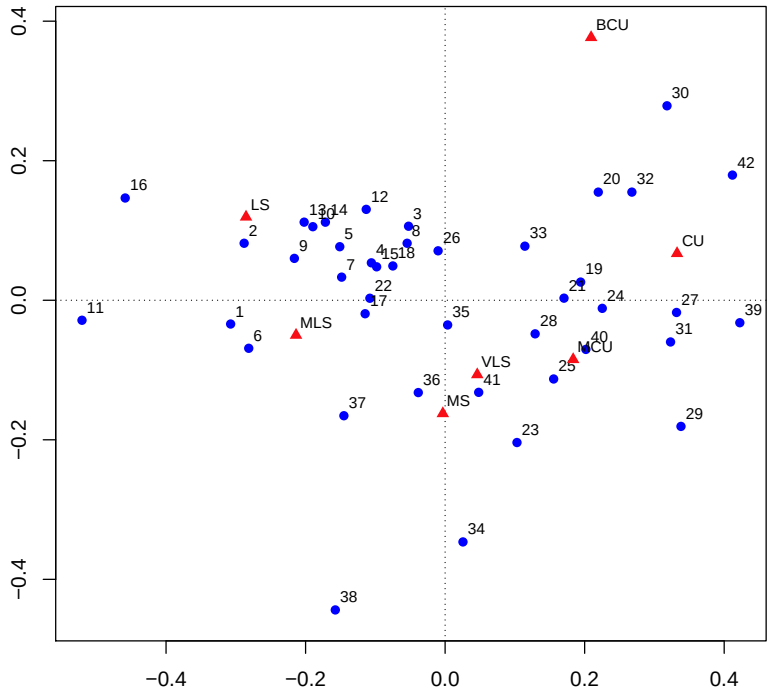

Figure 9.2: *A basic correspondence analysis of the shot-scale data from Table 9.1.*

There are some caveats about interpretation, discussed shortly, that need to be understood but, fundamentally, this is how a large number of applications of CA are presented and interpreted. Here, once the machinery has been set up, the analysis is executed by – in this example – typing one 'word' of 14 characters. For exploratory purposes the default output obtained using the `ca` function is often adequate[2]. All CA is doing, as applied here, is reducing a table of data to a picture; given its utility and ease of application the reasons for its popularity in a wide range of disciplines is obvious.

For publication purposes, or communicating results with others, the default CA plots produced by most software can usually be improved. Figure 9.2 is too 'busy' for my taste; duplicating points for films and labels doesn't seem necessary; information on whether the film is silent/sound, British/American needs to be recovered by referring the numbers on the plot back to the table, and so on.

Figure 9.3 is an alternative version of the plot. Coloured symbols replace the solid circles and labels; it is immediately apparent that the silent films form a comparatively tight cluster compared to the British and American sound films. The latter are rather spread out but, with the exception of four British sound films, do not overlap with other films. It is easily enough established that the overlapping British sound films are four of the five latest.

[2]Other packages that include functions for correspondence analysis include `MASS`, `vegan`, `ade4`, `anacor`, `FactoMineR` etc. They vary a little in ease of application and the way CA maps are presented, but the basics are similar.

Joining up the markers for shot-scales emphasises that there is a reasonably good 'stylistic gradient' of the kind discussed in connection with Figure 9.1, the position of the VLS marker being anomalous. The left-hand side of the plot is associated with the longer shots, and the right-hand side with the closer shots and, in broad terms, tends to differentiate earlier from later films.

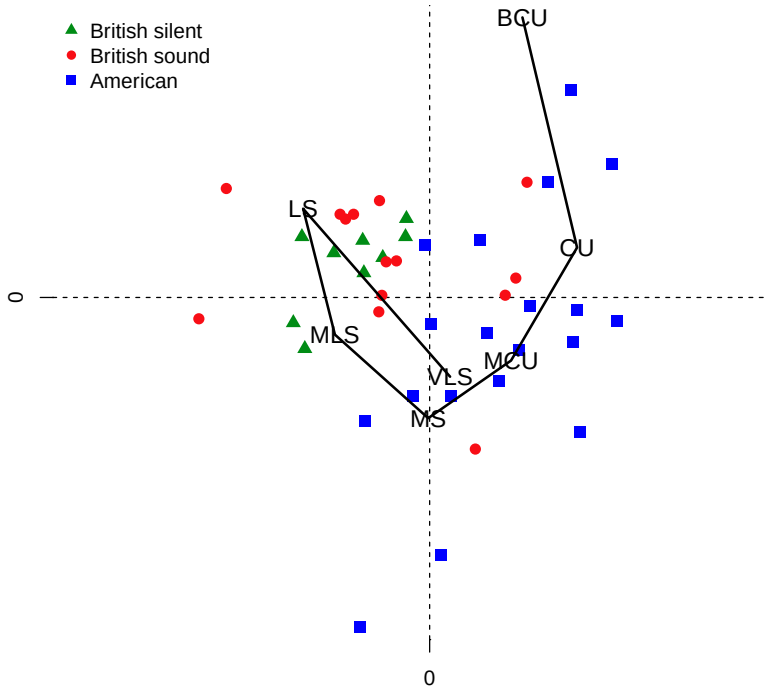

Figure 9.3: *An enhanced correspondence analysis of the shot-scale data from Table 9.1.*

9.2 What correspondence analysis does

While CA has earlier roots it is the (sometimes formidable) mathematical development of French mathematicians and statisticians, notably Jean-Paul Benzécri, that is usually cited as the important 'popularizing' development. Greenacre (1984), a student of Benzécri, published the first major English language text, also sufficiently mathematically demanding to prompt publication of Greenacre (2007), the second edition of a text first published in 1993, much cited and meriting the adjective popularizing without the use of inverted commas.

Since then CA has come to be widely used as a data-analytic tool across a range of disciplines in the sciences, social sciences and humanities. This has generated a large number of expository articles, often similar in kind but aimed at different audiences. My own contribution to this genre, Baxter and Cool (2010), was aimed at archaeologists; its relevance as far as this text is concerned is that there was a specific focus on `R`. The intended readers were assumed to have data they wanted to analyse, and knowledge that CA might be a useful tool, but no prior experience of `R` or carrying out their own CA[3].

[3]The paper can be found on the web, in more than one place, by Googling sensible terms.

The appeal of CA is down to both its utility and conceptual simplicity. As usually applied, ignoring the mathematics and 'deeper' theoretical issues, CA takes a (possibly) large table of non-negative data and reduces it to a picture with rows of the table represented by one set of points, and columns by another. Focusing on rows (films) to be specific, rows with similar profiles should be close to each other on the plot; rows with dis-similar profiles should be distant. (The row profile is given by the numbers in a row scaled to have a common value; the shot scale data collected by Salt are already in this form as they add to 500[4]).

The two previous figures are in two dimensions and can be thought of as maps of the data. The dimensionality of the data set is defined by the number of columns, seven in the example. The distance between any two rows, for example, can be defined precisely and mathematically but if there are more that three columns, can't be represented exactly using conventional plotting methods. The real data can be thought of as a 'cloud' of points, 'living' in seven dimensions that you can't visualise; to get a look at it the data needs to be 'squashed' (mathematically, of course) onto a two-dimensional plane that preserves as well as possible distances between points, and hence any patterns that characterise relationships between them.

This means that the end result is an *approximation* to the reality. The quality of this approximation can be judged in various ways, which will not be dealt with here[5]. The other technical point to note, but not worry about too much, is that mathematicians can define distance in different ways, and what is being approximated is *chi-squared distance*. It differs from what you are used to in daily life when you talk about physical 'distance', but not enough to worry about for the purpose of the present discussion.

Those who, quite rightly, are uneasy about accepting these blasé assurances and consult other literature need to be warned about some of what is around, particularly the more evangelical kind of exposition that can oversell CA. It is often emphasised that CA is appropriate for analysing tables of counted data, which is true but ignores the fact that CA can be applied to any table of non-negative numbers (with care) – that is, it actually has a wider range of application than some advocates imply. This does no harm; more misleading is the not uncommon suggestion that one of the defining features of CA is its ability to *jointly* represent row and column points on the same plot, each informing the others interpretation. Two comments can be made here; one is that this feature is shared by other statistical methods (Greenacre, 2010); the other is that there is absolutely no reason why just the row or column plot can't be reported separately. CA is a useful, decriptive, data-analytic tool to be applied as the user sees fit.

There are some other issues to be dealt with. These will arise naturally in what follows, where the plan is to show how Figure 9.3 can be constructed incrementally, by building separate plots for the rows and columns before overlaying them.

9.3 Producing enhanced CA plots

Remember that `plot(ca(shot))` was all that was needed to get Figure 9.2, the table of shot-scale data being held in `shot`. Proceed slightly differently as follows, where the first command carries out the CA without plotting and remaining commands extract the coordinates needed for plotting (note the use of the semi-colon to put several commands on the same line).

```
CAH <- ca(shot)
x <- CAH$rowcoord[,1:2]
y <- CAH$colcoord[,1:2]
x1 <- x[,1]; x2 <- x[,2]; y1 <- y[,1]; y2 <- y[,2]
```

At this point `plot(x1, x2)` would give a reasonably uninformative plot of the rows (films). Adding symbols can be done in various ways, one of which is

[4]It is not necessary that data be collected in this format; an early mathematical step in CA, invisible to the user, is to effect this scaling.

[5]Some authors are quite stern about this, and quantities called *inertias* can be used both to investigate how good the overall approximation is, and how well individual rows or columns are represented. The literature cited provides an entry point.

```
Symbol <- c(rep(17,9), rep(16, 14), rep(15, 19))
Colour <- c(rep("green4",9), rep("red", 14), rep("blue", 19))

library(MASS)  # Loads package MASS needed for eqscplot

eqscplot(x1, x2, pch = Symbol, col = Colour)
abline(v = 0, lty = 2); abline(h = 0, lty = 2)
```

The `rep(a, b)` function produces `b` copies of `a`. The `c(x, y, z)` structure strings together the contents of `x,y` and `z`; 17, 16 and 15 are the plotting symbols for solid triangles, circles and squares. The `eqscplot()` requires `MASS` and ensures equal scaling of the axes, important for the distance interpretation of the map. The arguments `pch` and `col` specify plotting symbols and colours; axis labelling arguments have been omitted. The `abline` commands add dashed vertical and horizontal lines at 0, to provide a reference grid. The outcome is the left-hand plot in Figure 9.4.

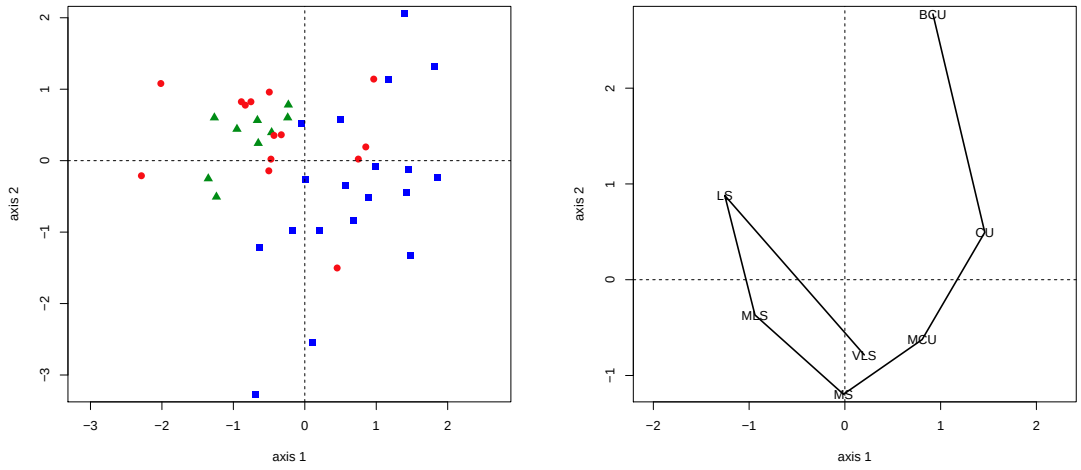

Figure 9.4: *Separate row and column plots for a CA of Table 9.1.*

For the column plot to the right, assuming `MASS` is loaded the following code was used, where labelling arguments are omitted again. Legends can be added to the plots, as desired.

```
Names <- names(shot)
eqscplot(y1,y2, type = "n")
text(y1, y2, Names)
lines(y1,y2, lwd = 2)
abline(v = 0, lty = 2); abline(h = 0, lty = 2)
```

This is a perfectly good joint representation of the data, and you can readily interpret it in the same way as Figure 9.3 (which is where the grid lines are useful). Some authors (myself included) often prefer to present results in this way since it avoids some of the clutter that can arise with large tables where the overlaid plots can become difficult to read. Nevertheless code for overlaying the plots will be illustrated. This is partly to introduce features of `R` which may be of interest, but also because it raises an important interpretational issue.

In Section 2.2.6 an analogy was suggested whereby the film markers could be thought of as settlments in a landscape whose terrain was defined by the column markers. This is achieved by overlaying the two maps of Fig 9.4. The problem is that it is not obvious how to do this since, before overlaying the maps, it is legitimate to stretch one map relative to another. Suppose, for example, that the row map is fixed. Imagine that the column map is on a rubber sheet and pin it

to the row map at the origin (the (0,0) point) with the grid (the dashed lines) aligned. Now grab the rubber sheet at each corner, pull out, and stretch the sheet uniformly, so that eventually the joint map, defined by the stretched sheet, will have column markers on its periphery, with row markers squeezed up in the middle. The role of the maps can be reversed, so you can get a plot with column markers so squeezed up.

There are other ways of stretching and how its done can be expressed quite precisely in mathematical terms. Greenacre's texts may be consulted for technical details; the 'correct' way of stretching (in the view of some) leads to precisely the squeezing effect just described, and because of it the resultant map can be difficult to decipher. Accordingly, many practitioners opt for a kind of compromise, which is the *symmetric* map of the sort shown in Figures 9.2 and 9.3. Roughly, the overlaying is done to give equal emphasis to both row and column markers, with a more interpretable map resulting.

The better implementations of CA allow a choice of mapping. The symmetric map (which can also be defined in a precise mathematical way) is the default in the `ca` function, but this is not always the case for other implementations. Incidentally, plotting row and column maps separately avoids too much agonizing about the issue.

There is an important consequence of all this, which is that you *cannot* interpret the positioning of a row marker relative to a column marker in terms of 'distance' in the way that the positioning of two row markers can be interpreted as distance. For example, in Figure 9.3 the green triangle that sits almost on the LS marker is for *Downhill*. The film is 'near' or 'close' to the marker on this particular joint map, but if the column map was stretched this would no longer apply. With a symmetric map interpretation can often be carried out by thinking in terms of *relatively* 'nearer' or 'further', but do the thinking in inverted commas.

The plotting coordinates saved when using the `ca` function do not need modifying to get a symmetric map when the separate plots are overlaid. In the code to follow `z` combines plotting coordinates for both rows and columns, so that everything is included in the plot. The `type = "n"` argument produces a blank plot to which points and lines will be added. The numerical values on the axes, as in Figure 9.2, don't add much to interpretation (and vary according to how scaling is effected in different packages and functions) so are eliminated by the `axes = F` argument. The subsequent `axis` commands add labels for the grid produced by the `abline` commands.

```
z <- rbind(x,y)
library(MASS)
eqscplot(z[,1], z[,2], type = "n", xlab = "", ylab = "", axes = F)
abline(v = 0, lty = 2); abline(h = 0, lty = 2)
points(x1, x2, pch = Symbol, col = Colour, cex = 1.2)
text(y1, y2, Names, cex = 1.2)
lines(y1,y2, lwd = 2)
axis(1, at = 0, labels = 0)
axis(2, at = 0, labels = 0)
legend("topleft", legend = c("British silent", "British sound", "American"), pch = c(17,16,15),
col = c("green4", "red", "blue") ,bty = "n", cex = 1)
```

The `points` function adds the symbols and colours for the row markers at the coordinates defined by `x1` and `x2`. Note that these points are added outside the original `plot` command; for the individual row plot they were specified within the `plot` command. The `cex` aregument is the character expansion, with a default of 1 and used to control the size of plotting symbols. The `text` and `lines` commands add the column marker information, the lines being an optional add-on, and `Names` in the `text` command is the previously defined text `Names <- names(shot)` plotted at the coordinates `y1` and `y2`. Figure 9.3 is the end result.

Time has been taken over this to illustrate the kind of control that can be exercised if you want to go beyond the default plot – and this is not compulsory. Identifying individual films is straightforward, either by labelling the row plot with numbers rather than symbols, or even adding film titles, though this produces an unreadable plot. To illustrate, with embellishments, Figure 9.5 shows a plot for rows only, labelled by date (year since 1900).

It is clear that the earliest of the British sound films have a shot-scale structure close to those of the silents, whereas some of the later ones have a structure more akin to the American ones.

This is emphasised by constructing enclosures that separate out 'early' films, defined here as films up to and including 1932, and 'late' American ones from 1940 on. Four of the five latest British sound films from 1936-39 sit comfortably within the body of American films; *Jamaica Inn* (1939) is the exception, which lies between the two groups highlighted. The two 1935 films are also intermediate between the highlighted early and late groups – it should be emphasised that the positioning of films is only influenced by date to the extent that date and shot-scale structure are associated.

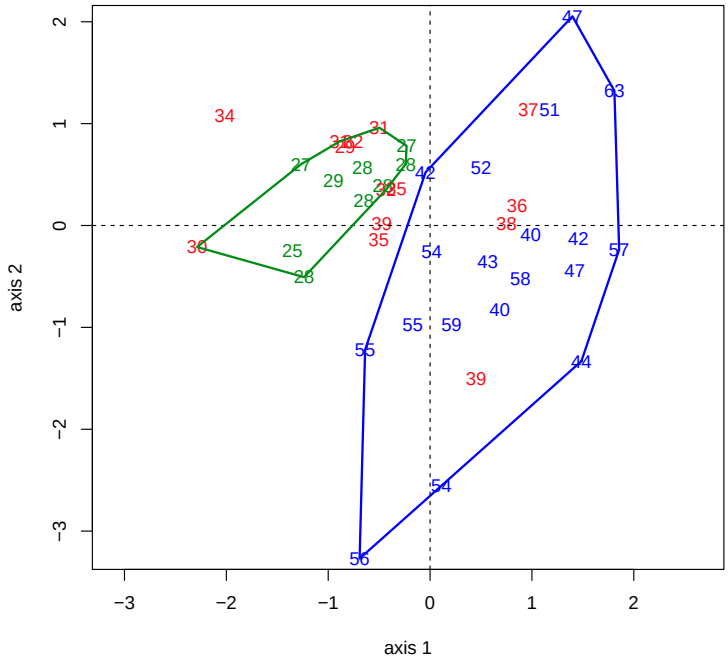

Figure 9.5: *The row plot from a CA of Table 9.1, labelled by date; green for silent, red for British sound and blue for American films. See the text for details of how the convex hulls are constructed.*

The 1934 film *Waltzes from Vienna* is an obvious outlier in the context of patterns displayed elsewhere in the plot. It jumps away from the main body of early films but in an opposite direction to those coming not long after (in the sample used here). The other Britsh sound film that behaves in the same way is *Juno and the Paycock* (1930), classified here as 'early' but an outlier *within* that group (it is the film furthest to the left in the relevant plots, with *Waltzes from Vienna* a bit above it). Both films are distinguished from all others by a relative paucity of close-ups, and *Waltzes from Vienna* is additionally distinguished by an excess of LSs[6].

The plot with dates colour coded by 'provenance' is very simply obtained.

```
date <- Hitchcock[,2] - 1900
eqscplot(x1, x2, type= "n", xlab = "axis 1", ylab = "axis 2")
text(x1, x2, date, col = Colour)
```

[6]If Hitchcock, in his conversations with Truffaut (1985, pp. 85-87), is to be believed *Waltzes from Vienna* is not, for other reasons, typical. He variously describes the film, made when his career was at a low ebb, as 'made cheaply', bearing no relation to his 'usual work' and 'very bad'. Later on (page 314) he does not demur when Truffaut suggests the film was 'not in your genre, you dismissed [it] as an out-and-out waste of time'. Truffaut fails to ask why this affected the shot-scale structure. Hitchcock wasn't too happy about *Juno and the Paycock* either saying (page 69) that he 'didn't feel like making the picture' because of the difficulty of narrating 'in cinematic form' what he describes as an 'excellent play'. He ends by saying that when the film got good notices he was 'ashamed because it had nothing to do with cinema'.

```
abline(v = 0, lty = 2); abline(h = 0, lty = 2)
```

Next, define two bodies of film `silent`, for the first 15 films in Table 9.1 that includes all the silents and British sound films to `Rich and Strange` in 1932; and `american` for all the American films from 1940 on. The eight British sound films from the 1934-39 period are accorded no special treatment.

```
silent <- x[1:15,1:2]
hpts <- chull(silent)
  hpts <- c(hpts, hpts[1])
  lines(silent[hpts, ], col = "green4", lwd = 2)
american <- x[24:42,1:2]
hpts <- chull(american)
  hpts <- c(hpts, hpts[1])
  lines(american[hpts, ], col = "blue", lwd = 2)
```

Added to the preceding code the bodies of early and late films defined above are enclosed, to produce Figure 9.5[7].

9.4 CA and supplementary points

If Figure 9.3 is compared with Figure 2.9 it will be seen that the configuration of those Hitchcock films common to both analyses is not identical. This is a consequence of the fact that the map produced is determined by all the data entered into an analysis. The main potential drawback of this, though it need not be serious, is that if there is an interest in comparing subsets of films to see if they are distinct, including them all in the CA analysis as has been done so far may blur distinctions.

There is a way round this. Suppose that there are two bodies of films one wishes to effect a comparison between. Construct the CA map using one body of data, then add points to the map, corresponding to films in the second body, as *supplementary* points. The idea is that they are brought along to a pre-existing landscape and then fitted into the part of the terrain that their shot-structure best suits them to. They cannot influence the terrain; they have to live with what's there.

For illustration 18 films of Max Ophuls will be used (Table 9.2). These will be fitted onto the map defined by the Hitchcock films that have been the subject of this chapter so far.

Title	Year	BCU	CU	MCU	MS	MLS	LS	VLS
Verliebte Firma, Die	1931	8	28	51	132	86	148	46
Lachende Erben	1932	8	14	45	114	145	125	49
Liebelei	1932	10	18	53	112	136	126	42
Verkaufte Braut, Die	1932	6	21	61	103	105	113	53
Signora di Tutti, La	1934	65	40	81	110	84	80	40
Komedie om Geld	1935	25	36	69	99	81	127	61
Tendre Ennemie, La	1936	12	46	58	119	90	108	66
Yoshiwara	1937	28	34	80	87	80	133	56
Werther	1938	10	34	43	72	91	180	71
Sans Lendemain	1939	20	69	86	124	90	85	26
De Mayerling à Sarajevo	1940	29	39	74	134	65	111	48
Exile, The	1948	7	15	49	109	140	143	37
Letter from an Unknown Woman	1948	16	50	71	116	116	108	24
Caught	1949	38	18	71	156	88	111	22
Reckless Moment, The	1949	22	37	76	118	86	113	49
Ronde, La	1950	5	66	41	163	127	71	25
Plaisir, Le	1952	0	7	46	87	120	153	87
Madame de ...	1953	4	22	71	120	161	88	29
Lola Montès	1955	1	20	55	114	108	135	61

Table 9.2: *Shot scale data (%) for films of Max Ophuls.*

If Figure 2.9 is examined it will be seen that with two exceptions there is little overlap between the Ophuls films and American Hitchcock, but rather more with the British Hitchcock sound films.

[7]The coding is not especially transparent. The 'enclosures' are what are called convex hulls and the coding is based on that given as an example in the help facilities `?chull`.

Plotting the Ophuls films as supplementary points on a map determined by *all* the Hitchcock films produces Figure 9.6.

There is still overlap but the distinction between the works of the two directors is sharper than was previously the case. This could be described in different ways, but it would be possible, for example, to construct a convex hull for 16/18 of the Ophuls films that enclosed only one Hitchcock in addition. The cluster of eight Ophuls films (one slightly astray) in the middle of the lower-left quadrant includes the four earliest Ophuls films in Table 9.2, German from the years 1931-32 and clearly distinct from the Hitchcock films of that date. Also in the group are the three last, French, films in the table dating from 20 years or more later. The films are distinguished by the relatively low number of close-ups and large numbers of MLSs, characteristics shared by *The Exile*, the only other film in this cluster[8].

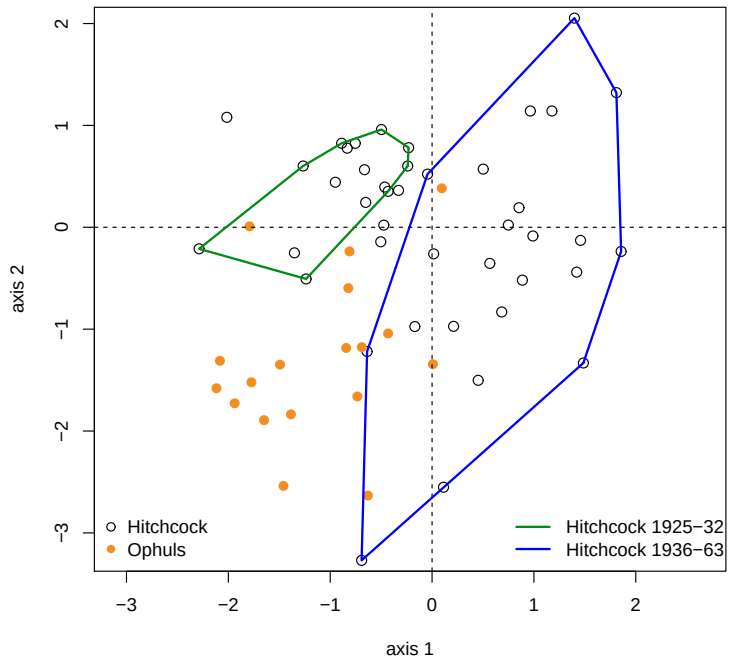

Figure 9.6: *The row plot from a CA of Hitchcock films from Table 9.1 (open circles) with supplementary points (closed circles) for the films of Ophuls from Table 9.2.*

The bones of the code needed are given below. It is assumed that a file `Ophuls`, similar to that of `Hitchcock` has been created; details of convex hull and legend construction are omitted.

```
sup <- Ophuls[, 3:9]
shotsup <- rbind(shot, sup)
CAso <- ca(shotsup, subsetrow = 1:42, suprow = 43:60)
x <- CAso$rowcoord[,1:2]
y <- CAso$colcoord[,1:2]
x1 <- x[,1]; x2 <- x[,2]; y1 <- y[,1]; y2 <- y[,2]
Symbol1 <- c(rep(1,42), rep(16,18))
Colour1 <- c(rep("black",42), rep("darkorange",18))
eqscplot(x1, x2, pch = Symbol1, col = Colour1, xlab = "axis 1", ylab = "axis 2", cex= 1.2)
abline(v = 0, lty = 2); abline(h = 0, lty = 2)
```

[8]In his chapter on the stylistic analysis of the films of Ophuls Salt (2009, p. 394) comments, in his discussion of *La Ronde*, on the similarity of shot-scale distribution of the late French films to the earlier sound films. Salt's analysis goes beyond what is attempted here, where the body of Ophuls films is being used simply to illustrate the idea of supplementary points.

9.5 Other plotting methods

9.5.1 Analysis of ranks

The utility of CA for looking at shot-scale data is unquestionable. It does what presentation in the form of bar charts does, does it more concisely, and – I would argue – more informatively. As with all techniques of this kind it is always possible to go back to the raw data, either in tabular form or as a bar chart, to check any inferences you may wish to draw. The CA directs attention to where it might be useful to look.

Redfern (2010b, 2010c), and in other posts on his research blog, has developed a method based on ranked shot-scale data. The first paper uses the same body of Hitchcock films used here. The focus is more on examining 'hypotheses' about differences (or their lack) between the periods of silent, British sound and American films, rather than descriptive graphical analysis. That is, whereas knowledge of period informs the interpretation of CA plots but not their construction, the same knowledge is integral to the construction of Redfern's method.

The method is best explained by describing the construction (this is not quite how Redfern presents it).

For each category:

1. Re-arrange the numbers in each row from largest to smallest.
2. Obtain the mean of each column in the new table so obtained.
3. Plot the means against the numbers 1 to 7 (the ranks of the data).
4. Fit a (regression) straight line through the data.

Once this is done compare the fitted lines. The row numbers need to be the same for this to work and, rounding error apart, they should be. Redfern converts to mean relative frequencies (MRFs) by dividing row numbers by row totals. Do all this for the data of Table 9.1 and Figure 9.7 can be constructed.

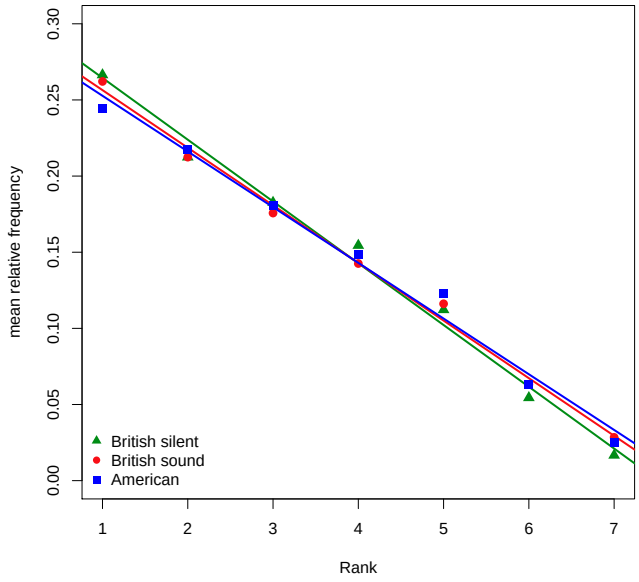

Figure 9.7: *Mean relative frequency plotted against shot scale ranks for three categories of Hitchcock films. See the text for an explanation.*

There is little difference in the fitted lines. Redfern concludes that 'no one shot scale dominates Hitchcock's films ...' and that 'it is clear that Hitchcock's style did not change with either the introduction of sound technology or the move from Britain to America'. You need to read the papers carefully to work out what this means, since 'shot scale' seems to be used to mean two different things. For example, in the sentence following the above conclusion it is stated that 'if we turn our attention to the shot scales themselves we can see that there is a clear change in Hitchcock's style'. On first reading this I thought it was a direct contradiction of the preceding sentence.

The second usage is straightforward. It is just saying something along the lines that the extent to which Hitchcock used some shot-scale *types* varied in his different periods. The first usage is based on the ranking of shot-scale types, and refers to the ranks, the data becoming divorced from types in the process. Thus, and hypothetically, any column of the rearranged table of data could consist of numbers for a single shot type, or of six examples for each of the seven shot types – the method doesn't distinguish between these possibilities.

In plotting the mean of each column of the data after it has been rearranged by row rank order, 'shot scale' simply refers to the rank order. I think the aim in comparison is to see if different bodies of data exhibit linear patterns in the plot of means of the ranked data and, if so, whether the slopes and intercepts of the fitted lines differ much. Departures from linearity are of particular interest; for example, and as I've understood it, if the mean for the data ranked 1 is noticeably greater that that predicted from a linear model based on all the means, that shot scale is said to 'dominate'. Hypothetically, had this occurred for one of Hitchcock's periods, it would be saying that within that period individual films tended to have a higher proportion of one type that the linear model based on that film predicts. No information about the distribution of shot-scale types that dominate the films is used.

The procedure is – in my opinion – rather arcane and the interpretation is not easy to understand. The detachment of the analysis from shot-scale types is a (severe) limitation. As much is acknowledged by Redfern, both explicitly (without the use of the bracketed adjective) and implicitly in the use of analyses that do recognise type differences and produce more readily understood conclusions. It is questionable whether the analysis of the ranked data adds much to what can be done by simpler means.

The analyses that allow for shot-scale type differences are of two kinds. One is a statistical hypothesis testing approach where, for each shot-scale type in turn, the hypothesis tested is that there is no difference between periods in the typical proportion of each type. The broad conclusion is that the American films differ from the British films, with a move towards 'tighter framing', but that the British silent and sound films do not differ. Ultimately shot-scale types are examined at the level of individual films, leading to the suggestion that the later British sound films were evolving towards the American style (see, also, Salt, 2009, p.243).

There are limitations to the hypothesis testing approach since it ignores the possiblity of temporal evolution of style *within* a period. That is, there may be a noticeable increase or decrease in the use of a particular shot-scale type within a period that does not sufficiently affect the 'typical' value (be it mean or median) for it to be statistically significantly different from the typical value in another period with a different pattern. An alternative approach is sketched in the next section.

The code to obtain a plot of the kind illustrated for a single body of data follows (style and labelling arguments omitted). Call the body of interest `body` and assume, as in Salt's database, rows have been scaled to sum to 500.

```
MRF <- apply(apply(body/500, 1, sort), 1, mean)
MRF <- sort(MRF, decreasing = T)
Rank <- 1:7
plot(Rank, MRF, ylim = c(0,.3))
abline(lm(MRF ~ Rank))
```

The first line produces the means that need to be plotted, but in the reverse rank order needed, which the second line corrects.

9.5.2 Analysis of individual shot-scale types

The idea is both simple and obvious. For each shot-scale type plot its occurrence (percentage in the plots to follow) against date. This is done in Figure 9.8 with the added embellishments of loess (3/4, s, 2) smooths complete with two standard error limits, and vertical lines at 1929.5 and 1939,5. The loess fit is designed to be a smooth one, ignoring very local variation, and the robust version that downweights extremes is also used. The vertical lines separate out the three periods except that the sound version of *Blackmail* is a 1929 film. The hypothesis testing approach of Redfern (2010b) effectively assumes 'flat' distributions within each period, apart from random variation, and tests the hypothesis that 'flat' lines fitted within each period do not differ significantly in their 'level'.

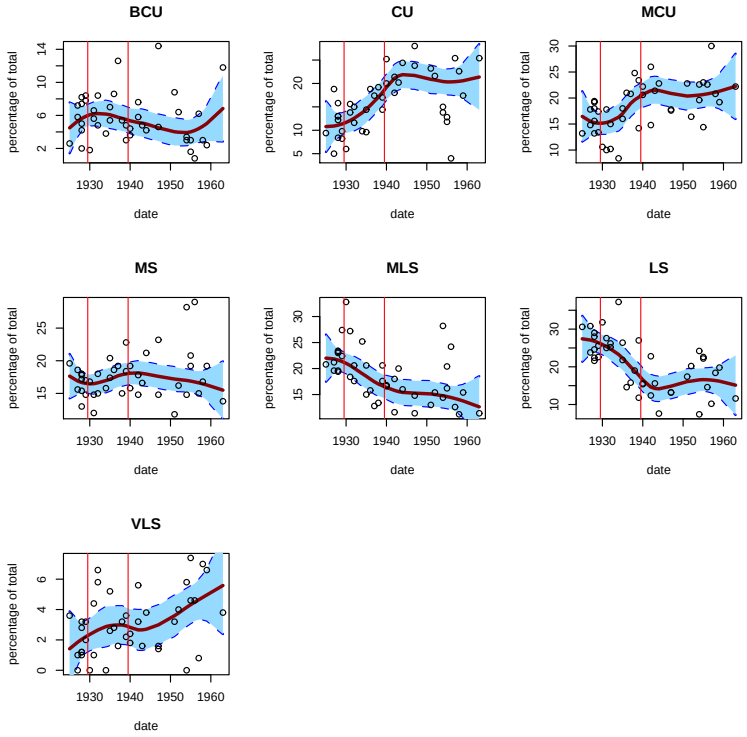

Figure 9.8: *Plots of shot-scale occurrence (%) against date for individual films and each shot-scale type, with added loess (3/4, s, 2) smooths.*

The usual caveats concerning sample sizes, dependence of the loess curve on the degree of smoothing, and so on, can be entered. There is little doubt that Hitchcock's style, as reflected in the shot scale distributions, does change over the time period involved, so interpretation can concentrate more on an 'holistic' examination of 'how'.

Given the shortness of the period involved, not too much can be said about developments within the silent period. There is the clear suggestion of evolution within the British sound period, with the use of CUs and MCUs increasing and that of MLSs and LSs decreasing. The change for MCUs possibly starts a little later than for the others. For the American period MLSs continue to decline while the use of VLSs shows an increase, though this is one of the least commonly used types. Otherwise, for types where some evolution in use in the British sound period was suggested there tends to be a levelling out.

It needs to be re-emphasised that this kind of interpretation does not really pay attention to the fact that there is a lot of variation in the data, evident from the CA graphs, particularly as the robust loess smooth downplays this. Nevetheless, the temporal variation exhibited in the

British sound period suggests that conclusions derived from hypothesis testing, that there are no significant differences between the use of shot-scales in the two British periods, could be viewed as un-nuanced. The American period is hardly static – it shows both some change in the extent to which some shot-scale types are used and is hignly variable – but in very broad terms it can be argued that the period was characterised by a fairly stable pattern of usage.

In the following code `y` is the shot-scale data for a single type and `x` is the date. This would produce a single one of the plots shown in Figure 9.8. I've left in the code that produces the shading between the 2-standard error limits, but not explained it (the shading is not really necessary, but it looks nice).

```
y <- y/5
fit <- loess(y ~ x, family = "s")
pred <- predict(fit, se = T)

# plot 2-standard error limits

plot(x, y, xlab = "date", type = "n", ylab = "percentage of total")
lines(x, pred$fit - 2*pred$se.fit, lty = 2, col = "blue", lwd=2)
lines(x, pred$fit + 2*pred$se.fit, lty = 2, col = "blue", lwd=2)

# create polygon bounds and add to plot

y.polygon <- c((pred$fit + 2*pred$se.fit), (pred$fit - 2*pred$se.fit)[42:1])
x.polygon <- c(x, sort(x, decreasing = T))
polygon(x.polygon, y.polygon, col="lightblue", border = NA)

# add everything else in

lines(x, pred$fit, col = "darkred", lwd = 3)
points(x, y)
abline(v = 1929.5, col = "red")
abline(v = 1939.5, col = "red")
```

In previous uses of `loess` a different way of getting the fitted line was used. Here it's done using the `predict` function which allows standard errors to be saved. These are used to plot 2-standard error limits with the first two `lines` commands. The blank plot has these limits and the fill between them added before the fitted line and points, to avoid over-plotting by the fill.