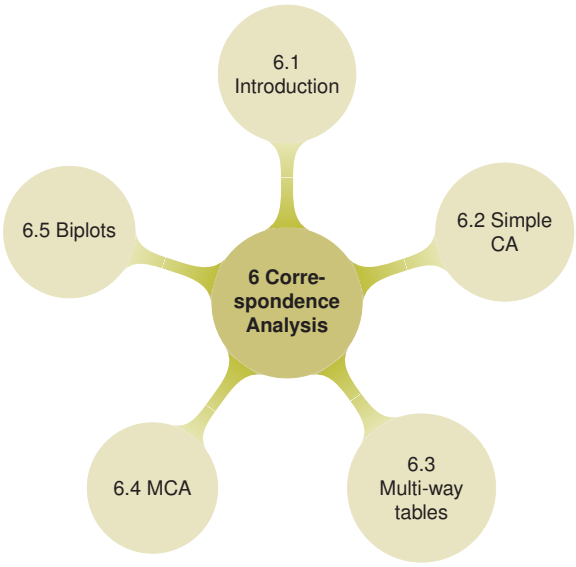


6

Correspondence Analysis

{ch:corresp}

Correspondence analysis provides visualizations of associations in a two-way contingency table in a small number of dimensions. Multiple correspondence analysis extends this technique to *n*-way tables. Other graphical methods, including mosaic matrices and biplots provide complementary views of loglinear models for two-way and *n*-way contingency tables.

6.1 Introduction

> Whenever a large sample of chaotic elements is taken in hand and marshalled in the order of their magnitude, an unsuspected and most beautiful form of regularity proves to have been latent all along.
>
> Sir Francis Galton (1822–1911)

Correspondence analysis (CA) is an exploratory technique which displays the row and column categories in a two-way contingency table as points in a graph, so that the positions of the points represent the associations in the table. Mathematically, correspondence analysis is related to the ***biplot***, to ***canonical correlation***, and to ***principal component analysis***.

This technique finds scores for the row and column categories on a small number of dimensions which account for the greatest proportion of the χ^2 for association between the row and column categories, just as principal components account for maximum variance of quantitative variables. But CA does more— the scores provide a quantification of the categories, and have the property that they maximize the correlation between the row and column variables. For graphical display two or three dimensions are typically used to give a reduced rank approximation to the data.

Correspondence analysis has a very large, multi-national literature and was rediscovered several times in different fields and different countries. The method, in slightly different forms, is also discussed under the names ***dual scaling***, ***optimal scaling***, ***reciprocal averaging***, ***homogeneity analysis***, and ***canonical analysis of categorical data***.

See Greenacre (1984) and Greenacre (2007) for an accessible introduction to CA methodology, or Gifi (1981), Lebart *et al.* (1984) for a detailed treatment of the method and its applications from the Dutch and French perspectives. Greenacre and Hastie (1987) provide an excellent discussion of the geometric interpretation, while van der Heijden and de Leeuw (1985) and van der Heijden *et al.* (1989) develop some of the relations between correspondence analysis and log-linear methods for three-way and larger tables. Correspondence analysis is usually carried out in an exploratory, graphical way. Goodman (1981, 1985, 1986) has developed related inferential models, the RC model and the canonical correlation model, with close links to CA.

One simple development of CA is as follows: For a two-way table the scores for the row categories, namely $\boldsymbol{X} = \{x_{im}\}$, and column categories, $\boldsymbol{Y} = \{y_{jm}\}$, on dimension $m = 1, \ldots, M$ are derived from a (generalized) ***singular value decomposition*** of (Pearson) residuals from independence, expressed as $d_{ij}/\sqrt{n}$, to account for the largest proportion of the χ^2 in a small number of dimensions. This decomposition may be expressed as

{eq:cadij}

$$\frac{d_{ij}}{\sqrt{n}} = \frac{n_{ij} - m_{ij}}{\sqrt{n\, m_{ij}}} = \boldsymbol{X}\, \boldsymbol{D}_\lambda\, \boldsymbol{Y}^{\mathsf{T}} = \sum_{m=1}^{M} \lambda_m\, x_{im}\, y_{jm} \ , \tag{6.1}$$

where m_{ij} is the expected frequency and where $\boldsymbol{D}_\lambda$ is a diagonal matrix with elements $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_M$, and $M = \min(I - 1,\ J - 1)$. In M dimensions, the decomposition Eqn. (6.1) is exact. For example, an $I \times 3$ table can be depicted exactly in two dimensions when $I \geq 3$. The useful result for visualization purposes is that a rank-d approximation in d dimensions is obtained from the first d terms on the right side of Eqn. (6.1). The proportion of the Pearson χ^2 accounted for by this approximation is

$$n \sum_{m}^{d} \lambda_m^2 / \chi^2 \ .$$

The quantity $\chi^2/n = \sum_i \sum_j d_{ij}^2/n$ is called the total ***inertia*** and is identical to the measure of association known as Pearson's mean-square contingency, the square of the ϕ coefficient.

Thus, correspondence analysis is designed to show how the data deviate from expectation when the row and column variables are independent, as in the sieve diagram, association plot and mosaic display. However, the sieve, association and mosaic plots depict every *cell* in the table, and for large tables it may be difficult to see patterns. Correspondence analysis shows only row and column *categories* as points in the two (or three) dimensions which account for the greatest proportion of deviation from independence. The pattern of the associations can then be inferred from the positions of the row and column points.

6.2 Simple correspondence analysis

{sec:ca-simple}

6.2.1 Notation and terminology

{sec:ca-notation}

Because Correspondence analysis grew up in so many homes, the notation, formulae and terms used to describe the method vary considerably. The notation used here generally follows Greenacre (1984, 1997, 2007).

The descriptions here employ the following matrix and vector definitions:

- $\boldsymbol{N} = \{n_{ij}\}$ is the $I \times J$ contingency table with row and column totals n_{i+} and n_{+j}, respectively. The grand total n_{++} is also denoted by n for simplicity.

- $\boldsymbol{P} = \{p_{ij}\} = \boldsymbol{N}/n$ is the matrix of joint cell proportions, called the ***correspondence matrix***.
- $\boldsymbol{r} = \sum_j p_{ij} = \boldsymbol{P}\mathbf{1}$ is the row margin of $\boldsymbol{P}$; $\boldsymbol{c} = \sum_i p_{ij} = \boldsymbol{P}^\mathsf{T}\mathbf{1}$ is the column margin. $\boldsymbol{r}$ and $\boldsymbol{c}$ are called the *row masses* and *column masses*.
- $\boldsymbol{D}_r$ and $\boldsymbol{D}_c$ are diagonal matrices with $\boldsymbol{r}$ and $\boldsymbol{c}$ on their diagonals, used as weights.
- $\boldsymbol{R} = \boldsymbol{D}_r^{-1}\boldsymbol{P} = \{n_{ij}/n_{+j}\}$ is the matrix of row conditional probabilities, called *row profiles*. Similarly, $\boldsymbol{C} = \boldsymbol{D}_c^{-1}\boldsymbol{P}^\mathsf{T} = \{n_{ij}/n_{i+}\}$ is the matrix of column conditional probabilities or *column profiles*.
- $\boldsymbol{S} = \boldsymbol{D}_r^{-1/2}(\boldsymbol{P} - \boldsymbol{r}\boldsymbol{c}^\mathsf{T})\boldsymbol{D}_c^{-1/2}$ is the matrix of standardized Pearson residuals from independence (denoted d_{ij} in the introduction).

Two types of coordinates, $\boldsymbol{X}$, $\boldsymbol{Y}$ for the row and column categories are defined, based on the singular value decomposition (SVD) of $\boldsymbol{S}$,

$$\boldsymbol{S} = \boldsymbol{U}\boldsymbol{D}_\lambda\boldsymbol{V}^\mathsf{T} \quad \text{where} \quad \boldsymbol{U}^\mathsf{T}\boldsymbol{U} = \boldsymbol{V}^\mathsf{T}\boldsymbol{V} = \boldsymbol{I} \ ,$$

and $\boldsymbol{D}_\lambda$ is the diagonal matrix of singular values $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_M$. $\boldsymbol{U}$ is the orthonormal $I \times M$ matrix of left singular vectors, and $\boldsymbol{V}$ is the $J \times M$ matrix of right singular vectors.

The SVD of $\boldsymbol{S}$ is related to the eigenvalue–eigenvector decomposition of a square symmetric matrix, in that $\boldsymbol{S}\boldsymbol{S}^\mathsf{T} = \boldsymbol{U}\boldsymbol{D}_\lambda^2\boldsymbol{U}$ and $\boldsymbol{S}^\mathsf{T}\boldsymbol{S} = \boldsymbol{V}\boldsymbol{D}_\lambda^2\boldsymbol{V}$, so the values λ^2 are the eigenvalues in both cases and the singular vectors are the corresponding eigenvectors. In correspondence analysis, these eigenvalues (squares of the singular values) are called the ***principal inertia***s, and are the values used in the decomposition of the Pearson χ^2 for the dimensions, $\chi^2 = n\sum_m \lambda_m^2$.

principal coordinates: The coordinates of the row ($\boldsymbol{F}$) and column ($\boldsymbol{G}$) profiles with respect to their own principal axes are defined so that the inertia along each axis is the corresponding eigenvalue value, λ_m,

$$\boldsymbol{F} = \boldsymbol{D}_r^{-1/2}\boldsymbol{U}\boldsymbol{D}_\lambda^2 \quad \text{scaled so that} \quad \boldsymbol{F}^\mathsf{T}\boldsymbol{D}_r\boldsymbol{F} = \boldsymbol{D}_\lambda^2 \tag{6.2}$$ {eq:pcoord1}

$$\boldsymbol{G} = \boldsymbol{D}_c^{-1/2}\boldsymbol{V}\boldsymbol{D}_\lambda^2 \quad \text{scaled so that} \quad \boldsymbol{G}^\mathsf{T}\boldsymbol{D}_c\boldsymbol{G} = \boldsymbol{D}_\lambda^2 \tag{6.3}$$ {eq:pcoord2}

The joint plot in principal coordinates, $\boldsymbol{F}$ and $\boldsymbol{G}$, is called the ***symmetric map*** because both row and column profiles are overlaid in the same coordinate system.

standard coordinates: The standard coordinates ($\boldsymbol{\Phi}, \boldsymbol{\Gamma}$) are a rescaling of the principal coordinates to unit inertia along each axis,

$$\boldsymbol{\Phi} = \boldsymbol{D}_r^{-1}\boldsymbol{U} \quad \text{scaled so that} \quad \boldsymbol{\Phi}^\mathsf{T}\boldsymbol{D}_r\boldsymbol{\Phi} = \boldsymbol{I} \tag{6.4}$$ {eq:scoord1}

$$\boldsymbol{\Gamma} = \boldsymbol{D}_c^{-1}\boldsymbol{V} \quad \text{scaled so that} \quad \boldsymbol{\Gamma}^\mathsf{T}\boldsymbol{D}_c\boldsymbol{\Gamma} = \boldsymbol{I} \tag{6.5}$$ {eq:scoord2}

These differ from the principal coordinates in Eqn. (6.2) and Eqn. (6.3) simply by the absence of the scaling factors, $\boldsymbol{D}_\lambda^2$. An ***asymmetric map*** shows one set of points (say, the rows) in principal coordinates and the other set in standard coordinates.

Thus, the weighted average of the squared principal coordinates for the rows or columns on a principal axis equals the squared singular value, λ^2 for that axis, whereas the weighted average of the squared standard coordinates equals 1. The relative positions of the row or column points along any axis is the same under either scaling, but the distances between points differ, because the axes are weighted differentially in the two scalings.

6.2.2 Geometric and statistical properties

{sec:ca-properties}

We summarize here some geometric and statistical properties of the Correspondence analysis solutions which are useful in interpretation.

nested solutions: Because they use successive terms of the SVD Eqn. (6.1), correspondence analysis solutions are *nested*, meaning that the first two dimensions of a three-dimensional solution will be identical to the two-dimensional solution.

centroids at the origin: In both principal coordinates and standard coordinates the points representing the row and column profiles have their centroids (weighted averages) at the origin. Thus, in CA plots, the origin represents the (weighted) average row profile and column profile.

reciprocal averages: CA assigns scores to the row and column categories such that the column scores are proportional to the weighted averages of the row scores, and vice-versa.

chi-square distances: In principal coordinates, the row coordinates may be shown equal to the row profiles $\boldsymbol{D}_r^{-1}\boldsymbol{P}$, rescaled by the inverse by the square-root of the column masses, $\boldsymbol{D}_c^{-1/2}$. Distances between two row profiles, $\boldsymbol{R}_i$ and $\boldsymbol{R}_{i'}$ are most sensibly defined as χ^2 distances, where the squared difference $[\boldsymbol{R}_{ij} - \boldsymbol{R}_{i'j}]^2$ is inversely weighted by the column frequency, to account for the different relative frequency of the column categories. The rescaling by $\boldsymbol{D}_c^{-1/2}$ transforms this weighted χ^2 metric into ordinary Euclidean distance. The same is true of the column principal coordinates.

interpretation of distances: In principal coordinates, the distance between two row points may be interpreted as described above, and so may the distance between two column points. The distance between a row and column point, however, does not have a clear distance interpretation.

residuals from independence: The distance between a row and column point do have a rough interpretation in terms of residuals or the difference between observed and expected frequencies, $n_{ij} - m_{ij}$. Two row (or column) points deviate from the origin (the average profile) when their profile frequencies have similar values. A row point appears in a similar direction away from the origin as a column point when $n_{ij} - m_{ij} > 0$, and in an opposite different direction from that column point when the residual is negative.

Because of these differences in interpretations of distances, there are different possibilities for graphical display. A joint display of principal coordinates for the rows and standard coordinates for the columns (or vice-versa), sometimes called an ***asymmetric map*** is suggested by Greenacre and Hastie (1987) and by Greenacre (1989) as the plot with the most coherent geometric interpretation (for the points in principal coordinates) and is sometimes used in the French literature.

Another common joint display is the ***symmetric map*** of the principal coordinates in the same plot. This is the default in the ca (Greenacre and Nenadic, 2014) package described below. In the authors' opinion, this produces better graphical displays, because both sets of coordinates are scaled with the same weights for each axis. Symmetric plots are used exclusively in this book, but that should not imply that these plots are universally preferred. Another popular choice is to avoid the possibility of misinterpretation by making separate plots of the row and column coordinates.

6.2.3 R software for correspondence analysis

{sec:ca-R}

Correspondence analysis methods for computation and plotting are available in a number of R packages including:

MASS (Ripley, 2015): `corresp()`; the plot method calls `biplot()` for a 2 factor solution, using a a symmetric biplot factorization that scales the row and column points by the square roots of the the singular values. There is also a `mca()` function for multiple correspondence analysis.

ca : `ca()`; provides 2D plots via the `plot.ca()` method and interactive (rgl (Adler and Murdoch, 2014)) 3D plots via `plot3d.ca()`. This package is the most comprehensive in terms of plotting options for various coordinate types, plotting supplementary points (see Section 6.3.2) and other features. It also provides `mjca()` for multiple and joint correspondence analysis of higher-way tables.

FactoMineR (Husson *et al.*, 2015): `CA()`; provides a wide variety of measures for the quality of the CA representation and many options for graphical display

These methods also differ in terms of the types of input they accept. For example, `MASS::corresp` handles matrices, data frames and "xtabs" objects, but not "table" objects. `ca()` is the most general, with methods for two-way tables, matrices, data frames, and "xtabs" objects. In the following, we largely use the ca package.

{ex:haireye3}

EXAMPLE 6.1: Hair color and eye color

The script below uses the two-way table `haireye` from the *`HairEyeColor`* data, collapsed over `Sex`. In this table, `Hair` colors form the rows, and `Eye` colors form the columns. By default, `ca()` produces a 2-dimensional solution. In this example, the complete, exact solution would have $M = \min((I-1),(J-1)) = 3$ dimensions, and you could obtain this using the argument `nd=3` in the call to `ca()`.

```
> haireye <- margin.table(HairEyeColor, 1:2)
> library(ca)
> (haireye.ca <- ca(haireye))


 Principal inertias (eigenvalues):
            1        2        3
Value       0.208773 0.022227 0.002598
Percentage 89.37%   9.52%    1.11%


 Rows:
            Black    Brown     Red  Blond
Mass      0.18243  0.48311  0.1199 0.2145
ChiDist   0.55119  0.15946  0.3548 0.8384
Inertia   0.05543  0.01228  0.0151 0.1508
Dim. 1   -1.10428 -0.32446 -0.2835 1.8282
Dim. 2    1.44092 -0.21911 -2.1440 0.4667


 Columns:
            Brown   Blue    Hazel    Green
Mass      0.37162 0.3632  0.15710  0.10811
ChiDist   0.50049 0.5537  0.28865  0.38573
Inertia   0.09309 0.1113  0.01309  0.01608
Dim. 1   -1.07713 1.1981 -0.46529  0.35401
Dim. 2    0.59242 0.5564 -1.12278 -2.27412
```

In the printed output, the table labeled "Principal inertias (eigenvalues)" indicates that nearly 99% of the Pearson χ^2 for association is accounted for by two dimensions, with most of that attributed to the first dimension.

The `summary` method for "ca" objects gives a more nicely formatted display, showing a ***scree plot*** of the eigenvalues, a portion of which is shown below.

```
> summary(haireye.ca)


Principal inertias (eigenvalues):

 dim    value      %   cum%   scree plot
 1      0.208773  89.4  89.4  **********************
 2      0.022227   9.5  98.9  **
 3      0.002598   1.1 100.0
        -------- -----
 Total: 0.233598 100.0
...
```

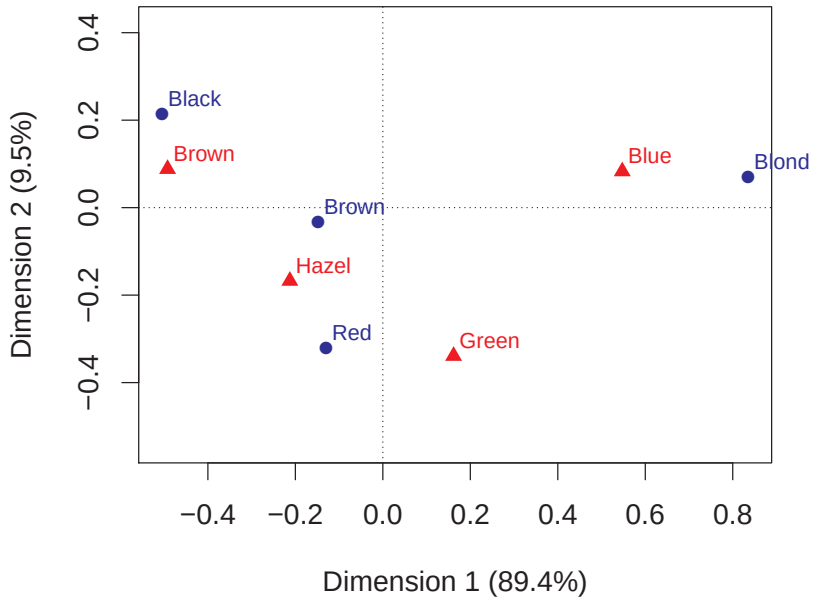
The Pearson χ^2 for this table (given by `chisq.test(haireye)`) is 138.29. This value is n (592) times the sum of the eigenvalues (0.2336) shown above.

The result returned by `ca()` can be plotted using the `plot.ca()` method. However, it is useful to understand that `ca()` returns the CA solution in terms of *standard coordinates*, Φ (`rowcoord`) and Γ (`colcoord`). We illustrate Eqn. (6.4) and Eqn. (6.5) using the components of the "ca" object `haireye.ca`.

```
> # standard coordinates Phi (Eqn 6.4) and Gamma (Eqn 6.5)
> (Phi <- haireye.ca$rowcoord)

          Dim1     Dim2     Dim3
Black -1.10428  1.44092 -1.08895
Brown -0.32446 -0.21911  0.95742
Red   -0.28347 -2.14401 -1.63122
Blond  1.82823  0.46671 -0.31809

> (Gamma <- haireye.ca$colcoord)

          Dim1     Dim2      Dim3
Brown -1.07713  0.59242 -0.423960
Blue   1.19806  0.55642  0.092387
Hazel -0.46529 -1.12278  1.971918
Green  0.35401 -2.27412 -1.718443

> # demonstrate orthogonality of std coordinates
> Dr <- diag(haireye.ca$rowmass)
> zapsmall(t(Phi) %*% Dr %*% Phi)

     Dim1 Dim2 Dim3
Dim1    1    0    0
Dim2    0    1    0
Dim3    0    0    1

> Dc <- diag(haireye.ca$colmass)
> zapsmall(t(Gamma) %*% Dc %*% Gamma)

     Dim1 Dim2 Dim3
Dim1    1    0    0
Dim2    0    1    0
Dim3    0    0    1
```

These standard coordinates are transformed internally within the plot function according to the `map` argument, which defaults to `map="symmetric"`, giving principal coordinates. The following call to `plot.ca()` produces Figure 6.1.

```
> op <- par(cex=1.4, mar=c(5,4,1,2)+.1)
> res <- plot(haireye.ca)
> par(op)
```



ig:ca-haireye-plot}

Figure 6.1: Correspondence analysis solution for the Hair color and Eye color data

For use in further customizing such plots (as we will see in the next example), the function `plot.ca()` returns (invisibly)[1] the coordinates for the row and column points actually plotted, which we saved above as `res`:

```
> res

$rows
           Dim1      Dim2
Black -0.50456  0.214820
Brown -0.14825 -0.032666
Red   -0.12952 -0.319642
Blond  0.83535  0.069579

$cols
           Dim1      Dim2
Brown -0.49216  0.088322
Blue   0.54741  0.082954
Hazel -0.21260 -0.167391
Green  0.16175 -0.339040
```

It is important to understand that in CA plots (and related biplots, Section 6.5), the interpretation of distances between points (and angles between vectors) is meaningful. In order to achieve this, the axes in such plots must be *equated*, meaning that the two axes are scaled so that the number of data units per inch are the same for both the horizontal and vertical axes, or an ***aspect ratio*** = 1.[2]

The interpretation of the CA plot in Figure 6.1 is then as follows:

- Dimension 1, accounting for nearly 90% of the association between hair and eye color corresponds to dark (left) vs. light (right) on both variables.

[1] This uses features incorporated in the ca package, version 0.54+.

[2] In base R graphics, this is achieved with the `plot()` option `asp=1`.

- Dimension 2 largely contrasts red hair and green eyes with the remaining categories, accounting for an additional 9.5% of the Pearson χ^2.
- With equated axes, and a symmetric map, the distances between row points and distances between column points are meaningful. Along Dimension 1, the eye colors could be considered roughly equally spaced, but for the hair colors, Blond is quite different in terms of its frequency profile.

△

{ex:mental3}

EXAMPLE 6.2: Mental impairment and parents' SES

In Example 4.3 we introduced the data set *Mental*, relating mental health status to parents' SES. As in Example 4.7, we convert this to a two-way table, `mental.tab` to conduct a correspondence analysis.

```
> data("Mental", package="vcdExtra")
> mental.tab <- xtabs(Freq ~ ses + mental, data=Mental)
```

We calculate the CA solution, and save the result in `mental.ca`:

```
> mental.ca <- ca(mental.tab)
> summary(mental.ca)


Principal inertias (eigenvalues):

 dim    value      %   cum%   scree plot
 1      0.026025  93.9  93.9  ***********************
 2      0.001379   5.0  98.9  *
 3      0.000298   1.1 100.0
        -------- -----
 Total: 0.027702 100.0
...
```

The scree plot produced by `summary(mental.ca)` shows that the association between mental health and parents' SES is almost entirely 1-dimensional, with 94% of the χ^2 (45.98, with 15 df) accounted for by Dimension 1.

We then plot the solution as shown below, giving Figure 6.2. For this example, it is useful to connect the row points and the column points by lines, to emphasize the pattern of these ordered variables.

```
> op <- par(cex=1.3, mar=c(5,4,1,1)+.1)
> res <- plot(mental.ca,  ylim=c(-.2, .2))
> lines(res$rows, col="blue", lty=3)
> lines(res$cols, col="red", lty=4)
> par(op)
```

The plot of the CA scores in Figure 6.2 shows that diagnostic mental health categories are well-aligned with Dimension 1. The mental health scores are approximately equally spaced, except that the two intermediate categories are a bit closer on this dimension than the extremes. The SES categories are also aligned with Dimension 1, and approximately equally spaced, with the exception of the highest two SES categories, whose profiles are extremely similar, suggesting that these two categories could be collapsed.

Because both row and column categories have the same pattern on Dimension 1, we may interpret the plot as showing that the profiles of both variables are ordered, and their relation can be explained as a positive association between high parents' SES and higher mental health status of

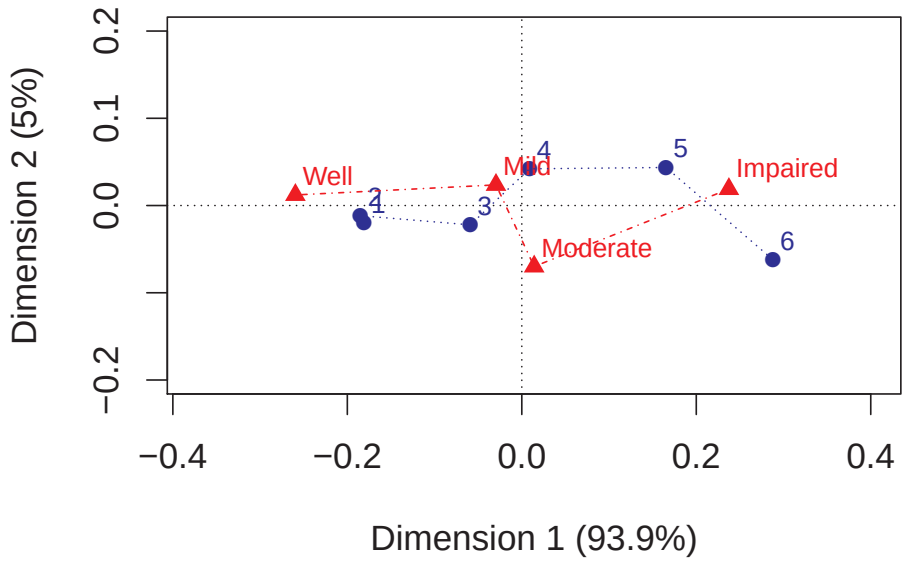


fig:ca-mental-plot}

Figure 6.2: Correspondence analysis solution for the Mental health data

children. A mosaic display of these data (Exercise 6.5) would show a characteristic opposite corner pattern of association.

From a modeling perspective, we might ask how strong is the evidence for the spacing of categories noted above. For example, we might ask whether assigning integer scores to the levels of SES and mental impairment provides a simpler, but satisfactory account of their association. Questions of this type can be explored in connection with loglinear models in Chapter 9.

△

{ex:victims2}

EXAMPLE 6.3: Repeat victimization

The data set `RepVict` in the vcd (Meyer *et al.*, 2015) package gives a 8×8 table (from Fienberg (1980, Table 2-8)) on repeat victimization for various crimes among respondents to a U.S. National Crime Survey. A special feature of this data set is that row and column categories reflect the *same* crimes, so substantial association is expected. Here we examine correspondence analysis results in a bit more detail and also illustrate how to customize the displays created by `plot(ca(...))`.

```
> data("RepVict", package="vcd")
> victim.ca <- ca(RepVict)
> summary(victim.ca)


Principal inertias (eigenvalues):

 dim    value      %    cum%   scree plot
 1      0.065456  33.8  33.8   ********
 2      0.059270  30.6  64.5   ********
 3      0.029592  15.3  79.8   ****
 4      0.016564   8.6  88.3   **
 5      0.011140   5.8  94.1   *
 6      0.007587   3.9  98.0   *
 7      0.003866   2.0 100.0
```

```
        -------- -----
 Total: 0.193474 100.0
...
```

The results above show that, for this 8×8 table, 7 dimensions are required for an exact solution, of which the first two account for 64.5% of the Pearson χ^2. The lines below illustrate that the Pearson χ^2 is n times the sum of the squared singular values, $n \sum \lambda_i^2$.

```
> chisq.test(RepVict)


Pearson's Chi-squared test

data:  RepVict
X-squared = 11131, df = 49, p-value < 2.2e-16

> (chisq <- sum(RepVict) * sum(victim.ca$sv^2))

[1] 11131
```

The default plot produced by `plot.ca(victim.ca)` plots both points and labels for the row and column categories. However, what we want to emphasize here is the relation between the *same* crimes on the first and second occurrence.

To do this, we label each crime just once (using `labels=c(2,0)`) and connect the two points for each crime by a line, using `segments()`, as shown in Figure 6.3. The addition of a `legend()` makes the plot more easily readable.

```
> op <- par(cex=1.3, mar=c(4,4,1,1)+.1)
> res <- plot(victim.ca, labels=c(2,0))
> segments(res$rows[,1], res$rows[,2], res$cols[,1], res$cols[,2])
> legend("topleft", legend=c("First", "Second"), title="Occurrence",
+        col=c("blue", "red"), pch=16:17, bg="gray90")
> par(op)
```

In Figure 6.3 it may be seen that most of the points are extremely close for the first and second occurrence of a crime, indicating that the row profile for a crime is very similar to its corresponding column profile, with Rape and Pick Pocket as exceptions.

In fact, if the table was symmetric, the row and column points in Figure 6.3 would be identcal, as can be easily demonstrated by analyzing a symmetric version.

```
> RVsym <- (RepVict + t(RepVict))/2
> RVsym.ca <- ca(RVsym)
> res <- plot(RVsym.ca)
> all.equal(res$rows, res$cols)

[1] TRUE
```

The first dimension appears to contrast crimes against the person (right) with crimes against property (left), and it may be that the second dimension represents degree of violence associated with each crime. The latter interpretation is consistent with the movement of Rape towards a higher position and Pickpocket towards a lower one on this dimension.

△

6.2.4 Corespondence analysis and mosaic displays

For a two-way table, CA and mosaic displays give complementary views of the pattern of association between the row and column variables, but both are based on the (Pearson) residuals from

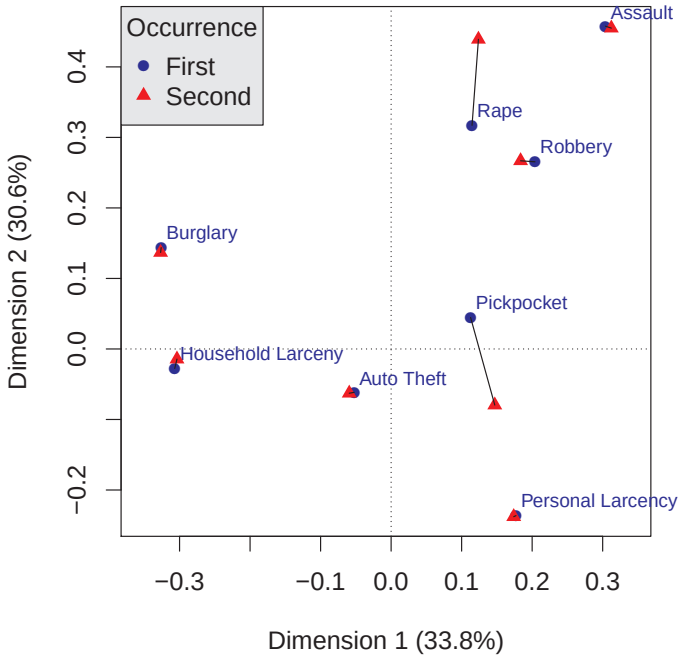


Figure 6.3: 2D CA solution for the repeat victimization data. Lines connect the category points for first and second occurrence to highlight these relations.

{fig:ca-victims-plot}

independence. CA shows the row and column categories as points in a 2D (or 3D) space accounting for the largest proportion of the Pearson χ^2, while mosaics show the association by the pattern of shading in the mosaic tiles. It is useful to compare them directly to see how associations can be interpreted from these graphs.

{ex:TV2}

EXAMPLE 6.4: TV viewing data

The data on television viewership from Hartigan and Kleiner (1984) was used as an example of manipulating complex categorical data in Section 2.9 and shown as a three-way mosaic plot in Figure **??**. From that figure, the main association concerned how viewership across days of the week varied by TV network, so we first collapse the *TV* data to a 5×3 two-way table.

```
> data("TV", package="vcdExtra")
> TV2 <- margin.table(TV, c(1,3))
> TV2

           Network
Day          ABC  CBS  NBC
  Monday    2847 2923 2629
  Tuesday   3110 2403 2568
  Wednesday 2434 1283 2212
  Thursday  1766 1335 5886
  Friday    2737 1479 1998
```

In this case, the 2D CA solution is exact, meaning that two dimensions account for 100% of the association.

```
> TV.ca <- ca(TV2)
> TV.ca
```

```
 Principal inertias (eigenvalues):
            1        2
Value       0.081934 0.010513
Percentage 88.63%   11.37%
...
```

The plot of this solution is shown in the left panel of Figure 6.4, using lines from the origin to the category points for the networks.

```
> res <- plot(TV.ca)
> segments(0, 0, res$cols[,1], res$cols[,2], col="red", lwd=2)
```

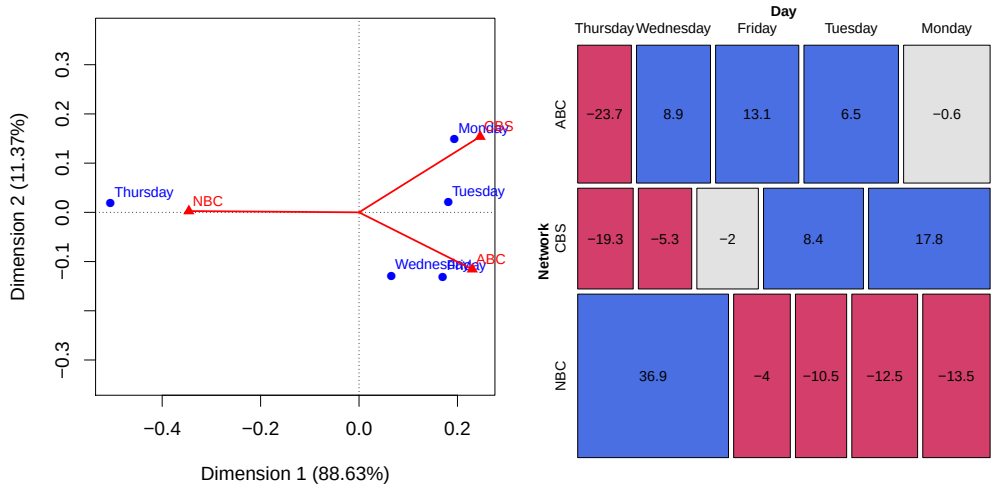


Figure 6.4: CA plot and mosaic display for the TV viewing data. The days of the week in the mosaic plot were permuted according to their order in the CA solution.

{fig:TV-mosaic-ca}

An analogous mosaic display, informed by the CA solution, is shown in the right panel of Figure 6.4. Here, the days of the week are reordered according to their positions on the first CA dimension, another example of effect ordering.

```
> days.order <- order(TV.ca$rowcoord[,1])
> mosaic(t(TV2[days.order,]), shade=TRUE, legend=FALSE,
+        labeling=labeling_residuals, suppress=0)
```

In the CA plot, you can see that the dominant dimension separates viewing on Thursday, with the largest share of viewers watching NBC, from the other weekdays. In the mosaic plot, Thursday stands out as the only day with a higher than expected frequency for NBC, and this is the largest residual in the entire table. The second dimension in the CA plot separates CBS, with its' greatest proportion of viewers on Monday, from ABC, with greater viewership on Wednesday and Friday.

Emerson (1998, Fig. 2) gives a table listing the shows in each half-hour time slot. Could the overall popularity of NBC on Thursday be due to *Friends* or *Seinfeld*? An answer to this and similar questions requires analysis of the three-way table (Exercise 6.8) and model-based methods for polytomous outcome variables described in Section 8.3.

△

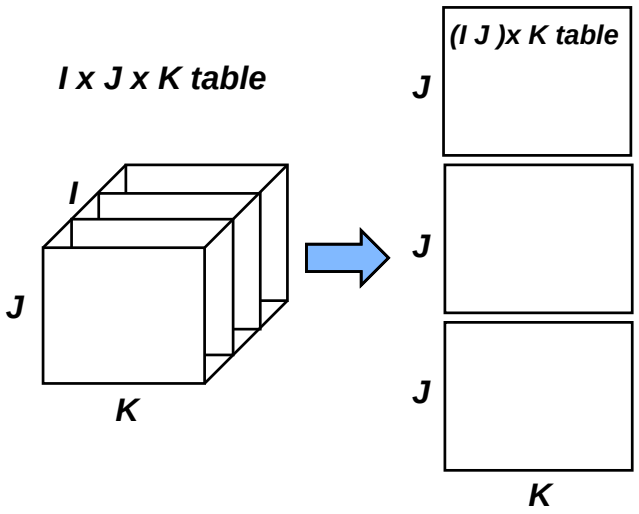


Figure 6.5: Stacking approach for a three-way table. Two of the table variables are combined interactively to form the rows of a two-way table. {fig:stacking}

6.3 Multi-way tables: Stacking and other tricks

{sec:ca-multiway}

A three- or higher-way table can be analyzed by correspondence analysis in several ways. Multiple correspondence analysis (MCA), described in Section 6.4, is an extension of simple correspondence analysis which analyzes simultaneously all possible two-way tables contained within a multiway table. Another approach, described here, is called ***stacking*** or ***interactive coding***. This is a bit of a trick, to force a multiway table into a two-way table for a standard correspondence analysis, but is a useful one.

A three-way table, of size $I \times J \times K$ can be sliced into I two-way tables, each $J \times K$. If the slices are concatenated vertically, the result is one two-way table, of size $(I \times J) \times K$, as illustrated in Figure 6.5. In effect, the first two variables are treated as a single composite variable with IJ levels, which represents the main effects and interaction between the original variables that were combined. Van der Heijden and de Leeuw (1985) discuss this use of correspondence analysis for multi-way tables and show how *each* way of slicing and stacking a contingency table corresponds to the analysis of a specified loglinear model. Like the mosaic display, this provides another way to visualize the relations in a loglinear model.

In particular, for the three-way table with variables A, B, C that is reshaped as a table of size $(I \times J) \times K$, the correspondence analysis solution analyzes residuals from the log-linear model $[AB][C]$. That is, for such a table, the $I \times J$ rows represent the joint combinations of variables A and B. The expected frequencies under independence for this table are

$$m_{[ij]k} = \frac{n_{ij+}\, n_{++k}}{n} \tag{6.6}$$ {eq:mij-k}

which are the ML estimates of expected frequencies for the log-linear model $[AB][C]$. The χ^2 that is decomposed by correspondence analysis is the Pearson χ^2 for this log-linear model. When the table is stacked as $I \times (J \times K)$ or $J \times (I \times K)$, correspondence analysis decomposes the residuals from the log-linear models $[A][BC]$ and $[B][AC]$, respectively, as shown in Table 6.1. In this approach, only the associations in separate $[\,]$ terms are analysed and displayed in the correspondence analysis maps. Van der Heijden and de Leeuw (1985) show how a generalized form of correspondence

analysis can be interpreted as decomposing the difference between two specific loglinear models, so their approach is more general than is illustrated here.

Table 6.1: Each way of stacking a three-way table corresponds to a loglinear model {tab:stacking}

Stacking structure	Loglinear model
$(I \times J) \times K$	$[AB][C]$
$I \times (J \times K)$	$[A][BC]$
$J \times (I \times K)$	$[B][AC]$

6.3.1 Interactive coding in R

ec:ca-interactiveR}

In the general case of an *n*-way table, the stacking approach is similar to that used by `ftable()` and `structable()` in vcd as described in Section 2.5 to flatten multiway tables to a two-way, printable form, where some variables are assigned to the rows and the others to the columns. Both `ftable()` and `structable()` have `as.matrix()` methods[3] that convert their result into a matrix suitable as input to `ca()`.

With data in the form of a frequency data frame, you can easily create interactive coding using `interaction()` or simply use `paste()` to join the levels of stacked variables together.

To illustrate, create a 4-way table of random Poisson counts (with constant mean, $\lambda = 15$) of types of Pet, classified by Age, Color and Sex.

```
> set.seed(1234)
> dim <- c(3, 2, 2, 2)
> tab <- array(rpois(prod(dim), 15), dim=dim)
> dimnames(tab) <- list(Pet=c("dog","cat","bird"),
+                       Age=c("young","old"),
+                       Color=c("black", "white"),
+                       Sex=c("male", "female"))
```

You can use `ftable()` to print this, with a formula that assigns `Pet` and `Age` to the columns and `Color` and `Sex` to the rows.

```
> ftable(Pet + Age ~ Color + Sex, tab)

             Pet   dog       cat       bird
             Age young old young old young old
Color Sex
black male          10  12    16  16    16  12
      female         8  12    13  15    11  13
white male          18  11    12  18    13  20
      female        13  13    16  15    12  15
```

Then, `as.matrix()` creates a matrix with the levels of the stacked variables combined with some separator character. Using `ca(pet.mat)` would then calculate the CA solution for the stacked table, analyzing only the associations in the loglinear model [Pet Age][Color Sex] .[4]

```
> (pet.mat <- as.matrix(ftable(Pet + Age ~ Color + Sex, tab), sep='.'))

             Pet.Age
Color.Sex     dog.young dog.old cat.young cat.old bird.young bird.old
  black.male         10      12        16      16         16       12
  black.female        8      12        13      15         11       13
  white.male         18      11        12      18         13       20
  white.female       13      13        16      15         12       15
```

[3] This requires at least R version 3.1.0 or vcd 1.3-2 or later.

[4] The result would not be at all interesting here. Why?

With data in a frequency data frame, a similar result (as a frequency table), can be obtained using `interaction()` as shown below. The result of `xtabs()` looks the same as `pet.mat`.

```
> tab.df <- as.data.frame(as.table(tab))
> tab.df <- within(tab.df,
+   {Pet.Age = interaction(Pet,Age)
+   Color.Sex = interaction(Color,Sex)
+   })
> xtabs(Freq ~ Color.Sex + Pet.Age, data=tab.df)
```

{ex:suicide1}

EXAMPLE 6.5: Suicide rates in Germany

To illustrate the use of correspondence analysis for the analysis for three-way tables, we use data on suicide rates in West Germany classified by sex, age, and method of suicide used. The data, from Heuer (1979, Table 1) have been discussed by Friendly (1991, 1994), van der Heijden and de Leeuw (1985) and others.

The original $2 \times 17 \times 9$ table contains 17 age groups from 10 to 90 in 5-year steps and 9 categories of suicide method, contained in the frequency data frame *Suicide* in vcd , with table variables `sex`, `age` and `method`. To avoid extremely small cell counts and cluttered displays, this example uses a reduced table in which age groups are combined in the variable `age.group`, a factor with 15 year intervals except for the last interval, which includes ages 70–90; the methods "toxic gas" and "cooking gas" were collapsed (in the variable `method2`) giving the $2 \times 5 \times 8$ table shown in the output below. These changes do not affect the general nature of the data or conclusions drawn from them.

In this example, we decided to stack the combinations of age and sex, giving an analysis of the loglinear model $[AgeSex][Method]$, to show how the age-sex categories relate to method of suicide.

In the case of a frequency data frame, it is quite simple to join two or more factors to form the rows of a new two-way table. Here we use `paste()` to form a new, composite factor, called `age_sex` here, abbreviating `sex` for display purposes.

```
> data("Suicide", package="vcd")
> # interactive coding of sex and age.group
> Suicide <- within(Suicide, {
+   age_sex <- paste(age.group, toupper(substr(sex,1,1)))
+   })
```

Then, use `xtabs()` to construct the two-way table `suicide.tab`:

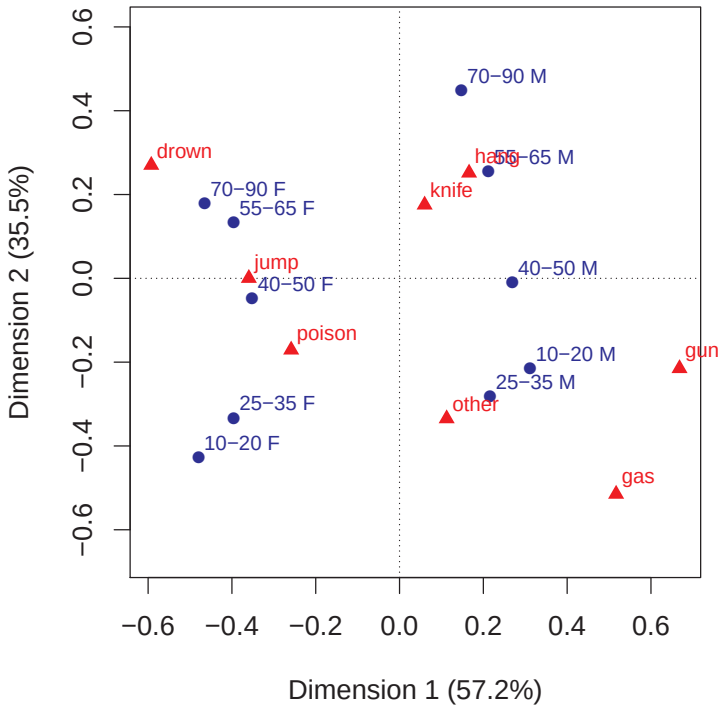
```
> suicide.tab <- xtabs(Freq ~ age_sex + method2, data=Suicide)
> suicide.tab

          method2
age_sex    poison  gas hang drown  gun knife jump other
  10-20 F     921   40  212    30   25    11  131   100
  10-20 M    1160  335 1524    67  512    47  189   464
  25-35 F    1672  113  575   139   64    41  276   263
  25-35 M    2823  883 2751   213  852   139  366   775
  40-50 F    2224   91 1481   354   52    80  327   305
  40-50 M    2465  625 3936   247  875   183  244   534
  55-65 F    2283   45 2014   679   29   103  388   296
  55-65 M    1531  201 3581   207  477   154  273   294
  70-90 F    1548   29 1355   501    3    74  383   106
  70-90 M     938   45 2948   212  229   105  268   147
```

The results of the correspondence analysis of this table are shown below:

```
> suicide.ca <- ca(suicide.tab)
> summary(suicide.ca)


Principal inertias (eigenvalues):

 dim    value      %    cum%   scree plot
 1      0.096151  57.2  57.2   **************
 2      0.059692  35.5  92.6   *********
 3      0.008183   4.9  97.5   *
 4      0.002158   1.3  98.8
 5      0.001399   0.8  99.6
 6      0.000557   0.3 100.0
 7      6.7e-050   0.0 100.0
        -------- -----
 Total: 0.168207 100.0
...
```

It can be seen that 92.6% of the χ^2 for this model is accounted for in the first two dimensions. Plotting these gives the display shown in Figure 6.6.

```
> plot(suicide.ca)
```



ig:ca-suicide-plot}

Figure 6.6: 2D CA solution for the stacked [AgeSex][Method] table of the suicide data

Dimension 1 in the plot separates males (right) and females (left), indicating a large difference between suicide profiles of males and females with respect to methods of suicide. The second dimension is mostly ordered by age with younger groups at the bottom and older groups at the top. Note also that the positions of the age groups are roughly parallel for the two sexes. Such a pattern indicates that sex and age do not interact in this analysis.

The relation between the age–sex groups and methods of suicide can be approximately interpreted in terms of similar distance and direction from the origin, which represents the marginal row and column profiles. Young males are more likely to commit suicide by gas or a gun, older males by hanging, while young females are more likely to ingest some toxic agent and older females by jumping or drowning. △

{ex:suicide2}

EXAMPLE 6.6: Suicide rates in Germany – mosaic plot

For comparison, it is useful to see how to construct a mosaic display showing the same associations for the loglinear model $[AS][M]$ as in the correspondence analysis plot. To do this, we first construct the three-way table, `suicide.tab3`,

```
> suicide.tab3 <- xtabs(Freq ~ sex + age.group + method2, data=Suicide)
```

As discussed in Chapter 5, mosaic plots are sensitive both to the order of variables used in successive splits, and to the order of levels within variables and are most effective when these orders are chosen to reflect the some meaningful ordering.

In the present example, `method2` is an unordered table factor, but Figure 6.6 shows that the methods of suicide vary systematically with both sex and age, corresponding to dimensions 1 and 2 respectively. Here we choose to reorder the table according to the coordinates on Dimension 1. We also delete the low-frequency `"other"` category to simplify the display.

```
> # methods, ordered as in the table
> suicide.ca$colnames

[1] "poison" "gas"    "hang"   "drown"  "gun"    "knife"
[7] "jump"   "other"

> # order of methods on CA scores for Dim 1
> suicide.ca$colnames[order(suicide.ca$colcoord[,1])]

[1] "drown"  "jump"   "poison" "knife"  "other"  "hang"
[7] "gas"    "gun"

> # reorder methods by CA scores on Dim 1
> suicide.tab3 <- suicide.tab3[, , order(suicide.ca$colcoord[,1])]
> # delete "other"
> suicide.tab3 <- suicide.tab3[,, -5]
> ftable(suicide.tab3)

                method2 drown jump poison knife hang  gas  gun
sex    age.group
male   10-20              67  189   1160    47 1524  335  512
       25-35             213  366   2823   139 2751  883  852
       40-50             247  244   2465   183 3936  625  875
       55-65             207  273   1531   154 3581  201  477
       70-90             212  268    938   105 2948   45  229
female 10-20              30  131    921    11  212   40   25
       25-35             139  276   1672    41  575  113   64
       40-50             354  327   2224    80 1481   91   52
       55-65             679  388   2283   103 2014   45   29
       70-90             501  383   1548    74 1355   29    3
```

To construct the mosaic display for the same model analysed by correspondence analysis, we use the argument `expected=~age.group*sex + method2` to supply the model formula. For this large table, it is useful to tweak the labels for the `method2` variable to reduce overplotting; the `labeling_args` argument provides many options for customizing `strucplot` displays.

```
> library(vcdExtra)
> mosaic(suicide.tab3, shade=TRUE, legend=FALSE,
+        expected=~age.group*sex + method2,
+        labeling_args=list(abbreviate_labs=c(FALSE, FALSE, 5)),
+                           rot_labels = c(0, 0, 0, 90))
```

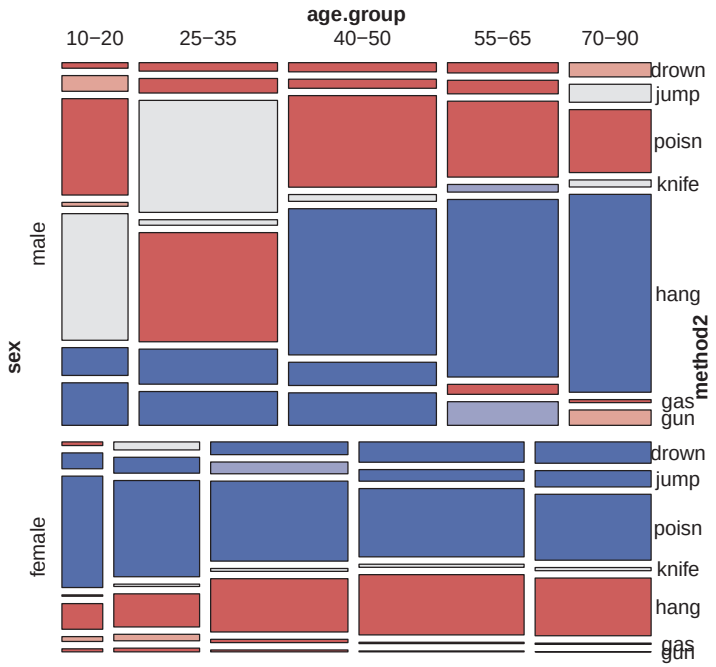


Figure 6.7: Mosaic display showing deviations from the model [AgeSex][Method] for the suicide data

:ca-suicide-mosaic}

This figure (Figure 6.7) again shows the prevalence of `gun` and `gas` among younger males and decreasing with age, whereas use of `hang` increases with age. For females, these three methods are used less frequently, whereas `poison`, `jump`, and `drown` occur more often. You can also see that for females the excess prevalence of these high frequency methods varies somewhat less with age than it does for males.

△

6.3.2 Marginal tables and supplementary variables

{ca:marginal}

An *n*-way table in frequency form or case form is automatically collapsed over factors which are not listed in the call to `xtabs()` when creating the table input for `ca()`. The analysis gives a ***marginal model*** for the categorical variables which *are* listed.

The positions of the categories of the omitted variables may nevertheless be recovered, by treating them as ***supplementary variables***, given as additional rows or columns in the two-way table. A supplementary variable is ignored in finding the CA solution, but its categories are then projected into that space. This is another useful trick to extend traditional CA to higher-way tables.

To illustrate, the code below list only the `age` and `method2` variables, and hence produces an

analysis collapsed over `sex`. This ignores not only the effect of sex itself, but also all associations of age and method with sex, which are substantial. We don't show the `ca()` result or the plot yet.

```
> # two way, ignoring sex
> suicide.tab2 <- xtabs(Freq ~ age.group + method2, data=Suicide)
> suicide.tab2

          method2
age.group poison  gas hang drown  gun knife jump other
    10-20   2081  375 1736    97  537    58  320   564
    25-35   4495  996 3326   352  916   180  642  1038
    40-50   4689  716 5417   601  927   263  571   839
    55-65   3814  246 5595   886  506   257  661   590
    70-90   2486   74 4303   713  232   179  651   253

> suicide.ca2 <- ca(suicide.tab2)
```

To treat the levels of `sex` as supplementary points, we calculate the two-way table of sex and method, and append this to the `suicide.tab2` as additional rows:

```
> # relation of sex and method
> suicide.sup <- xtabs(Freq ~ sex + method2, data=Suicide)
> suicide.tab2s <- rbind(suicide.tab2, suicide.sup)
```

In the call to `ca()`, we then indicate these last two rows as supplementary:

```
> suicide.ca2s <- ca(suicide.tab2s, suprow=6:7)
> summary(suicide.ca2s)


Principal inertias (eigenvalues):

 dim    value      %    cum%   scree plot
 1      0.060429  93.9  93.9   ***********************
 2      0.002090   3.2  97.1   *
 3      0.001479   2.3  99.4   *
 4      0.000356   0.6 100.0
        -------- -----
 Total: 0.064354 100.0

...
```

This CA analysis has the same total Pearson chi-square, $\chi^2(28) = 3422.5$ as the result of `chisq.test(suicide.tab2)`. However, the scree plot display above shows that the association between age and method is essentially one-dimensional, but note also that dimension 1 ("age-method") in this analysis has nearly the same inertia (0.0604) as the second dimension (0.0596) in the analysis of the stacked table. We plot the CA results as shown below (see Figure 6.8), and add a line connecting the supplementary points for sex.

```
> op <- par(cex=1.3, mar=c(4,4,1,1)+.1)
> res <- plot(suicide.ca2s, pch=c(16, 15, 17, 24))
> lines(res$rows[6:7,])
> par(op)
```

Comparing this graph with Figure 6.6, you can see that ignoring sex has collapsed the differences between males and females which were the dominant feature of the analysis including sex. The dominant feature in Figure 6.8 is the Dimension 1 ordering of both age and method. However, as in Figure 6.6, the supplementary points for sex point toward the methods that are more prevalent for females and males.

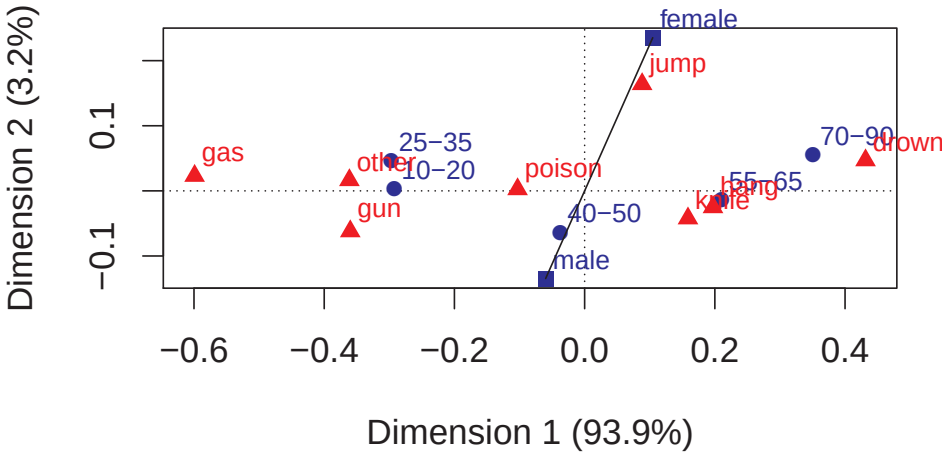


Figure 6.8: 2D CA solution for the [Age] [Method] marginal table. Category points for Sex are shown as supplementary points {fig:ca-suicide-sup}

6.4 Multiple correspondence analysis {sec:mca}

Multiple correspondence analysis (MCA) is designed to display the relationships of the categories of two or more discrete variables, but it is best used for multiway tables where the extensions of classical CA described in Section 6.3 do not suffice. Again, this is motivated by the desire to provide an *optimal scaling* of categorical variables, giving scores for the discrete variables in an n-way table with desirable properties and which can be plotted to visualize the relations among the category points.

The most typical development of MCA starts by defining indicator ("dummy") variables for each category and reexpresses the n-way contingency table in the form of a cases by variables indicator matrix, $\boldsymbol{Z}$. Simple correspondence analysis for a two-way table can, in fact, be derived as the canonical correlation analysis of the indicator matrix.

Unfortunately, the generalization to more than two variables follows a somewhat different path, so that simple CA does not turn out to be precisely a special case of MCA in some respects, particularly in the decomposition of an interpretable χ^2 over the dimensions in the visual representation.

Nevertheless, MCA does provide a useful graphic portrayal of the *bivariate* relations among any number of categorical variables, and has close relations to the mosaic matrix (Section 5.6). If its limitations are understood, it is helpful in understanding large, multivariate categorical data sets, in a similar way to the use of scatterplot matrices and dimension-reduction techniques (e.g., principal component analysis) for quantitative data.

6.4.1 Bivariate MCA {sec:mca-bi}

For the hair color, eye color data, the indicator matrix $\boldsymbol{Z}$ has 592 rows and $4 + 4 = 8$ columns. The columns refer to the eight categories of hair color and eye color and the rows to the 592 students in Snee's 1974 sample.

For simplicity, we show the calculation of the indicator matrix below in frequency form, using `model.matrix()` to compute the dummy (0/1) variables for the levels of hair color (`Hair1–Hair4`) and eye color (`Eye1–Eye4`).

```
> haireye.df <- cbind(
+     as.data.frame(haireye),
```

```
+     model.matrix(Freq ~ Hair + Eye, data=haireye,
+         contrasts.arg=list(Hair=diag(4), Eye=diag(4)))[,-1]
+     )
> haireye.df

    Hair   Eye Freq Hair1 Hair2 Hair3 Hair4 Eye1 Eye2 Eye3 Eye4
1  Black Brown   68     1     0     0     0    1    0    0    0
2  Brown Brown  119     0     1     0     0    1    0    0    0
3    Red Brown   26     0     0     1     0    1    0    0    0
4  Blond Brown    7     0     0     0     1    1    0    0    0
5  Black  Blue   20     1     0     0     0    0    1    0    0
6  Brown  Blue   84     0     1     0     0    0    1    0    0
7    Red  Blue   17     0     0     1     0    0    1    0    0
8  Blond  Blue   94     0     0     0     1    0    1    0    0
9  Black Hazel   15     1     0     0     0    0    0    1    0
10 Brown Hazel   54     0     1     0     0    0    0    1    0
11   Red Hazel   14     0     0     1     0    0    0    1    0
12 Blond Hazel   10     0     0     0     1    0    0    1    0
13 Black Green    5     1     0     0     0    0    0    0    1
14 Brown Green   29     0     1     0     0    0    0    0    1
15   Red Green   14     0     0     1     0    0    0    0    1
16 Blond Green   16     0     0     0     1    0    0    0    1
```

Thus, the first row in `haireye.df` represents the 68 individuals having black hair (`Hair1=1`) and brown eyes (`Eye1=1`). The indicator matrix $\boldsymbol{Z}$ is then computed by replicating the rows in `haireye.df` according to the `Freq` value, using the function `expand.dft`. The result has 592 rows and 8 columns.

```
> Z <- expand.dft(haireye.df)[,-(1:2)]
> vnames <- c(levels(haireye.df$Hair), levels(haireye.df$Eye))
> colnames(Z) <- vnames
> dim(Z)

[1] 592   8
```

Note that if the indicator matrix is partitioned as $\boldsymbol{Z} = [\boldsymbol{Z}_1, \boldsymbol{Z}_2]$, corresponding to the two sets of categories, then the contingency table is given by $\boldsymbol{N} = \boldsymbol{Z}_1^{\mathsf{T}} \boldsymbol{Z}_2$.

```
> (N <- t(as.matrix(Z[,1:4])) %*% as.matrix(Z[,5:8]))

      Brown Blue Hazel Green
Black    68   20    15     5
Brown   119   84    54    29
Red      26   17    14    14
Blond     7   94    10    16
```

With this setup, MCA can be described as the application of the simple correspondence analysis algorithm to the indicator matrix $\boldsymbol{Z}$. This analysis would yield scores for the rows of $\boldsymbol{Z}$ (the cases), usually not of direct interest and for the columns (the categories of both variables). As in simple CA, each row point is the weighted average of the scores for the column categories, and each column point is the weighted average of the scores for the row observations.[5]

Consequently, the point for any category is the centroid of all the observations with a response in that category, and all observations with the same response pattern coincide. As well, the origin reflects the weighted average of the categories for *each* variable. As a result, category points with low marginal frequencies will be located further away from the origin, while categories with high marginal frequencies will be closer to the origin. For a binary variable, the two category points will appear on a line through the origin, with distances inversely proportional to their marginal frequencies.

{ex:haireye4}

[5] Note that, in principle, this use of an indicator matrix could be extended to three (or more) variables. That extension is more easily described using an equivalent form, the ***Burt matrix***, described in Section 6.4.2.

EXAMPLE 6.7: Hair color and eye color

For expository purposes, we illustrate the analysis of the indicator matrix below for the hair color, eye color data using `ca()`, rather than the function `mjca()` which is designed for a more general approach to MCA.

```
> Z.ca <- ca(Z)
> res <- plot(Z.ca, what=c("none", "all"))
```

In the call to `plot.ca`, the argument `what` is used to suppress the display of the row points for the cases. The plot shown in Figure 6.9 is an enhanced version of this basic plot.

```
      Dim1     Dim2 factor levels
1 -0.94250  1.09220   Hair  Black
2 -0.27693 -0.16608   Hair  Brown
3 -0.24194 -1.62513   Hair    Red
4  1.56039  0.35376   Hair  Blond
5 -0.91933  0.44905    Eye  Brown
6  1.02254  0.42176    Eye   Blue
7 -0.39712 -0.85105    Eye  Hazel
8  0.30215 -1.72375    Eye  Green
```

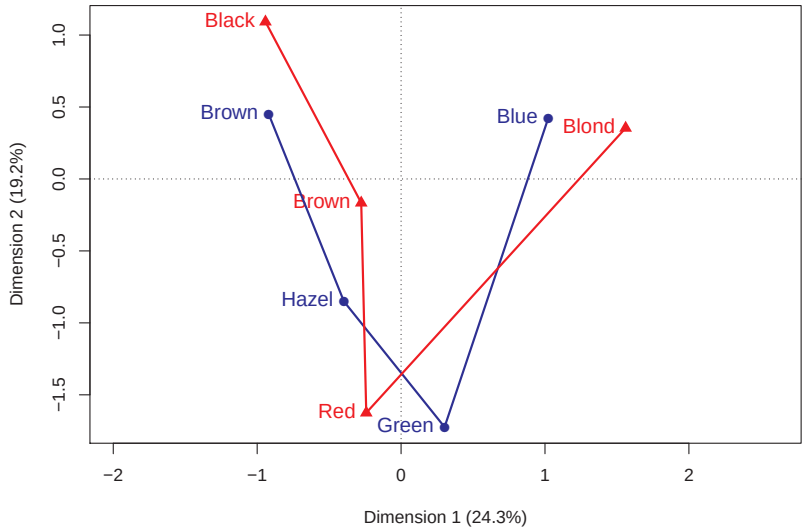


Figure 6.9: Correspondence analysis of the indicator matrix Z for the hair color, eye color data. The category points are joined separately by lines for the hair color and eye color categories.

{fig:mca-haireye1}

Comparing Figure 6.9 with Figure 6.1, we see that the general pattern of the hair color and eye color categories is the same in the analysis of the contingency table (Figure 6.1) and the analysis of the indicator matrix (Figure 6.9), except that the axes are scaled differently—the display has been stretched along the second (vertical) dimension. The interpretation is the same: Dimension 1 reflects a dark–light ordering of both hair and eye colors, and Dimension 2 reflects something that largely distinguishes red hair and green eyes from the other categories.

Indeed, it can be shown (Greenacre, 1984, 2007) that the two displays are identical, except for changes in scales along the axes. There is no difference at all between the displays in standard

coordinates. Greenacre (1984, pp. 130–134) describes the precise relations between the geometries of the two analyses.

△

Aside from the largely cosmetic difference in relative scaling of the axes, a major difference between analysis of the contingency table and analysis of the indicator matrix is in the decomposition of principal inertia and corresponding χ^2 contributions for the dimensions. The plot axes in Figure 6.9 indicate 24.3% and 19.2% for the contributions of the two dimensions, whereas Figure 6.1 shows 89.4% and 9.5%. This difference is the basis for the more general development of MCA methods and is reflected in the `mcja()` function illustrated later in this chapter. But first, we describe a second approach to extending simple CA to the multivariate case based on the ***Burt matrix***.

6.4.2 The Burt matrix

{sec:mca-burt}

The same solution for the category points as in the analysis of the indicator matrix may be obtained more simply from the so-called ***Burt matrix*** (Burt, 1950),

$$\boldsymbol{B} = \boldsymbol{Z}^{\mathsf{T}}\boldsymbol{Z} = \begin{bmatrix} \boldsymbol{N}_1 & \boldsymbol{N} \\ \boldsymbol{N}^{\mathsf{T}} & \boldsymbol{N}_2 \end{bmatrix} ,$$

where $\boldsymbol{N}_1$ and $\boldsymbol{N}_2$ are diagonal matrices containing the marginal frequencies of the two variables (the column sums of $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$). In this representation, the contingency table of the two variables, $\boldsymbol{N}$ appears in the off-diagonal block, $\boldsymbol{N}$ in this equation. This calculation is shown below.

```
> Burt <- t(as.matrix(Z)) %*% as.matrix(Z)
> rownames(Burt) <- colnames(Burt) <- vnames
> Burt

       Black Brown Red Blond Brown Blue Hazel Green
Black    108     0   0     0    68   20    15     5
Brown      0   286   0     0   119   84    54    29
Red        0     0  71     0    26   17    14    14
Blond      0     0   0   127     7   94    10    16
Brown     68   119  26     7   220    0     0     0
Blue      20    84  17    94     0  215     0     0
Hazel     15    54  14    10     0    0    93     0
Green      5    29  14    16     0    0     0    64
```

The standard coordinates from an analysis of the Burt matrix $\boldsymbol{B}$ are identical to those of $\boldsymbol{Z}$. (However, the singular values of $\boldsymbol{B}$ are the squares of those of $\boldsymbol{Z}$.) Then, the following code, using `Burt` produces the same display of the category points for hair color and eye color as shown for the indicator matrix `Z` in Figure 6.9.

```
> Burt.ca <- ca(Burt)
> plot(Burt.ca)
```

6.4.3 Multivariate MCA

{sec:mca-multi}

The coding of categorical variables in an indicator matrix and the relationship to the Burt matrix provides a direct and natural way to extend this analysis to more than two variables. If there are Q categorical variables, and variable q has J_q categories, then the Q-way contingency table, of size $J = \prod_{q=1}^{Q} J_q = J_1 \times J_2 \times \cdots \times J_Q$, with a total of $n = n_{++\ldots}$ observations may be represented by the partitioned $(n \times J)$ indicator matrix $[\boldsymbol{Z}_1\, \boldsymbol{Z}_2\, \ldots\, \boldsymbol{Z}_Q]$.

Then the Burt matrix is the symmetric partitioned matrix

$$\boldsymbol{B} = \boldsymbol{Z}^\mathsf{T}\boldsymbol{Z} = \begin{bmatrix} \boldsymbol{N}_1 & \boldsymbol{N}_{12} & \cdots & \boldsymbol{N}_{1Q} \\ \boldsymbol{N}_{21} & \boldsymbol{N}_2 & \cdots & \boldsymbol{N}_{2Q} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{N}_{Q1} & \boldsymbol{N}_{Q2} & \cdots & \boldsymbol{N}_Q \end{bmatrix} , \tag{6.7}$$

{eq:burt}

where again the diagonal blocks $\boldsymbol{N}_i$ contain the one-way marginal frequencies. The off-diagonal blocks $\boldsymbol{N}_{ij}$ contain the bivariate marginal contingency tables for each pair (i, j) of variables.

Classical MCA (see, e.g., Greenacre (1984), Gower and Hand (1996)) can then be defined as a singular value decomposition of the matrix $\boldsymbol{B}$ which produces scores for the categories of *all* variables so that the greatest proportion of the bivariate, pairwise associations in all blocks (including the diagonal blocks) is accounted for in a small number of dimensions.

In this respect, MCA resembles multivariate methods for quantitative data based on the joint bivariate correlation or covariance matrix ($\boldsymbol{\Sigma}$) and there is some justification to regard the Burt matrix as the categorical analog of $\boldsymbol{\Sigma}$.[6]

There is a close connection between this analysis and the bivariate mosaic matrix (Section 5.6): The mosaic matrix displays the residuals from independence for each pair of variables, and thus provides a visual representation of the Burt matrix. The one-way margins shown (by default) in the diagonal cells reflect the diagonal matrices $\boldsymbol{N}_i$ in Eqn. (6.7). The total amount of shading in all the individual mosaics portrays the total pairwise associations decomposed by MCA. See Friendly (1999) for further details.

For interpretation of MCA plots, we note the following relations (Greenacre, 1984, §5.2):[7]

- The inertia contributed by a given variable increases with the number of response categories.
- The centroid of the categories for each discrete variable is at the origin of the display.
- For a particular variable, the inertia contributed by a given category increases as the marginal frequency in that category *decreases*. Low frequency points therefore appear further from the origin.
- The category points for a binary variable lie on a line through the origin. The distance of each point to the origin is inversely related to the marginal frequency.

{ex:marital3}

EXAMPLE 6.8: Marital status and pre- and extramarital sex

The data on the relation between marital status and reported premarital and extramarital sex was explored earlier using mosaic displays in Example 5.9 and Example 5.13.

Using the ca package, an MCA analysis of the `PreSex` data is carried out using `mjca()`. This function typically takes a data frame in *case form* containing the factor variables, but converts a table to this form. This example analyzes the Burt matrix calculated from the `PreSex` data, specified as `lambda="Burt"`

```
> data("PreSex", package="vcd")
> PreSex <- aperm(PreSex, 4:1)    # order variables G, P, E, M
> presex.mca <- mjca(PreSex, lambda="Burt")
> summary(presex.mca)


Principal inertias (eigenvalues):
```

[6] For multivariate normal data, however, the mean vector and covariance matrix are sufficient statistics, so all higher-way relations are captured in the covariance matrix. This is not true of the Burt matrix. Moreover, the covariance matrix is typically expressed in terms of mean-centered variables, while the Burt matrix involves the marginal frequencies. A more accurate statement is that the uncentered covariance matrix is analogous to the Burt matrix.

[7] This book, now out of print, is available for free download at `http://www.carme-n.org/`

```
 dim    value      %   cum%   scree plot
 1      0.149930  53.6  53.6  *************
 2      0.067201  24.0  77.6  ******
 3      0.035396  12.6  90.2  ***
 4      0.027365   9.8 100.0  **
        -------- -----
 Total: 0.279892 100.0
...
```

The output from `summary()` seems to show that 77.6% of the total inertia is accounted for in two dimensions. A basic, default plot of the MCA solution is provided by the `plot()` method for "mjca" objects.

```
> plot(presex.mca)
```

This plotting method is not very flexible in terms of control of graphical parameters or the ability to add additional annotations (labels, lines, legend) to ease interpretation. Instead, we use the plot method to create an empty plot (with no points or labels), and return the calculated plot coordinates (`res`) for the categories. A bit of processing of the coordinates provides the customized display shown in Figure 6.10.

```
> # plot, but don't use point labels or points
> res <- plot(presex.mca, labels=0, pch='.', cex.lab=1.2)
>
> # extract factor names and levels
> coords <- data.frame(res$cols, presex.mca$factors)
> nlev <- presex.mca$levels.n
> fact <- unique(as.character(coords$factor))
>
> cols <- c("blue", "red", "brown", "black")
> points(coords[,1:2], pch=rep(16:19, nlev), col=rep(cols, nlev), cex=1.2)
> text(coords[,1:2], label=coords$level, col=rep(cols, nlev), pos=3,
+      cex=1.2, xpd=TRUE)
> lwd <- c(2, 2, 2, 4)
> for(i in seq_along(fact)) {
+   lines(Dim2 ~ Dim1, data = coords, subset = factor==fact[i],
+         lwd = lwd[i], col = cols[i])
+ }
>
> legend("bottomright", legend = c("Gender", "PreSex", "ExtraSex", "Marital"),
+   title = "Factor", title.col = "black",
+   col = cols, text.col = cols, pch = 16:19,
+   bg = "gray95", cex = 1.2)
```

As indicated above, the category points for each factor appear on lines through the origin, with distances inversely proportional to their marginal frequencies. For example, the categories for No premarital and extramarital sex are much larger than the corresponding Yes categories, so the former are positioned closer to the origin. In contrast, the categories of gender and marital status are more nearly equal marginally.

Another aspect of interpretation of Figure 6.10 concerns the alignment of the lines for different factors. The positions of the category points on Dimension 1 suggest that Women are less likely to have had pre-marital and extra-marital sex and that still being married is associated with the absence of pre- and extra-marital sex. As well, the lines for gender and marital status are nearly at right angles, suggesting that these variables are unassociated. This interpretation is more or less correct, but it is only approximate in this MCA scaling of the coordinate axes. An alternative scaling, based on a ***biplot*** representation is described in Section 6.5.

If you compare the MCA result in Figure 6.10 with the mosaic matrix in Figure 5.23, you will see that they are both showing the bivariate pairwise associations among these variables, but in different ways. The mosaic plots show the details of marginal and joint frequencies together with residuals from independence for each 2×2 marginal subtable. The MCA plot using the Burt matrix

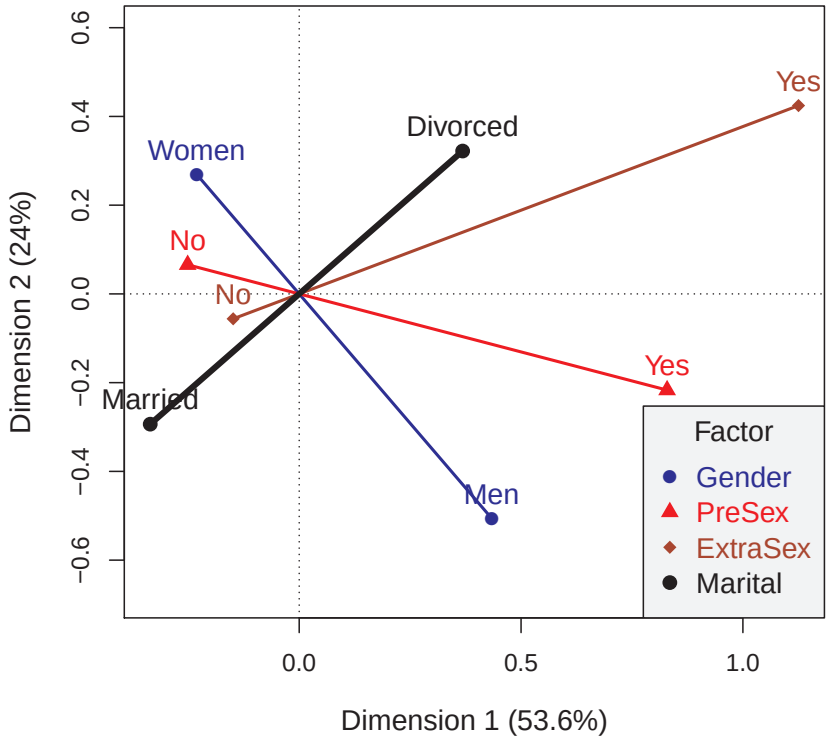


Figure 6.10: MCA plot of the Burt matrix for the PreSex data. The category points are joined separately by lines for the factor variables. {fig:presex-mca-plot}

summarizes each category point in terms of a 2D representation of contributions to total inertia (association). △

6.4.3.1 Inertia decomposition

The transition from simple CA to MCA is straight-forward in terms of the category scores derived from the indicator matrix $\boldsymbol{Z}$ or the Burt matrix, $\boldsymbol{B}$. It is less so in terms of the calculation of total inertia, and therefore in the chi-square values and corresponding percentages of association accounted for in some number of dimensions.

In simple CA, the total inertia is χ^2/n, and it therefore makes sense to talk of percentage of association accounted for by each dimension. But in MCA of the indicator matrix the total inertia, $\sum \lambda^2$, is simply $(J - Q)/Q$, because the inertia of each subtable, $\boldsymbol{Z}_i$ is equal to its dimensionality, $J_i - 1$, and the total inertia of an indicator matrix is the average of the inertias of its subtables. Consequently, the average inertia per dimension is $1/Q$, and it is common to interpret only those dimensions that exceed this average (analogous to the use of 1 as a threshold for eigenvalues in principal components analysis).

To more adequately reflect the percentage of association in MCA, Greenacre (1990), revising an earlier proposal by Benzécri (1977), suggested the calculation of ***adjusted inertia***, which ignores the contributions of the diagonal blocks in the Burt matrix,

$$(\lambda_i^\star)^2 = \left[\frac{Q}{Q-1} (\lambda_i^Z - \frac{1}{Q}) \right]^2 \tag{6.8}$$ {eq:benzecri}

as the principal inertia due to the dimensions with $(\lambda^Z)^2 > 1/Q$. This adjustment expresses the

contribution of each dimension as $(\lambda_i^\star)^2 / \sum(\lambda_i^\star)^2$, with the summation over only dimensions with $(\lambda^Z)^2 > 1/Q$.

A related method, also handled by `mjca()`, is ***joint correspondence analysis*** (Greenacre, 1994, Greenacre, 2007, Chapter 19) an iterative method that replaces the diagonal blocks of the Burt matrix with values that minimize their impact on inertia. Unlike MCA, solutions in JCA are not nested, however.

{ex:titanic2}

EXAMPLE 6.9: Survival on the *Titanic*

An MCA analysis of the `Titanic` data is carried out using `mjca()` as shown below.

```
> titanic.mca <- mjca(Titanic)
```

`mjca()` allows different scaling methods for the contributions to inertia of the different dimensions. The default (`lambda="adjusted"`), used here, is the adjusted inertias as in Eqn. (6.8).

```
> summary(titanic.mca)


Principal inertias (eigenvalues):

 dim    value      %   cum%   scree plot
 1      0.067655  76.8  76.8  ************************
 2      0.005386   6.1  82.9  **
 3      00000000   0.0  82.9
        -------- -----
 Total: 0.088118
...
```

Using similar code to that used in Example 6.8, Figure 6.11 shows an enhanced version of the default plot that connects the category points for each factor by lines using the result returned by the `plot()` function.

In this plot, the points for each factor have the property that the sum of coordinates on each dimension, weighted inversely by the marginal proportions, equals zero. Thus high frequency categories (e.g., Adult and Male) are close to the origin.

The first dimension is perfectly aligned with gender, and also strongly aligned with Survival. The second dimension pertains mainly to Class and Age effects. Considering those points which differ from the origin most similarly (in distance and direction) to the point for Survived, gives the interpretation that survival was associated with being female or upper class or (to a lesser degree) being a child.

△

6.5 Biplots for contingency tables

{sec:biplot}

Like correspondence analysis, the ***biplot*** (Bradu and Gabriel, 1978, Gabriel, 1971, 1980, 1981, Gower *et al.*, 2011) is a visualization method which uses the SVD to display a matrix in a low-dimensional (usually 2-dimensional) space. They differ in the relationships in the data that are portrayed, however:

- In correspondence analysis the (weighted, χ^2) *distances* between row points and distances between column points are designed to reflect *differences* between the row profiles and column profiles.
- In the biplot, on the other hand, row and column points are represented by *vectors* from the

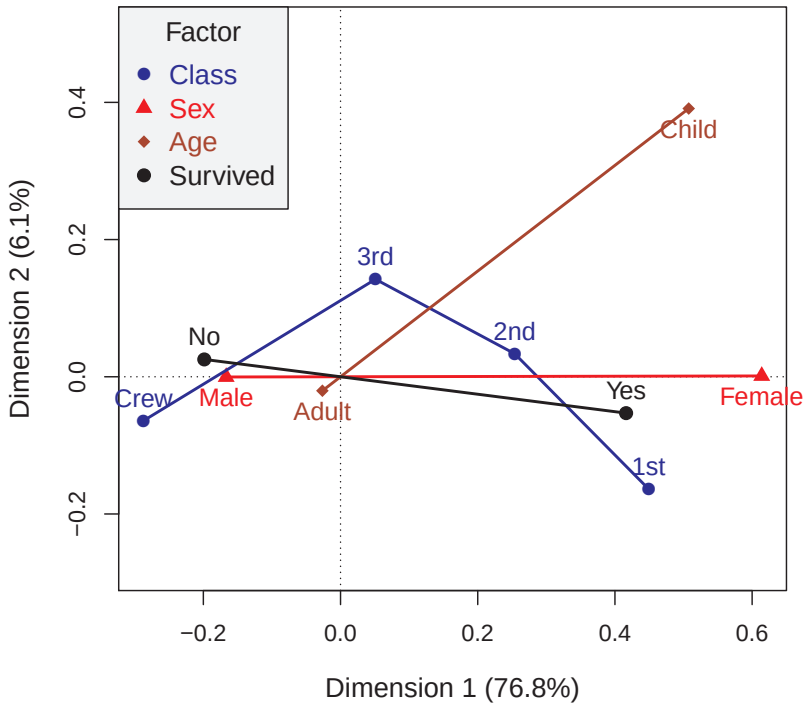


Figure 6.11: MCA plot of the Titanic data. The category points are joined separately by lines for the factor variables. {fig:titanic-mca-plot}

origin such that the projection (inner product) of the vector $\boldsymbol{a}_i$ for row i on $\boldsymbol{b}_j$ for column j approximates the data element y_{ij},

$$\boldsymbol{Y} \approx \boldsymbol{A}\boldsymbol{B}^{\mathsf{T}} \iff y_{ij} \approx \boldsymbol{a}_i^{\mathsf{T}} \boldsymbol{b}_j \ . \tag{6.9}$$ {eq:biplot1}

Geometrically, Eqn. (6.9) may be described as approximating the data value y_{ij} by the projection of the end point of vector $\boldsymbol{a}_i$ on $\boldsymbol{b}_j$ (and vice-versa), as shown in Figure 6.12.

6.5.1 CA bilinear biplots

As in CA, there are a number of different representations of coordinates for row and column points for a contingency table within a biplot framework. One set of connections between CA and the biplot can be be seen through the *reconstitution formula*, giving the decomposition of the correspondence matrix $\boldsymbol{P} = \boldsymbol{N}/n$ in terms of the standard coordinates $\boldsymbol{\Phi}$ and $\boldsymbol{\Gamma}$ defined in Eqn. (6.4) and Eqn. (6.5) as:

$$p_{ij} = r_i c_j \left(1 + \sum_{m=1}^{M} \sqrt{\lambda_m} \phi_{im} \gamma_{jm} \right) \tag{6.10}$$ {eq:reconstitution1}

or, in matrix terms,

$$\boldsymbol{P} = \boldsymbol{D}_r (\boldsymbol{1}\boldsymbol{1}^{\mathsf{T}} + \boldsymbol{\Phi} \boldsymbol{D}_\lambda^{1/2} \Gamma^{\mathsf{T}}) \boldsymbol{D}_c \tag{6.11}$$ {eq:reconstitution2}

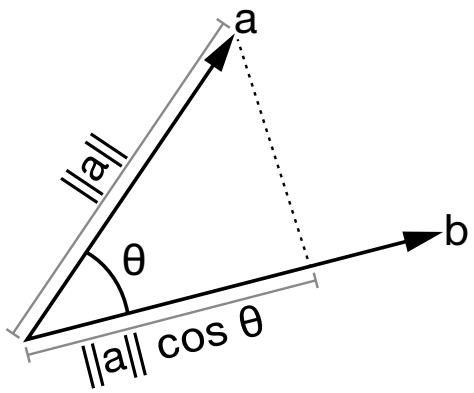


Figure 6.12: The scalar product of vectors of two points from the origin is the length of the projection of one vector on the other.

{fig:Scalarproduct}

The CA solution approximates this by a sum over $d \ll M$ dimensions, or by using only the first d (usually 2) columns of $\mathbf{\Phi}$ and $\mathbf{\Gamma}$.

Eqn. (6.10) can be re-written in biplot scalar form as

$$\left(\frac{p_{ij}}{r_i c_j}\right) - 1 \approx \sum_{m=1}^{d} (\sqrt{\lambda_m}\phi_{im})\gamma_{jm} = \sum_{m=1}^{d} f_{im}\gamma_{jm} \tag{6.12}$$

{eq:rowprincipal}

where $f_{im} = (\sqrt{\lambda_m}\phi_{im})$ gives the principal coordinates of the row points. The left-hand side of Eqn. (6.12) contains the ***contingency ratios***, $p_{ij}/r_i c_j$ of the observed cell probabilities to their expected values under independence. This shows that an ***asymmetric CA plot*** of row principal coordinates $\boldsymbol{F}$ and the column standard coordinates $\mathbf{\Gamma}$ is a biplot that approximates the deviations of the contingency ratios from their values under independence.

In the ca package, this plot is obtained by specifying `map="rowprincipal"` in the call to `plot()`, or `map="colprincipal"` to plot the column points in principal coordinates. It is typical in such biplots to display one set of coordinates as points and the other as vectors from the origin, as controlled by the `arrows` argument, so that one can interpret the data values represented as approximated by the projections of the points on the vectors.

Two other types asymmetric "maps" are also defined with different scalings that turn out to have better visual properties in terms of representing the relations between the row and column categories, particularly when the strength of association (inertia) in the data is low.

- The option `map="rowgab"` (or `map="colgab"`) gives a biplot form proposed by Gabriel and Odoroff (1990) with the rows (columns) shown in principal coordinates and the columns (rows) in standard coordinates multiplied by the mass c_j (r_i) of the corresponding point.
- The *contribution biplot* for CA (Greenacre, 2013), with the option `map="rowgreen"` (or `map="colgreen"`) provides a reconstruction of the standardized residuals from independence, using the points in standard coordinates multiplied by the square root of the corresponding masses. This has the nice visual property of showing more directly the contributions of the vectors to the low-dimensional solution.

{ex:suicide3}

EXAMPLE 6.10: Suicide rates in Germany – biplot

To illustrate the biplot representation, we continue with the data on suicide rates in Germany

from Example 6.5 using the stacked table `suicide.tab` comprised of the age–sex combinations as rows and methods of suicide as columns.

```
> suicide.tab <- xtabs(Freq ~ age_sex + method2, data=Suicide)
> suicide.ca <- ca(suicide.tab)
```

Using this result, `suicide.ca`, in the call to `plot()` below, we use `map="colgreen"` and vectors represent the methods of suicide, as shown in Figure 6.13.

```
> plot(suicide.ca, map="colgreen", arrows=c(FALSE, TRUE))
```

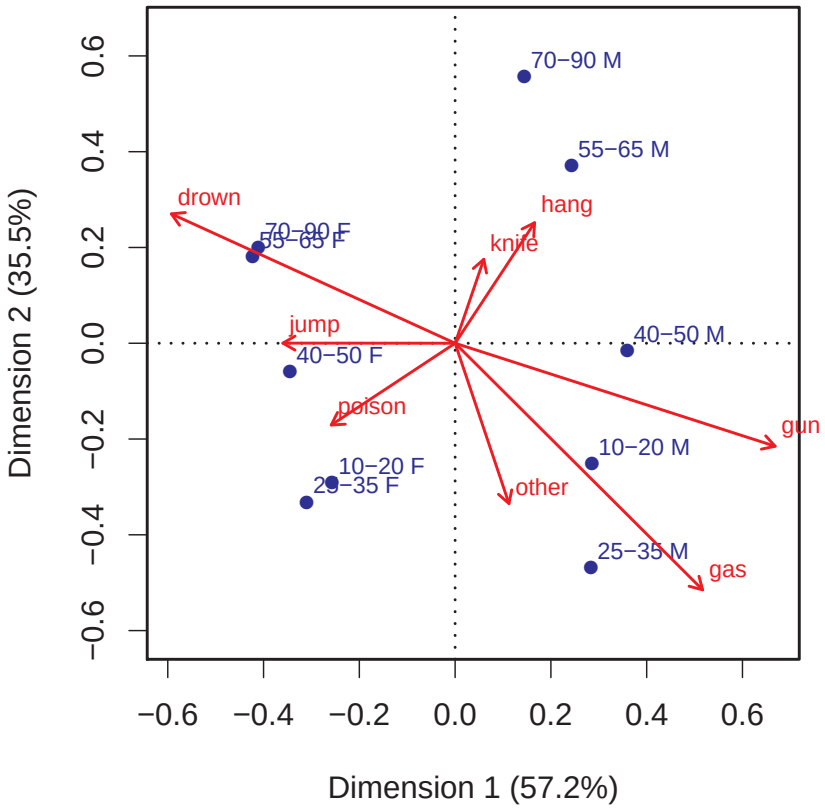


Figure 6.13: CA biplot of the suicide data using the contribution biplot scaling. Associations between the age-sex categories and the suicide methods can be read as the projections of the points on the vectors. The lengths of the vectors for the suicide categories reflect their contributions to this representation in a 2D plot.

:ca-suicide-biplot}

The interpretation of the row points for the age–sex categories is similar to what we saw earlier in Figure 6.6. But now, the vectors for the suicide categories reflect the contributions of those methods to the representation of association. Thus, the methods `drown`, `gun` and `gas` have large contributions, while `knife`, `hang`, and `poison` are relatively small. Moreover, the projections of the points for the age–sex combinations on the method vectors reflect the standardized residuals from independence.

The most comprehensive modern treatment of biplot methodology is the book *Understanding Biplots* (Gower *et al.*, 2011). Together with the book, they provide an R package, UBbipl (le Roux

and Lubbe, 2013), that is capable of producing an astounding variety of high-quality plots. Unfortunately, that package is only available on their publisher's web site[8] and you need the book to be able to use it because all the documentation is in the book. Nevertheless, we illustrate the use of the `cabipl()` function to produce the version of the CA biplot shown in Figure 6.14.

```
> library(UBbipl)
> cabipl(as.matrix(suicide.tab),
+     axis.col = gray(.4), ax.name.size=1,
+     ca.variant = "PearsonResA",
+     markers = FALSE,
+     row.points.size = 1.5,
+     row.points.col = rep(c("red", "blue"), 4),
+     plot.col.points = FALSE,
+     marker.col = "black", marker.size=0.8,
+     offset = c(2, 2, 0.5, 0.5),
+     offset.m = rep(-0.2, 14),
+     output=NULL)
```

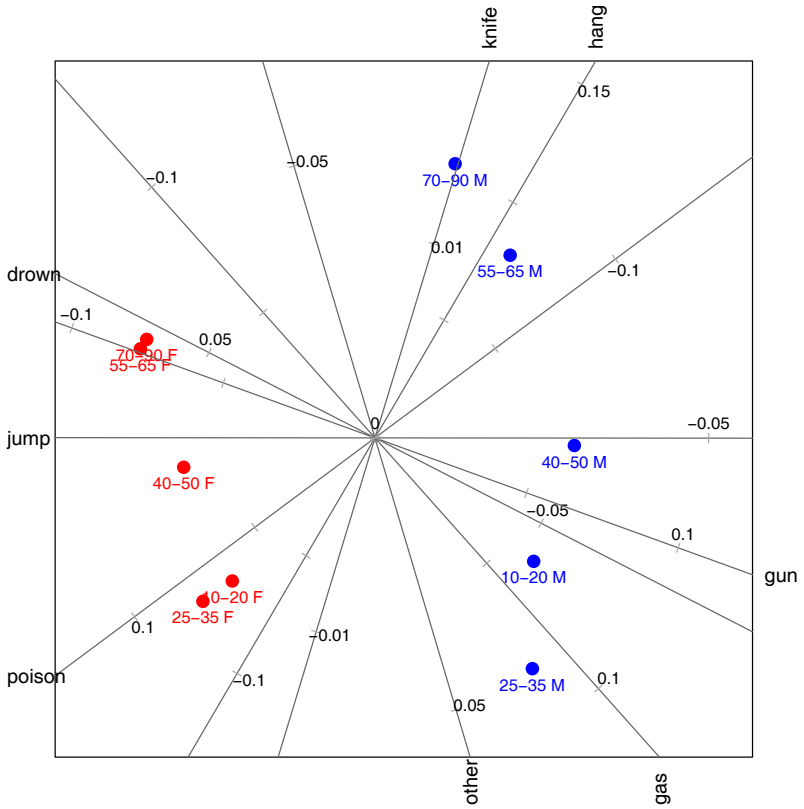


Figure 6.14: CA biplot of the suicide data, showing calibrated axes for the suicide methods. {fig:cabipl-suicide}

This plot uses `ca.variant = "PearsonResA"` to specify that the biplot is to approximate the standardized Pearson residuals by the inner product of each row point on the vector for the column point for the suicide methods, as also in Figure 6.13. However, Figure 6.14 represents the methods calibrated axis lines, designed to be read as scales for the projections of the row points

[8] `http://www.wiley.com/legacy/wileychi/gower/material.html`

(age–sex) on the methods. The UBbipl package has a huge number of options for controlling the details of the biplot display. See (Gower *et al.*, 2011, Ch. 2) for all the details.

△

6.5.2 Biadditive biplots

A different use of biplots for contingency tables stems from the close analogy between additive relations for a quantitative response when there is no interaction between factors, and the multiplicative relations for a contingency table when there is no association.

For quantitative data Bradu and Gabriel (1978) show how the biplot can be used to diagnose additive relations among rows and columns. For example, when a two-way table is well-described by a two-factor ANOVA model with no interaction,

$$y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij} \iff \boldsymbol{Y} \approx \boldsymbol{a}\mathbf{1}^\mathsf{T} + \mathbf{1}\boldsymbol{b}^\mathsf{T}$$

then, the row points, $\boldsymbol{a}_i$, and the column points, $\boldsymbol{b}_j$, will fall on two straight lines at right angles to each other in the biplot. For a contingency table, the multiplicative relations among frequencies under independence become additive relations in terms of log frequency, and Gabriel *et al.* (1997) illustrate how biplots of log frequency can be used to explore associations in two-way and three-way tables.

That is, For a two-way table, independence, $A \perp B$, implies that ratios of frequencies should be proportional for any two rows, i, i' and any two columns, j, j'. Equivalently, this means that the log odds ratio for all such sets of four cells should be zero:

$$A \perp B \iff \log \theta_{ii',jj'} = \log \left(\frac{n_{ij} n_{i'j'}}{n_{i'j} n_{ij'}} \right) = 0$$

Now, if the log frequencies have been centered by subtracting the grand mean, Gabriel *et al.* (1997) show that $\log \theta_{ii',jj'}$ is approximated in the biplot (of $\log(n_{ij}) - \overline{\log(n_{ij})}$)

$$\log \theta_{ii',jj'} \approx \boldsymbol{a}_i^\mathsf{T} \boldsymbol{b}_j - \boldsymbol{a}_{i'}^\mathsf{T} \boldsymbol{b}_j - \boldsymbol{a}_i^\mathsf{T} \boldsymbol{b}_{j'} + \boldsymbol{a}_i^\mathsf{T} \boldsymbol{b}_{j'} = (\boldsymbol{a}_i - \boldsymbol{a}_{i'})^\mathsf{T} (\boldsymbol{b}_i - \boldsymbol{b}_{i'})$$

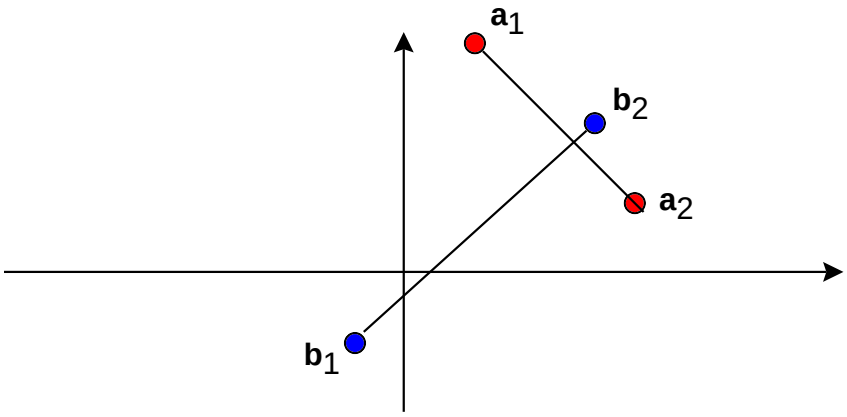


Figure 6.15: Independence implies orthogonal vector differences in a biplot of log frequency. The line joining $\boldsymbol{a}_1$ to $\boldsymbol{a}_2$ represents $(\boldsymbol{a}_1 - \boldsymbol{a}_2)$. This line is perpendicular to the line $(\boldsymbol{b}_1 - \boldsymbol{b}_2)$ under independence.

{fig:bidemo}

Therefore, this biplot criterion for independence in a two-way table is whether $(\boldsymbol{a}_i - \boldsymbol{a}_{i'})^\mathsf{T} (\boldsymbol{b}_i -$

$\boldsymbol{b}_{i'}) \approx 0$ for all pairs of rows, i, i', and all pairs of columns, j, j'. But $(\boldsymbol{a}_i - \boldsymbol{a}_{i'})$ is the vector connecting $\boldsymbol{a}_i$ to $\boldsymbol{a}_{i'}$ and $(\boldsymbol{b}_j - \boldsymbol{b}_{j'})$ is the vector connecting $\boldsymbol{b}_j$ to $\boldsymbol{b}_{j'}$, as shown in Figure 6.15, and the inner product of any two vectors equals zero *iff* they are orthogonal. Hence, this criterion implies that all lines connecting pairs of row points are orthogonal to lines connecting pairs of column points, as illustrated in Figure 6.15.

{ex:soccer3}

EXAMPLE 6.11: UK Soccer scores

We examined the data on UK Soccer scores in Example 5.5 and saw that the number of goals scored by the home and away teams were largely independent (see Figure 5.10). This data set provides a good test of the ability of the biplot to diagnose independence.

```
> data("UKSoccer", package="vcd")
> dimnames(UKSoccer) <- list(Home=paste0("H", 0:4),
+                            Away=paste0("A", 0:4))
```

Basic biplots in R are provided by `biplot()` that works mainly with the result calculated by `prcomp()` or `princomp()`. Here, we use `prcomp()` on the log frequencies in the *`UKSoccer`* table, adding 1, because there is one cell with zero frequency.

```
> soccer.pca <- prcomp(log(UKSoccer+1), center=TRUE, scale.=FALSE)
```

The result is plotted using a customized plot based on `biplot()` as shown in Figure 6.16.

```
> biplot(soccer.pca, scale=0, var.axes=FALSE,
+   col=c("blue", "red"), cex=1.2, cex.lab=1.2,
+   xlab="Dimension 1", ylab="Dimension 2")
```

To supplement this plot and illustrate the orthogonality of row and column category points under independence, we added horizontal and vertical lines as calculated below, using the results returned by `prcomp()`. The initial version of this plot showed that two points, A2 and H2 did not align with the others, so these were excluded from the calculations.

```
> # get the row and column scores
> rscores <- soccer.pca$x[,1:2]
> cscores <- soccer.pca$rotation[,1:2]
> # means, excluding A2 and H2
> rmean <- colMeans(rscores[-3,])[2]
> cmean <- colMeans(cscores[-3,])[1]
>
> abline(h=rmean, col="blue", lwd=2)
> abline(v=cmean, col="red", lwd=2)
> abline(h=0, lty=3, col="gray")
> abline(v=0, lty=3, col="gray")
```

You can see that all the A points (except for A2) and all the H points (except for H2) lie along straight lines, and these lines are indeed at right angles, signifying independence. The fact that these straight lines are parallel to the coordinate axes is incidental, and unrelated to the independence interpretation.

△

6.6 Chapter summary

{sec:ca-summary}

- Correspondence analysis is an exploratory technique, designed to show the row and column categories in a two- (or three-) dimensional space. These graphical displays, and various extensions, provide ways to interpret the patterns of association and explore visually the adequacy of certain loglinear models.

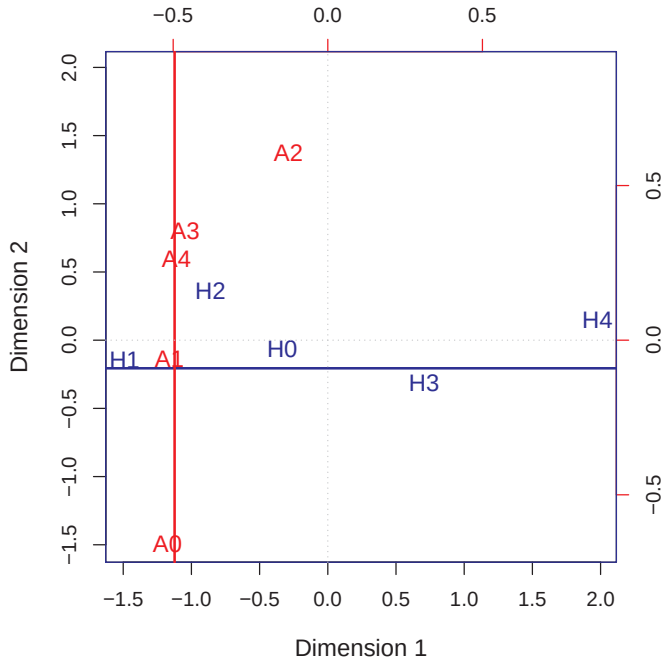


Figure 6.16: Biplot for the biadditive representation of independence for the UK Soccer scores. The row and column categories are independent in this plot when they appear as points on approximately orthogonal lines.

{fig:biplot-soccer-plot

- The scores assigned to the categories of each variable are optimal in several equivalent ways. Among other properties, they maximize the (canonical) correlations between the quantified variables (weighted by cell frequencies), and make the regressions of each variable on the other most nearly linear, for each CA dimension.

- Multi-way tables may be analyzed in several ways. In the "stacking" approach, two or more variables may be combined interactively in the rows and/or columns of an *n*-way table. Simple CA of the restructured table reveals associations between the row and column categories of the restructured table, but hides associations between the variables combined interactively. Each way of stacking corresponds to a particular loglinear model for the full table.

- Multiple correspondence analysis is a generalization of CA to two or more variables based on representing the data as an indicator matrix, or the Burt matrix. The usual MCA provides an analysis of the joint, bivariate relations between all pairs of variables.

- The biplot is a related technique for visualizing the elements of a data array by points or vectors in a joint display of their row and column categories. A standard CA biplot represents the contributions to lack of independence as the projection of the points for rows (or columns) on vectors for the other categories.

- Another application of the biplot to contingency table data is described, based on analysis of log frequency. This analysis also serves to diagnose patterns of independence and partial independence in two-way and larger tables.

6.7 Lab exercises

{lab:6.1}
{sec:ca-lab}

Exercise 6.1 The *JobSat* data in vcdExtra (Friendly, 2015) gives a 4×4 table recording job satisfaction in relation to income.

(a) Carry out a simple correspondence analysis on this table. How much of the inertia is accounted for by a one-dimensional solution? How much by a two-dimensional solution?
(b) Plot the 2D CA solution. To what extent can you consider the association between job satisfaction and income "explained" by the ordinal nature of these variables?

{lab:6.2}

Exercise 6.2 Refer to Exercise 1 in Chapter 5. Carry out a simple correspondence analysis on the 4×5 table *criminal* from the logmult (Bouchet-Valat, 2015) package.

(a) What percentages of the Pearson χ^2 for association are explained by the various dimensions?
(b) Plot the 2D correspondence analysis solution. Describe the pattern of association between year and age.

{lab:6.3}

Exercise 6.3 The data set *caith* in MASS gives a classic table tabulating hair color and eye color of people in Caithness, Scotland, originally from Fisher (1940).

(a) Carry out a simple correspondence analysis on this table. How many dimensions seem necessary to account for most of the association in the table?
(b) Plot the 2D solution. The interpretation of the first dimension should be obvious; is there any interpretation for the second dimension?

{lab:6.4}

Exercise 6.4 The same data, plus a similar table for Aberdeen, are given as a three-way table as *HairEyePlace* in vcdExtra .

(a) Carry out similar correspondence analysis to the last exercise for the data from Aberdeen. Comment on any differences in the placement of the category points.
(b) Analyze the three-way table, stacked to code hair color and place interactively, i.e., for the loglinear model [Hair Place][Eye] . What does this show?

{lab:6.5}{Gilby}

Exercise 6.5 The data set *Gilby* in vcdExtra gives a classic (but now politically incorrect) 6×4 table of English school boys classified according to their clothing and their teachers rating of "dullness" (lack of intelligence).

(a) Compute and plot a correspondence analysis for this data. Write a brief description and interpretation of these results.
(b) Make an analogous mosaic plot of this table. Interpret this in relation to the correspondence analysis plot.

{lab:6.6}

Exercise 6.6 For the mental health data analyzed in Example 6.2, construct a shaded sieve diagram and mosaic plot. Compare these with the correspondence analysis plot shown in Figure 6.2. What features of the data and the association between SES and mental health status are shown in each?

{lab:6.7}

Exercise 6.7 Simulated data is often useful to help understand the connections between data, analysis methods and associated graphic displays. Section 6.3.1 illustrated interactive coding in R, using a simulated 4-way table of counts of pets, classified by age, color and sex, but with no associations because the counts had a constant Poisson mean, $\lambda = 15$.

(a) Re-do this example, but in the call to `rpois()`, specify a non-negative vector of Poisson means to create some associations among the table factors.

(b) Use CA methods to determine if and how the structure you created in the data appears in the results.

{lab:TV8}

Exercise 6.8 The *TV* data was analyzed using CA in Example 6.4, ignoring the variable `Time`. Carry out analyses of the 3-way table, reducing the number of levels of `Time` to three hourly intervals as shown below.

```
> data("TV", package="vcdExtra")
> # reduce number of levels of Time
> TV.df <- as.data.frame.table(TV)
> levels(TV.df$Time) <- rep(c("8", "9", "10"), c(4, 4, 3))
> TV3 <- xtabs(Freq ~ Day + Time + Network, TV.df)
> structable(Day ~ Network + Time, TV3)

             Day Monday Tuesday Wednesday Thursday Friday
Network Time
ABC     8            536     861       744      735   1119
        9           1401    1205      1022      682    907
        10           910    1044       668      349    711
CBS     8           1167     646       550      680    509
        9            967     959       409      385    544
        10           789     798       324      270    426
NBC     8            858    1090       512     1927    823
        9            946     890       831     1858    590
        10           825     588       869     2101    585
```

(a) Use the stacking approach (Section 6.3) to perform a CA of the table with `Network` and `Time` coded interactively. You can create this using the `as.matrix()` method for a "structable" object.

```
> TV3S <- as.matrix(structable(Day ~ Network + Time, TV3), sep=":")
```

(b) What loglinear model is analyzed by this approach?
(c) Plot the 2D solution. Compare this to the CA plot of the two-way table in Figure 6.4.
(d) Carry out an MCA analysis using `mjca()` of the three-way table `TV3`. Plot the 2D solution, and compare this with both the CA plot and the solution for the stacked three-way table.

{lab:PreSex2}

Exercise 6.9 Refer to the MCA analysis of the *PreSex* data in Example 6.8. Use the stacking approach to analyze the stacked table with the combinations of premarital and extramarital sex in the rows and the combinations of gender and marital status in the columns. As suggested in the exercise above, you can use `as.matrix(structable()` to create the stacked table.

(a) What loglinear model is analyzed by this approach? Which associations are included and which are excluded in this analysis?
(b) Plot the 2D CA solution for this analysis. You might want to draw lines connecting some of the row points or column points to aid in interpretation.
(c) How does this analysis differ from the MCA analysis shown in Figure 6.10?

{lab:Vietnam}

Exercise 6.10 Refer to Exercise 5.9 for a description of the *Vietnam* data set in vcdExtra .

(a) Using the stacking approach, carry out a correspondence analysis corresponding to the loglinear model [R][YS] , which asserts that the response is independent of the combinations of year an sex.
(b) Construct an informative 2D plot of the solution, and interpret in terms of how the response varies with year for males and females.
(c) Use `mjca()` to carry out an MCA on the three-way table. Make a useful plot of the solution and interpret in terms of the relationship of the response to year and sex.

{lab:ca-accident}

Exercise 6.11 Refer to Exercise 5.8 for a description of the *Accident* data set in vcdExtra . The data set is in the form of a frequency data frame, so first convert to table form.

```
> accident.tab <- xtabs(Freq ~ age + result + mode + gender, data=Accident)
```

(a) Use `mjca()` to carry out an MCA on the four-way table `accident.tab`.
(b) Construct an informative 2D plot of the solution, and interpret in terms of how the variable `result` varies in relation to the other factors.

References

Adler, D. and Murdoch, D. (2014). *rgl: 3D visualization device system (OpenGL)*. R package version 0.95.1201.

Benzécri, J.-P. (1977). Sur l'analyse des tableaus binaries associés a une correspondense multiple. *Cahiers de l'Analyse des Données*, 2, 55–71.

Bouchet-Valat, M. (2015). *logmult: Log-Multiplicative Models, Including Association Models*. R package version 0.6.1.

Bradu, D. and Gabriel, K. R. (1978). The biplot as a diagnostic tool for models of two-way tables. *Technometrics*, 20, 47–68.

Burt, C. (1950). The factorial analysis of qualitative data. *British Journal of Statistical Psychology*, 3, 166–185.

Emerson, J. W. (1998). Mosaic displays in S-PLUS: A general implementation and a case study. *Statistical Graphics and Computing Newsletter*, 9(1), 17–23.

Fienberg, S. E. (1980). *The Analysis of Cross-Classified Categorical Data*. Cambridge, MA: MIT Press, 2nd edn.

Fisher, R. A. (1940). The precision of discriminant functions. *Annals of Eugenics*, 10, 422–429.

Friendly, M. (1991). *SAS System for Statistical Graphics*. Cary, NC: SAS Institute, 1st edn.

Friendly, M. (1994). Mosaic displays for multi-way contingency tables. *Journal of the American Statistical Association*, 89, 190–200.

Friendly, M. (1999). Extending mosaic displays: Marginal, conditional, and partial views of categorical data. *Journal of Computational and Graphical Statistics*, 8(3), 373–395.

Friendly, M. (2015). *vcdExtra: vcd Extensions and Additions*. R package version 0.6-7.

Gabriel, K. R. (1971). The biplot graphic display of matrices with application to principal components analysis. *Biometrics*, 58(3), 453–467.

Gabriel, K. R. (1980). Biplot. In N. L. Johnson and S. Kotz, eds., *Encyclopedia of Statistical Sciences*, vol. 1, (pp. 263–271). New York: John Wiley and Sons.

Gabriel, K. R. (1981). Biplot display of multivariate matrices for inspection of data and diagnosis. In V. Barnett, ed., *Interpreting Multivariate Data*, chap. 8, (pp. 147–173). London: John Wiley and Sons.

Gabriel, K. R., Galindo, M. P., and Vincente-Villardón, J. L. (1997). Use of biplots to diagnose independence models in three-way contingency tables. In M. Greenacre and J. Blasius, eds., *Visualization of Categorical Data*, chap. 27, (pp. 391–404). San Diego, CA: Academic Press.

Gabriel, K. R. and Odoroff, C. L. (1990). Biplots in biomedical research. *Statistics in Medicine*, 9, 469–485.

Gifi, A. (1981). *Nonlinear Multivariate Analysis*. The Netherlands: Department of Data Theory, University of Leiden.

Goodman, L. A. (1981). Association models and canonical correlation in the analysis of cross-classifications having ordered categories. *Journal of the American Statistical Association*, 76(374), 320–334.

Goodman, L. A. (1985). The analysis of cross-classified data having ordered and/or unordered categories: Association models, correlation models, and asymmetry models for contingency tables with or without missing entries. *Annals of Statistics*, 13(1), 10–69.

Goodman, L. A. (1986). Some useful extensions of the usual correspondence analysis approach and the usual log-linear models approach in the analysis of contingency tables. *International Statistical Review*, 54(3), 243–309. With a discussion and reply by the author.

Gower, J., Lubbe, S., and Roux, N. (2011). *Understanding Biplots*. Wiley.

Gower, J. C. and Hand, D. J. (1996). *Biplots*. London: Chapman & Hall.

Greenacre, M. (1984). *Theory and Applications of Correspondence Analysis*. London: Academic Press.

Greenacre, M. (1989). The Carroll-Green-Schaffer scaling in correspondence analysis: A theoretical and empirical appraisal. *Journal of Marketing Research*, 26, 358–365.

Greenacre, M. (1990). Some limitations of multiple correspondence analysis. *Computational Statistics Quarterly*, 3, 249–256.

Greenacre, M. (1994). Multiple and joint correspondence analysis. In M. J. Greenacre and B. Jörg, eds., *Correspondence Analysis in the Social Sciences*. London: Academic Press.

Greenacre, M. (1997). Diagnostics for joint displays in correspondence analysis. In J. Blasius and M. Greenacre, eds., *Visualization of Categorical Data*, (pp. 221–238). Academic Press.

Greenacre, M. (2007). *Correspondence analysis in practice*. Boca Raton: Chapman & Hall/CRC.

Greenacre, M. (2013). Contribution biplots. *Journal of Computational and Graphical Statistics*, 22(1), 107–122.

Greenacre, M. and Hastie, T. J. (1987). The geometric interpretation of correspondence analysis. *Journal of the American Statistical Association*, 82, 437–447.

Greenacre, M. and Nenadic, O. (2014). *ca: Simple, Multiple and Joint Correspondence Analysis*. R package version 0.58.

Hartigan, J. A. and Kleiner, B. (1984). A mosaic of television ratings. *The American Statistician*, 38, 32–35.

Heuer, J. (1979). *Selbstmord Bei Kinder Und Jugendlichen*. Stuttgard: Ernst Klett Verlag. [Suicide by children and youth.].

Husson, F., Josse, J., Le, S., and Mazet, J. (2015). *FactoMineR: Multivariate Exploratory Data Analysis and Data Mining*. R package version 1.29.

le Roux, N. and Lubbe, S. (2013). *UBbipl: Understanding Biplots: Data Sets And Functions*. R package version 3.0.4.

Lebart, L., Morineau, A., and Warwick, K. M. (1984). *Multivariate Descriptive Statistical Analysis: Correspondence Analysis and Related Techniques for Large Matrices*. New York: John Wiley and Sons.

Meyer, D., Zeileis, A., and Hornik, K. (2015). *vcd: Visualizing Categorical Data*. R package version 1.3-3.

Ripley, B. (2015). *MASS: Support Functions and Datasets for Venables and Ripley's MASS*. R package version 7.3-40.

Snee, R. D. (1974). Graphical display of two-way contingency tables. *The American Statistician*, 28, 9–12.

van der Heijden, P. G. M., de Falguerolles, A., and de Leeuw, J. (1989). A combined approach to contingency table analysis using correspondence analysis and log-linear analysis. *Applied Statistics*, 38(2), 249–292.

van der Heijden, P. G. M. and de Leeuw, J. (1985). Correspondence analysis used complementary to loglinear analysis. *Psychometrika*, 50, 429–447.