MONASH University

MONASH BUSINESS SCHOOL

Correspondence Analysis

High Dimensional Data Analysis

Anastasios Panagiotelis

Lecture 6

Motivation

Non-metric data

- So far we have looked at dimension reduction methods such as PCA and MDS where:
  - The number of variables is large
  - The data are (mostly) metric data
- Today we cover tools for understanding the relationships between nominal/categorical data.
- We focus on the case where there are only two variables, but a potentially large number of categories for each variable.

Outline

- First we revise the cross tabulation, a useful summary for nominal data.
- We then cover ways to visualise the information in a cross tab.
- Ultimately we will discuss Correspondence Analysis which can be applied to large tables.
- We cover applications of Correspondence Analysis in the real world including with text data.

A basic analysis

Beer example

- A cross tab can be created in R using the `table` function. The input is either
  - A single matrix or data frame with 2 columns/variables
  - One vector for each variable
- The output is a *table* object.
- Let's try it with the Beer data which can be found on Moodle.

Beer example

- We look at two categorical variables
  - Availabilty
  - Light
- The number of categories for availability is 2 (National/Regional)
- The number of categories for light is 2 (Light/non-light).

Doing it in R

```
load('Beer.RData')
Beer %>%
  select(light,avail)%>%
  table%>% #Creates Tables
  addmargins()-> #Includes totals
  crosstab
```

The table

```
print(crosstab)
```

```
##               avail
## light          National Regional Sum
##    NONLIGHT           7       21  28
##    LIGHT              5        2   7
##    Sum               12       23  35
```

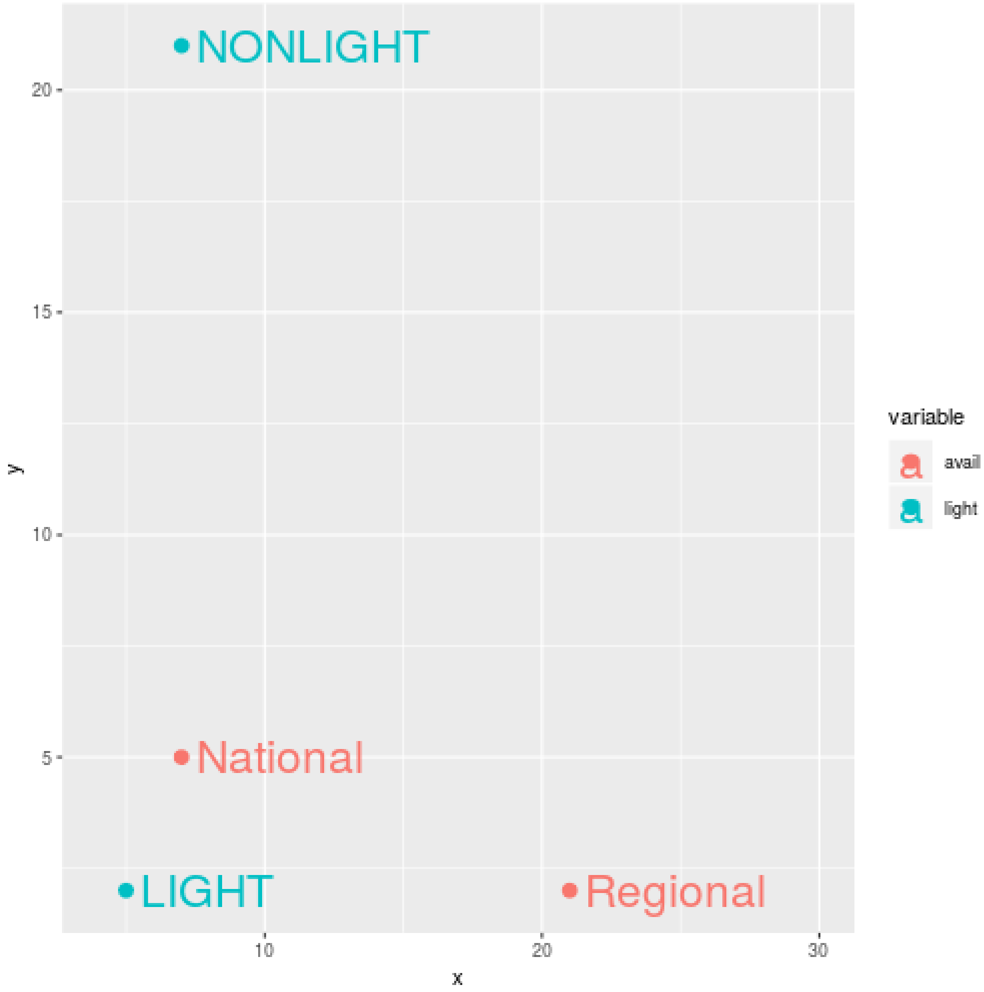
What do we see?

- There are more beers available at a regional level.
- The nationally available beers are just as likely to be light or non-light.
- Regional beers are overwhelmingly non-light.
- But is there a way we can visualise this?

How to visualise?

- In this small example we can think of four sets of coordinates
  - Coordinates for national
  - Coordinates for regional
  - Coordinates for non-light
  - Coordinates for light
- Let's plot these

Plot



NONLIGHT
National
LIGHT
Regional
20
15
10
5
10
20
30
x
y
variable
avail
light

Summary

- Even on this very basic plot we can see an association between light beers and national availability
- However what do we do with
  - Large cross tabulations
  - Non-square cross tabulations
- To solve these issues **Correspondence Analysis** can be used. It is more complicated than simply plotting rows and columns in the cross tab

A Bigger Cross Tab

Breakfast example

- The table on the next slide is reproduced from Bendixen, M., (2003).
  - Different *breakfast* foods (e.g. CER=cereal, MUE=muesli), with a total of 8 categories.
  - Different *attributes* of those foods (‘Healthy’, ‘Economical’, ‘Tasteless’) with a total of 14 categories.
- Survey asked to match attributes to breakfasts.
- A cross tab shows the frequency with which each food was matched to each attribute.

Breakfast example

	BE	CER	FRF	MUE	POR	STF	TT	Y
Economical	3	24	7	3	20	3	16	
Expensive	27	6	9	33	5	18	3	
Family Favourite	31	14	7	4	10	2	5	
Healthy	18	14	31	38	25	28	8	
Long Prepare	35	0	0	0	9	10	1	
Nutritious	25	14	32	28	25	26	7	

Visualising this

We can visualise this using Correspondence Analysis which requires the `ca` package.

```
library(ca)
caout<-ca(breakfastct)
plot(caout)
```

You need to install the `ca` package first

Visualising this

What can we see?

- Towards the top of the plot are categories like *Expensive*, *Healthy* and *Nutritious*. There are associated with *Muesli (MUE)* and *Fresh Fruit(FRF)*.
- The left of the plot has the catgeory *Long Prepare*, with *Bacon and Eggs (BE)* closest to this point.
- *Cereal (CER)* is associated with *Weekdays*.
- What else?

Correspondence Analysis

- The plot is easy to interpret. Categories that are close to one another on the plot have a strong association with one another.
- This is the case when we compare
  - Two categories in the rows of the table,
  - Two categories in the column of the table,
  - A category in the row of the cross tab with a category in the column of a cross tab
- What about the remaining output?

Other output

```
summary(caout,row=FALSE,column=FALSE)
```

```
##
## Principal inertias (eigenvalues):
##
##  dim    value      %   cum%   scree plot
##  1      0.193095  52.5  52.5  ************
##  2      0.077731  21.1  73.6  *****
##  3      0.043854  11.9  85.6  ***
##  4      0.032804   8.9  94.5  **
##  5      0.012257   3.3  97.8  *
##  6      0.005687   1.5  99.4
```

Connection to PCA/MDS

- There are similarities with material covered in PCA and MDS
  - We visualise with a biplot.
  - Terms such as **eigenvalues** and **scree plot** reappear.
- In PCA/MDS the aim was to maximise variance or minimise strain.
- In CA the aim is to maximise **inertia**.

Inertia

- Categorical data are not ordinal.
  - We cannot measure dependence in categorical data by seeing whether 'large' values of one variable coincide with 'large' values of the other variable.
  - We cannot use correlation.
- Inertia is a measure of the dependence in categorical data, closely related to the chi square statistic from a test of independence between two categorical variables.
- Let us quickly revise this.

Chi Square test

- Suppose we have two variables
  - Variable 1 has two categories A and B
  - Variable 2 has two categories X and Y
- Assume Variable 1 and 2 are independent
- On the next slide we will have an incomplete cross tab

Cross Tab

V1 \ V2	X	Y	Total
A			50
B			50
Total	20	80	100

If variable 1 and variable 2 are independent then what numbers do you expect to be in the empty cells?

Cross Tab

V1 \ V2	X	Y	Total
A	10	40	50
B	10	40	50
Total	20	80	100

Under independence

- $\Pr(A, X) = \Pr(A)\Pr(X)$
- $\Pr(B, X) = \Pr(B)\Pr(X)$
- $\Pr(A, Y) = \Pr(A)\Pr(Y)$
- $\Pr(B, Y) = \Pr(B)\Pr(Y)$

Independence is boring

- Independence is not interesting.
- We cannot draw any conclusions about association between categories across different variables.
- If we were to do the crude plot from the beer example, all points would lie in the same direction.
- In correspondence analysis, for perfect independence all row and column categories fall on a single point.

Random variation

- Even for independence, due to randomness we may actually get a table like this:

V1 \ V2	X	Y	Total
A	12	38	50
B	8	42	50
Total	20	80	100

- How do we know whether the variables are truly independent and not due to random variation?

The chi square test

For the chi square test, in each cell we compute

$$\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

where O_i is the observed count in each cell and E_i is the expected count in each cell.

Chi Square Statistic

The chi square statistic is

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

where r and c are the number of rows and columns in the cross tab respectively.

The chi square test

- If the variables are truly independent then it is unlikely that one would observe large values of χ^2
- In this case we reject the null and conclude the variables are dependent.
- However, we can also think of the χ^2 stat as a measure of dependence where:
  - Small values indicate low dependence
  - Large values indicate high dependence

Inertia

- Correspondence analysis is based on a similar idea.
- However the counts in each cell O_i and E_i are replaced with probabilities $o_{ij} = \frac{O_{ij}}{n}$ and $e_{ij} = E_{ij}/n$.
- Each count is dividided by n which is the total of all cell counts (i.e. $r \times c$).
- Instead of the χ^2 we get inertia defined as

$$\text{Inertia} = \frac{\chi^2}{n}$$

Correspondence Analysis

- Correspondence analysis is about explaining as much inertia as possible with a small number of dimensions.
- Instead of the original rows and columns in the cross tab, a small number of linear combinations of these rows and columns are formed.
- A good approximation to the original cross tab could be be reconstructed from these linear combinations.

Geometric Interpretation

- Each column category can be plotted in r-dimensions.
- Each row category can be plotted in c-dimensions.
- Correspondence Analysis rotates both of these to provide the most interesting 'optimal' 2D visualisation
- Here 'optimal' refers to maximising inertia.

Back to the output

```r
summary(caout,row=FALSE,column=FALSE)
```

```
##
## Principal inertias (eigenvalues):
##
##  dim    value      %   cum%   scree plot
##  1      0.193095  52.5  52.5  *************
##  2      0.077731  21.1  73.6  *****
##  3      0.043854  11.9  85.6  ***
##  4      0.032804   8.9  94.5  **
##  5      0.012257   3.3  97.8  *
##  6      0.005687   1.5  99.4
```

How to intepret this

- Eigenvalues previously told us:
  - The variance explained by each principal component in PCA.
  - Give some indication of the Goodness of fit for MDS.
- In CA the eigenvalues tell us the proportion of inertia explained by the solution.
- A 2D solution is usually used for visualisation.
- In the breakfast example the visualisation explains 73.6% of the inertia.

Example: Hotel Reviews

- For an interesting example related to marketing consider hotel reviews.
- Many websites provide user reviews of hotels.
- The words in each review can be scraped from the web using a number of software packages
- In the following example eight hotels in Melbourne were considered
  - Four that were highly rated: Crown Towers, Adelphi, Larwill and QT

Example: Hotel Reviews

- For each hotel, 100 reviews were scraped.
- So called *stop words* ('the', 'a', 'is') were removed as were the names of the hotels.
- The 20 most frequent words used for each hotel.
- Combining these lists for 8 hotels led to 63 words (some words appear on multiple top 20 lists)

Example: Hotel Reviews

- Jaccard similarity could be used to do MDS
- However there are two interesting things that will not be captured by such an analysis
  - The frequency with which words appear is important.
  - The association between the hotels and words.

Example: Hotel Reviews

- On Moodle you will find a cross tab featuring the frequency with which each word appeared on each review

	Adelphi	Citiclub	CrownTowe
amazing	24	1	2
bar	13	4	
bathroom	3	7	
	Adelphi	**Citiclub**	**CrownTowe**

Example: Hotel Reviews

The data can be loaded and correspondence analysis can be carried out using

```
load('hotels.RData')
hoteltable%>%ca%>%plot
```

Example: Hotel Reviews

Conclusions

- Towards the bottom left of the plot are words like *wonderful*, *amazing* and *fantastic*.
  - The more highly rated hotels *Crown Towers*, *QT* and *Adelphi* are closer towards the bottom left
- Towards the top of the plot the words *noise* and *club* appear together with the *Citiclub* hotel
  - This suggests that there may be complaints about noise from a night club.

Conclusions

- Towards the right of the plot the word *old* appears as does *Hotel Sophia* and *Flagstaff*
  - These are lower rated hotels, the age of the hotels may be a problem.
- Can you see anything else?

Example: Hotel Reviews

```
##
## Principal inertias (eigenvalues):
##
##  dim    value      %   cum%   scree plot
##  1      0.199004  27.1  27.1  *******
##  2      0.184450  25.1  52.2  ******
##  3      0.155095  21.1  73.3  *****
##  4      0.075622  10.3  83.6  ***
##  5      0.058206   7.9  91.5  **
##  6      0.038432   5.2  96.7  *
##  7      0.024115   3.3 100.0  *
##         -------- -----
```

Example: Hotel Reviews

Critique of the Analysis

- Together the first two dimensions only explain slightly more than half of the inertia (52.2%)
  - This suggests a large proportion of the dependence is not well explained by the plot
- Counting the frequency of words can be problematic.
  - Consider *clean* v *not clean*.
- Also some aspects of the analysis are quite crude. Why use top 20 words? Why not 10?

Summary

- Main things to know
  - CA used for categorical data.
  - Used to visualise two variable with many categories.
  - Aim is to maximise proportion of explained inertia.
  - Know how can it be used in practice.