More on Correspondence Analysis

(not discussed in the seminar)

Other Issues

Example – CA of Proximity Data

It may appear more of a novelty application, but correspondence analysis can be used to graphically depict proximity, or distance, data.

Suppose we have a square matrix of distances, N. These distances could be as the "bird fly's" (Euclidean) or some other distance measure between two towns, cities, locations, etc. Generally, for **classical CA**, the problem is that **large** cell frequencies will lead to a **small** distance between its row/column.

For **proximity data**, **large** distances between two towns, cities, locations, etc need to be reflected in the plot so there is a large **distance** between their in the plot. So Weller and Romney (1990, pg 75 – 76) consider that

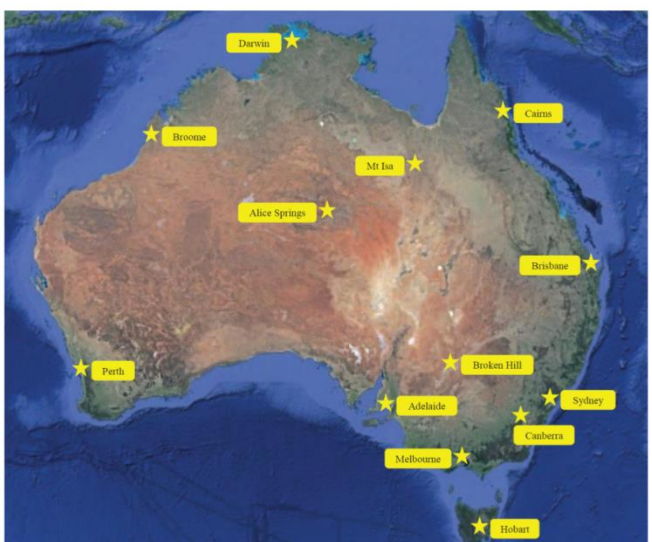
$$N_{prox} = N - \max(N) + 1$$

and the classical CA is performed in N_{prox}.

Other Issues

Distance, in kilometres, of 13 major capital and metropolitan Australian cities

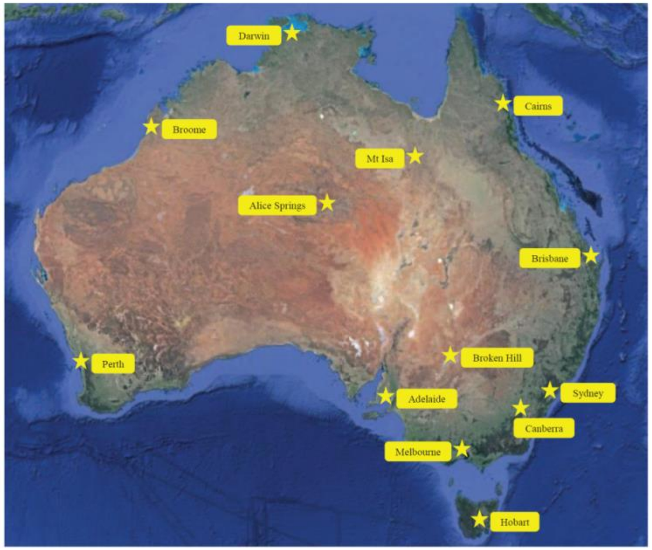
	Sydney	Perth	Adelaide	Melbourne	Hobart	Brisbane	Darwin	Canberra	Alice Springs	Cairns	Broken Hill	Broome	Mt Isa
Sydney	0	3289	1161	713	1057	732	3146	248	2026	1959	934	3374	1860
Perth	3289	0	2130	2721	2006	3604	2651	3087	1990	3439	2407	1681	2655
Adelaide	1161	2130	0	654	1161	1600	2616	958	1328	2124	422	2484	1581
Melbourne	713	2721	654	0	597	1374	3147	465	1890	2323	725	3122	1971
Hobart	1057	3006	1161	597	0	1790	3733	858	2462	2888	1318	3638	2568
Brisbane	732	3604	1600	1374	1790	0	2846	945	1962	1388	1223	3317	1562
Darwin	3146	2651	2616	3147	3733	2846	0	3133	1289	1679	2423	1106	1300
Canberra	248	3087	958	465	858	945	3133	0	1952	2069	801	3275	1873
Alice Springs	2026	1990	1328	1890	2462	1962	1289	1952	0	1449	1181	1366	665
Cairns	1959	3439	2124	2323	2888	1388	1679	2069	1449	0	1727	2496	784
Broken Hill	934	2407	422	725	1318	1233	2423	801	1181	1727	0	2476	1264
Broome	3374	1681	2484	3122	3638	3317	1106	3275	1366	2496	2476	0	1834
Mt Isa	1860	2655	1581	1971	2568	1562	1300	1873	665	784	1264	1834	0

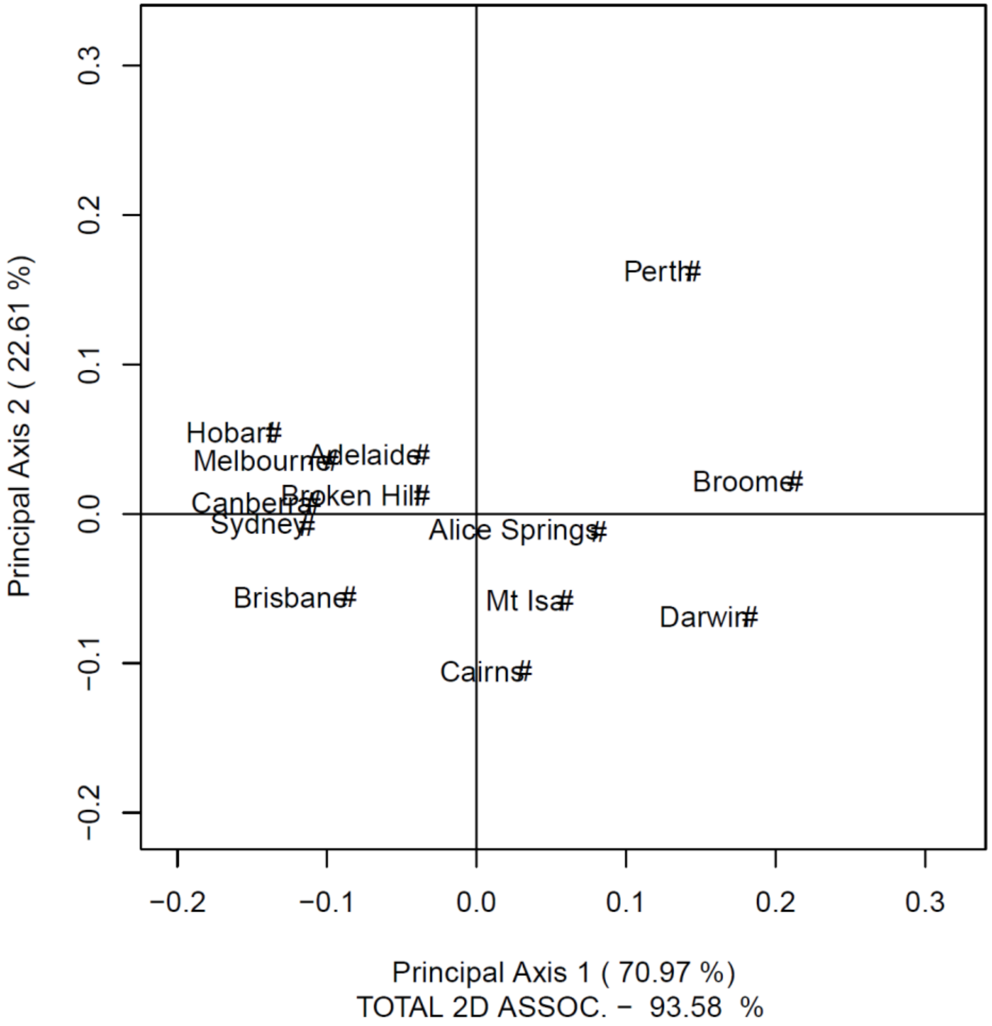


From Beh & Lombardo (2014)

Other Issues

The correspondence plot shows very similar relative positions when compared with their geographical position, except for some rotation.





Confidence Regions

There has been a lot of recent attention given to the inferential aspects in correspondence analysis. In particular, there are two ways in which one can monitor the statistical significance of a row or column point to the association between the variables

- Simple algebraic expressions – Lebart, Morineau & Warwick (1984), Beh (2010), Beh and Lombardo (2014)
- Bootstrap techniques - Ringrose (1992, 2012), Greenacre (2007, pp 196 – 197)

Here we will consider this issue, but focus our attention on the first way.

The algebraic expressions we can consider derive

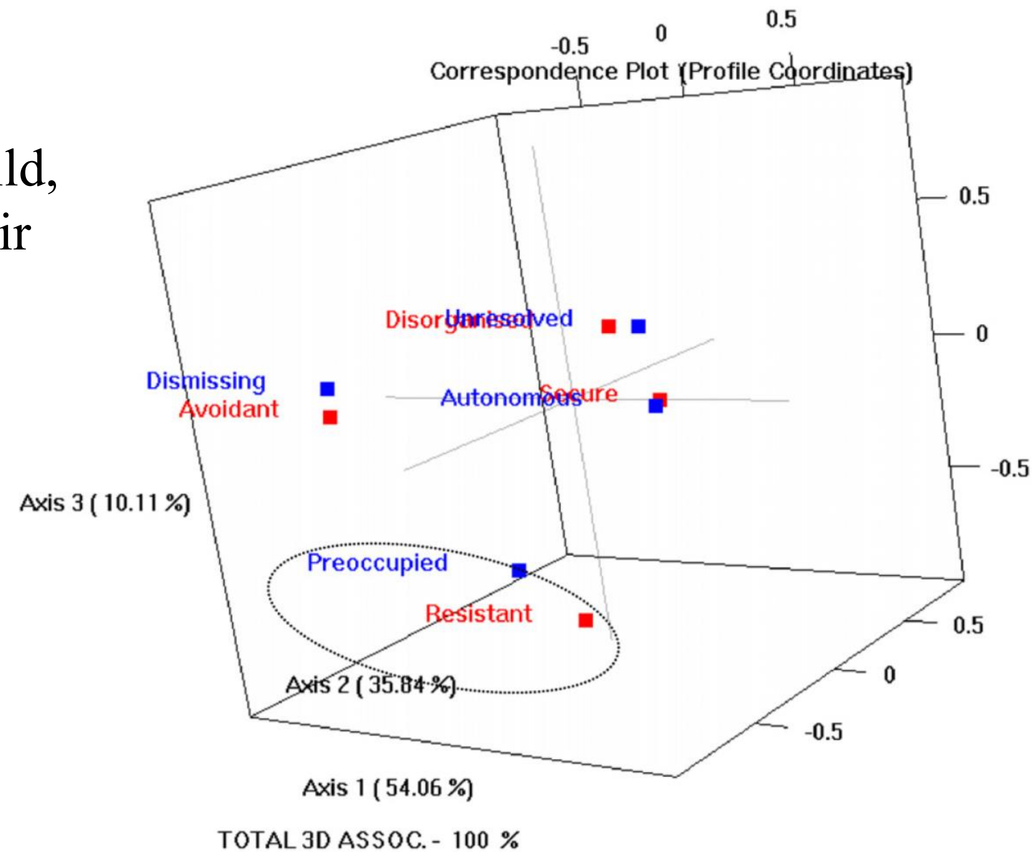
- Confidence circles for each point
- Confidence ellipses for each point
- The p-value of each points contribution to the association between the categorical variables (Beh & Lombardo, 2014)

Confidence Regions

Infant response	Mother's attachment classification			
	Dismissing	Autonomous	Preoccupied	Unresolved
Avoidant	62	29	14	11
Secure	24	210	14	39
Resistant	3	9	10	6
Disorganised	19	26	10	62

Two-way table that looks at a mother's attachment to her child, and the child's response to their mother's level of attachment - van IJzendoorn (1995)

A 2-D plot would miss that the profiles of *Preoccupied* and *Resistant* are very different from the other categories



Confidence Regions

100(1-α)% Confidence Circle

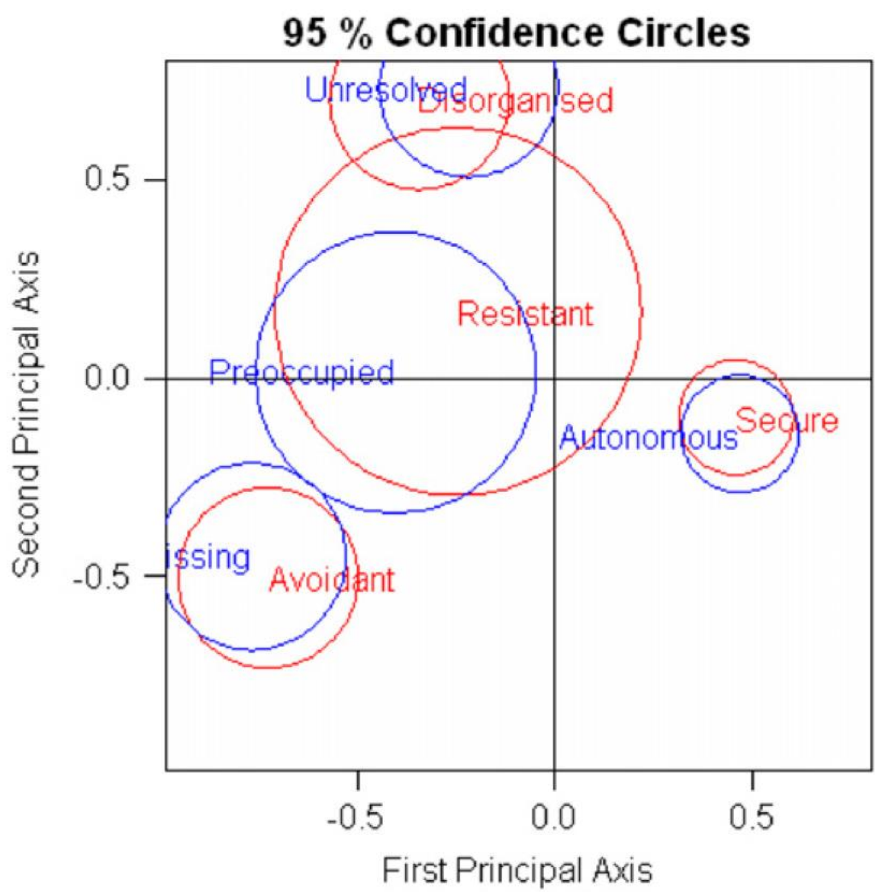
Radius length

$$r_{i(\alpha)} = \sqrt{\frac{\chi_{\alpha}^{2}}{np_{i\bullet}}}$$

Lebart, Morineau & Warwick (1984)

But

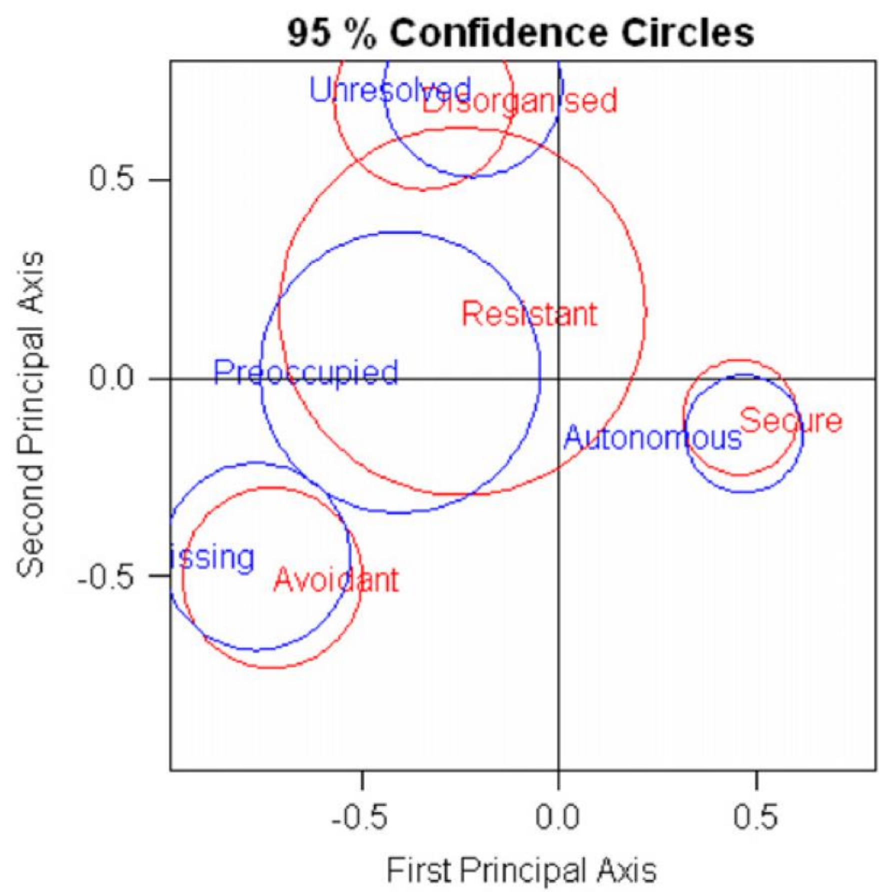
- Assumes each axis is equally weighted
- Ignores configuration in dimensions higher than the third



Confidence Regions

If the region does not overlap the origin that that category is a statistically significant contributor to the association.

However, despite what the three-dimensional correspondence plot suggests, the confidence circles of Lebart *et al*. (1984) show that *Resistant* does not provide a statistically significant contribution to the association. *Preoccupied* is not far behind.



Also, remember that the principal inertia (weight) of each axis is not the same ($\lambda_1^2 = 0.489$, $\lambda_2^2 = 0.089$)

Confidence Regions

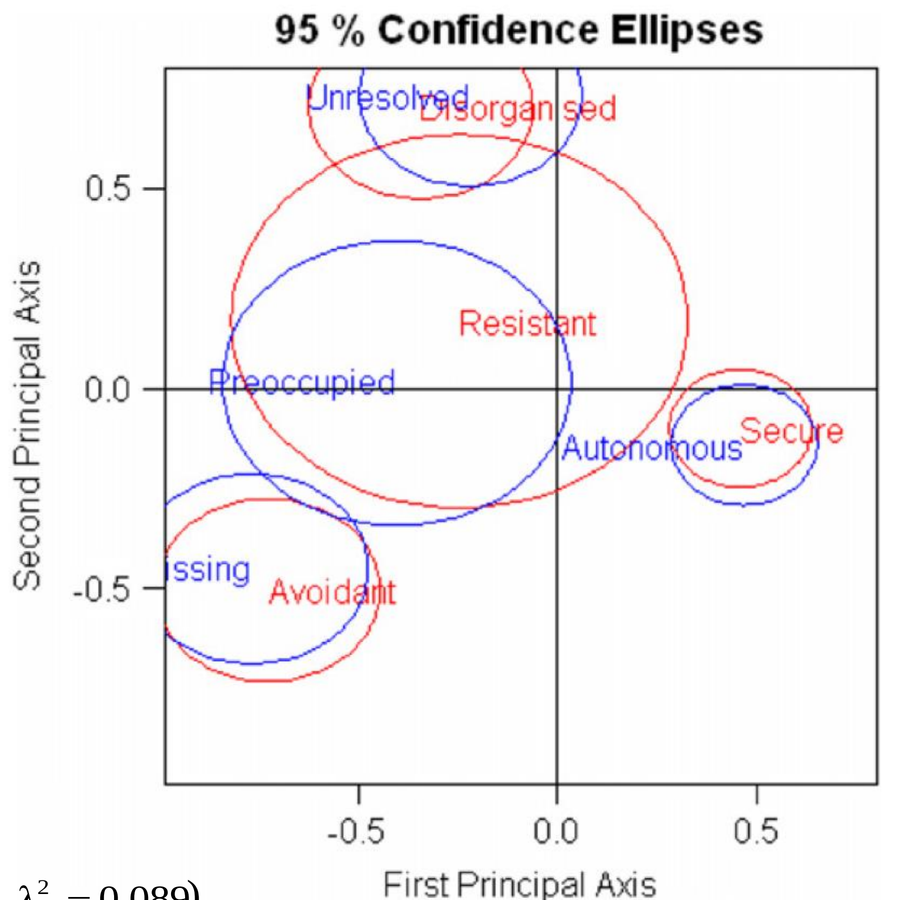
100(1-α)% Confidence Ellipse

Semi major & minor axis length

$$x_i = \lambda_1 \sqrt{\frac{\chi^2_\alpha}{X^2 p_{i\bullet}}}$$

$$y_i = \lambda_2 \sqrt{\frac{\chi^2_\alpha}{X^2 p_{i\bullet}}}$$

Beh (2010)



Takes into account

- Unequally weighted axes ($\lambda_1^2 = 0.489$, $\lambda_2^2 = 0.089$)
- Only when the principal inertia values are the same along each dimension will these regions be circular . . . but . . . what about Resistant and Preoccupied?

Confidence Regions

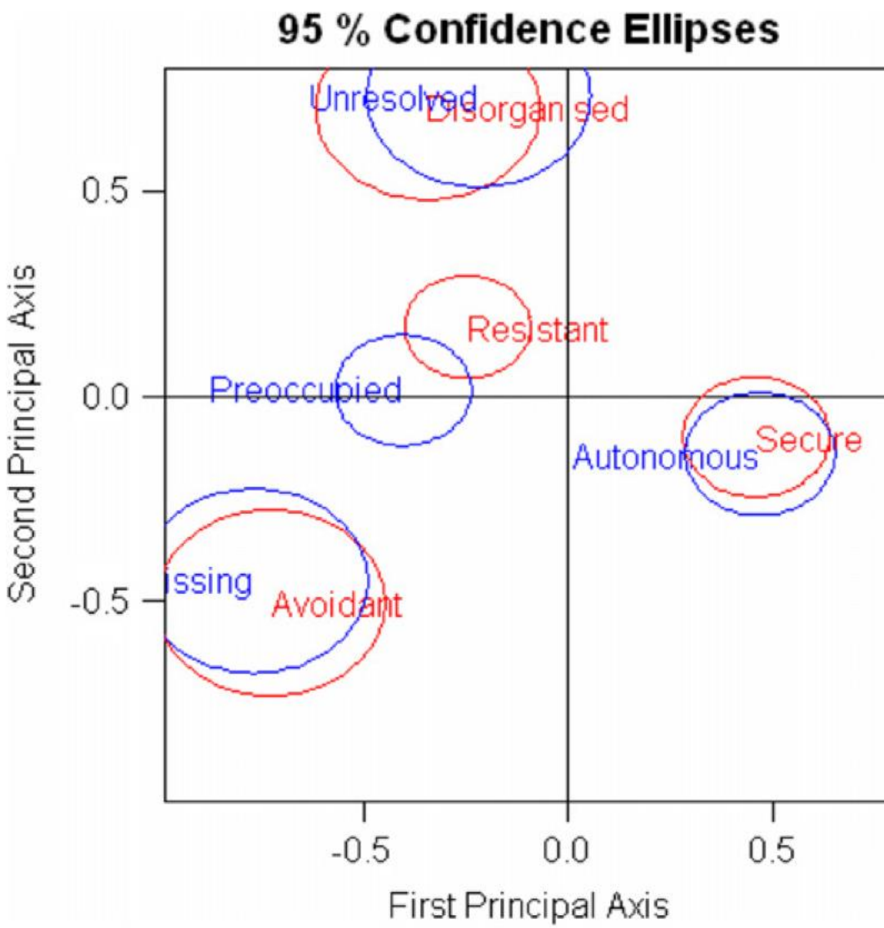
100(1-α)% Confidence Ellipse

Semi major & minor axis length

$$x_i = \lambda_1 \sqrt{\frac{\chi_\alpha^2}{X^2}\left(\frac{1}{p_{i\bullet}} - \sum_{m=3}^{M}\left(\frac{f_{im}}{\lambda_m}\right)^2\right)}$$

$$y_i = \lambda_2 \sqrt{\frac{\chi_\alpha^2}{X^2}\left(\frac{1}{p_{i\bullet}} - \sum_{m=3}^{M}\left(\frac{f_{im}}{\lambda_m}\right)^2\right)}$$

Beh (2010)



Takes into account

- Unequally weighted axes
- Takes into consider the information contained in dimensions higher than the second

Thank you