(Multiple) Correspondence Analysis

Measurement, Scaling, and Dimensional Analysis
2019 ICPSR Summer Program
Prof. Adam M. Enders

What is Correspondence Analysis?

- Very similar to both multidimensional scaling and PCA
- In this case, we have rectangular proximity data, usually in the form of a cross-tabulation table
  - Usually used when data is nominal level, rather than ordinal or metric
- Question: can we represent the information in the rows and columns of cross-tabulation table spatially, in a low-dimension space (such as a 2D plane)?
  - In the simplest case, we have one categorical (nominal, ordinal) row variable, and one categorical column variable
  - Basically, a relationship you'd normally analyze using a cross-tab, rather than a correlation coefficient
- Correspondence Analysis can also be used to do social network analysis (*basically* the same thing)

More Formally...

- CA decomposes the χ^2 statistic associated with the test for independence of rows and columns into orthogonal factors
  - Like with PCA and MDS, two factors/dimensions is best because we can easily plot and visualize the resultant information
  - Results in χ^2 distances between row objects and column objects, similar to the Euclidean distances produced by MDS analyses
- Generally speaking, if the χ^2 test is not statistically significant (row and column variables are not related), then there's no reason to proceed to a CA
  - However, CA can be used if the χ^2 test isn't "appropriate" (i.e., small cell values)

An Example (from Mair 2018)

Band	Fan: Horst	Helga	Klaus
Slayer	9	13	15
Iron Maiden	12	1	4
Metallica	8	6	23
Judas Priest	1	20	18

- Three individuals' preferences for four bands
- Entries are the number of times they saw each band
- Could make individuals the rows...doesn't matter
- This is a cross-tabulation/contingency table – two categorical variables (bands, people) with counts in the cells

An Example (from Mair 2018)

Band	Fan: Horst	Helga	Klaus
Slayer	9	13	15
Iron Maiden	12	1	4
Metallica	8	6	23
Judas Priest	1	20	18

- An obvious first step: Pearson χ^2 test of independence
  - Are these two variables independent from one another?
  - Alternatively, is there any relationship between individuals and preferences (as measured by concert attendance)?
- Test statistic: $\chi^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$
  - o_{ij}: observed frequencies
  - e_{ij}: expected frequencies, under null hypothesis of statistical independence
  - $e_{ij} = \frac{o_{i.} 0_{.j}}{N}$

An Example (from Mair 2018)

Expected:

```
                 Fan
Band                 Horst     Helga     Klaus
  Slayer          8.538462 11.384615 17.076923
  Iron Maiden     3.923077  5.230769  7.846154
  Metallica       8.538462 11.384615 17.076923
  Judas Priest    9.000000 12.000000 18.000000

> fit.chisq <- chisq.test(superfan)
> fit.chisq

Pearson's Chi-squared test

data:  superfan
X-squared = 39.523, df = 6, p-value = 5.653e-07
```

An Example (from Mair 2018)

- Small *p*-values signifies that the two variables are statistically dependent (related)
- χ^2 tests don't really tell us much more – where's the dependence? how much?
- In other words, we now want to move on to deciphering *structure* in the data
- Can visualize the structure via correspondence analysis
- In a sense, doing the same thing as MDS, except we have similarity data rather than dissimilarities
- Want to represent categories in low-dimensional space so we can see how/where categories are related

General Procedure

1. Preprocessing
   - Create matrix of proportions by dividing each cell by *n*
   - Create vectors of row and column "masses" – the within row/column sums
   - Transform proportion matrix into new matrix **S** where columns and rows are weighted by column and row masses
   - (We'll see that **S** is equal to the residuals – difference between observed and expected – from a Person's χ^2 test of cross-tabulation table)
2. Conduct (generalized) SVD on transformed/preprocessed table
   - This step is just as we've done with other techniques
3. Compute "factor scores" just as we would with EFA or PCA
   - Use the scores of variable categories (the individuals rows and columns) along the *m* factors as coordinates

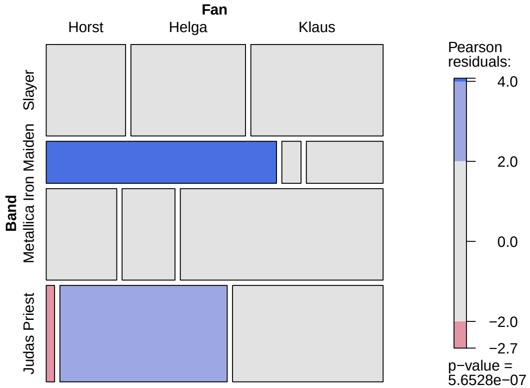
An Example (from Mair 2018)

```
> S <- fit.chisq$residuals
> round(S, 3)
              Fan
Band            Horst  Helga  Klaus
  Slayer        0.158  0.479 -0.503
  Iron Maiden   4.078 -1.850 -1.373
  Metallica    -0.184 -1.596  1.433
  Judas Priest -2.667  2.309  0.000
```

- Retain residuals from the χ^2 test (the function stores these in the object with results)
- Then do a SVD

```
svd.fans <- svd(S)
```

An Example (from Mair 2018)



This mosaic plot of residuals tells us where most of the deviations from independence occur

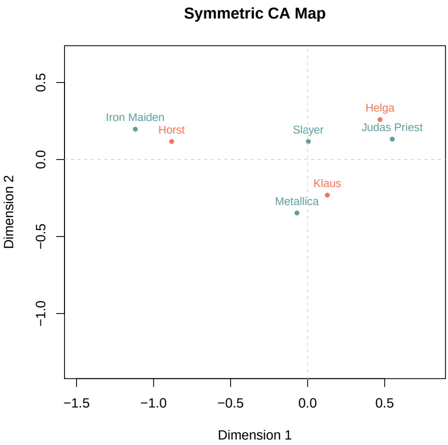
Generating New Coordinates

- As with PCA, we need to weight the right and left singular vectors by the variance explained/captured by each new dimension
- Two different "types" of coordinates:
  1. Standardized: used for most accurately representing either rows or columns in a shared space (kind of...) as points
     - Can't represent both elements in standardized coordinates
     - Usually pick one that we're most interested in
     - Produces an *asymmetric map*
  2. Principal: slightly distorted version of "true" coordinates that can be used to map both row and column elements in the same space
     - Most popular type of standardization/representation
     - Produces an *symmetric map*, the most common (and default in most software)

Generating New Coordinates

- Row standardized: $\mathbf{U^s} = \mathbf{R}^{-.5}\mathbf{U}$
- Row principal: $\mathbf{U^p} = \mathbf{R}^{-.5}\mathbf{UD}$
- Column standardized: $\mathbf{V^s} = \mathbf{C}^{-.5}\mathbf{V}$
- Column principal: $\mathbf{V^p} = \mathbf{C}^{-.5}\mathbf{VD}$
- Where:
  - $\mathbf{R}^{-.5}$: (square root of) row masses (sum of elements across rows)
  - $\mathbf{C}^{-.5}$: (square root of) columns masses (sum of elements down columns)
  - **U**, **V**, and **D**: left (row) and right (column) singular vectors, and singular values
- Plot standardized row and column coordinate and interpret!

An Example (from Mair 2018)



Two-dimensional plot of row and column principal coordinates: a symmetric map (sometimes called a “joint plot”)

Constructing a Plot

- A little trickier than with MDS
  - Can't, strictly speaking, interpret distance between row and column objects
  - χ^2 distance, rather than Euclidean distance ("as the crow flies")
- Option 1: Asymmetric map
  - Either row or columns in standard coordinates, the other in principal coordinates
  - Perfectly preserves distances between row or columns objects, but not between them
  - Biplots are asymmetric maps – usually column vectors "matter" more
  - Oftentimes one set of points is bunched up in the center of the plot, the other more spread out
  - Only useful if you only care about row OR column objects (not usually the case)

Constructing a Plot

- <u>Option 2</u>: Symmetric map ("joint plot")
  - Row and column objects in principal coordinates
  - Allows us to visualize relative distance between row objects and distance between column objects
  - Points more evenly spread out, more aesthetic – this is certainly the most popular type of plot
  - The better two dimensions account for variation, the better the within column/row interpoint distances will approximate the true χ^2 distances (just like a biplot)
  - Still cannot technically compare interpoint distance between row and column objects
    - The column and row points occupy different spaces – we're just projecting them into the same space so we can look at them
    - Imagine having two distinct clouds of points in some high dimensional space and then shining a large light such that the shadows of the points are projected onto the same surface

Model Fit: Variance Explained

- In the language of CA, sometimes variance explained is called "inertia"
- This is a disciplinary difference
  - Psychometricians in the Dutch tradition (much of what we've seen) just talk about variance explained and singular values/eigenvalues
  - The French oftentimes talk about "inertia" (even when it comes to PCA, FA, etc.)
- Regardless, the question is how well we approximate the multidimensional data in a lower dimensional space
- In other words, the goal is the same as previous techniques, as is the general procedure
- Can use a scree plot to assess "appropriate" dimensionality, though 2 is always best (just like MDS, because we can draw a picture)

Strategies for Interpreting Output

- Always begin by simply looking at the cloud of points – both row and column objects
  - Are there clusters of points? Can you make sense of what seems to be going with what?
  - Interesting "directions"? Are some points on one side of the map, others on the other?
- Embedding external information
  - (Roughly) quantitative variables can be used just like with MDS
  - Can construct MCA biplots where external variables are regressed into the configuration
  - Can also embed qualitative information as new points, differentially coloring existing points by certain characteristics, etc.

Multiple Correspondence Analysis

- CA is fairly limited in that units are categories of variables, rather than more "traditional" units of analysis (e.g., people, countries, schools)
    - Because of this, CA most frequently applied in a very descriptive way
    - Not much opportunity to employ results in subsequent analyses – the CA is the analysis of interest
- MCA is an extension of CA for cases when more than two categorical variables are being considered
- Basically, MCA is a nonlinear (categorical) PCA where the input variables are (theoretically) measured at the nominal level
    - Most applications can handle a combination of nominal and ordinal variables
    - We'll see later that several options exist for data of "mixed" measurement level

Goals of MCA

- Understand inter-individual variability with respect to the set of categorical variables
  - Ultimately we want to plot individuals in points in some low dimensional space that allows us to visual inter-individual variability
  - Can understand structure and use row scores in subsequent analyses
- Understand the association between categories of our variables
  - Again, we want to plot these categories in a low dimensional space
  - We also want to increase our level of measurement: nominal $\longrightarrow$ interval
- These goals are very similar to those of PCA, except we're using qualitative variables
- Oftentimes MCA is treated as PCA for nominal variables

General Procedure

1. Re-express data so that each category of each variable is its own column
   - Two methods for doing this: indicator matrix, Burt matrix
   - Both methods accomplish the same goal, but result in different amounts of variance explained
   - Both break apart nominal-level (and ordinal) variables with k categories into k new dichotomous variables
2. Do a singular value decomposition on the re-expressed dataset
3. Use the information from the SVD to construct row and column coordinates along the new dimensions

Seem familiar?

Method: Indicator Matrix

- Re-express data so that columns correspond to each of the categories of each of the original categorical variables
  - Rows are unchanged
- In principle, we never calculate this – software does it for us
- Next, the data is centered, just like in PCA
  - Centering: $x_{ik} = y_{ik}/p_k - 1$
- Finally, we do a SVD and calculate standard and/or principal coordinates just as in simple CA

Method: Indicator Matrix

	A	B	C	D	sex	age	$education$
1	2	3	4	3	2	2	3
2	3	4	2	3	1	3	4
3	2	3	2	4	2	3	2
4	2	2	2	2	1	2	3
5	3	3	3	3	1	5	2
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
871	1	2	2	2	2	3	6

(From Nenadić and Greenacre)

Method: Indicator Matrix

	A					*B*					*C*					*D*				
	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
1	0	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	1	0	0
2	0	0	1	0	0	0	0	0	1	0	0	1	0	0	0	0	0	1	0	0
3	0	1	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	1	0
4	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0
5	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
871	1	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0

(From Nenadić and Greenacre)

Method: Burt Matrix

- Table of all possible two-way cross tabulation tables between pairs of variables
  - Sort of like a correlation matrix for qualitative variables
- Calculate χ^2 residuals just like you would with a two-way contingency table
- Then do eigendecomposition of residual matrix
- Results are substantively identical
  - Eigenvalues from the Burt method are the squared eigenvalues from the complete disjunctive table method
- Again, this is very similar to PCA
  - We saw that we could do PCA on a correlation matrix using an eigenvalue decomposition
  - That's analogous to doing MCA with a Burt table

Method: Burt Matrix

		A						D				
		1	2	3	4	5	...	1	2	3	4	5
A	1	119	0	0	0	0		15	25	17	34	28
	2	0	322	0	0	0		22	102	76	68	54
	3	0	0	204	0	0	...	10	44	68	58	24
	4	0	0	0	178	0		9	52	28	54	35
	5	0	0	0	0	48		4	9	13	12	10
	⋮			⋮			⋱			⋮		
D	1	15	22	10	9	4		60	0	0	0	0
	2	25	102	44	52	9		0	232	0	0	0
	3	17	76	68	28	13	...	0	0	202	0	0
	4	34	68	58	54	12		0	0	0	226	0
	5	28	54	24	35	10		0	0	0	0	151

(From Nenadić and Greenacre)

CA vs. MDS

- Both represent the (dis)similarities between objects as distances in some low dimensional space
- MDS focuses on dissimilarities, CA on similarities
  - This makes sense: contingency tables are designed to assess the strength of relationships
- In MDS, output distances are Euclidean (our standard understanding of distance), whereas CA depicts χ^2 distances
  - Don't need to make too big a deal about this
  - BUT, some of the geometric properties of MDS do not apply to CA
- Ultimately, CA and MDS usually produce similar plots (especially if a two-dimensional solution seems to capture most of the variance in the data)