Coincidence Analysis

Principal components, correspondence, network analysis

M. Escobar M.(modesto@usal.es)

Universidad de Salamanca

10th October 2013

About Stata
Pros and Cons

Stata is a cool statistics software which has been well accepted by the communities of

- Economist (and econometricians)
- Biomedical researchers

Recently, psychologists have also been adopting it, because of the implementation of SEM.
However, social scientists have been less willing to adapt it. To increase the use of Stata by them, it should improve in key areas:

- Increase in table capacities and allow multiple response analysis.
- Offer more user-friendly graph options.
- Implement data-mining including CART, CHAID, and neural networks.
- Introduce social-network analysis

The aims of this presentation are

- To connect social networks with *coincidence analysis*, which is a statistical framework to study concurrence of events in large sets of scenarios.
- To announce an ado program that is able to perform this new statistical framework with Stata.
- We applied previous versions of this analysis to
  - Content analysis of media (Linz Archive)
  - Content analysis of textbooks (Representation of science in scholar texts)
  - Multiresponse analysis in questionnaires (Readers of newspapers)
- I will present today an analysis of characters in picture albums.

- Coincidence analysis is a set of techniques whose object is to detect which characters, subjects, objects, attributes or events tend to appear at the same time in different delimited spaces.
- These delimited spaces are called n scenarios, and are considered as units of analysis (i).
- In each scenario a number of J incidences X_j may occur (1) or may not (0) occur.
- We call incidence matrix ($\mathbf{X}$) an $n \times J$ matrix composed by 0 and 1, according to the incidence or not of every X_j.
- In order to make comparative analysis of coincidences, these scenarios may be classified in H sets

- Two concrete incidences (x_{ij} and x_{ik}) are defined as coincident if they occur in the same scenario:

$$(x_{ij} = 1 \wedge x_{ik} = 1) \Longrightarrow f_{ijk} = 1$$

- From the incidence matrix ($\mathbf{X}$), the coincidence matrix ($\mathbf{F}$) can be obtained by

$$\mathbf{F} = \mathbf{X}'\mathbf{X}$$

- where each element f_{jk} represents the number of scenarios where X_j and X_k are both 1, that is to say, the two characters, subjects, objects, attributes or events coincide
- As may be imagined, there is a special case (f_{jj}), which represents the number of incidences of X_j in the n scenarios.

Independence

Definition

- By definition two incidences (X_j and X_k) are independent in a collection of n scenarios if and only if

$$f_{jk} = \frac{f_{jj} f_{kk}}{n}$$

- where f_{jj} and f_{kk} are the number of scenarios amongst n where each of them can be observed.

Dependence

Definition

- Therefore, in a broad sense two incidences are coincident (or dependent) if they meet the following criterion:

$$f_{jk} > \frac{f_{jj} f_{kk}}{n}$$

- However, as long as it is usual to work with samples of scenarios, the Haberman residuals (r_{jk}), with normal distribution, may be used:

$$r_{jk} = \frac{f_{jk} - \frac{f_{jj} f_{kk}}{n}}{\sqrt{\frac{1 - f_{jj}}{n} \frac{1 - f_{kk}}{n}}}$$

Adjacency
Definition

- Two incidents can be considered adjacent according to the following rule:

$$[p(r_{jk} <= 0) < c] \wedge j \neq k$$

- Therefore, a $J \times J$ matrix **A** may be elaborated with 0 valued diagonal elements and 1 in the case where r_{jk} is significantly below the level c. Other elements should also be 0.
- From **A** the $J \times J$ distance matrix **D**, with geodesics (shortest paths between elements), can be obtained.

Analysis

Five alternatives

- Principal component analysis (Pearson, 1901) with tetrachoric correlations (Everitt 1910).
- Correspondence analysis (Benzecri 1973), using matrix **X** as input and obtaining only column coordinates (incidents).
- Network (Moreno 1934) and coincidence (Escobar 2009) analyses, based on Haberman residuals of **F**.
- Other analysis, based on Haberman residuals, can be used: multidimensional scaling (MDS) and cluster analysis.

Principal components
Definition

- "Principal components analysis (PCA) is a statistical technique used for data reduction. The leading eigenvectors from the eigen decomposition of the correlation or covariance matrix variables describe a series of uncorrelated linear combinations of the variables that contain most of the variance. In addition to data reduction, the eigenvectors from a PCA are often inspected to learn more about the underlying structure of the data"(Stata-MV, 2013: 573)
- In a coincidence analysis it can be used to obtain the factors that structure the appearance of events, subjects or objects.

- The source of PCA is a correlation matrix. However, our material are J X_J dichotomous variables. A way to adequate this nature of nominal variable to correlation measures is using tetrachoric correlations, assuming "a latent bivariate normal distribution (X_j, X_k) for each pair of variables"... this correlation matrix "can be used to perform a factor analysis or a principal component analysis of binary variables using factormat or pcmat commands" (Stata-R, 2013: 2366).
- So, in the coincidence analysis, a tetrachoric correlation matrix must be performed using the $\mathbf{X}$ matrix, and its results (r(Rho)) should be used as input for the principal components analysis.

Correspondence analysis
Definition

- “Correspondence analysis (CA) offers a geometric representation of the rows and columns of a two-way frequency table that is helpful in understanding the similarities between the categories of variables and the association between the variables.” ...“In some respects, CA can be thought of as an analogue to principal components for nominal variables” (Stata-MV, 2013:32)
- With this analysis, similar categories (incidences according to its pattern of concurrence) can be represented together in an euclidean space of $\min((R-1)(C-1))$ dimensions.

- We could use **F** as a contingency table introduced in the CA. However, that is not appropriate as long as the X_j are not independent as the categories of X and Y variables are in this sort of analysis.
- So, instead of **F**, it is possible to introduce the matrix **X**, and consider only column coordinates for the analysis to perform a CA of indicator matrix. (See help mkmat)

Social network analysis

Definition

- Social network analysis (SNA) is “a set of methods for the analysis of social structures, methods that specifically allow an investigation of the relational aspects of these structures"(Scott, 2000: 38).
- Its main mathematical source is topology, especially graph theory.
- A graph G is a collection of points or vertices $x_1, x_2, ..., x_n$ (denoted by the set X), and a collection of lines $a_1, a_2, ..., a_m$ (denoted by the set A) joining all or some of these points. The graph G is then fully described and denoted by the doublet (X, A). (Christofides, 1975)

- This doublet (X, A) can be represented just by a $n \times n$ matrix $\mathbf{A}$ with whose elements a_{ij} represent if the point i is connected to the point j.
- There are, however, two other ways of storing graphs:
  - Adjacency list where n rows are points, and columns are only neighbors points
  - Edge list where rows are m connections with a first column indicating the origin point and a second indicating the target point. (Mihura, 2011:7)

- Ostensibly, Stata doesn't have tools for network analysis.
- Although there are no tools for SNA in Stata, some advanced users have begun to write some routines. I wish to highlight the following works:
  - Corten (2010) wrote a routine to visualize social networks [netplot]
  - Mihura (2012) created routines (SGL) to calculate networks centrality measures, including two Stata commands [netsis and netsummarize]
  - Recently, White (2013) presented a suite of Stata programs for network meta-analyis which includes the network_graphs of Anna Chaimani in the UK users group meeting. And Grund (2013) announced a presentation on plotting and analyzing social networks in the Nordic and Baltic Stata Users Group.

Multidimensional scaling

Definition

- Multidimensional scaling (MDS) is "dimension-reduction and visualization technique. Dissimilarities (for instance, Euclidean distances between observations in a high-dimensional space are represented in a lower-dimensional space (typically two dimensions) so that the Euclidian distance in the lower-dimensional space approximates the dissimilarities in the higher-dimensional space" (Stata-MV, 2013: 450).
- This is the classical method to represent nodes in network analysis (Corten, 2010). Its coordinates are used to position nodes in a bi-dimensional representation.

Cluster analysis
Definition

- Cluster analysis is aimed to determinate the natural groupings of observations (Stata-MV, 2013: 94).
- There are numerous procedures to aggregate cases: single, complete, average, median, Ward, ...
- In the coincidence analysis, clustering could be useful to rank the concurrence of events, using the Haberman residuals (r_{jk}) or the distance matrix ($\mathbf{D}$) as inputs to cluster.

Study of picture collections
Approach

- The aim is to analyze the set of characters in the collection.
- The first step is to quantify the number of pictures of every character.
- However, it is not only important how many times they appear, but also with whom.
- These ideas are based on the interactionist theory of the self in G. H. Mead.
- The pictures are going to be considered as scenarios.
- The characters are going to be considered as incidences (variables). Do they or don't they appear?

Data
Sources

- Turina's Archive consists of 1,438 photographs from the family album, plus over 1,800 other photos stored in folders, and a collection of postcards acquired by Turina himself. The file set adds up to a total of 5,271 graphic materials.
- The photos and Turina's Archive documentation come from the Spanish Library of Contemporary Music and Theatre (Juan March Foundation in Madrid).

Joaquín Turina
Biography

Joaquín Turina Pérez was born in Seville in December 1882, He studied music from an early age and completed his training in Madrid and Paris, where he met artists such as Isaac Albéniz and Manuel de Falla.
He returned to Madrid at the beginning of World War I.
In 1910 he was responsible for the management of the theater Eslava in Madrid and from 1919 he served as the conductor of the Teatro Real.
In 1931 he became Professor of Composition at Conservatory of Madrid and in 1935 was appointed as a member of the Real Academia de Bellas Artes de San Fernando.
He died in 1949 leaving behind musicals like *Fantastic Dances* and *Fancy Clock*.
He also published academic works like *A Treatise on Musical Composition* (1946).

Pictures of Turina

Turina (1882-1949)

Public pictures
Turina

Turina's nuclear family

Turina-Garzón's family

Turina's nuclear family

Examples of codification

The $\mathbf{X}$ matrix is constructed with i rows representing pictures, and the j columns representing people:

$$\mathbf{X} = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 \end{bmatrix}$$

coin

What is it?

- coin is an ado program in its development phase, which is capable of performing coincidence analysis
- Its input is a dataset with scenarios as rows and events as columns.
- Its outputs are:
  - Different matrices (frequencies, percentages, residuals (3), edges and adjacencies)
  - Several network graphs (circle, mds, pca, ca) and dendrograms (single, average, waverage, complete, wards, median, centroid)
  - Measures of centrality
- Its syntax is simple, but flexible. Many options (output, bonferroni, p value, minimum, special event, graph control and options, ...)

Frequencies of coincidences

coin Turina-Valle, f

	Turina	Garzon	Joaquin	Maria	Concha	JoseLuis	Obdulia	Valle
Joaquin Turina	291							
Obdulia Garzón	42	216						
Joaquín	42	71	262					
María	25	62	124	222				
Concha	20	39	68	100	134			
José Luis	18	30	40	60	64	101		
Obdulia	13	27	33	54	60	58	86	
Josefa Valle	2	9	15	10	3	0	0	21

Haberman residuals

coin Turina-Valle, n

	Turina	Garzon	Joaquin	Maria	Concha	JoseLuis	Obdulia	Valle
Joaquin Turina	30.2							
Obdulia Garzón	-4.5	30.2						
Joaquín	-6.5	1.6	30.2					
María	-7.6	1.7	10.3	30.2				
Concha	-4.5	1.6	6.1	14.7	30.2			
José Luis	-3.2	1.5	2.6	8.7	14.7	30.2		
Obdulia	-3.5	1.8	2.1	8.8	15.2	17.5	30.2	
Josefa Valle	-2.2	2.1	4.4	2.5	0	-1.6	-1.5	30.2

List of significant Haberman residuals

coin Turina-Valle, I

Haberman	Edge
17.5	José Luis <-> Obdulia
15.2	Concha <-> Obdulia
14.7	María <-> Concha
14.7	Concha <-> José Luis
10.3	Joaquín <-> María
8.8	María <-> Obdulia
8.7	María <-> José Luis
6.1	Joaquín <-> Concha
4.4	Joaquín <-> Josefa Valle
2.6	Joaquín <-> José Luis
2.5	María <-> Josefa Valle
2.1	Obdulia Garzón <-> Josefa Valle
2.1	Joaquín <-> Obdulia
1.8	Obdulia Garzón <-> Obdulia
1.7	Obdulia Garzón <-> María

Adjacencies matrix

coin Turina-Valle, a

	Turina	Garzon	Joaquin	Maria	Concha	JoseLuis	Obdulia	Valle
Joaquin Turina	0							
Obdulia Garzón	0	0						
Joaquín	0	0	0					
María	0	1	1	0				
Concha	0	0	1	1	0			
José Luis	0	0	1	1	1	0		
Obdulia	0	1	1	1	1	1	0	
Josefa Valle	0	1	1	1	0	0	0	0

Tetrachoric matrix

coin Turina-Valle, t

	Turina	Garzon	Joaquin	Maria	Concha	JoseLuis	Obdulia	Valle
Joaquín Turina	1.00							
Obdulia Garzón	-0.28	1.00						
Joaquín	-0.39	0.10	1.00					
María	-0.49	0.12	0.55	1.00				
Concha	-0.34	0.12	0.38	0.76	1.00			
José Luis	-0.21	0.08	0.14	0.46	0.65	1.00		
Obdulia	-0.25	0.10	0.12	0.47	0.67	0.87	1.00	
Josefa Valle	-0.25	0.17	0.37	0.17	-0.10	-0.71	-0.70	1.00

Distances matrix

coin Turina-Valle, d

	Turina	Garzon	Joaquin	Maria	Concha	JoseLuis	Obdulia	Valle
Joaquin Turina	0							
Obdulia Garzón	0	0						
Joaquín	0	2	0					
María	0	1	1	0				
Concha	0	2	1	1	0			
José Luis	0	2	1	1	1	0		
Obdulia	0	1	1	1	1	1	0	
Josefa Valle	0	1	1	1	2	2	2	0

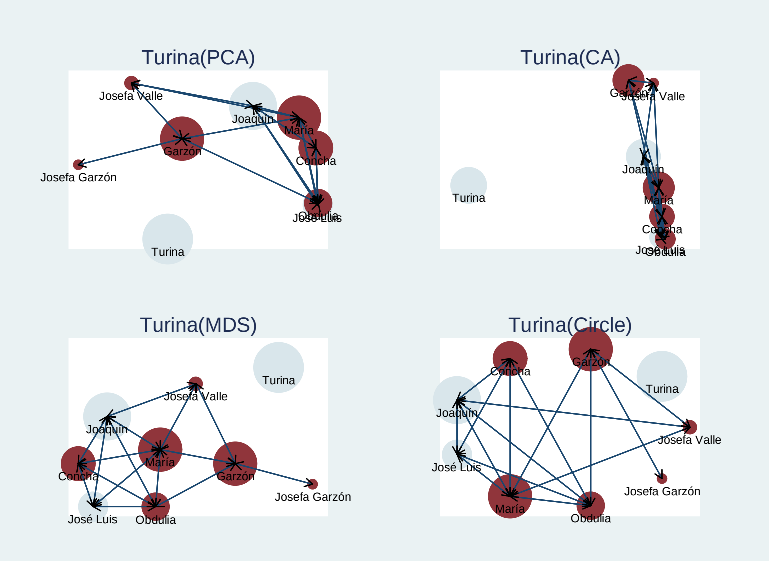
Centrality measures

coin Turina-Valle, c alpha(.10)

	Degree	Closeness	Betweeness	Eigenvector	Power(0.1)	Power(-0.1)
Joaquín Turina	-	-	-	-	-	-
Obdulia Garzón	0.38	0.67	0.02	0.26	1.55	0.80
Joaquín	0.63	0.86	0.09	0.42	1.90	0.65
María	0.75	1.00	0.18	0.47	2.04	0.56
Concha	0.50	0.75	0.00	0.38	1.76	0.74
José Luis	0.50	0.75	0.00	0.38	1.76	0.74
Obdulia	0.63	0.86	0.09	0.42	1.90	0.65
Josefa Valle	0.38	0.67	0.02	0.26	1.55	0.80

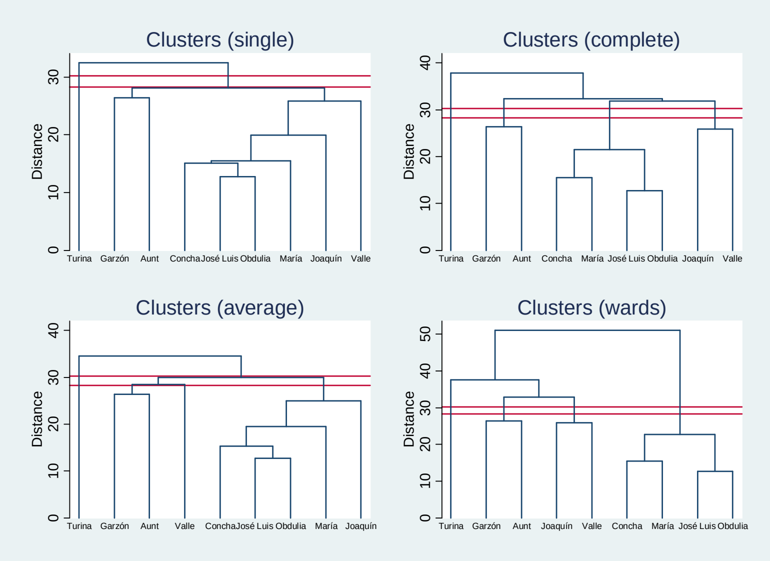
Graph comparisons

Principal components,Correspondence, MDS and circle.



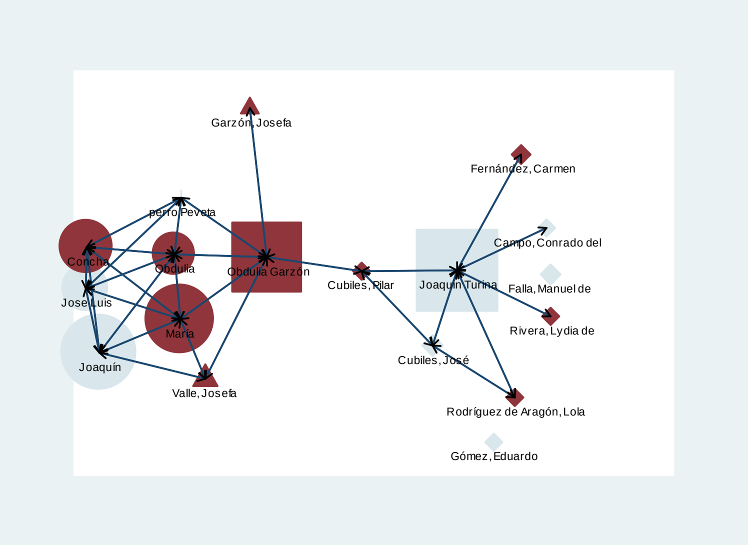
Graph comparisons

Single, complete, average, Ward dendrograms



Characters in Turina's Album

Family and colleagues



Remarks

About coincidence analysis

I've proposed a manner of analyzing coincidences using the existing statistical tools.

I think that the novelty of coincidence analysis is putting together different techniques in order to represent reality.

This may also be useful in comparing different kinds of analysis with dichotomous variables.

We think that the above approach could be extensively used with the aid of the coin program.

Final slide

Thankfulness

Thanks for your attention and comments!