# An Application of Correspondence Analysis to Social Psychology

Mercy Munemo nee Marimo Dzikiti, email: mercy.marimo@yahoo.com

School of Statistics and Actuarial Science
University of the Witwatersrand, Johannesburg, South Africa

## Abstract

Correspondence Analysis (CA) is a versatile multivariate exploratory data analysis technique. It is widely used across various research domains. In this study, CA is applied to Social Psychology using the Spanish women data set of 2002. We investigate the issue of missing data in CA by creating an additional missingness category for each variable. We further formulate data with missingness due to Missing Completely at Random (MCAR) and Missing Not at Random (MNAR) and compare with complete cases. We also compute Factor Analysis to identify latent variables. The results show that, the missing categories of all variables are isolated in the second quadrant of the correspondence map and the rest of the categories remain concealed and clustered around the centroid. Correspondence map orientates ordinal categories in their natural order. Results obtained from Factor Analysis were in complete agreement with results obtained from CA. We conclude that CA maps can be distorted by missing data and CA is not sensitive to MCAR and MNAR mechanisms. CA can be used to further explain latent variables in Factor Analysis. There are some challenges associated with CA technique, therefore, our recommendations relate to further researches to improve CA.

**Key Words**: column profile, inertia, mass, row profile, primitive matrix

## 1. A brief Introduction to Correspondence Analysis

Correspondence Analysis is an exploratory multivariate data analysis technique. It is a valuable tool in the interpretation of categorical variables given in contingency tables. CA transforms a table of numerical information into dual displays called correspondence maps. A Correspondence map is a symmetric plot of rows and columns on the same space. In symmetric maps, both rows and columns are in principal coordinates. Distances between row (column) profiles are not exactly but they are estimated. Both rows and columns are presented as profiles. CA is used to analyze research questions across many domains of study (Greenacre, 2007). These include geology, education, marketing, medicine, sociology and psychology.

## 2. Background of Study

A working woman is a female who is regularly occupied in gainful activities usually outside the home. (Clear, 2003) gave some insight into the background of working women. Traditionally, women were considered intellectually and physically inferior to men, hence, there existed distinct and strict sex roles. A woman's existence was limited to the interior life of the home. Over the past 150 years, women around the world have successfully organized political movements to be legally accorded equal opportunities in all aspects of life that men have traditionally enjoyed. According to a panel study by Thorton A and Freedman D (1979), there has been incredible achievements women have made towards more unrestricted sex roles.

## 3. Aims and Objectives

This study employs Correspondence Analysis technique to reveal the structure of data relating to attitudes toward working women. We seek to investigate the impact of missing data mechanisms and small cell entries on Correspondence Maps. We explore the relationship between CA and Factor Analysis.

## 4. Research Data

The data set was obtained from the International Social Survey Program (ISSP). It is a multinational program with 48 member states. It conducts social scientific surveys on annual basis. This study explores a data set in the area of Family and Changing Gender Roles III entitled “women”. The survey was conducted in 2002 in Spain involving 2471 respondents. It consists of 8 substantial variables, A to H, outlined in section7 (Factor Analysis), recorded on a 5 point scale, where 1=Strongly Agree, 2=Somewhat Agree, 3=Neither Agree nor Disagree, 4=Somewhat Agree, 5=Strongly Disagree. Four demographic variables, gender (g), age (a), marital status (m) and education (e) were included in the survey. We have some limitations in this study. A lot has happened between 2002 and now. Changes have not been incorporated in this study. Spain alone may not necessarily represent the whole world in terms of attitudes towards working women, cultures and beliefs globally, are too diverse to be represented by a single nation.

## 5. Response Pattern

| **Substantial Variables** | | | | | | | |
|---|---|---|---|---|---|---|---|
| Variable | 1 | 2 | 3 | 4 | 5 | 6 | Missing |
| A | 435 | 1097 | 106 | 681 | 104 | - | 48 |
| B | 142 | 1111 | 265 | 748 | 135 | - | 70 |
| C | 189 | 1142 | 273 | 698 | 106 | - | 63 |
| D | 102 | 882 | 333 | 800 | 230 | - | 124 |
| E | 144 | 847 | 272 | 840 | 254 | - | 114 |
| F | 567 | 1371 | 172 | 271 | 27 | - | 63 |
| G | 111 | 490 | 217 | 1034 | 583 | - | 36 |
| H | 1244 | 1074 | 59 | 34 | 2 | - | 58 |
| **Supplementary Variables** | | | | | | | |
| gender (g) | 1192 | 1279 | - | - | - | - | - |
| m. status (m) | 1380 | 212 | 50 | 74 | 751 | - | 4 |
| education (e) | 304 | 617 | 664 | 496 | 186 | 196 | 8 |
| age (a) | 384 | 507 | 432 | 374 | 296 | 478 | - |

Table 1: Responses

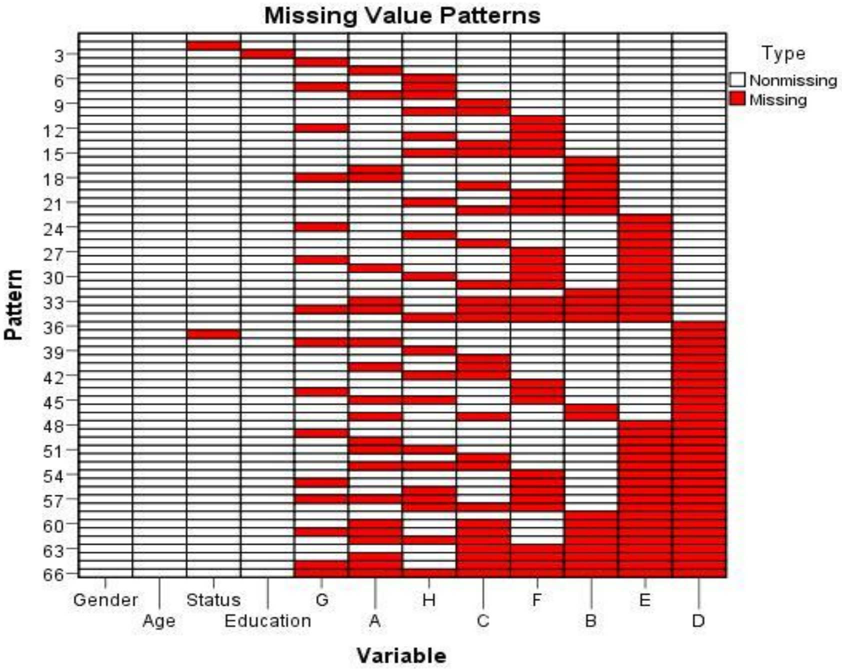

Figure 1: Missingness Pattern

**6. Missing Values Pattern**

Figure 1 exhibits 66 different missingness patterns inherent in women data set. Pattern 1 is the condition of complete cases, pattern 2 is the scenario whereby the respondents missed marital status and so on. The final, 66th pattern is where only demographics were collected and therefore the responses are invalid for CA. Variables D and E suggest a woman's place is in the home and both have the most nonresponses. The patterns are graphical representation of missing data given in Table 1.

**7. Factor Analysis Results**

In Factor Analysis, we explore the “women” data set and investigate which variables go together and it helps us identify unobservable constructs contained in the data.

| | | **Components** | | | |
|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 |
| A | A working mother can establish a warm relationship | -.745 | | | |
| B | A pre-school child suffers if mother works | .771 | | | |
| C | When a woman works, the family life suffers | .779 | | | |
| D | What women really want is a home and kids | | .728 | | |
| E | Running a household is just as satisfying | | .802 | | |
| F | Work is best for a woman's independence | | | | .955 |
| G | A man's job is to work; a woman's job is the home | | .664 | | |
| H | Working women should get paid maternity leave | | | .986 | |

Table 2: Rotated Component Matrix

**8. Factor 1 - Instability in a home.**

High loadings appear on variables B, C, and A(negative). These relate to discomfort a family endures if a mother works. An unemployed woman stabilizes the home.

**9. Factor 2 - Women belong in the interior of the home.**

High loadings appear on variables D, E, and G. These variables suggest that women should be restricted to the interior life of the home. Working life is strictly for men.

**10. Factors 3 and 4**

Components 3 and 4 loaded one variable each. Both relate to positive attitudes about women working. Women are able to stand on their own. They are able to handle both work and family concurrently. Women should be treated fairly at work.

**11. Formulation of a Nonresponse Category and Missingness Mechanisms**

We create a new category 98 which captures all non responses from the original data. We run Correspondence Analysis on complete cases and compare with original data. On the full data set, we sample and retain 1782 cases using simple random sampling. Fifteen percent of cases are discarded. We call this procedure Missing Completely At Random (MCAR). Secondly, we chose a category from the full data set with 325 cases which is age group category 4, 46 - 55 years and completely delete the category. That leaves us with 1782 cases. We call this process Missing Not At Random (MNAR). We compare MCAR and MNAR with complete cases.

**12. Correspondence Analysis Results**

We compare CA output of the original data with complete cases and interpret how CA responds to missing data. Total inertia is 20,75 percent higher in original data, compared to complete cases (numerical output not shown here). The scree plot (not shown here) elbows at 5 dimensions in original data and at 4 dimensions in complete cases. The total inertia is overestimated in data containing missing values.

## 13. Correspondence Map of Original Data

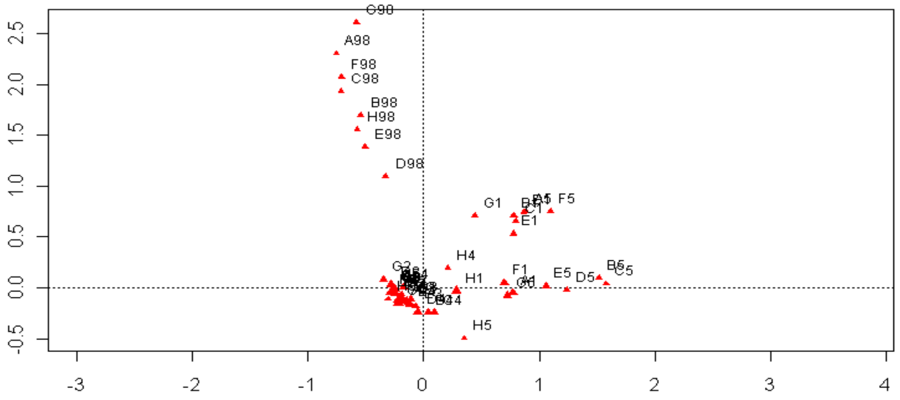

Figure 2: Correspondence Map of Original Data

Figure 2 gives a 2 dimensional symmetric plot of the original data. The 98 is the missing value category. The horizontal line represents, the first dimension and the vertical the second. Dimension 1 separates out extreme categories, 1 and 5 to the right and the even categories as well as the missing appear on the left. Dimension 2 describes the up and down spread. H5 is isolated on the bottom and nonresponses separate out diagonally at the top right corner of the second quadrant. Variable D has the highest number of missing values followed by E and so on with G having the least. The same order is exhibited along the second axis with D98 closer to the centroid and G98 at the furthest. All missing points lie diagonally and isolated in the second quadrant of the plane. There is a cloud of points at the center, which is still concealed owing to the missingness category. The spread of clustered cloud of points should be revealed once we delete the "missing" category.

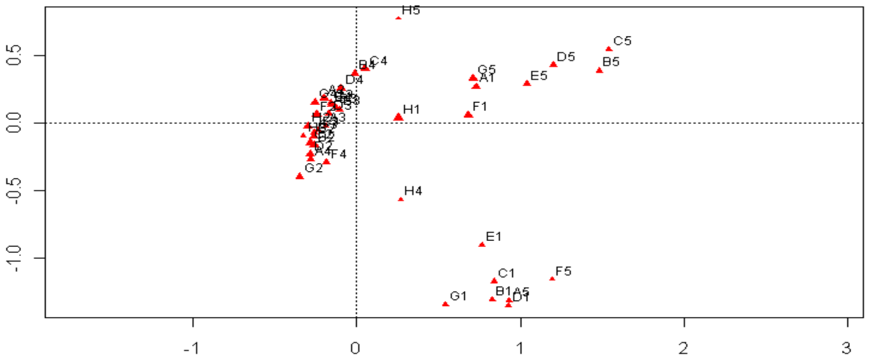

Figure 3: Correspondence Map of Complete Cases

## 14. Correspondence Map of Complete Cases

In Figure 3, we present the correspondence map of complete cases. The first dimension separates extreme categories 1 and 5. Extreme categories generally have smaller masses. They have smaller cell entries in the contingency table. Each variable should have a horseshoe pattern formed by its categories. The point H5 is isolated at the top with an insignificant mass, it does not seem to follow the arch. The small cell causes unnecessary variation. Hence, we collapse H5 into H4. Figure 4 provides the CA map of complete cases, including demographics with collapsed H5. Comparing with Figure 3, before the collapse, there has been a shift in positions of categories from the 1$^{st}$ to the 4$^{th}$ quadrant. For all variables now the horseshoe effect in ordering categories has been achieved.

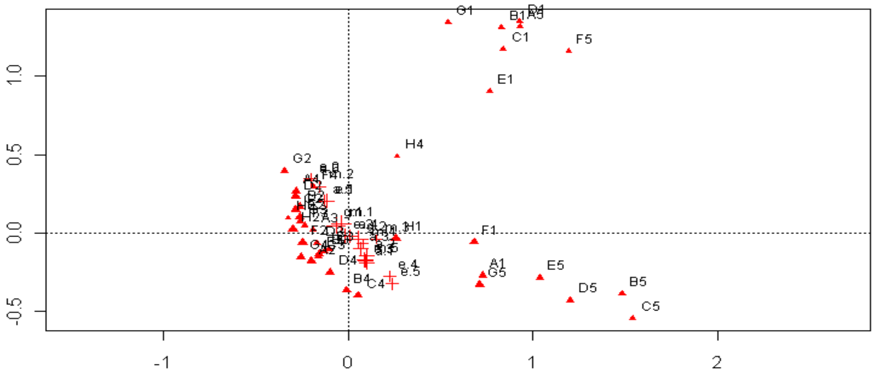

Figure 4: Correspondence Map of all Factors and Demographics – H5 collapsed

**15. Correspondence Analysis versus Factor Analysis**

The orientation of variables in Figure 5 plane conforms to the Factor Analysis results. Variables making up Factor 1, B and C have their “strongly agree” categories lying in the same region as A's “strongly disagree”. The other extreme categories were placed on the fourth quadrant. The same goes for variables making Factor 2 which are D,E and G. Factor 3, variable H has its “strongly agree” category lying along dimension 1 and very close to the centroid. H2, H3 and H4 are not far from the centroid as well. Factor 4 which accommodates only variable F also has a unique orientation of profile points. F1 lies close to the horizontal dimension in the first quadrant and F5 stands out as the last in the fourth quadrant. For further analysis, we shall group the variables in subsets of factors from the Factor Analysis and investigate how different combinations are related to demographics. Demographics (+) are diagonally ordered in the plane.

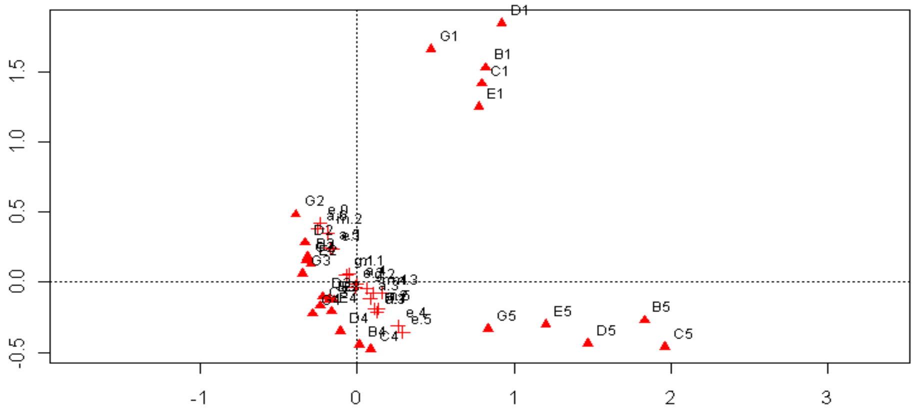

Figure 5: Correspondence Map of Factors 1 and 2

**16. Correspondence Map of Factors 1 and 2**

From the Factor Analysis results, Factors 1 and 2 support traditional notion about women. Figure 5 is a map of Factors 1 and 2 combined. All members follow the same general trend. Education groups e4 and e5 constituting the highly educated citizens strongly disagree with statements about women belonging entirely in the home. On the other hand, education group e0 somewhat agreed to Factors 1 and 2 statements. This implies that the uneducated and the lowly educated people still believe a woman belongs in the interior of the home and should leave work for men. Factors 3 and 4 (map not shown here) are modern approaches to working women. Its elements move in an opposite direction to Factors 1 and 2. The single and those never married, m5 are in support of the modern approach to working women and support independence of women. The middle aged people somewhat agree to statements of Factors 3 and 4. These are also the most educated participants.

## 17. Conclusions

Missing data is not desirable in Correspondence Analysis. Deletion of cases with missing items reduces precision but it allows a better spread of points on a correspondence map. However the missingness mechanism does not influence CA. The total inertia obtained from complete cases versus MCAR and MNAR were compared using the chi-square test. At 5% significance level there is no significant difference between inertia from complete cases and that obtained from MCAR and MNAR. Complete Cases versus MCAR, p-value = 0.684353, Complete Cases versus MNAR, p-value = 0.673517. CA complements factor analysis and it allows for subset analysis if desired. The construction of a correspondence map is highly influenced by the value of cell entries of the original primitive matrix. If the entry is too small, it distorts the orientation of the rest of the cloud. According to this study, collapsing of small cells gives a more precise interpretation of a correspondence map.

## 18. Recommendations

According to this study, there is no difference between MCAR and MNAR impact on Correspondence Analysis. Thus, there is no restriction to a particular imputation method. Since we used deletion in this case, we recommend further investigation into imputation method of data handling and determine how the CA maps differ from the original data set and complete cases. Small cell count distorts geometrical orientation of profile points. One way to deal with such points is to collapse the small count cells as done in this study. It interests to investigate how the maps change if we use deletion instead. We therefore recommend further studies into comparison between collapsing and deletion and find out the best method to deal with small cell counts. Correspondence Analysis complements Factor Analysis and the two methods may be used concurrently to enrich data interpretation.

## References

1. Clear, C. (2003) Hardship, Help, Happiness in Oral History Narratives of Women's Lives in Ireland, 1921-1961 *Oral History Society* vol.31, no.2 pp33-42

2. Greenacre, M. (2007) *Correspondence Analysis in Practice*, 2$^{nd}$ ed., Chapman and Hall/CRC: Taylor and Francis Group, 280 pages.

3. Thornton, A. and Freedman, D. (1979) Changes in the sex role attitude of Women, 1962-1977: Evidence from a panel Study. *American Sociological Association*, vol.44, no. 5, pp.831-842

4. The R Foundation for Statistical Computing (2011), R Version 2.13.2 [2011-09-30]