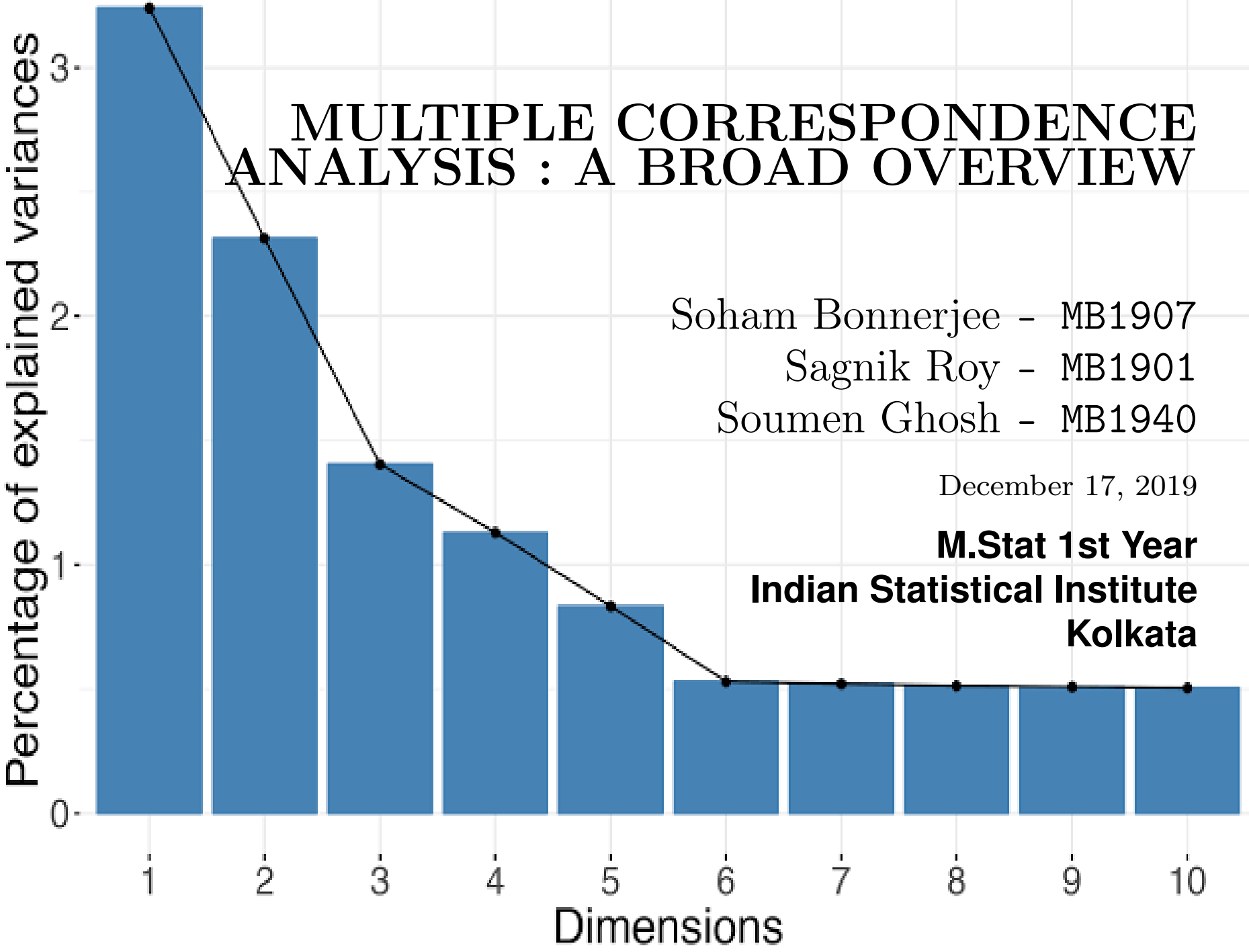


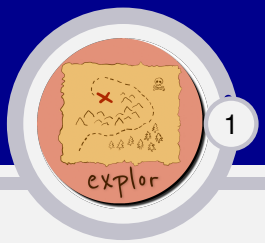
MULTIPLE CORRESPONDENCE ANALYSIS : A BROAD OVERVIEW

Soham Bonnerjee - MB1907
Sagnik Roy - MB1901
Soumen Ghosh - MB1940

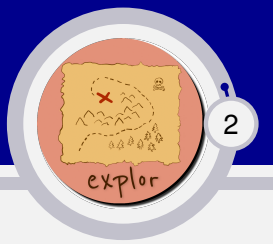
December 17, 2019

M.Stat 1st Year
Indian Statistical Institute
Kolkata

Content

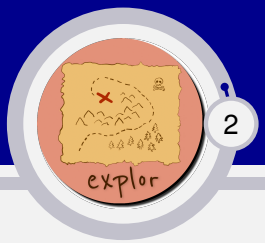


What is MCA



In statistics, Multiple Correspondence Analysis (MCA) is a data analysis technique for nominal categorical data, used to detect and represent underlying structures in a data set. It does this by representing data as points in a low-dimensional Euclidean space.

What is MCA

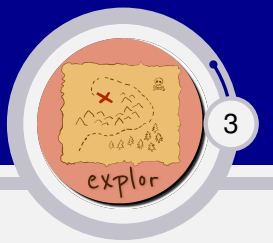


In statistics, Multiple Correspondence Analysis (MCA) is a data analysis technique for nominal categorical data, used to detect and represent underlying structures in a data set. It does this by representing data as points in a low-dimensional Euclidean space.

Goal

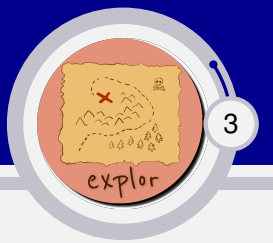
To briefly discuss the motivation and the theories behind MCA, and also explore the different realms of Statistics this technique is applicable in.

A Brief History



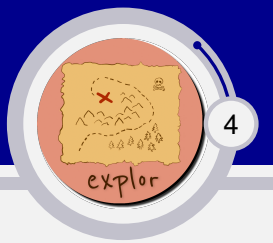
Multiple Correspondence Analysis was developed in France by the mathematician Jean-Paul Benzécri, and popularized by Bourdieu outside France.

A Brief History



Multiple Correspondence Analysis was developed in France by the mathematician Jean-Paul Benzécri, and popularized by Bourdieu outside France. In essence, It is a generalization of Simple Correspondence Analysis, for the latter is primarily applicable to 2-way contingency tables, whereas MCA tackles the more general problem of relation between categories in higher-dimensional tables.

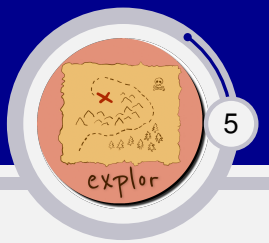
MCA Algorithm



Consider a contingency table with n observation and Q categories.

The i-th category has k_i levels. Define $J = \sum_{i=1}^{Q} k_i$. MCA can be simply performed by applying the CA algorithm to either a *Super-Indicator matrix* (also called complete disjunctive table – CDT) or a *Burt Matrix* formed from these variables.

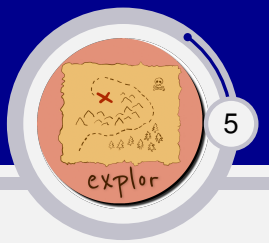
MCA Algorithm
Using Super-Indicator Matrix



- For level i, form a $n \times k_i$ indicator matrix X_i as follows:

  $(X_i)_{pq} = \mathbb{I}\,(p\text{-th observation belongs to } q\text{-th level of the category})$

MCA Algorithm
Using Super-Indicator Matrix



- For level i, form a $n \times k_i$ indicator matrix X_i as follows:

  $(X_i)_{pq} = \mathbb{I}\,(p\text{-th observation belongs to } q\text{-th level of the category})$

- Form the $n \times J$ super indicator matrix X as follows:

$$X = (X_1 : X_2 : \cdots : X_Q)$$

MCA Algorithm
Using Super-Indicator Matrix

- For level i, form a $n \times k_i$ indicator matrix X_i as follows:

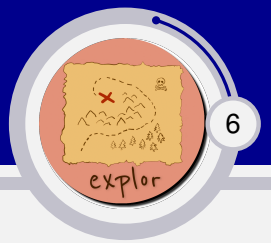
  $(X_i)_{pq} = \mathbb{I}\,(p\text{-th observation belongs to } q\text{-th level of the category})$

- Form the $n \times J$ super indicator matrix X as follows:

  $$X = (X_1 : X_2 : \cdots : X_Q)$$

- Let $N = \mathbf{1}^T X \mathbf{1} = nQ$ be the sum of all the entries in X. Define $Z = \frac{X}{N}$. The row sum of Z is $\mathbf{r^X} = \frac{1}{n}\mathbf{1}$. Let the vector of marginals of category i be $\mathbf{c_i}$. Then the vector of column sums of Z is $\mathbf{c^X} = \frac{1}{N}\left(\mathbf{c_1} : \mathbf{c_2} : \cdots : \mathbf{c_Q}\right)^T$. Also let $\mathbf{D_r^X} = \mathrm{Diag}(\mathbf{r^X}) = \frac{1}{n}\mathbf{I}_{n\times n}$ and $\mathbf{D_c^X} = \mathrm{Diag}(\mathbf{c^X})$

MCA Algorithm
Using Super-Indicator Matrix



- Perform a Singular Value Decomposition of $C = \mathbf{D_r^X}^{-\frac{1}{2}} \left(Z - \mathbf{r^X}(\mathbf{c^X})^T \right) \mathbf{D_c^X}^{-\frac{1}{2}} = \sqrt{n} \left(Z - \mathbf{r^X}(\mathbf{c^X})^T \right) \mathbf{D_c^X}^{-\frac{1}{2}}$:

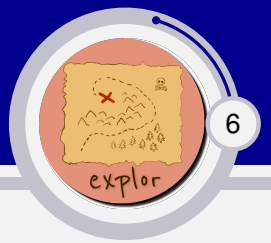
$$C = \Gamma \Lambda \Delta^T$$

.

- The row and respectively column factor scores are obtained by:

$$\mathbf{F} = \mathbf{D_r^X}^{-\frac{1}{2}} \Gamma \Lambda \qquad \mathbf{G} = \mathbf{D_c^X}^{-\frac{1}{2}} \Delta \Lambda$$

MCA Algorithm
Using Super-Indicator Matrix



- Perform a Singular Value Decomposition of $C = \mathbf{D_r^X}^{-\frac{1}{2}} \left(Z - \mathbf{r^X}(\mathbf{c^X})^T \right) \mathbf{D_c^X}^{-\frac{1}{2}} = \sqrt{n} \left(Z - \mathbf{r^X}(\mathbf{c^X})^T \right) \mathbf{D_c^X}^{-\frac{1}{2}}$:

$$C = \Gamma \Lambda \Delta^T$$

.

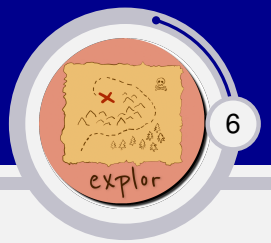
- The row and respectively column factor scores are obtained by:

$$\mathbf{F} = \mathbf{D_r^X}^{-\frac{1}{2}} \Gamma \Lambda \qquad \mathbf{G} = \mathbf{D_c^X}^{-\frac{1}{2}} \Delta \Lambda$$

These are called *Principal Coordinates*. Note that the number of factors would be $s = J - Q$. The *Standard Coordinates* are given by:

$$\mathbf{F_s} = \mathbf{D_r^X}^{-\frac{1}{2}} \Gamma \qquad \mathbf{G_s} = \mathbf{D_c^X}^{-\frac{1}{2}} \Delta$$

MCA Algorithm
Using Super-Indicator Matrix



- Perform a Singular Value Decomposition of $C = \mathbf{D_r^X}^{-\frac{1}{2}} \left(Z - \mathbf{r^X}(\mathbf{c^X})^T \right) \mathbf{D_c^X}^{-\frac{1}{2}} = \sqrt{n} \left(Z - \mathbf{r^X}(\mathbf{c^X})^T \right) \mathbf{D_c^X}^{-\frac{1}{2}}$:

$$C = \Gamma \Lambda \Delta^T$$

.

- The row and respectively column factor scores are obtained by:

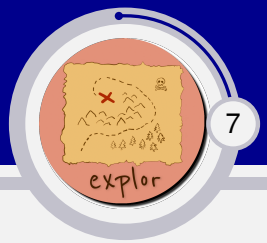
$$\mathbf{F} = \mathbf{D_r^X}^{-\frac{1}{2}} \Gamma \Lambda \qquad \mathbf{G} = \mathbf{D_c^X}^{-\frac{1}{2}} \Delta \Lambda$$

These are called *Principal Coordinates*. Note that the number of factors would be $s = J - Q$. The *Standard Coordinates* are given by:

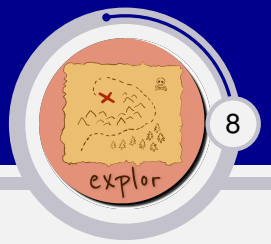
$$\mathbf{F_s} = \mathbf{D_r^X}^{-\frac{1}{2}} \Gamma \qquad \mathbf{G_s} = \mathbf{D_c^X}^{-\frac{1}{2}} \Delta$$

- Let the singular values be $\lambda_1, \cdots, \lambda_s$. Then the squared singular values, i.e the eigen values of CC^T : $\lambda_1^2, \cdots, \lambda_s^2$ are called Principal Inertia of the matrix X.

MCA Algorithm
Using Burt Matrix

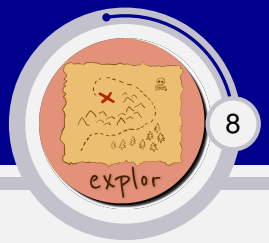


Consider the Burt Matrix $B = X^T X$. Note that this matrix is symmetric, and has the vector of row sums= vector of column sums = $NQ\mathbf{c}^{\mathbf{x}}$. We apply the similar algorithm as above to get row and column factor scores.



Properties

- *Since* B *is symmetric, the solution for the rows and for the columns is identical*.
- *The analysis of* B *only gives a solution for the response categories (that is, what were previously the columns of* **X**.
- *The singular values in the analysis of* **B** *are also eigenvalues.*



Properties

- *Since* B *is symmetric, the solution for the rows and for the columns is identical*.
- *The analysis of* B *only gives a solution for the response categories (that is, what were previously the columns of* **X**.
- *The singular values in the analysis of* **B** *are also eigenvalues.*
- *The standard coordinates of the row (equivalent to columns) of* B*, are identical to the standard coordinates of the columns of* **X**, *and the principal inertias of* **B** *are the squares of those of* **X** .

Relation of MCA with CA

Consider a $j_1 \times j_2$ Bivariate Contingency table $\mathbf{M}$.Form the super-indicator matrix $\mathbf{X} = (\mathbf{X_1} : \mathbf{X_2})$. Note that

$$\mathbf{M} = \mathbf{X_1}^T \mathbf{X_2}$$

We apply both CA on $\mathbf{M}$ and MCA on $\mathbf{X}$.

Relation of MCA with CA

Consider a $j_1 \times j_2$ Bivariate Contingency table $\mathbf{M}$.Form the super-indicator matrix $\mathbf{X} = (\mathbf{X_1} : \mathbf{X_2})$. Note that

$$\mathbf{M} = \mathbf{X_1}^T \mathbf{X_2}$$

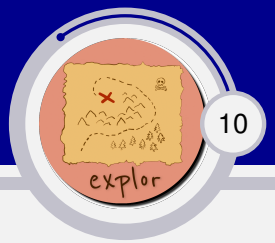
We apply both CA on $\mathbf{M}$ and MCA on $\mathbf{X}$.

Claim

If Λ_M denote the diagonal matrix of singular values for CA on $\mathbf{M}$, and Λ_X and Λ_B denote the corresponding quantities for the MCA using Super-indicator matrix and Burt matrix respectively, then

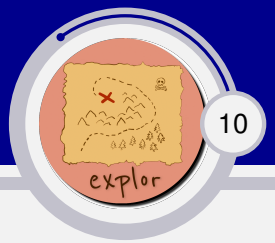
$$\Lambda_M^2 = (2\Lambda_X^2 - \mathbf{I})^2 = (2\Lambda_B - \mathbf{I})^2$$

A Real Life Application



We analyze the `Cars93` data from **MASS** package in R. This dataset is from 93 Cars on Sale in the USA in 1993.

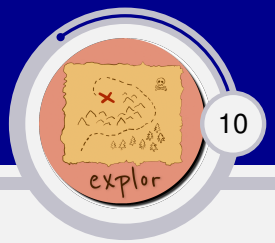
A Real Life Application



We analyze the `Cars93` data from **MASS** package in R. This dataset is from 93 Cars on Sale in the USA in 1993. The data has 4 qualitative variables:

- `Type`: Describes the car-type. One of 6 types: Compact, Large, Midsize, Small, Sporty, Van.
- `AirBags`: Indicates the existence of air-bags. One of 3 types: Driver and Passenger, Driver, None.
- `DriveTrain`: Drive Train type. One of 3 factors: 4WD, Front Wheel, Rear Wheel.
- `Origin`: Country of origin. One of two factors: USA, non-USA.

A Real Life Application



We analyze the `Cars93` data from **MASS** package in R. This dataset is from 93 Cars on Sale in the USA in 1993. The data has 4 qualitative variables:

- `Type`: Describes the car-type. One of 6 types: Compact, Large, Midsize, Small, Sporty, Van.
- `AirBags`: Indicates the existence of air-bags. One of 3 types: Driver and Passenger, Driver, None.
- `DriveTrain`: Drive Train type. One of 3 factors: 4WD, Front Wheel, Rear Wheel.
- `Origin`: Country of origin. One of two factors: USA, non-USA.

We apply MCA on the super-indicator matrix. Note that there will be $14 - 4 = 10$ factors.

A Real Life Application

Biplot

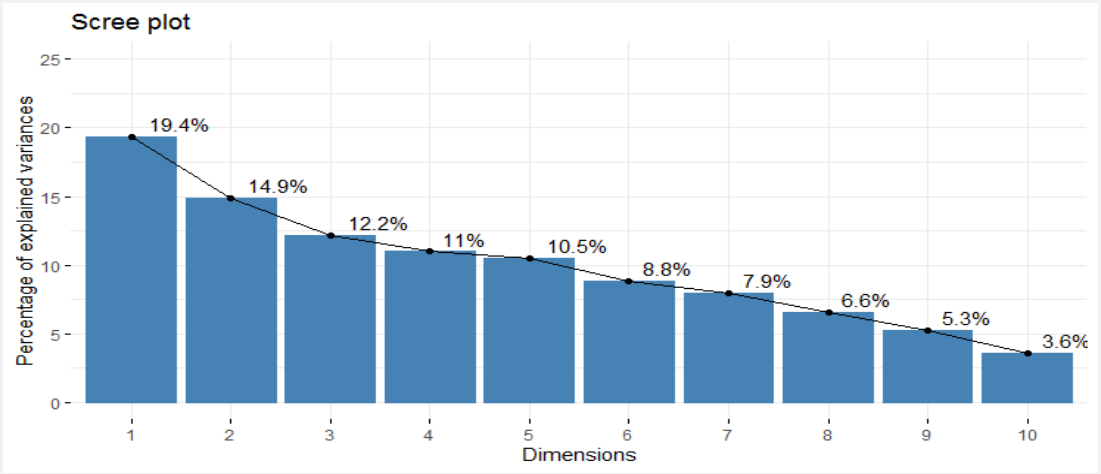


Figure: Variances Explained by Dimensions

A Real Life Application
Biplot

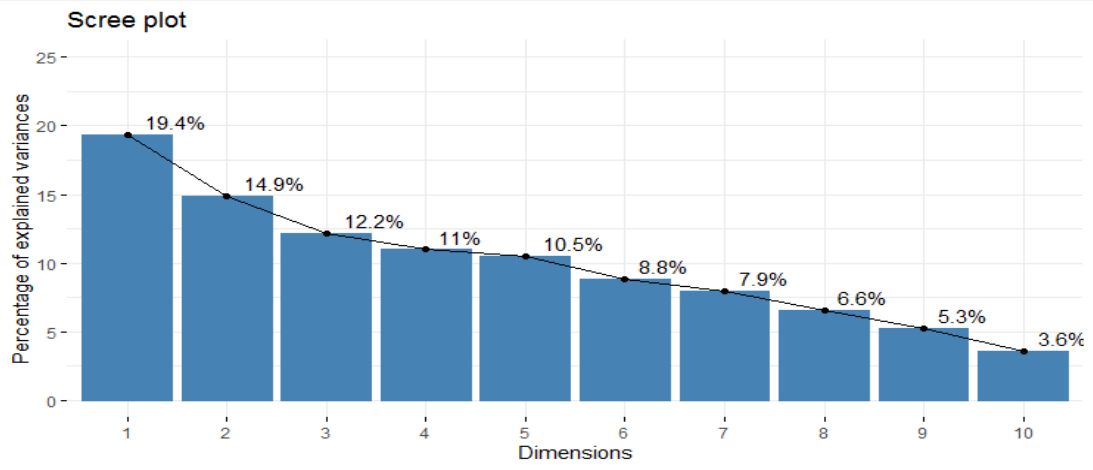


Figure: Variances Explained by Dimensions

Are this percentages of inertia explained correct?

A Real Life Application
Biplot

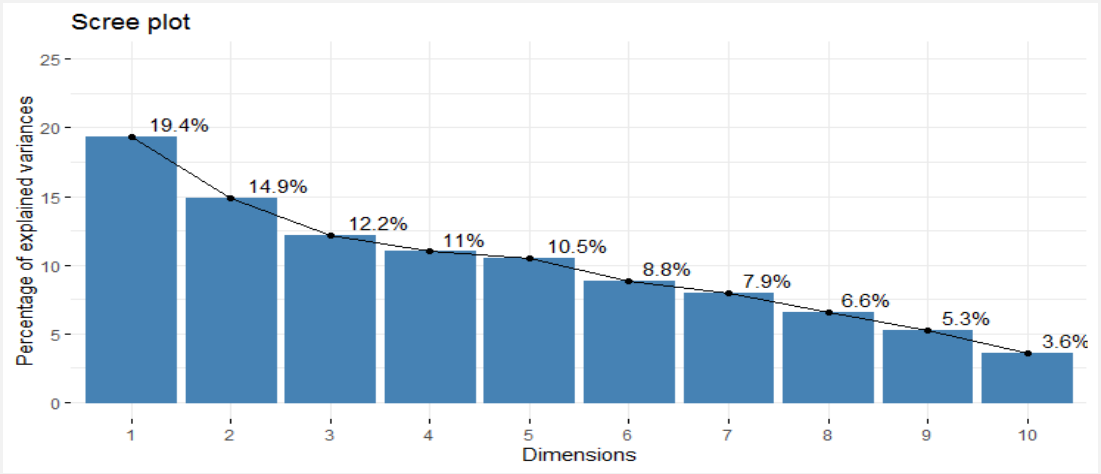
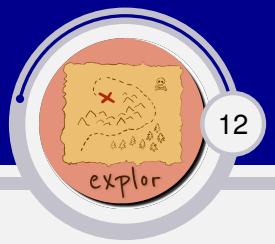


Figure: Variances Explained by Dimensions

Are this percentages of inertia explained correct?

No,percentage explained by the the first two dimensions is severely *underestimated*.

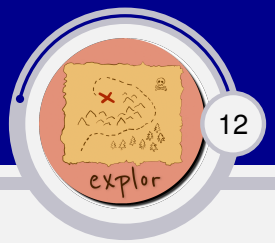


Greenacre(1993) suggested the following formulae :

Formula for Adjusted Inertia

$$I = \frac{Q}{Q-1}\left(\sum_l \lambda_l^4 - \frac{J-Q}{Q^2}\right)$$

A Real Life Application
Eigen Value Correction



Greenacre(1993) suggested the following formulae :

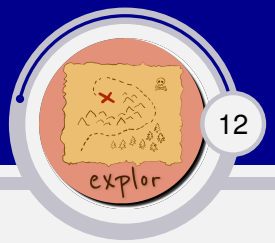
Formula for Adjusted Inertia

$$I = \frac{Q}{Q-1}\left(\sum_l \lambda_l^4 - \frac{J-Q}{Q^2}\right)$$

Formula for corrected eigen values

If λ_l^2 denote the eigen values (square of the singular values) obtained from the analysis of indicator matrix, then, the corrected eigen values in the analysis of Burt Matrix, λ_c^2 is given by:

$$\lambda_c^2 = \left[\frac{Q}{Q-1}\left(\lambda_l^2 - \frac{1}{Q}\right)\right]^2 \mathbb{I}\left(\lambda_l^2 > \frac{1}{Q}\right)$$



Greenacre(1993) suggested the following formulae :

Formula for Adjusted Inertia

$$I = \frac{Q}{Q-1}\left(\sum_l \lambda_l^4 - \frac{J-Q}{Q^2}\right)$$

Formula for corrected eigen values

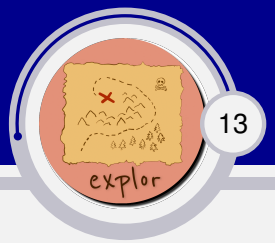
If λ_l^2 denote the eigen values (square of the singular values) obtained from the analysis of indicator matrix, then, the corrected eigen values in the analysis of Burt Matrix, λ_c^2 is given by:

$$\lambda_c^2 = \left[\frac{Q}{Q-1}\left(\lambda_l^2 - \frac{1}{Q}\right)\right]^2 \mathbb{I}\left(\lambda_l^2 > \frac{1}{Q}\right)$$

The percentage of inertia explained by each corrected eigen-value is given by:

$$\tau_c = \frac{\lambda_c^2}{I}$$

A Real Life Application



Using this correction, we see that about 58.85% of inertia was explained by first dimension and 15.88% of inertia is explained by the second dimension. Thus we use only two dimensions to represent our data as they together explain about 75% of the data.

A Real Life Application
Biplot

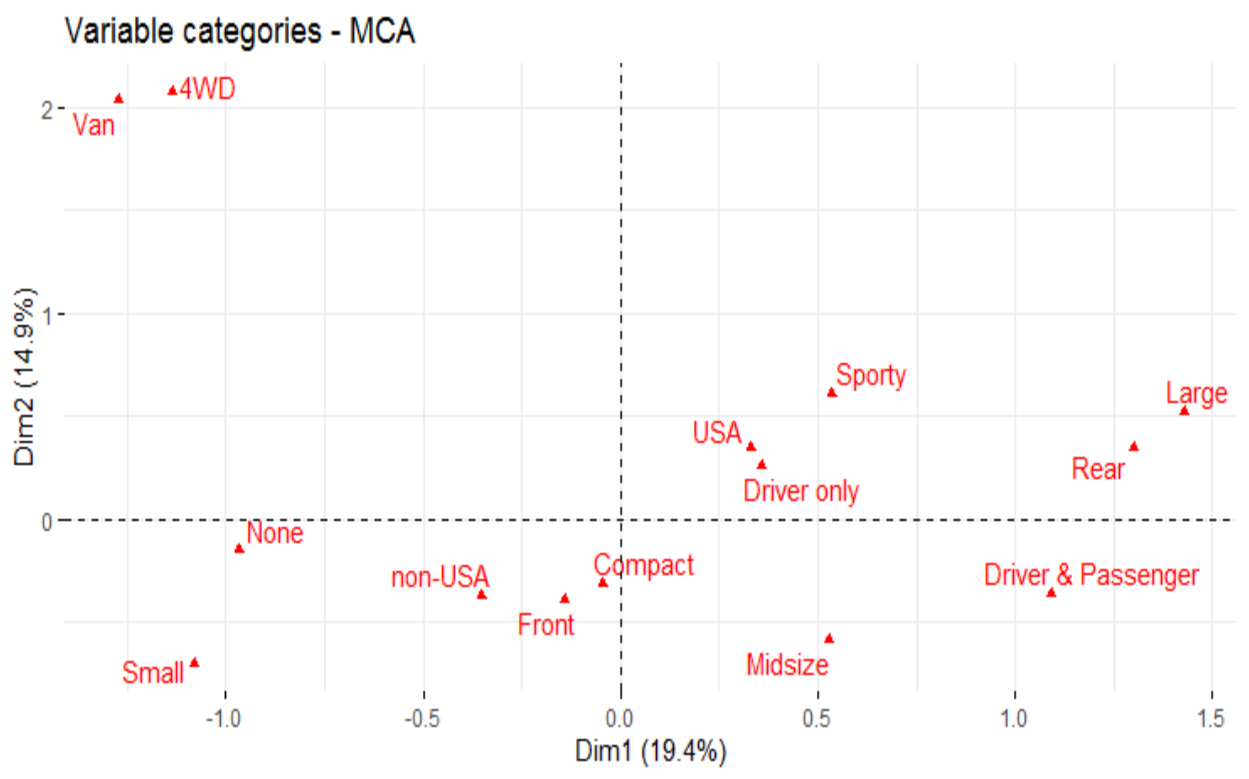
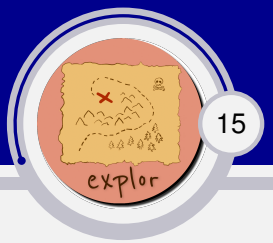


Figure: Biplot

A Real Life Application
Biplot

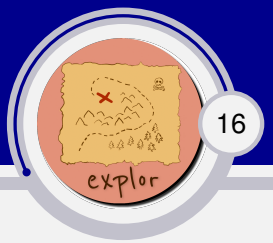


From the biplot, we can say the following things about the sample:

- With reference to the principal plane, Car-type Van and Drive-Train 4WD have occurred together.
- Occupying the lower left corner are "Small" cars with No-airbags.
- Cars with airbags for both driver and passenger are mostly midsize in this sample.
- Cars with rear wheel drive-train tend to be large.

Conjoint Analysis using MCA

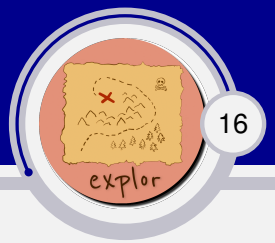
What is Conjoint Analysis



'Conjoint analysis' is a survey-based statistical technique used mainly in market research that helps determine how people value different attributes (feature, function, benefits) that make up an individual product or service. It is also really prevalent is psychological studies.

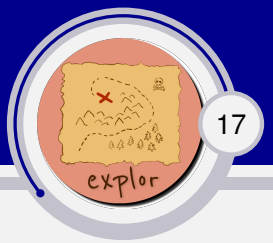
Conjoint Analysis using MCA

What is Conjoint Analysis



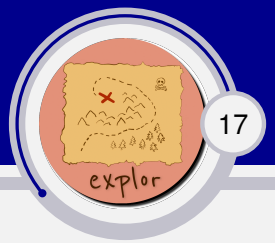
'Conjoint analysis' is a survey-based statistical technique used mainly in market research that helps determine how people value different attributes (feature, function, benefits) that make up an individual product or service. It is also really prevalent is psychological studies. In this section, we take up the `immigrationconjoint` data from **cjoint** package.

Conjoint Analysis using MCA
Data Description



1396 Americans were presented with five questions to gauge their outlook to immigration. Each question described 5 features (sex, reason for application, job experience, job plan, language skills) each of two possible immigrants, and the respondents has to choose one of them for allowed immigration.

Conjoint Analysis using MCA
Data Description



1396 Americans were presented with five questions to gauge their outlook to immigration. Each question described 5 features (sex, reason for application, job experience, job plan, language skills) each of two possible immigrants, and the respondents has to choose one of them for allowed immigration. The 5 features are:

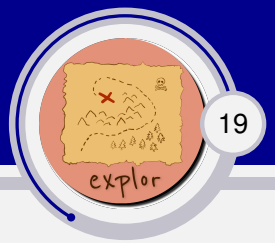
- Sex: Male or Female
- Reason of Applicaion: **Reunite** with family, seek better **Job**, Escape **Persecution**.
- Job Experience: None ("Exp0"), 1-2 Yrs ("Exp2"), 3-5 Yrs ("Exp5"), 5+ Yrs ("Exp10").
- Job Plans: None ("PlanN"), Will Look for Work ("PlanW"), Contract with Employer ("PlanC"), Interview with employer ("PlanI").
- Language Skills: Fluent English ("LF"), Broken English ("LB"), Tried English but unable ("LU"), Used Interpreter ("LI").

Conjoint Analysis using MCA

Data Description

	CaseID	contest_no	Gender	Reason for Application	Job Experience	Job Plans	Language Skills	Chosen_Immigrant
1	4	1	male	seek better job	5+ years	contract with employer	tried English but unable	1
2	4	1	female	seek better job	3-5 years	interviews with employer	used interpreter	0
3	4	2	female	escape persecution	3-5 years	contract with employer	fluent English	0
4	4	2	female	reunite with family	5+ years	interviews with employer	fluent English	1
5	4	3	female	seek better job	5+ years	interviews with employer	broken English	1
6	4	3	male	seek better job	none	contract with employer	fluent English	0
7	4	4	male	escape persecution	5+ years	will look for work	used interpreter	0
8	4	4	female	seek better job	1-2 years	interviews with employer	fluent English	1
9	4	5	female	seek better job	3-5 years	interviews with employer	used interpreter	1
10	4	5	male	seek better job	5+ years	no plans to look for work	used interpreter	0

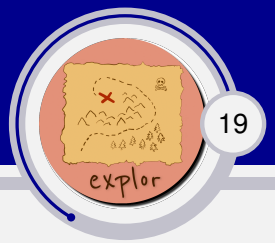
Figure: Data



We follow procedures followed by E.Kaciak (1990). To ensure preliminary comparability among all variables in the analysis, we divided the totals by 5 to obtain fuzzy- coded pairs of attribute levels. We denote the resulting by 1396×17 pseudo-indicator matrix as $\mathbf{X}$. Obviously, all row totals of $\mathbf{X}$ are equal to 5.

Will applying MCA algorithm on Pseudo-indicatr matrix be meaningful?

For a fuzzy-coded indicator matrix, the total inertia is no longer equal to $(J/Q) - 1$.



We follow procedures followed by E.Kaciak (1990). To ensure preliminary comparability among all variables in the analysis, we divided the totals by 5 to obtain fuzzy- coded pairs of attribute levels. We denote the resulting by 1396×17 pseudo-indicator matrix as $\mathbf{X}$. Obviously, all row totals of $\mathbf{X}$ are equal to 5.

Will applying MCA algorithm on Pseudo-indicatr matrix be meaningful?

For a fuzzy-coded indicator matrix, the total inertia is no longer equal to $(J/Q) - 1$.
But, as Greenacre(1987) suggested, we can still interpret the biplot and obtain meaningful relations about the variables.

Conjoint Analysis using MCA

Biplot

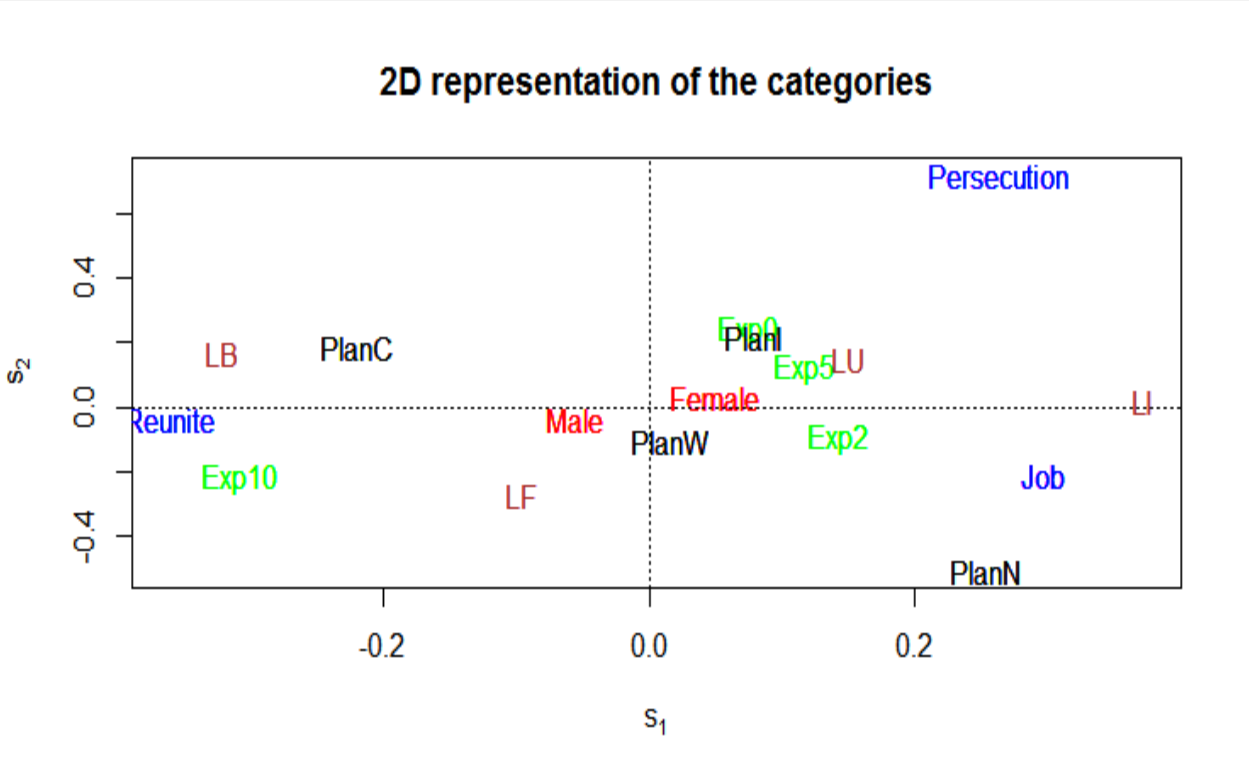
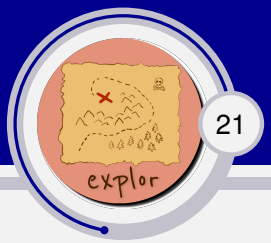


Figure: Biplot

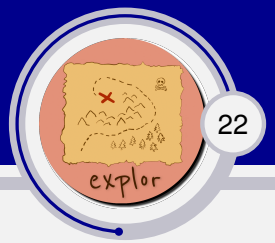
Conjoint Analysis using MCA

Biplot



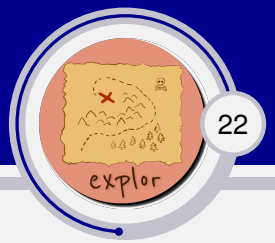
From this biplot, evidently respondents who found immigration for reuniting with family to be acceptable, also tended to choose immigrants with broken English, 5+ years of experience and contract with employer. This shows that this respondents did not care much about language as long as the immigrant was guaranteed to contribute to the country. Negatively correlated with the above are the immigrants who came to America for jobs but with no Job plan. Immigrannts using an interpreter also occupy similar spot. Also, the group of Immigrants escaping Persecution is not really correlated with any other categories.

Validation using MCA



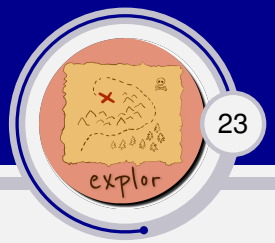
Finding the statistical properties of the category factors in MCA can be theoretically hard to find, as the distribution of eigen values is more complex than that of CA, and their sum does not have the same meaning.

Validation using MCA



Finding the statistical properties of the category factors in MCA can be theoretically hard to find, as the distribution of eigen values is more complex than that of CA, and their sum does not have the same meaning.
In this section, we use *Internal Validation* based on re-sampling techniques such as bootstrapping to find the confidence ellipsoids of the factors of the variables in 2-dimension, and see if the sample is a good representative of the population.

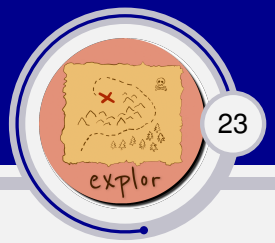
Validation using MCA
Data Description



Our data is the `wg93` data from **FactomineR** package in R.This dataset is from the International Social Survey Program (ISSP 1993). Each of the 871 respondents were asked the following four questions:

A. We believe too often in science, and not enough in feelings and faith
B. Overall modern science does more harm than good.
C. Any change humans cause in nature-no matter how scientific-is likely to make things worse.
D. Modern Science will solve our environmental problems with little change to our way of life.

Validation using MCA
Data Description



Our data is the `wg93` data from **FactomineR** package in R.This dataset is from the International Social Survey Program (ISSP 1993). Each of the 871 respondents were asked the following four questions:

A. We believe too often in science, and not enough in feelings and faith
B. Overall modern science does more harm than good.
C. Any change humans cause in nature-no matter how scientific-is likely to make things worse.
D. Modern Science will solve our environmental problems with little change to our way of life.

Each question has five possible response categories:

1. Agree Strongly
2. Agree
3. Neither agree nor disagree
4. Disagree
5. Disagree strongly

Validation using MCA

Biplot

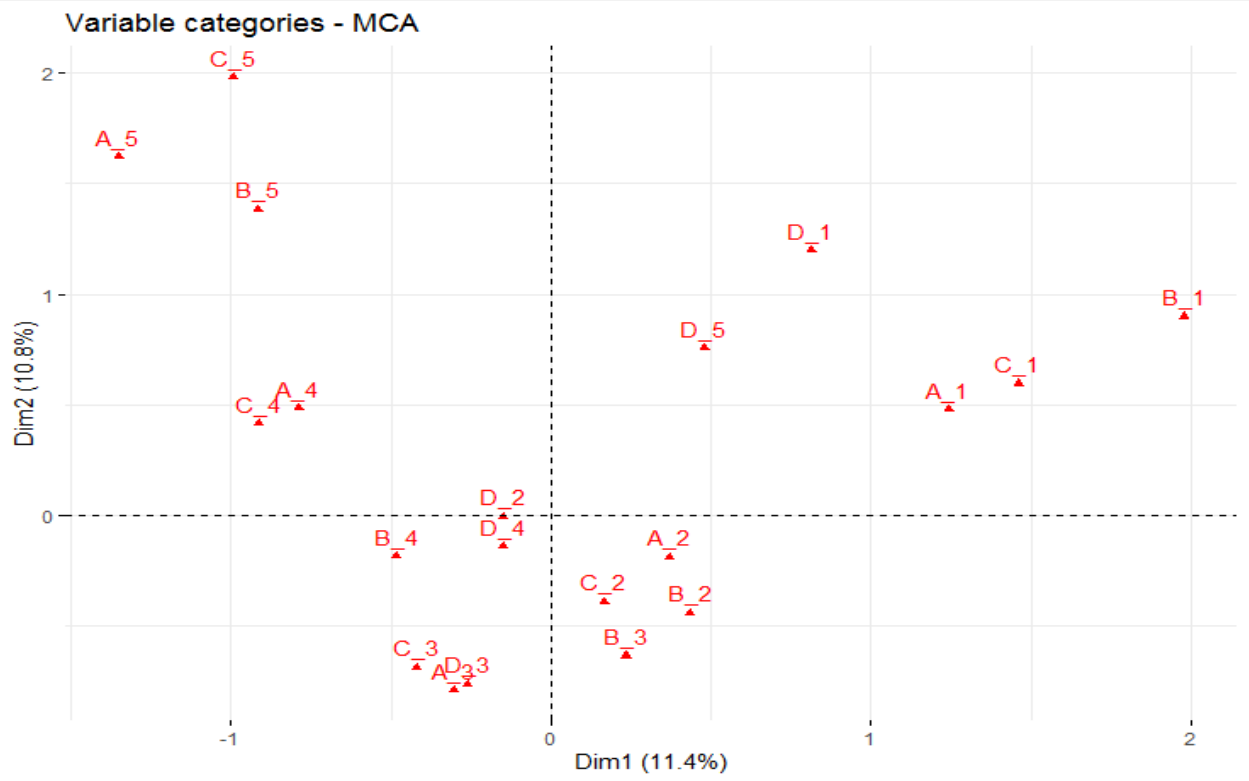
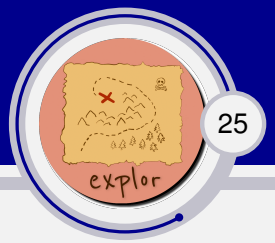


Figure: Biplot

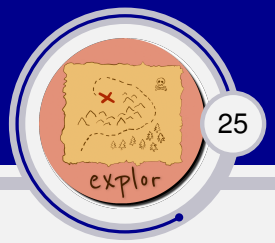


Horseshoe Effect

When the factor points are on or close to a convex function, with shape similar to a parabola. This can happen when there is no strong 2nd dimension in the data such that a folded version of the first axis, which satisfies the orthogonality requirement of the CA axes, explains more "inertia" than another direction in the data. We can do two-step MCA or non-linear PCA to get rid of this.

Validation using MCA

Horseshoe and Battery Effect



Horseshoe Effect

When the factor points are on or close to a convex function, with shape similar to a parabola. This can happen when there is no strong 2nd dimension in the data such that a folded version of the first axis, which satisfies the orthogonality requirement of the CA axes, explains more "inertia" than another direction in the data. We can do two-step MCA or non-linear PCA to get rid of this.

Battery Effect

This is often observed in survey analysis. When several questions have the same response categories, the respondent often chooses identical answers without sufficiently considering the content of the questions. That could account for the grouping for "Neither Agree nor Disagree" and "Disagree" in the bottom part of the plot.

Validation using MCA
Total Bootstrapping

- A specific replication is generated by a drawing with replacement of n individuals from the rows of $\mathbf{X}$. Each replication k leads to a separate MCA.

Validation using MCA
Total Bootstrapping

- A specific replication is generated by a drawing with replacement of n individuals from the rows of $\mathbf{X}$. Each replication k leads to a separate MCA.
- We plot the factor scores of all the MCAs in the same principal plane, corresponding to the principal axes of the original MCA.

Validation using MCA
Total Bootstrapping

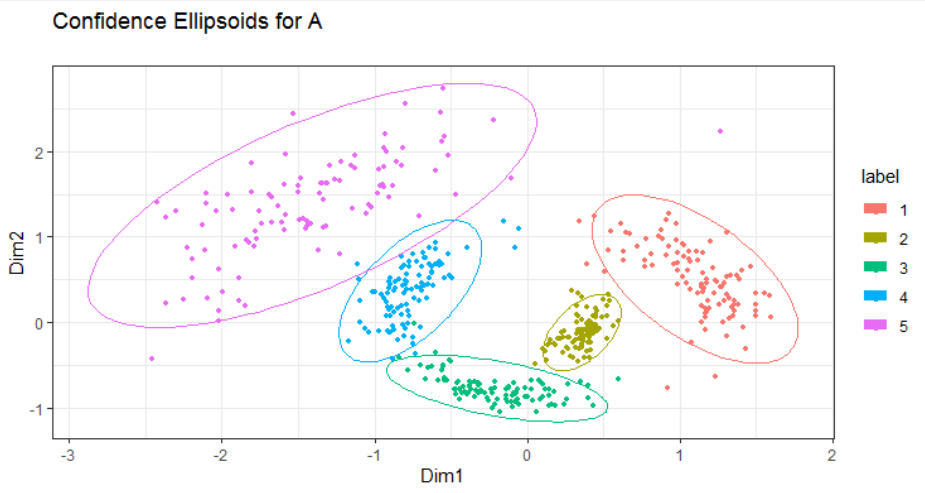
- A specific replication is generated by a drawing with replacement of n individuals from the rows of $\mathbf{X}$. Each replication k leads to a separate MCA.
- We plot the factor scores of all the MCAs in the same principal plane, corresponding to the principal axes of the original MCA.
- To remedy the arbitrariness of the signs of the axes in each replication, the orientation of the replicated axes are *a posteriori* changed (where necessary) to maintain a positive correlation with the axes of the original MCA.

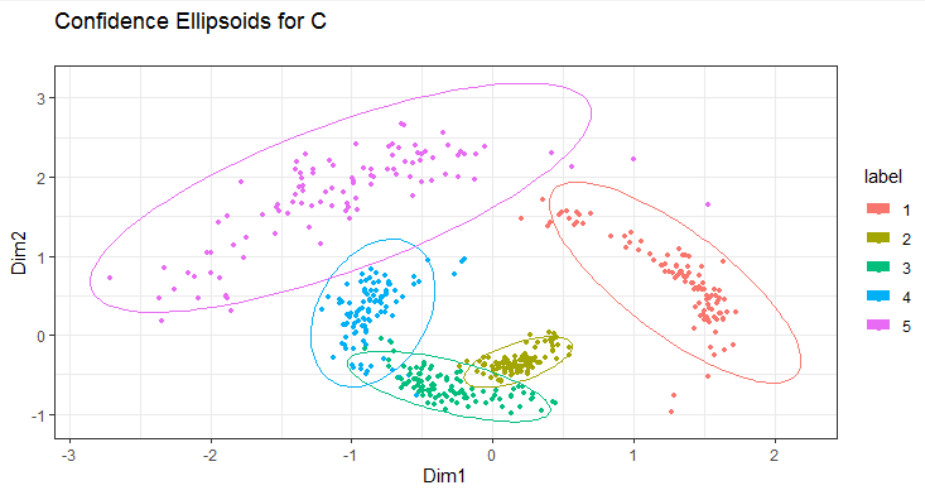
Validation using MCA
Total Bootstrapping

- A specific replication is generated by a drawing with replacement of n individuals from the rows of $\mathbf{X}$. Each replication k leads to a separate MCA.
- We plot the factor scores of all the MCAs in the same principal plane, corresponding to the principal axes of the original MCA.
- To remedy the arbitrariness of the signs of the axes in each replication, the orientation of the replicated axes are *a posteriori* changed (where necessary) to maintain a positive correlation with the axes of the original MCA.

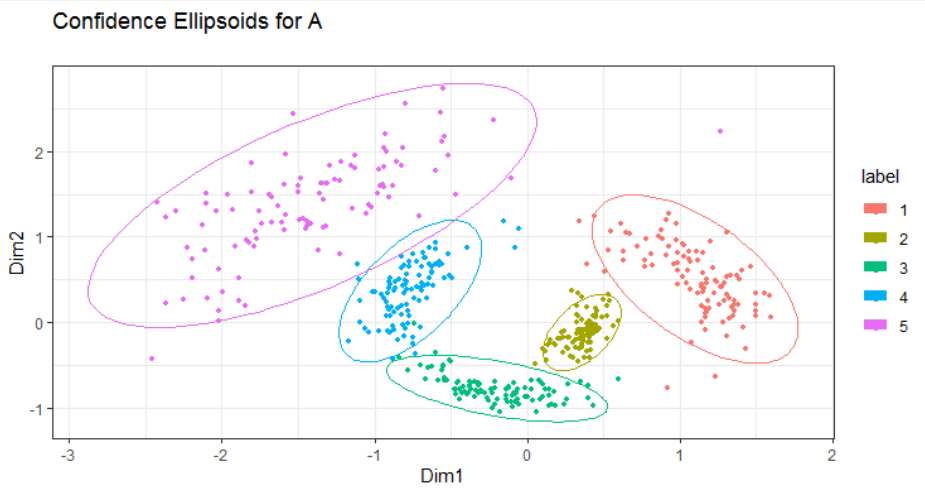
We perform $K = 100$ bootstraps, and form the 95% confidence ellipsoids for the levels of A, C and D.

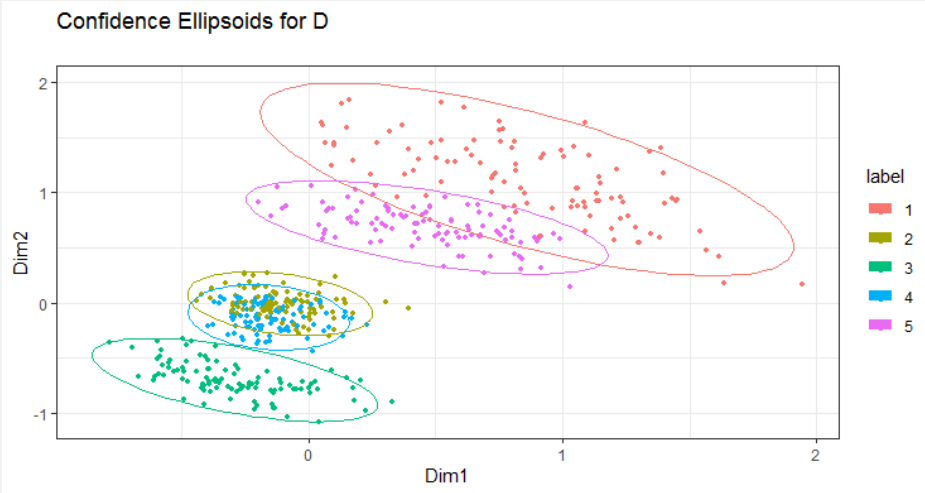
Validation using MCA
Confidence Ellipsoids





Validation using MCA
Confidence Ellipsoids





Validation using MCA

In the case of A and C (and also B, which is not displayed here), the response categories belonging to a specific question appear to be distinct. Also, the confidence areas of homologous categories in A and C (and B) have similar sizes, with almost all sets of identical categories relating to A,B and C overlap. For example we can see A5 and C5 largely overlap. But for question D, the confidence ellipsoid for "Agree Strongly" and "Disagree Strongly" has a significant intersection, whereas the categories Agree and Disagree almost overlap. This hints at the misinterpretation of this question, and also battery effect for the option "Disagree".

Classification using MCA

Classifying observations described by Categorical Predictors into one of the k classes can be done using log-linear graphical model, but these model suffer from Curse of Dimensionality and are difficult to apply when the number of categories is large.

Classification using MCA

Classifying observations described by Categorical Predictors into one of the k classes can be done using log-linear graphical model, but these model suffer from Curse of Dimensionality and are difficult to apply when the number of categories is large.
Saporta(1976) suggested the following for tackling this problem:

1. An MCA is performed on the predictors.
2. A Fisher's Linear Discriminant Analysis is done using factor coordinates as predictors.

Classification using MCA

Classifying observations described by Categorical Predictors into one of the *k* classes can be done using log-linear graphical model, but these model suffer from Curse of Dimensionality and are difficult to apply when the number of categories is large.
Saporta(1976) suggested the following for tackling this problem:

1. An MCA is performed on the predictors.
2. A Fisher's Linear Discriminant Analysis is done using factor coordinates as predictors.

This method is known as *Disqual*.

Classification using MCA
Data Description

We use the `Poison` data from the **FactomineR** package for this classification purpose. This data appeared in Box, G. E. P. and D. R. Cox, *An Analysis of Transformations (with discussion)*, Journal of the Royal Statistical Society, Series B, Vol. 26, No. 2, pp. 211–254. The data used here refer to a survey carried out on a sample of children of primary school who suffered (and did not suffer) from food poisoning. They were asked about their symptoms and about what they ate.

Classification using MCA
Data Description

We use the `Poison` data from the **FactomineR** package for this classification purpose. This data appeared in Box, G. E. P. and D. R. Cox, *An Analysis of Transformations (with discussion)*, Journal of the Royal Statistical Society, Series B, Vol. 26, No. 2, pp. 211–254. The data used here refer to a survey carried out on a sample of children of primary school who suffered (and did not suffer) from food poisoning. They were asked about their symptoms and about what they ate.
The predictors are:

- Fish.
- Mayo.
- Courgate.
- Cheese.
- Ice-cream.

Classification using MCA
Data Description

We use the `Poison` data from the **FactomineR** package for this classification purpose. This data appeared in Box, G. E. P. and D. R. Cox, *An Analysis of Transformations (with discussion)*, Journal of the Royal Statistical Society, Series B, Vol. 26, No. 2, pp. 211–254. The data used here refer to a survey carried out on a sample of children of primary school who suffered (and did not suffer) from food poisoning. They were asked about their symptoms and about what they ate. The predictors are:

- Fish.
- Mayo.
- Courgate.
- Cheese.
- Ice-cream.

Each of the predictor have two levels-Yes or No. The response variable is "sick" with categories "Yes" or "No".

Classification using MCA

Data Analysis

First we perform an MCA and obtain the individual factor scores. Then we divide the whole dataset i two parts: Training and Test set, and perform a Fisher's LDA on Training set and run it on Test set. We do this bi-partition of data 10 times in a 10-fold cross validation.

Classification using MCA

Data Analysis

First we perform an MCA and obtain the individual factor scores. Then we divide the whole dataset i two parts: Training and Test set, and perform a Fisher's LDA on Training set and run it on Test set. We do this bi-partition of data 10 times in a 10-fold cross validation.

	Observed Not Sick	Observed Sick
Predicted Not Sick	16.84%	11.58%
Predicted Sick	15.26%	56.32%

Table: LDA Confusion Matrix

see that almost 83% of sick students have been predicted to be sick and almost 52% of not-sick students have been predicted to be not sick. Now, we had in total only 55 data points, and more samples would have resulted in a clearer reflection of thee efficiency of the method. Further to improve its efficiency, we can use Logistic Regression/Support Vector Machine. Thus using MCA we can classify the data.

THANK YOU!