Definition of Multiple Correspondence Analysis

Multiple correspondence analysis is an extension of binary correspondence analysis.

It makes it possible to analyze at a glance a multi-way contingency table.

A classical example of multi-way contingency table is an array containing the answers provided by respondents to a multiple choice exam comprising Q questions.

Multiple correspondence analysis is particularly useful to visualize the results of a survey and to attribute scores in order to segment the respondents in homogeneous groups.

Example

To illustrate multiple correspondence analysis, consider the following fictitious example:

ID	Type of Employee					Smoking			Total
	1	2	3	4	5	A	B	C	Q
1	0	0	1	0	0	0	1	0	2
2	0	1	0	0	0	1	0	0	2
3	1	0	0	0	0	0	1	0	2
4	0	0	0	0	1	0	0	1	2
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
193	0	1	0	0	0	1	0	0	2
			Z_1				Z_2		

Coding

In this example, there are $Q = 2$ questions. The table is thus of the form

$$\mathbf{Z} = [\mathbf{Z}_1 \mid \mathbf{Z}_2].$$

For a questionnaire with Q questions, the table would be of the form

$$\mathbf{Z} = [\mathbf{Z}_1 \mid \cdots \mid \mathbf{Z}_Q].$$

Notation

$$Q = \text{number of questions,}$$

$$n = \text{number of respondents,} \quad \text{\# rows}$$

$$p_q = \text{number of modalities (choices of answers) for question } q,$$

$$p = p_1 + \cdots + p_Q. \quad \text{\# cols}$$

Empty Cells

The larger Q, the larger the number of empty cells.

Indeed, the proportion of non-empty cells is

$$\frac{nQ}{np} = \frac{Q}{p}.$$

If all the questions have the same number of possible answers, then

$$p_1 = \cdots = p_Q = \frac{p}{Q},$$

and hence

$$\frac{Q}{p} = \frac{1}{p_1} \to 0 \quad \text{as } p_1 \to \infty.$$

Condensed Table

It is an $n \times Q$ table which identifies the answer provided by a respondent to each of the Q questions.

For example, in the table below the first respondent is an employee of Category 3 who is a smoker of Category 2.

ID	Type of Employee	Smoking
1	3	2
2	2	1
3	1	2
4	5	3
$\vdots$	$\vdots$	$\vdots$
193	2	1

Burt Table

A Burt table is another method for coding a contingency table involving more than two variables.

Given a table of responses

$$\mathbf{Z} = [\underbrace{\mathbf{Z}_1}_{}^{n\times p_1} \mid \cdots \mid \mathbf{Z}_Q^{\,n\times p_Q}],$$

the corresponding Burt table is the $p \times p$ matrix given by

$$\mathbf{B} = \mathbf{Z}\mathbf{Z}^\top,$$

viz.

$$\begin{array}{c} \scriptstyle p_1\times n \;\; n\times p_1 \\ \mathbf{B} = \begin{bmatrix} \mathbf{Z}_1^\top \mathbf{Z}_1 & \mathbf{Z}_1^\top \mathbf{Z}_2 & \cdots & \mathbf{Z}_1^\top \mathbf{Z}_Q \\ \mathbf{Z}_2^\top \mathbf{Z}_1 & \mathbf{Z}_2^\top \mathbf{Z}_2 & \cdots & \mathbf{Z}_2^\top \mathbf{Z}_Q \\ \vdots & \vdots & \cdots & \vdots \\ \mathbf{Z}_Q^\top \mathbf{Z}_1 & \mathbf{Z}_Q^\top \mathbf{Z}_2 & \cdots & \mathbf{Z}_Q^\top \mathbf{Z}_Q \end{bmatrix}. \end{array}$$

Cyril Burt (1883–1971)

Sir Cyril L. Burt was an English educational psychologist and geneticist who also made contributions to statistics. He is known for his studies on the heritability of IQ.

Shortly after he died, his studies of inheritance of intelligence were discredited after evidence emerged indicating he had falsified research data, inventing correlations in separated twins which did not exist.

You can read about "The Burt Affair" on Wikipedia.

Characteristics of $Z_q^\top Z_{q'}$

- ✓ $\mathbf{Z}_q^\top \mathbf{Z}_q$ is a $p_q \times p_q$ diagonal matrix containing the answers to question q.
- ✓ The (j, j) element of $\mathbf{Z}_q^\top \mathbf{Z}_q$ is equal to the number d_{jj} of individuals who chose category j for question q.
- ✓ $\mathbf{Z}_q^\top \mathbf{Z}_{q'}$ is a contingency table providing the number of answers to questions q and q'.
- ✓ The (j, j') element of matrix $\mathbf{Z}_q^\top \mathbf{Z}_{q'}$ is equal to the number $d_{jj'}$ of individuals who chose category j for question q and category j' for question q'.

Example (cont'd)

193	1	2	3	4	5	A	B	C
1	11	0	0	0	0	4	5	2
2	0	18	0	0	0	4	10	4
3	0	0	51	0	0	25	22	4
4	0	0	0	88	0	18	57	13
5	0	0	0	0	25	10	13	2
A	4	4	25	18	10	61	0	0
B	5	10	22	57	13	0	107	0
C	2	4	4	13	2	0	0	25

193

This matrix is $(5+3) \times (5+3)$.

Additional Notation

For each $i \in \{1, \ldots, Q\}$, set $\mathbf{D}_i = \mathbf{Z}_i^\top \mathbf{Z}_i$ and

$$\mathbf{D} = \begin{bmatrix} \mathbf{D}_1 & 0 & \cdots & 0 \\ 0 & \mathbf{D}_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbf{D}_Q \end{bmatrix} = \begin{bmatrix} \mathbf{Z}_1^\top \mathbf{Z}_1 & 0 & \cdots & 0 \\ 0 & \mathbf{Z}_2^\top \mathbf{Z}_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbf{Z}_Q^\top \mathbf{Z}_Q \end{bmatrix}.$$

Again, these matrices are $p \times p$.

Multiple Correspondence Analysis

A multiple correspondence analysis is a binary correspondence analysis performed either on the matrix $\mathbf{Z}$ or on the Burt table $\mathbf{B}$.

It will be shown that the result of the analysis is the same, whether it is performed on $\mathbf{Z}$ or on $\mathbf{B}$.

For standard binary correspondence analysis, one starts with a matrix

$$\mathbf{F} = (f_{ij}).$$

To carry out a multiple correspondence analysis on the matrix **Z**, one has

$$\mathbf{F} = \frac{\mathbf{Z}}{nQ}$$

1's.

people

questions

with

$$\sum_{i=1}^{n}\sum_{j=1}^{p} f_{ij} = \sum_{i=1}^{n}\sum_{j=1}^{p} \frac{Z_{ij}}{nQ} = 1.$$

Analysis based on **Z** (2–3)

In standard binary correspondence analysis, one has

$$\mathbf{D}_{n} = \operatorname{diag}\left(f_{i\bullet}\right) \quad \text{and} \quad \mathbf{D}_{p} = \operatorname{diag}\left(f_{\bullet j}\right).$$

#people (#rows) #choices (#cols)

To carry out a multiple correspondence analysis on the matrix $\mathbf{Z}$, the sum of each row equals Q, and hence

$$\mathbf{D}_n = \frac{Q}{nQ}\,\mathbf{I}_n = \frac{\mathbf{I}_n}{n}\,.$$

Moreover,

$$\mathbf{D}_p = \frac{\mathbf{D}}{nQ} = \frac{1}{nQ}\,\operatorname{diag}(\mathbf{Z}_i^\top \mathbf{Z}_i).$$

Analysis based on $\mathbf{Z}$ (3–3)

As a result, the factors $\varphi_j = \mathbf{D}_p^{-1} u_j$ are such that

$$\mathbf{F}^\top \mathbf{D}_n^{-1} \mathbf{F} \mathbf{D}_p^{-1} u_j = \lambda_j u_j.$$

Therefore,

$$\mathbf{D}_p^{-1} \mathbf{F}^\top \mathbf{D}_n^{-1} \mathbf{F} \varphi_j = \lambda_j \varphi_j$$

ou equivalently,

$$\frac{1}{Q} \mathbf{D}^{-1} \mathbf{Z}^\top \mathbf{Z} \varphi_j = \lambda_j \varphi_j.$$

Analysis based on $\mathbf{B}$ (1–3)

For correspondence analysis with Burt's table, one has

$$\mathbf{F} = \frac{\mathbf{B}}{nQ^2}$$

because each of the Q blocks in $\mathbf{B}$ consists of integers whose sum is equal to n.

Furthermore, $\mathbf{B}$ is a symmetric matrix. Multiple correspondence analysis on Burt's table is thus performed in the case $n = p$.

In this special case, one has

$$\mathbf{D}_n = \mathbf{D}_p = \frac{\mathbf{D}}{nQ}.$$

Analysis based on **B** (2–3)

The factors in the multiple correspondence analysis of the Burt table are given by

$$\varphi_j^* = \mathbf{D}_n^{-1} v_j = nQ\mathbf{D}^{-1} v_j,$$

where

$$\mathbf{F}\mathbf{D}_p^{-1}\mathbf{F}^\top \mathbf{D}_n^{-1} v_j = \lambda_j^* v_j.$$

Equivalently, one has

$$\frac{1}{Q^2}\,\mathbf{B}\mathbf{D}^{-1}\mathbf{B}^\top \mathbf{D}^{-1} v_j = \lambda_j^* v_j.$$

Analysis based on $\mathbf{B}$ (3–3)

The factor φ_j^* is the solution to the equation

$$\frac{1}{Q^2}\,\mathbf{B}\mathbf{D}^{-1}\mathbf{B}^\top\varphi_j^* = \lambda_j^*\mathbf{D}\varphi_j^*.$$

Upon multiplication on both sides by $\mathbf{D}^{-1}$, the same factor φ_j^* is seen to be a solution to

$$\frac{1}{Q^2}\,\mathbf{D}^{-1}\mathbf{B}\mathbf{D}^{-1}\mathbf{B}^\top\varphi_j^* = \lambda_j^*\varphi_j^*.$$

Comparison between MCA on $\mathbf{Z}$ and $\mathbf{B}$

For the analysis based on $\mathbf{Z}$, one has

$$\frac{1}{Q}\,\mathbf{D}^{-1}\mathbf{B}^\top \varphi_j = \lambda_j \varphi_j,$$

so that upon multiplication on both sides by $\mathbf{D}^{-1}\mathbf{B}/Q$, one finds

$$\frac{1}{Q^2}\,\mathbf{D}^{-1}\mathbf{B}\mathbf{D}^{-1}\mathbf{B}^\top \varphi_j = \lambda_j \frac{\mathbf{D}^{-1}\mathbf{B}\varphi_j}{Q} = \lambda_j^2 \varphi_j.$$

It follows that, for all $j \in \{1, \ldots, p\}$, one has

$$\lambda_j^* = \lambda_j^2 \quad \text{and} \quad \varphi_j^* = \varphi_j.$$

Remark

In the case $Q = 2$, multiple correspondence analysis on $\mathbf{Z}$ is equivalent to binary analysis on the matrix $\mathbf{Z}_2^\top \mathbf{Z}_1$.

In fact the jth vector Φ_j of the multiple correspondence analysis on $\mathbf{Z} = [\mathbf{Z}_1 \mid \mathbf{Z}_2]$ can be expressed in the form

$$\Phi_j = (\varphi_j, \psi_j)^\top,$$

where φ_j and ψ_j are respectively the jth direct or dual factor of the analysis performed on $\mathbf{Z}_2^\top \mathbf{Z}_1$. Furthermore, given that

$$\lambda_j^* = j\text{th eigenvalue of } \mathbf{Z}_2^\top \mathbf{Z}_1,$$

then, for all $j \in \{1, \ldots, p\}$, one has

$$\lambda_j = \left(1 + \sqrt{\lambda_j^*}\right)/2.$$

Interpretation of the Resulting Graph

A factor map is created in the same way as in binary correspondence analysis.

However, the distance between points and the global geometry of the map can no longer be interpreted as in binary correspondence analysis.

Of particular interest are

- ✓ points which are close to one another or in the same quadrant;
- ✓ points which are in the same direction with respect to the origin.

Example 1: Land Use (1–3)

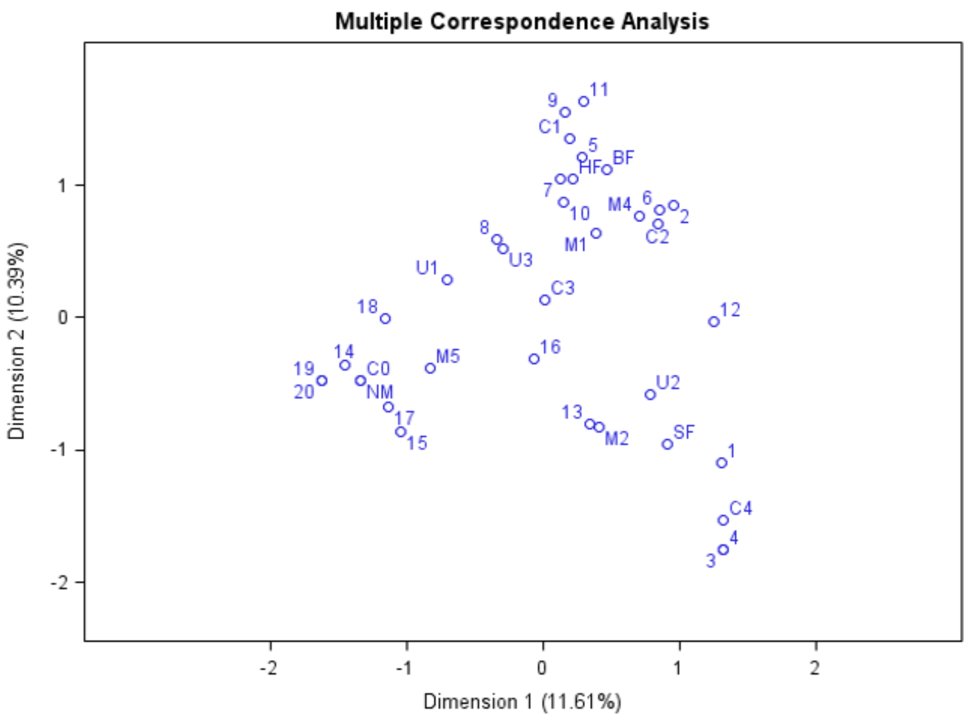
Information was collected about 20 farms in the Netherlands.

```
Humidity
     Ground humidity level (1, 2, 4, 5)

Management
     Land management type (SF = Standard Farm,
     BF = Biological Farm, HF = Holiday Farm,
     NM = Nature Conservation Management)

Production
     Type of production (U1 = Hay,
     U2 = Intermediate Production, U3 = Pasture)

Manure
     Manure use intensity level (C0, C1, C2, C3, C4)
```

Example 1: Land Use (2–3)

Farm	Humidity	Management	Production	Manure
1	M1	SF	U2	C4
2	M1	BF	U2	C2
3	M2	SF	U2	C4
4	M2	SF	U2	C4
5	M1	HF	U1	C2
6	M1	HF	U2	C2
7	M1	HF	U3	C3
8	M5	HF	U3	C3
9	M4	HF	U1	C1
10	M2	BF	U1	C1
11	M1	BF	U3	C1
12	M4	SF	U2	C2
13	M5	SF	U2	C3
14	M5	NM	U3	C0
15	M5	NM	U2	C0
16	M5	SF	U3	C3
17	M2	NM	U1	C0
18	M1	NM	U1	C0
19	M5	NM	U1	C0
20	M5	NM	U1	C0

Example 1: Land Use (3–3)



Example 2: Client Retention (1–2)

Businesses routinely try to understand which factors are associated with clientele mobility. The IBM Watson website provides an example of data analysis on client retention in telecoms.

```
No gender SeniorCitizen Partner Dependents      tenure PhoneService
1 Female              No      Yes         No ShortTenure           No
MultipleLines InternetService OnlineSecurity OnlineBackup
No              DSL               No            Yes
  DeviceProtection TechSupport StreamingTV StreamingMovies Contract
1              No          No          No               No       MtM
PaperlessBilling  PaymentMethod MonthlyCharges Churn
Yes Electronicheck      LowCharge    No
```

Example 2: Client Retention (2–2)

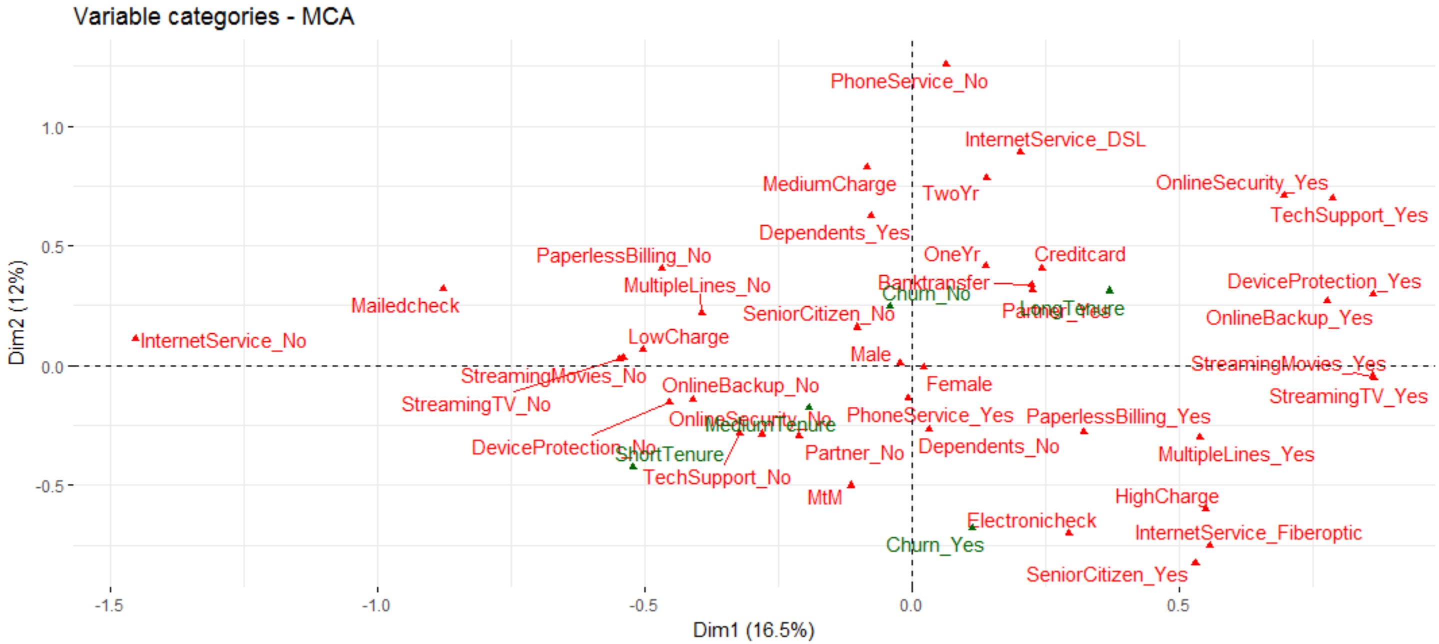


Illustration: Vehicles Sold (1–4)

Data are available on various models of cars sold in the USA in 1993.

`Manufacturer`
    `Car manufacturer`

`Type`
    `Type of vehicle`

`Airbags`
    `Position of the airbags`

`Traction`
    `Front-wheel drive, Rear-wheel drive`

Illustration: Vehicles Sold (2–4)

```
Manufacturer Category Airbags Traction
Acura Small None Front
Acura Midsize DriPas Front
Audi Compact Driver Front
Audi Midsize DriPas Front
BMW Midsize Driver Rear
Buick Midsize Driver Front
Buick Large Driver Front
Buick Large Driver Rear
Buick Midsize Driver Front
Cadillac Large Driver Front
Cadillac Midsize DriPas Front
Chevrolet Compact None Front
Chevrolet Compact Driver Front
Chevrolet Van None 4WD
Chevrolet Large Driver Rear
Chevrolet Sporty Driver Rear
Chrysler Large DriPas Front
Chrysler Compact DriPas Front
Chrysler Large Driver Front
Dodge Small None Front
Dodge Small Driver Front
Dodge Compact Driver Front
.         .       .     .
.         .       .     .
.         .       .     .
```

Without the manufacturers

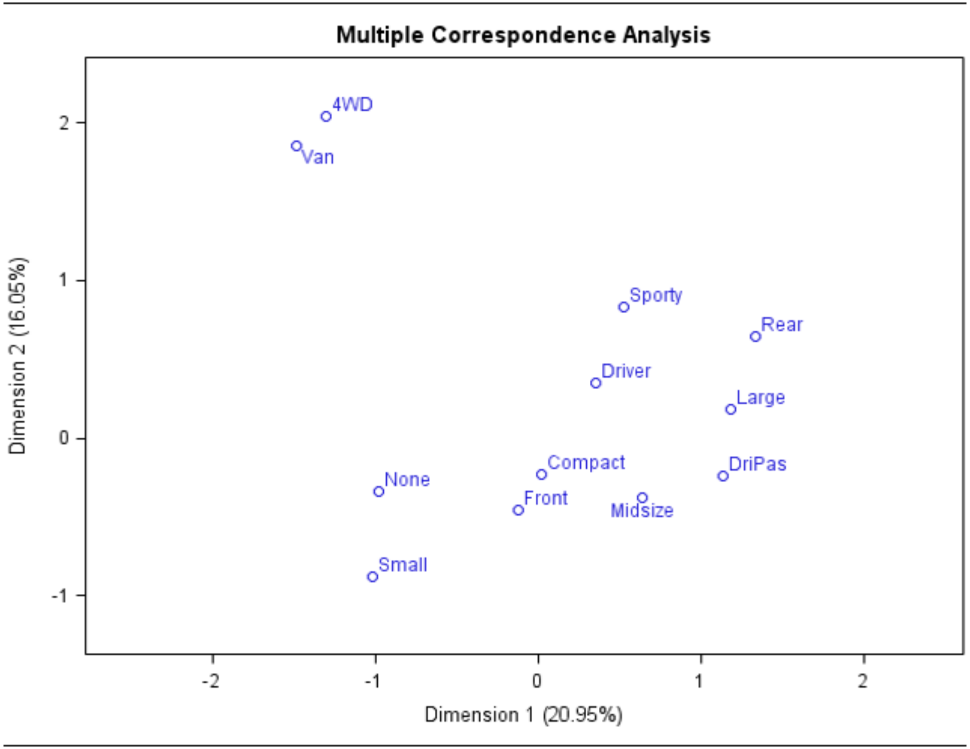


Illustration: Vehicles Sold (4–4)

With the manufacturers