MCA - Applications

Objectives

Study of individuals, variables and categories.

1.Individuals'study: two individuals are close to each other if they answered the questions the same way. We will not be so interested in single individuals but rather in populations: are there groups of individuals?

2.Variables and categories'study: the questions are the same as for a PCA. First, we want to see the relationship between variables and the associations between categories. Two categories are close to each other if they are often taken together. Second, we are also interested in looking for one or several continuous synthetical variables to summarize categorical ones. Third, we want to characterize groups of individuals by categories.

MCA is used to analyze a database from survey. The goal is to identify:

- A group of individuals with similar profile in their answers to the questions.
- The associations between categories.

POISON DATA SET →

According to MCA terminology, our data contains :

- *Active individuals* (rows 1:55): Individuals that are used in the multiple correspondence analysis.
- *Active variables* (columns 5:15) : Variables that are used in the MCA.
- *Supplementary variables*: They don't participate to the MCA. The coordinates of these variables will be predicted.
    - *Supplementary quantitative variables* (quanti.sup): Columns 1 and 2 corresponding to the columns *age* and *time*, respectively.
    - *Supplementary qualitative variables* (quali.sup): Columns 3 and 4 corresponding to the columns *Sick* and *Sex*, respectively. This factor variables will be used to color individuals by groups.

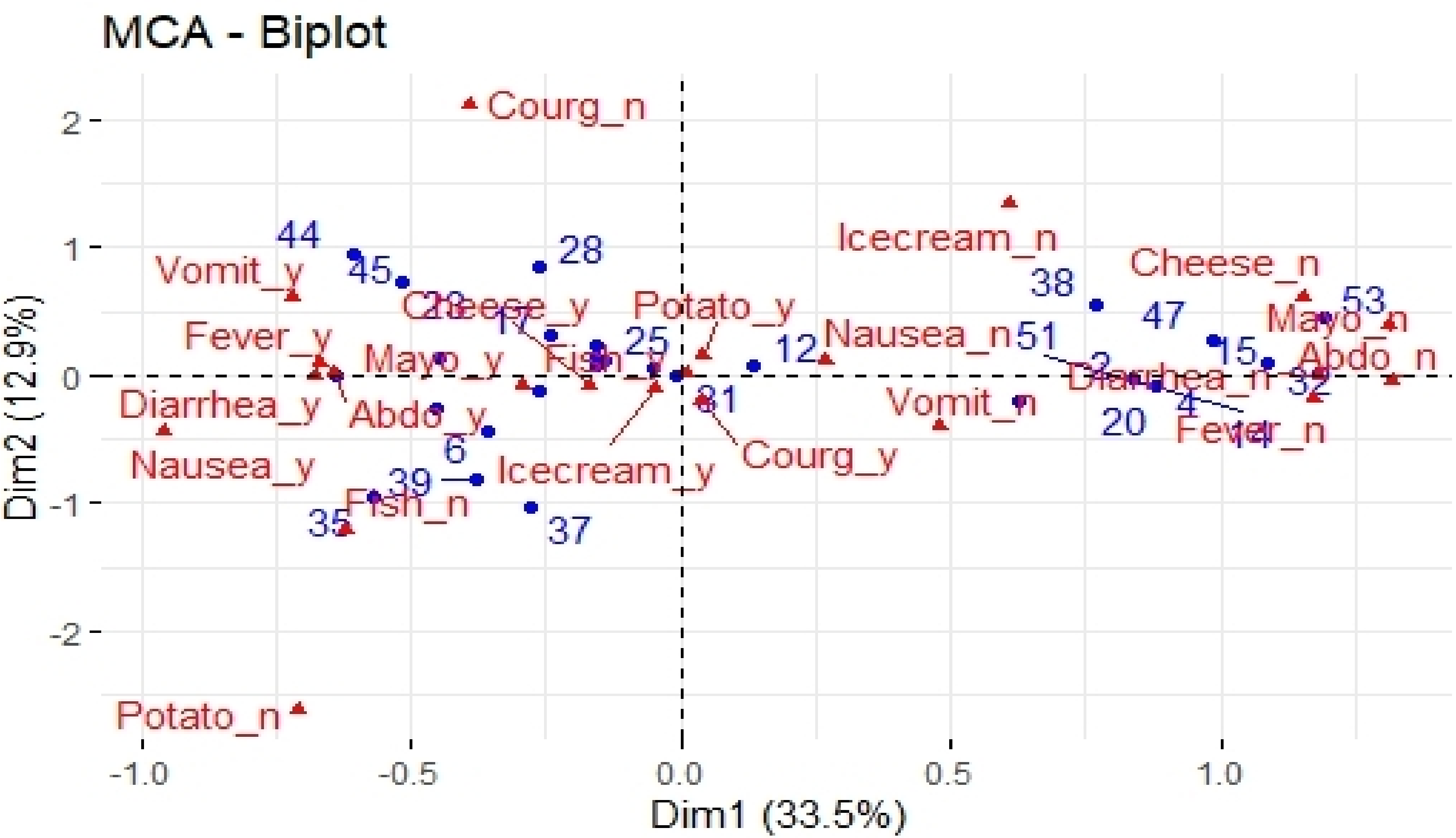
ESSENTIALS
→

R code

The function MCA()[FactoMiner package] can be used. A simplified format is :

```
MCA(X, ncp = 5, graph = TRUE)
```

X: a data frame with n rows (individuals) and p columns (categorical variables)
ncp: number of dimensions kept in the final results.
graph: a logical value. If TRUE a graph is displayed.



The plot above shows a global pattern within the data. Rows (individuals) are represented by blue points and columns (variable categories) by red triangles.

The distance between any row points or column points gives a measure of their similarity (or dissimilarity). Row points with similar profile are closed on the factor map. The same holds true for column points.

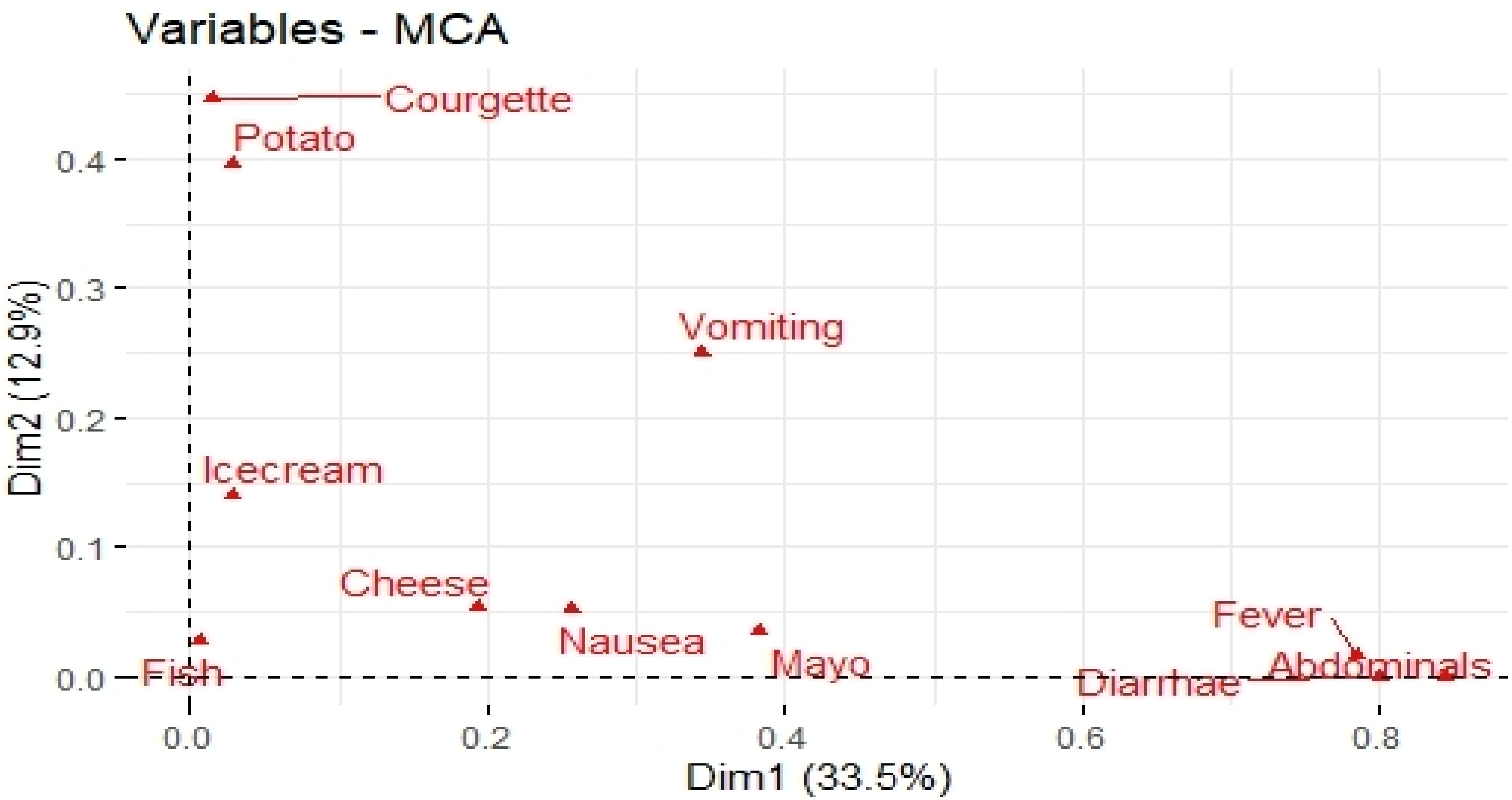
Var →

```
## Multiple Correspondence Analysis Results for variables
##  ===========================================================
##   Name       Description
## 1 "$coord"   "Coordinates for categories"
## 2 "$cos2"    "Cos2 for categories"
## 3 "$contrib" "contributions of categories"
```

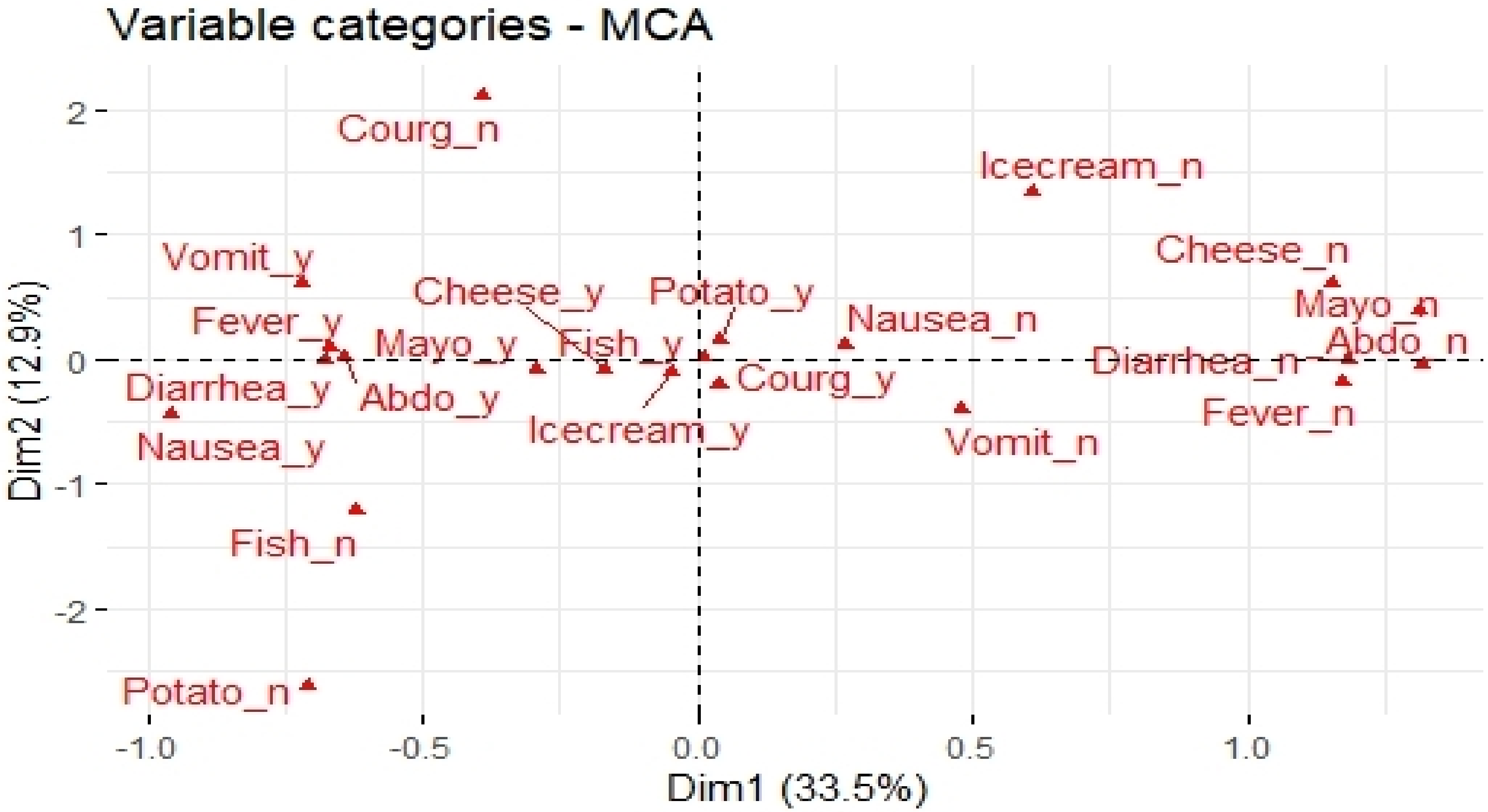
The components of the get_mca_var() can be used in the plot of rows as follow:

var$coord: coordinates of variables to create a scatter plot
var$cos2: represents the quality of the representation for variables on the factor map.
var$contrib: contains the contributions (in percentage) of the variables to the definition of the dimensions.

Note that, it’s possible to plot variable categories and to color them according to either i) their quality on the factor map (cos2) or ii) their contribution values to the definition of dimensions (contrib).

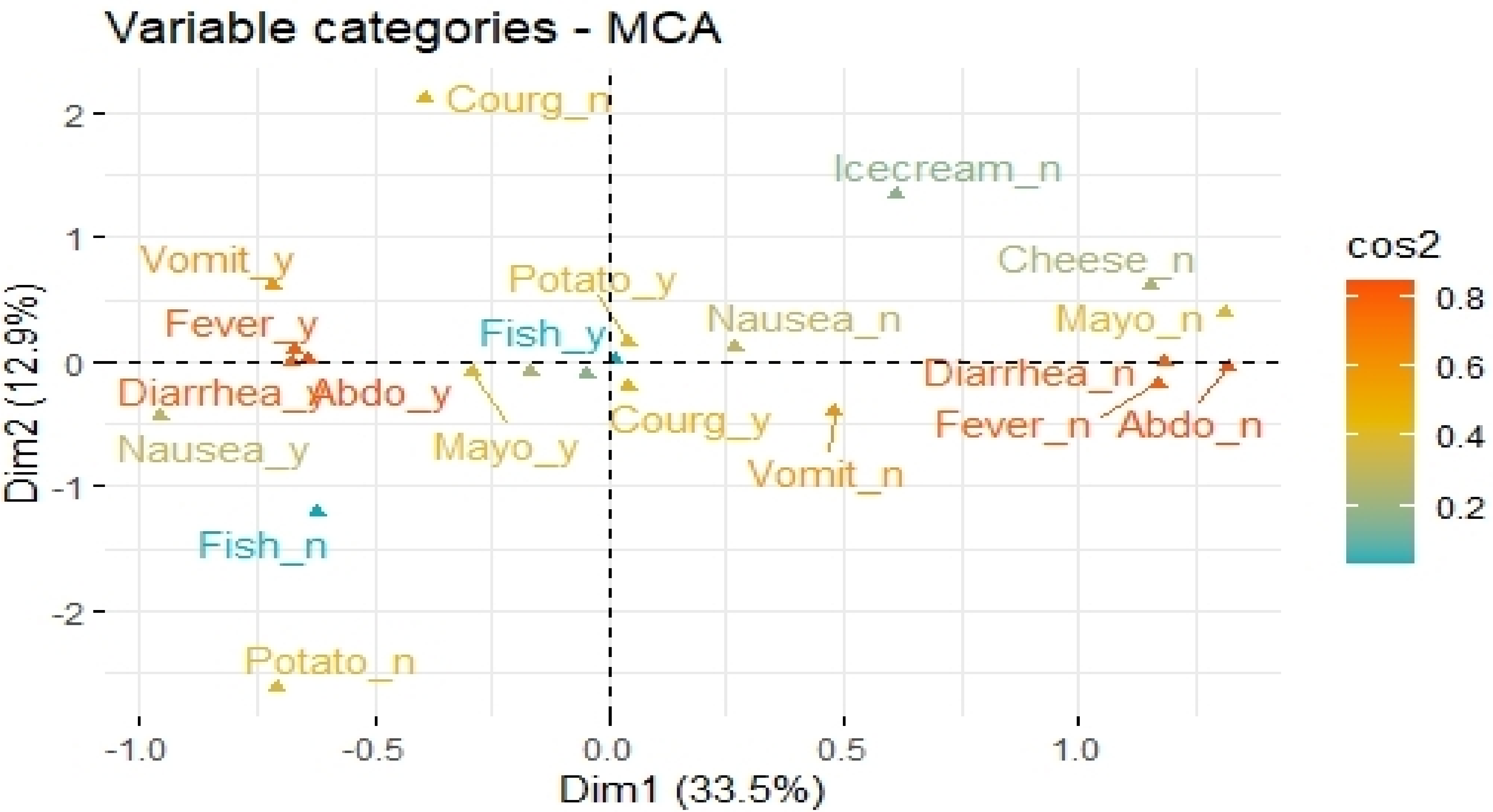


- The plot identifies variables that are the most correlated with each dimension. The squared correlations between variables and the dimensions are used as coordinates.
- It can be seen that, the variables Diarrhae, Abdominals and Fever are the most correlated with dimension 1. Similarly, the variables Courgette and Potato are the most correlated with dimension 2.

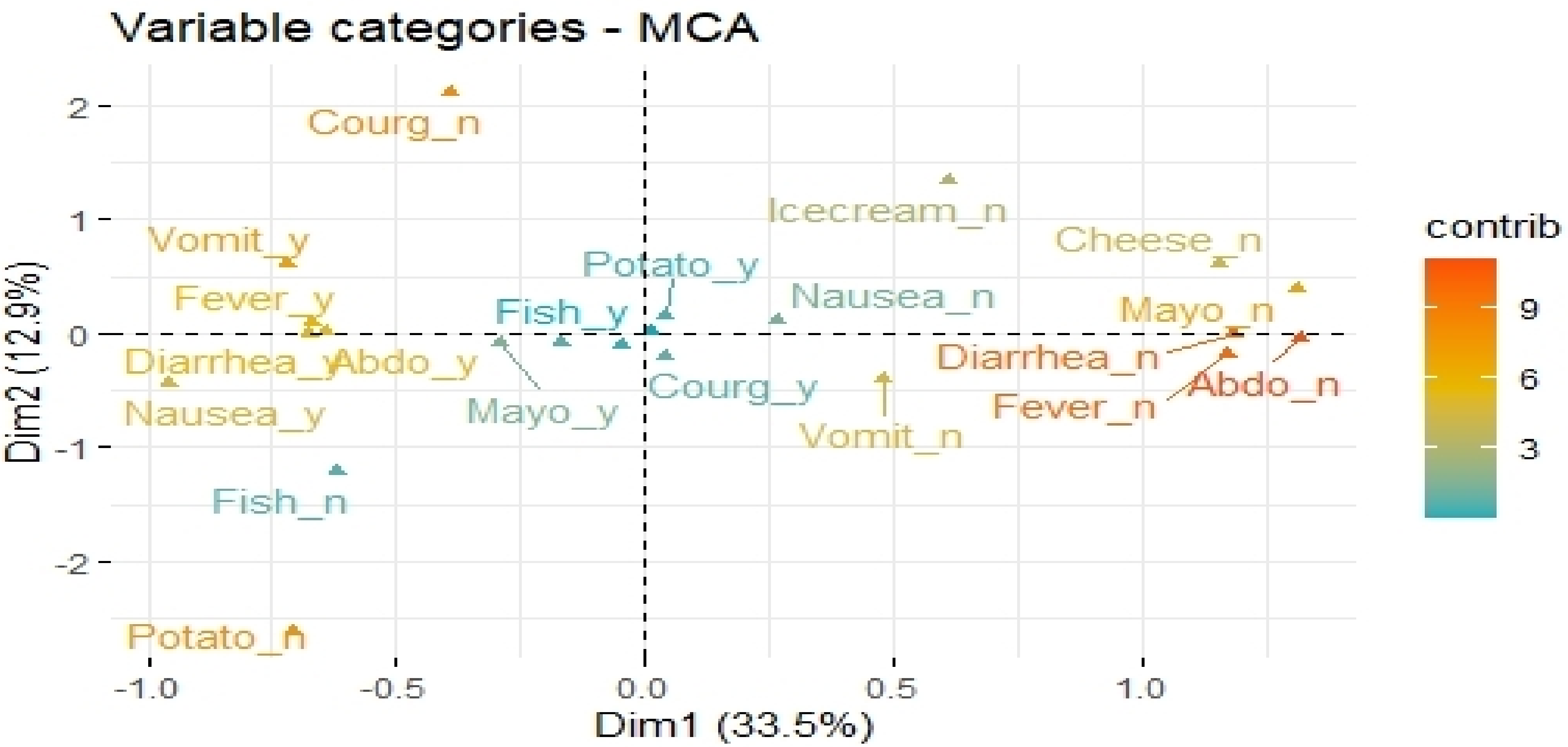


The plot above shows the relationships between variable categories. It can be interpreted as follow:

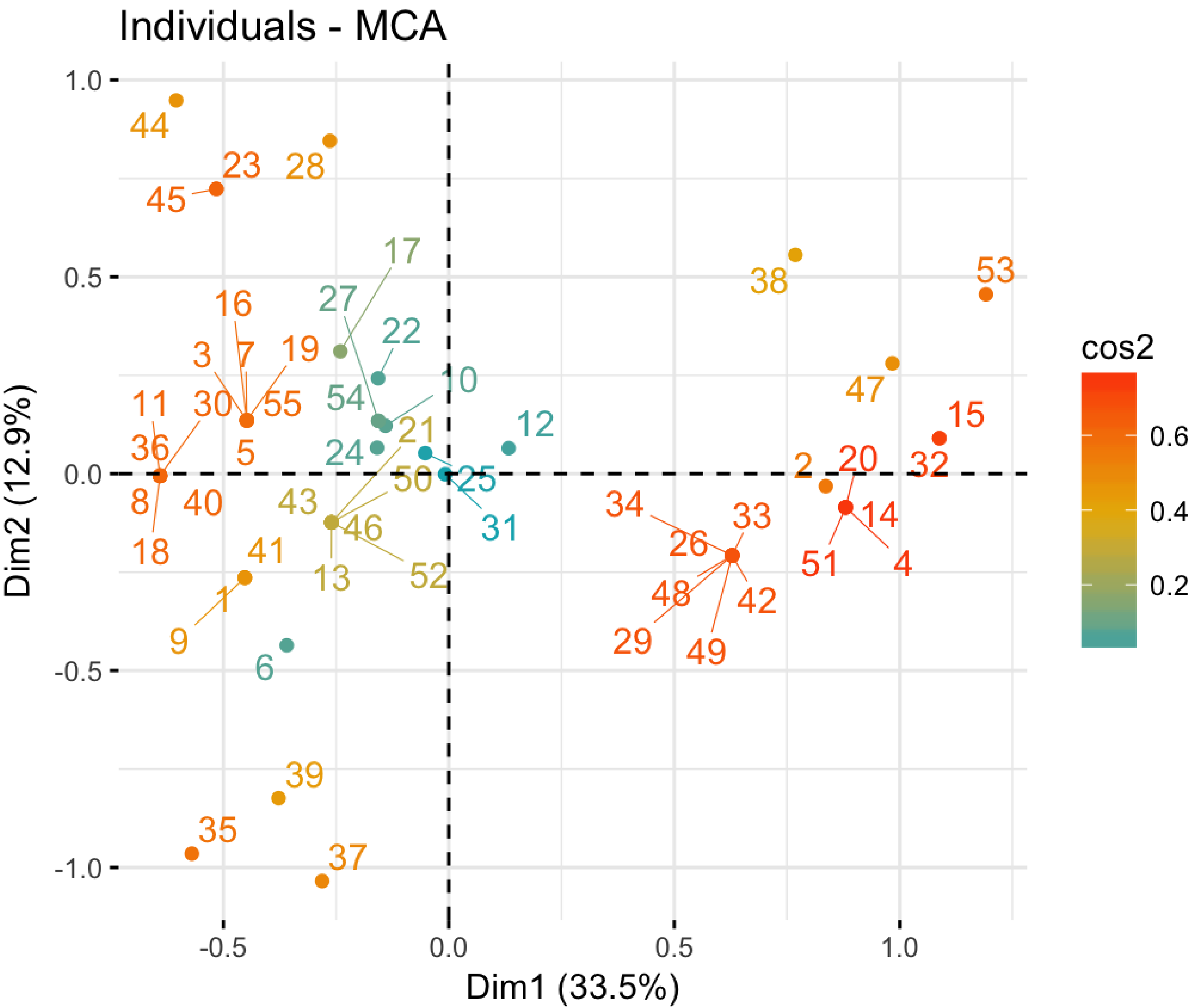
- •Variable categories with a similar profile are grouped together.
- •Negatively correlated variable categories are positioned on opposite sides of the plot origin (opposed quadrants).
- •The distance between category points and the origin measures the quality of the variable category on the factor map. Category points that are away from the origin are well represented on the factor map.



Variable categories - MCA
Courg_n
Icecream_n
Cheese_n
Vomit_y
Potato_y
Nausea_n
Mayo_n
Fever_y
Fish_y
Diarrhea_n
Diarrhea_y
Abdo_y
Fever_n
Abdo_n
Courg_y
Vomit_n
Nausea_y
Mayo_y
Fish_n
Potato_n
Dim2 (12.9%)
Dim1 (33.5%)
2
1
0
-1
-2
-1.0
-0.5
0.0
0.5
1.0
cos2
0.8
0.6
0.4
0.2

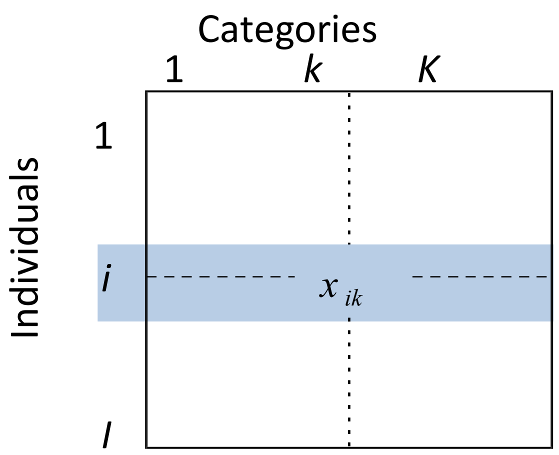


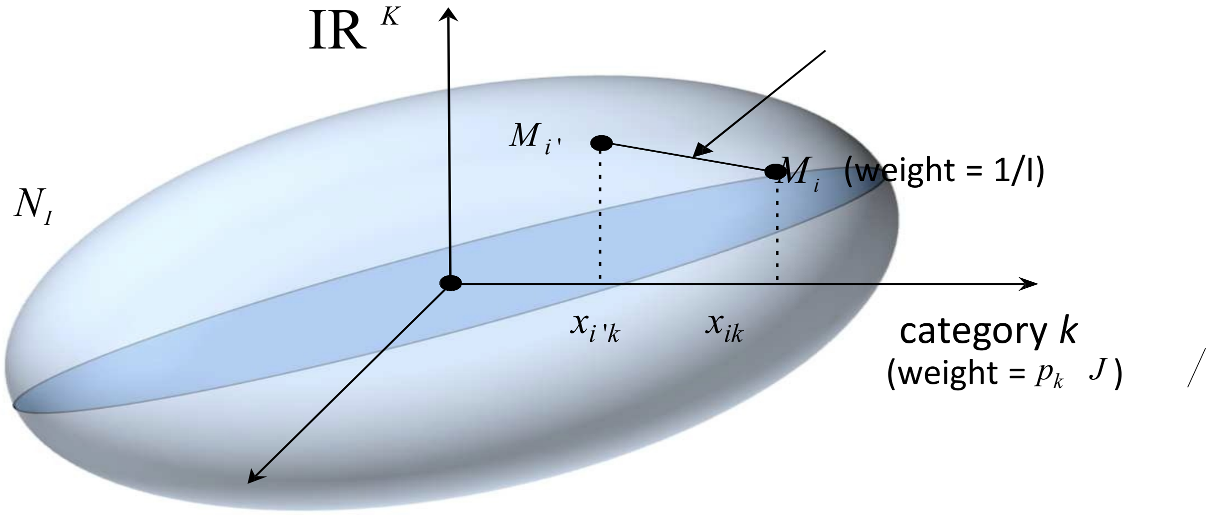
The plot gives us an idea of what pole of the dimensions the categories are contributing to. It is evident that the categories Abdo_n, Diarrhea_n, Fever_n and Mayo_n have an important contribution to the positive pole of the first dimension, while the categories Fever_y and Diarrhea_y have a major contribution to the negative pole of the first dimension; etc, ….


Individuals - MCA
Dim1 (33.5%)
Dim2 (12.9%)
-0.5
0.0
0.5
1.0
-1.0
-0.5
0.0
0.5
1.0
cos2
0.6
0.4
0.2
44
23
28
45
17
27
22
16
3
7
19
10
54
11
30
55
21
12
36
24
5
50
25
8
40
43
46
31
18
41
13
52
1
9
6
38
53
47
15
2
20
32
34
33
14
26
51
4
48
42
29
49
39
35
37

Point cloud of individuals

Indicator matrix

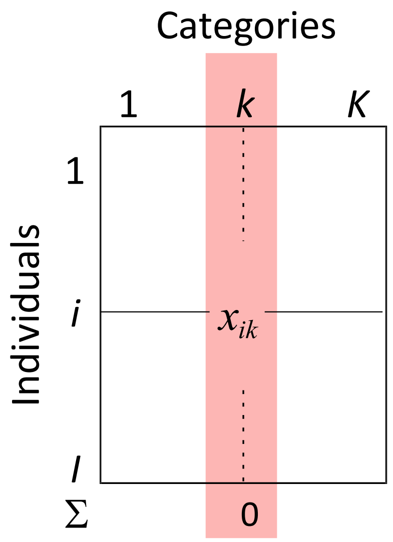


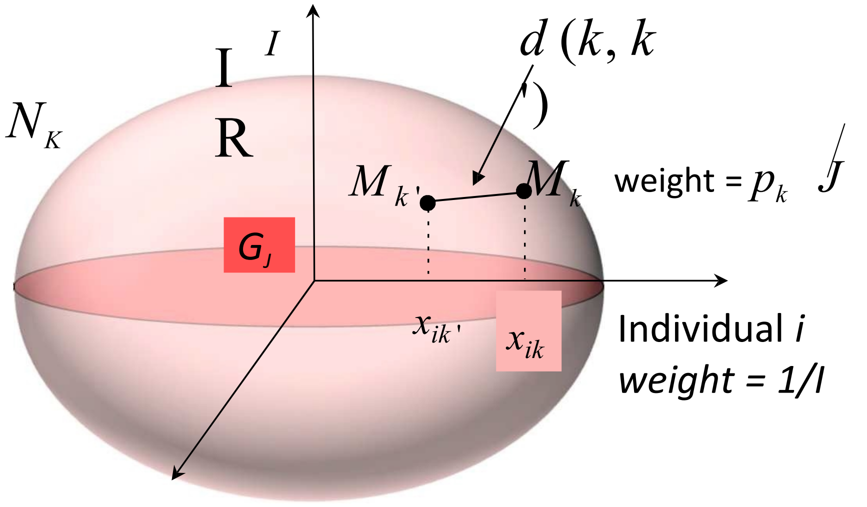


$$d^2_{i,i'} = \sum_{k=1}^{K} \frac{p_k}{J}(x_{ik}-x_{i'k})^2 = \sum_{k=1}^{K}\frac{p_k}{J}\left(\frac{y_{ik}}{p_k}-\frac{y_{i'k}}{p_k}\right)^2 = \frac{1}{J}\sum_{k=1}^{K}\frac{1}{p_k}(y_{ik}-y_{i'k})^2$$

- 2 individuals with same categories : distance = 0
- 2 individuals with many shared categories : small distance
- 2 individuals, only 1 with a rare category : large distance to indicate this
- 2 individuals share rare category : small distance to indicate this shared specificity

Point cloud of categories





$$Var(k) = d^2(k, O) = \sum_{i=1}^{I} \frac{1}{I} x_{ik}^2 = \sum_{i=1}^{I} \left(\frac{y_{ik}}{p_k} - 1 \right)^2 = \frac{1}{p_k} - 1$$

	p_k	$1/2$	$1/5$	$1/10$	$1/101$
	$d(k, O)$	1	2	3	10
$(\text{si } J = 10)$	$Inertia(k)$	0.05	0.08	0.09	0.099

$$Inertia(k) = \frac{p_k}{J} d^2(k, O) = \frac{1 - p_k}{J}$$

$$d^2(k, k') = \sum_{i=1}^{I} \left(\frac{y_{ik}}{p_k} - \frac{y_{ik'}}{p_{k'}} \right)^2 = \frac{p_k + p_{k'} - 2p_{kk'}}{p_k p_{k'}}$$

Conclusion

- MCA is the best factor analysis method for tables of individuals with qualitative variables
- Eigenvalues represent the means of squared correlation ratios
- The values of these squared links are particularly important when there are lots of variables
- MCA can be use to pre-treat data before doing classification