Correspondence Analysis and Text Data

François Husson

Department of Applied Mathematics - Rennes Agrocampus

husson@agrocampus-ouest.fr

From text to a word frequency table

PCA	; This week we have for you three videos which together present the main details of principal component analysis Principal component analysis is a set of tools which allow us to study and visualize large data sets We method from a theoretical as well as a practical point of view The outline of this week's work is as follows We will first define the types of data we can use principal component analysis on then we'll look at examples of principal component analysis can be performed Then we'll define some useful notation Next we'll focus on the individuals then on the variables At the end we will spend some time looking at how to interpret the results a
CA	; This week we have for you 5 course videos on correspondence analysis In the videos we will see the following first we start by describing the data giving a little notation and considering questions to ask when runnin analysis We'll see that the main point of correspondence analysis is studying the links between pairs of qualitative variables This really means looking at the difference between the given data and what it would be like i We're therefore going to see how the analysis captures deviation from independence Our reasoning will mainly be geometrical creating point clouds for the rows and point clouds for the columns These clouds will be rep factor analysis In practice this means projecting onto planes We will also have a look at percentages of inertia From this point of view correspondence analysis is no different from other methods of factor analysis like pr
MCA	; This week we have four videos for you on multiple correspondence analysis MCA for short We'll have a look at the main features of the method using a specific example to guide us along the way The videos look at th First we describe the types of data MCA can be used for With this data in mind we will look at what our goals are and what issues we may have This will lead us to ways to manipulate the data table In multiple correspo any principal component methods we are going to build point clouds including point clouds of the rows and point clouds of the columns In the MCA context we are going to have a point cloud of individuals and a point cl
Clustering	; This week we're going to look at classification methods including hierarchical classification and a partitioning method called k means The course videos for this week get into the following things After a brief introductio for classification and the goals of classification we are going to have a look at some general principles of classification and in particular hierarchical classification We'll have questions like what criteria to use Which algor take a close look at a partitioning method the well known k means algorithm Following this we'll get into how we can use classification and k means at the same time and how to do classification with high dimensional d

From text to a word frequency table

PCA ; This week we have for you three videos which together present the main details of principal component analysis Principal component analysis is a set of tools which allow us to study and visualize large data sets We method from a theoretical as well as a practical point of view The outline of this week's work is as follows We will first define the types of data we can use principal component analysis on then we'll look at examples of principal component analysis can be performed Then we'll define some useful notation Next we'll focus on the individuals then on the variables At the end we will spend some time looking at how to interpret the results a

CA ; This week we have for you 5 course videos on correspondence analysis In the videos we will see the following first we start by describing the data giving a little notation and considering questions to ask when runnin analysis We'll see that the main point of correspondence analysis is studying the links between pairs of qualitative variables This really means looking at the difference between the given data and what it would be like i We're therefore going to see how the analysis captures deviation from independence Our reasoning will mainly be geometrical creating point clouds for the rows and point clouds for the columns These clouds will be rep factor analysis In practice this means projecting onto planes We will also have a look at percentages of inertia From this point of view correspondence analysis is no different from other methods of factor analysis like pr

MCA ; This week we have four videos for you on multiple correspondence analysis MCA for short We'll have a look at the main features of the method using a specific example to guide us along the way The videos look at th First we describe the types of data MCA can be used for With this data in mind we will look at what our goals are and what issues we may have This will lead us to ways to manipulate the data table In multiple correspo any principal component methods we are going to build point clouds including point clouds of the rows and point clouds of the columns In the MCA context we are going to have a point cloud of individuals and a point cl

Clustering ; This week we're going to look at classification methods including hierarchical classification and a partitioning method called k means The course videos for this week get into the following things After a brief introductio for classification and the goals of classification we are going to have a look at some general principles of classification and in particular hierarchical classification We'll have questions like what criteria to use Which algo take a close look at a partitioning method the well known k means algorithm Following this we'll get into how we can use classification and k means at the same time and how to do classification with high dimensional d

1852 rows

	PCA	CA	MCA	Clustering
able	2	1	1	2
above	0	0	0	1
absolute	1	0	0	5
absolutely	0	1	0	0
acceptable	0	0	0	1
access	1	0	0	0
accident	0	0	1	0
accord	0	0	1	0
according	3	0	2	0
account	0	0	2	1
…	…	…	…	…

Some data pre-treatment steps

To obtain the final word frequency table to analyze, we :

- remove connecting words like : for example, then, therefore, and, etc.
- group words with the same root or the same conjugations together (e.g., reduced, reduction, reduces)
- group singular and plurals together
- remove words used nine times or less

Some data pre-treatment steps

To obtain the final word frequency table to analyze, we :

- remove connecting words like : for example, then, therefore, and, etc.
- group words with the same root or the same conjugations together (e.g., reduced, reduction, reduces)
- group singular and plurals together
- remove words used nine times or less

Distributional equivalence is very useful in text analysis

Some data pre-treatment steps

To obtain the final word frequency table to analyze, we :

- remove connecting words like : for example, then, therefore, and, etc.
- group words with the same root or the same conjugations together (e.g., reduced, reduction, reduces)
- group singular and plurals together
- remove words used nine times or less

Distributional equivalence is very useful in text analysis

$\Longrightarrow$ 246 words, $n = 8821$ total occurrences

Word frequency table

246 words $\times$
4 methods

	PCA	CA	MCA	Clustering
variables	132	20	93	49
individuals	76	7	110	98
between	62	63	50	48
dimension	73	51	45	3
data	54	41	32	40
inertia	9	65	43	47
variable	46	13	65	38
first	50	40	38	30
point	32	53	53	10
analysis	16	77	38	14
categories	1	22	107	7
class	1	0	2	118
table	18	43	47	7
cloud	43	34	32	0
...	...	...	...	...

Word frequency table

246 words × 4 methods

	PCA	CA	MCA	Clustering
variables	132	20	93	49
individuals	76	7	110	98
between	62	63	50	48
dimension	73	51	45	3
data	54	41	32	40
inertia	9	65	43	47
variable	46	13	65	38
first	50	40	38	30
point	32	53	53	10
analysis	16	77	38	14
categories	1	22	107	7
class	1	0	2	118
table	18	43	47	7
cloud	43	34	32	0
...	...	...	...	...

Simple idea : look for high-frequency words ?
Example : the word *variables* is used 132 times in PCA text
BUT *variables* is the most used word overall (294 times), so is it really representative of PCA ?

Analysis of word frequency tables using CA : some history

- First applications of CA (early 1960s)
- Jean-Paul Benzécri, University professor in Rennes

- Ph.D. thesis of Brigitte Escofier (1965) : transition formulas, reconstitution formulas, etc.
- Characters from the play Phèdre, verb-nous associations, rhyme associations, etc.

Inertia and percentage of inertia

$n = 8821\,;\ \chi^2 = 6985.026$ p-value $= <\ 10^{-160}$

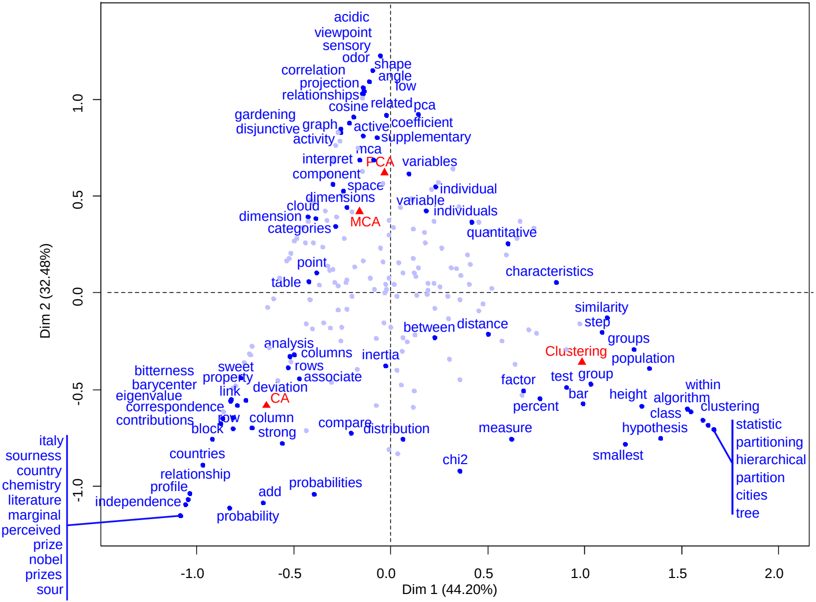
$\Phi^2 = \frac{6985.026}{8821} = 0.792$ high Φ^2 (maximum possible $\Phi^2 = 3$)

$\Longrightarrow$ strong association between words and methods (very far from independence)

	eigenvalue	% inertia
dim 1	0.35	44.20
dim 2	0.26	32.48
dim 3	0.18	23.32

Fairly large eigenvalues

Simultaneous representation of methods and words



Interpreting the results

Terms exclusive to certain methods are superposed

1st axis (inertia $= 0.35$) :

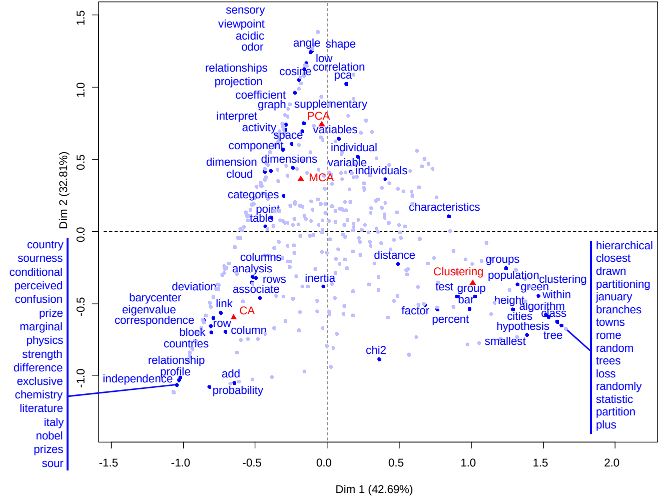
- clear division between the factor methods and clustering
- specific words used in factor analysis are to the left
- specific words used in clustering are to the right

2nd axis (inertia $= 0.26$) :

- separates the 3 factor analysis methods
- CA uses terms common to PCA and MCA

Stability of results with resect to cut-off

Words included : used at least 5 times



$\Longrightarrow$ Stable representation