Analyzing categorization data

Marine Cadoret, Sébastien Lê, Jérôme Pagès

AGROCAMPUS OUEST, France

International Federation of Classification Societies

March 15th 2009, Dresden

Introduction

- Categorization consists in grouping objects in function of their resemblances.
- Following this task, a verbalization task can also be asked to describe the groups (“qualified” categorization).

Data

- 98 consumers carried out a “qualified” categorization on 12 luxury perfumes:


Angel


Lolita Lempicka

L’instant


Cinéma


J’adore (ET)


J’adore (EP)


Shalimar


Aromatics Elixir


Coco Mademoiselle

Chanel n°5

Pure Poison


Pleasures

« gourmand,
vanilla, wooded »

« spicy, aldehyde »

« white flower,
vanilla, orange »

« oriental,
showy,
wooded,
Patchouli oil »

« flower,
floral,
green »

Data table (1)

	Shalimar	Aromatics Elixir	Chanel n°5	Angel	Lolita Lempicka	Cinéma	L'instant	Pure Poison	Coco Mademoiselle	Pleasures	J'adore (EP)	J'adore (ET)
Shalimar	98	42	30	21	9	10	13	11	9	6	6	7
Aromatics Elixir	42	98	51	27	6	8	13	12	12	11	12	7
Chanel n°5	30	51	98	15	8	9	10	21	11	14	12	14
Angel	21	27	15	98	36	18	14	10	10	11	11	12
Lolita Lempicka	9	6	8	36	98	42	22	18	21	18	18	18
Cinéma	10	8	9	18	42	98	26	28	30	22	23	24
L'instant	13	13	10	14	22	26	98	25	20	23	28	22
Pure Poison	11	12	21	10	18	28	25	98	33	30	29	28
Coco Mademoiselle	9	12	11	10	21	30	20	33	98	28	28	38
Pleasures	6	11	14	11	18	22	23	30	28	98	38	48
J'adore (EP)	6	12	12	11	18	23	28	29	28	38	98	56
J'adore (ET)	7	7	14	12	18	24	22	28	38	48	56	98

Data usually gathered in a cooccurrences (or dissimilarities) matrix and analyzed by non-metric MDS

Data table (2)

produit	juge 12	juge 13	juge 14	juge 15	juge 16
Angel	1	4	1	5	2
Aromatic Elixir	3	3	5	2	1
Chanel n°5	4	3	4	1	3
Cinéma	2	5	6	4	2
Coco Mademoiselle	1	5	2	4	3
J'adore (EP)	1	6	2	3	3
J'adore (ET)	2	6	2	3	3
L'instant	1	4	6	2	4
Lolita Lempicka	1	5	1	5	2
Pleasures	3	4	6	3	4
Pure Poison	1	1	2	4	4
Shalimar	2	2	3	2	1

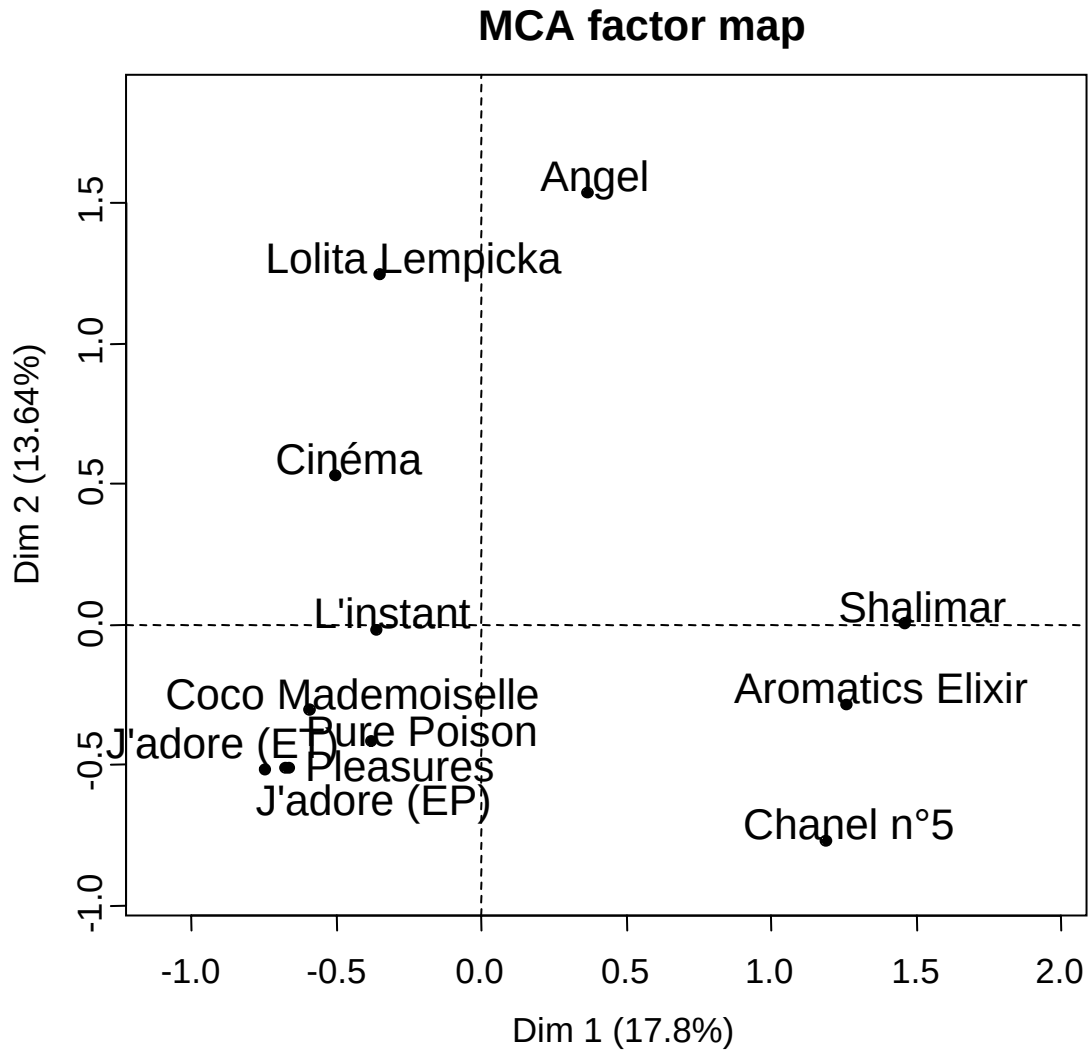
Each consumer can also be considered as a categorical variable

Data table (2)

produit	juge 12	juge 13	juge 14	juge 15	juge 16
Angel	fleuri doux	fruité fort	vanillé épicé esprit des îles	à manger sucré	nourriture épice
Aromatic Elixir	fort homme	capiteux grand-mère	rude fort	le vieux	ménager cire
Chanel n°5	Gr 4	capiteux grand-mère	toilettes	savon	connu classique
Cinéma	fleuri artificiel herbe	fruité moyen	sucré	doux	nourriture épice
Coco Mademoiselle	fleuri doux	fruité moyen	douceur fleuri	doux	connu classique
J'adore (EP)	fleuri doux	sucré faible	douceur fleuri	fleuri	connu classique
J'adore (ET)	fleuri artificiel herbe	sucré faible	douceur fleuri	fleuri	connu classique
L'instant	fleuri doux	fruité fort	sucré	le vieux	fleuri
Lolita Lempicka	fleuri doux	fruité moyen	vanillé épicé esprit des îles	à manger sucré	nourriture épice
Pleasures	fort homme	fruité fort	sucré	fleuri	fleuri
Pure Poison	fleuri doux	acidulé désodorisant	douceur fleuri	doux	fleuri
Shalimar	fleuri artificiel herbe	fort lavande eau de cologne	renfermé agressif	le vieux	ménager cire

Let’s run MCA on this data table!

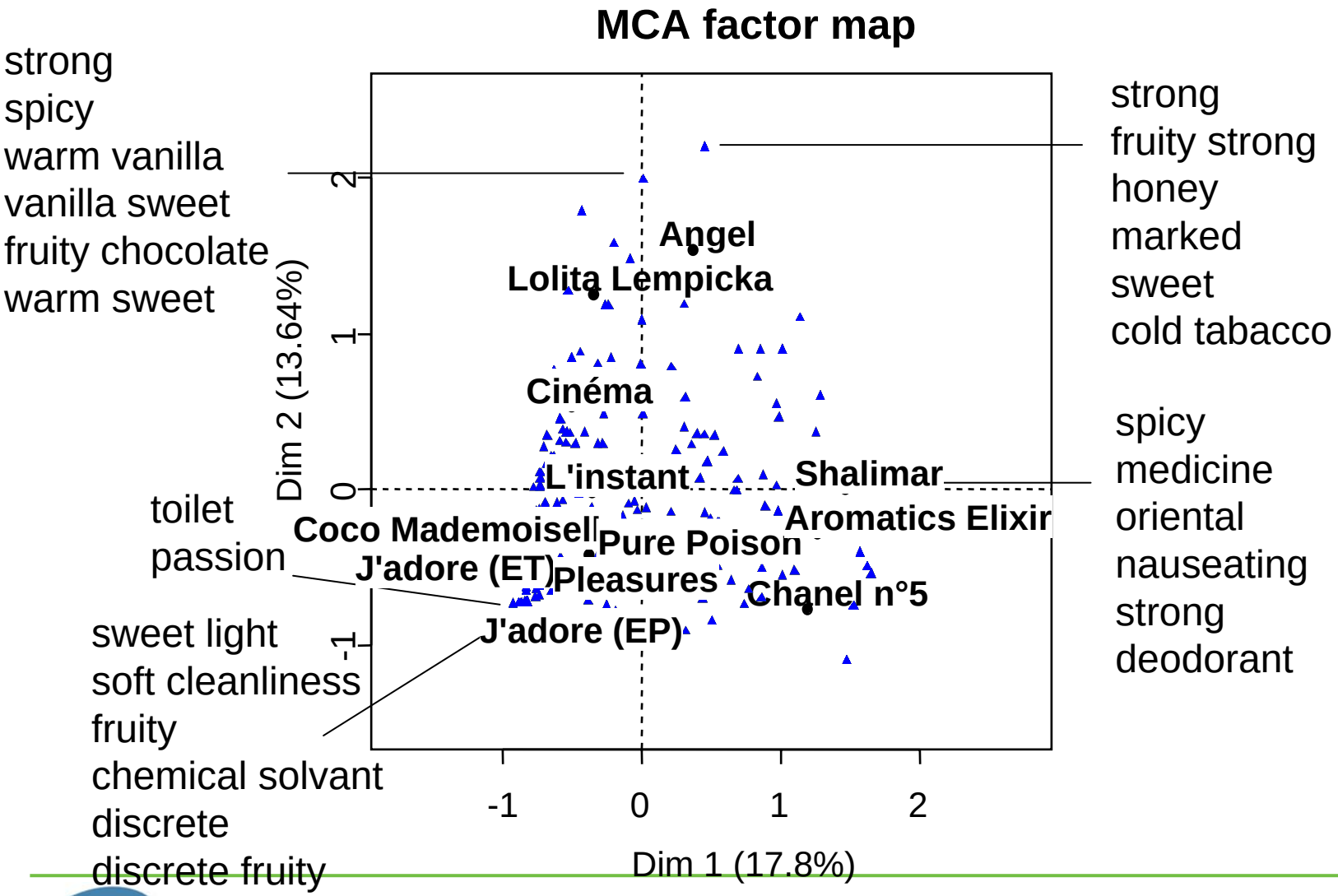
Representation of the perfumes



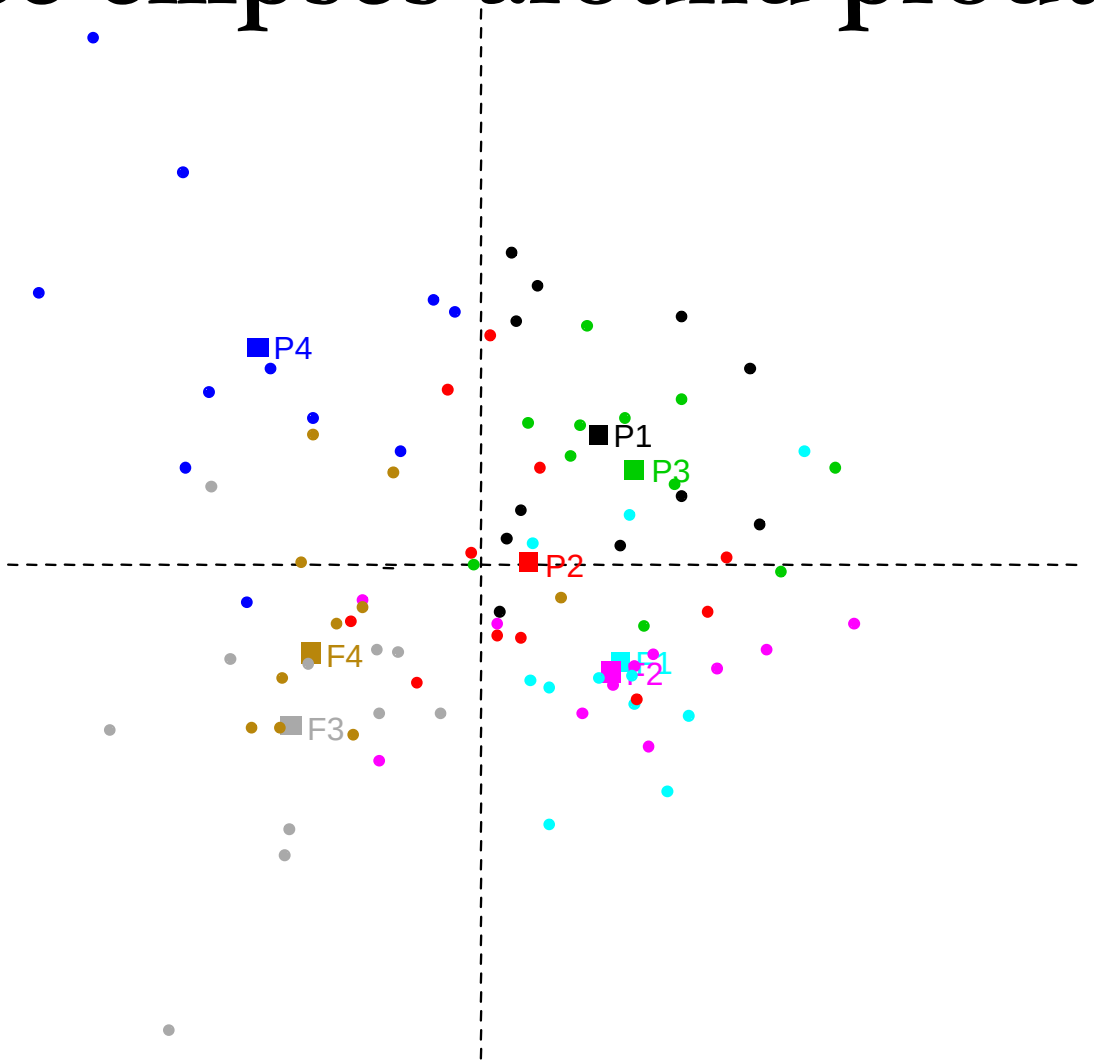
Co-occurrences matrix

	Shalimar	Aromatics Elixir	Chanel n°5	Angel	Lolita Lempicka	Cinéma	L'instant	Pure Poison	Coco Mademoiselle	Pleasures	J'adore (EP)	J'adore (ET)
Shalimar	98	42	30	21	9	10	13	11	9	6	6	7
Aromatics Elixir	42	98	51	27	6	8	13	12	12	11	12	7
Chanel n°5	30	51	98	15	8	9	10	21	11	14	12	14
Angel	21	27	15	98	36	18	14	10	10	11	11	12
Lolita Lempicka	9	6	8	36	98	42	22	18	21	18	18	18
Cinéma	10	8	9	18	42	98	26	28	30	22	23	24
L'instant	13	13	10	14	22	26	98	25	20	23	28	22
Pure Poison	11	12	21	10	18	28	25	98	33	30	29	28
Coco Mademoiselle	9	12	11	10	21	30	20	33	98	28	28	38
Pleasures	6	11	14	11	18	22	23	30	28	98	38	48
J'adore (EP)	6	12	12	11	18	23	28	29	28	38	98	56
J'adore (ET)	7	7	14	12	18	24	22	28	38	48	56	98

Representation of the words

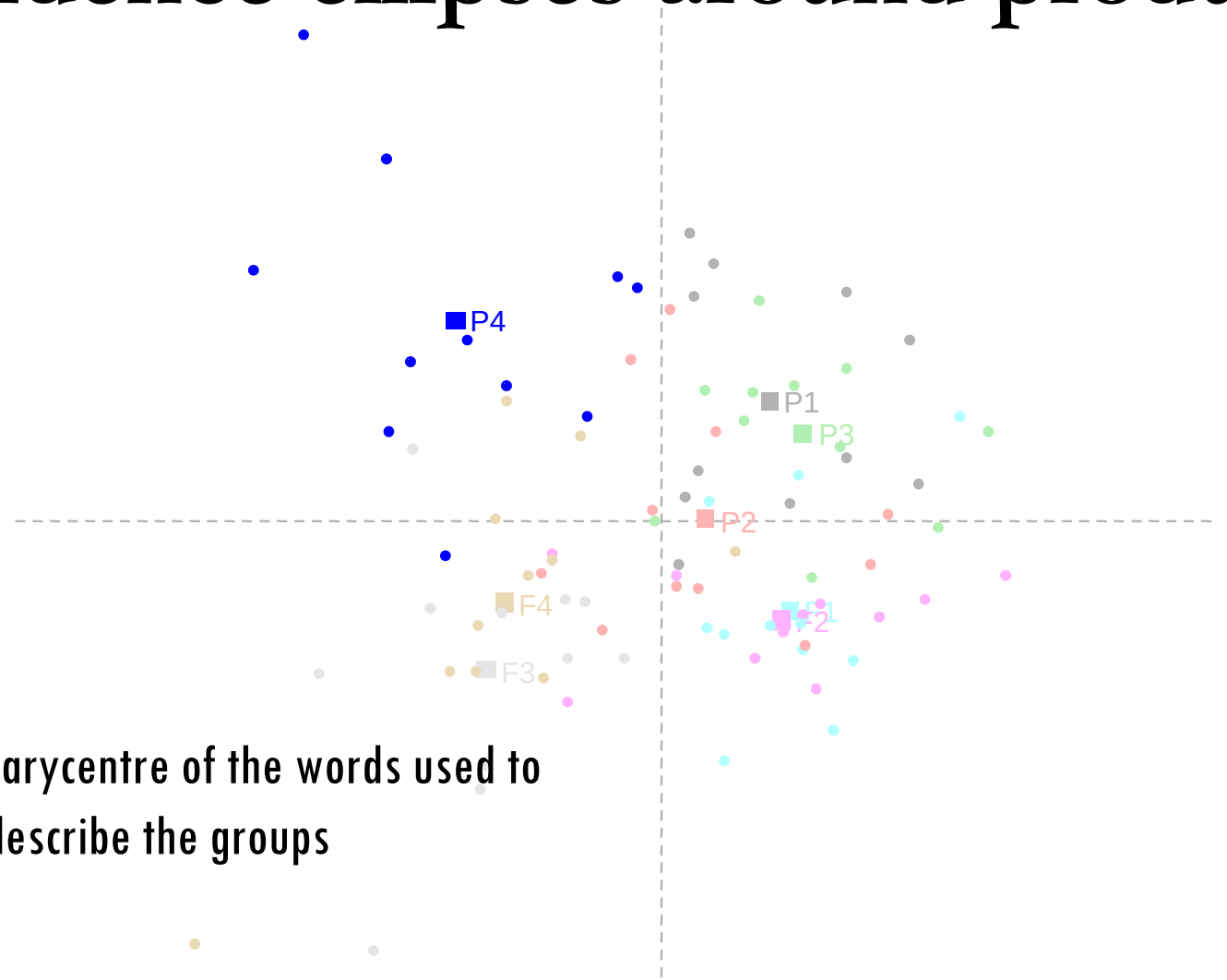


Confidence ellipses around products



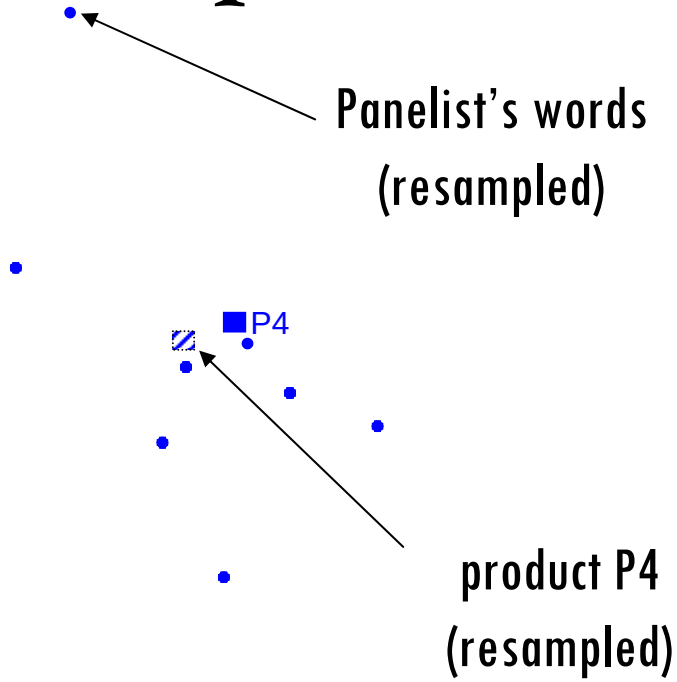
Superimposed representation of the products and their descriptions

Confidence ellipses around products

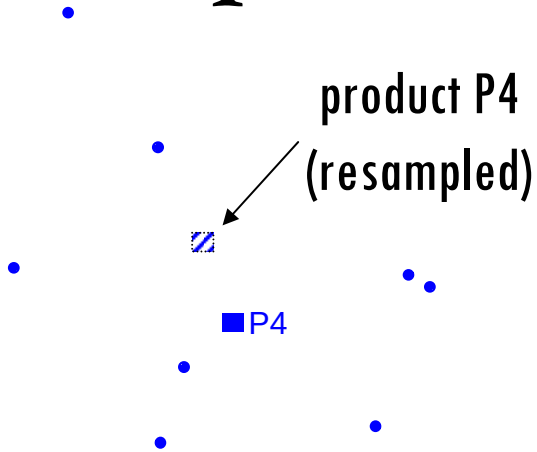


P4 is at the barycentre of the words used to describe the groups

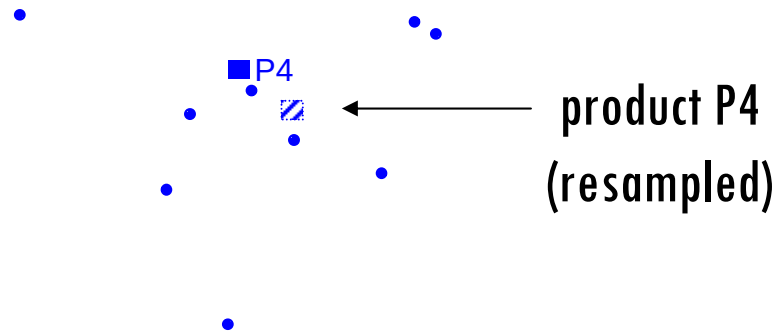
Confidence ellipses around products



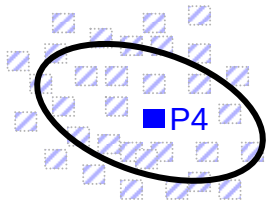
Confidence ellipses around products



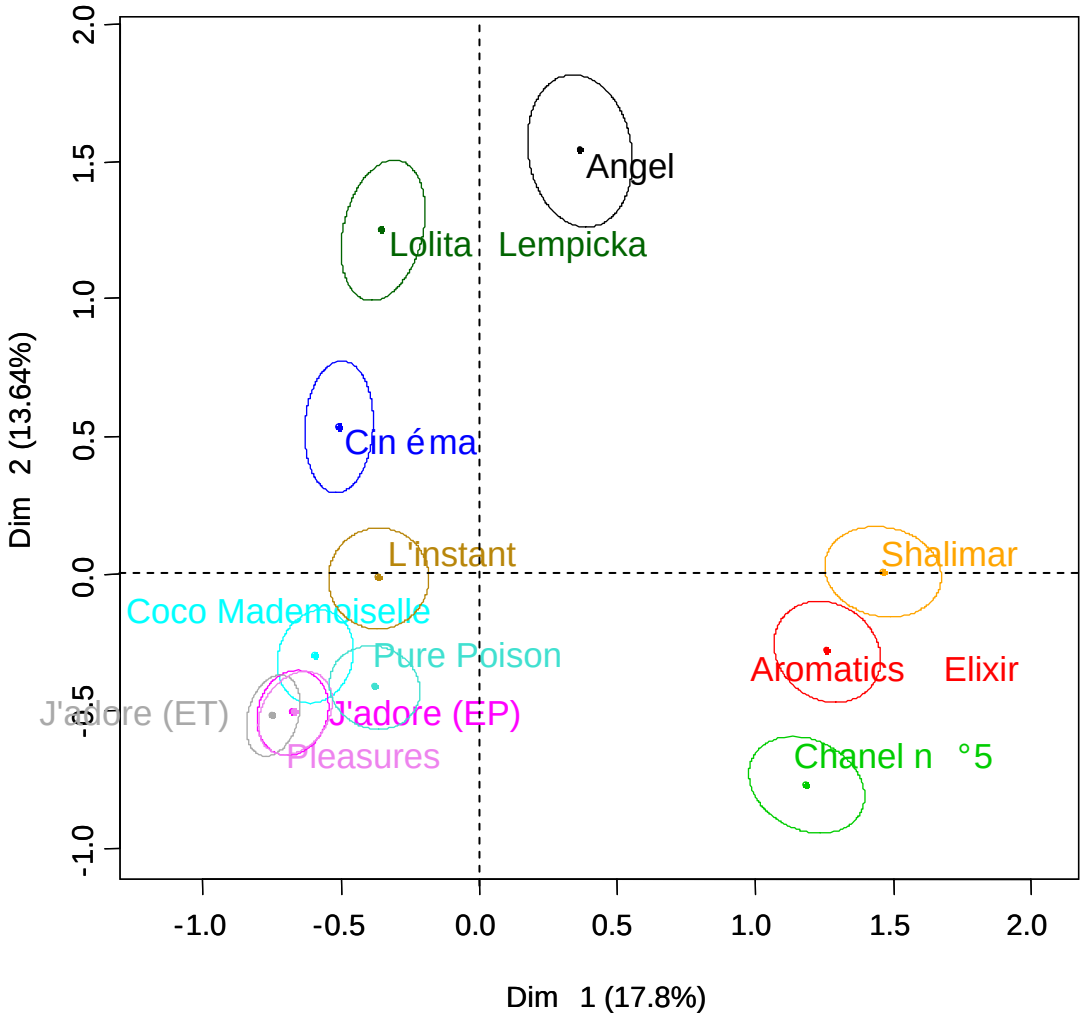
Confidence ellipses around products



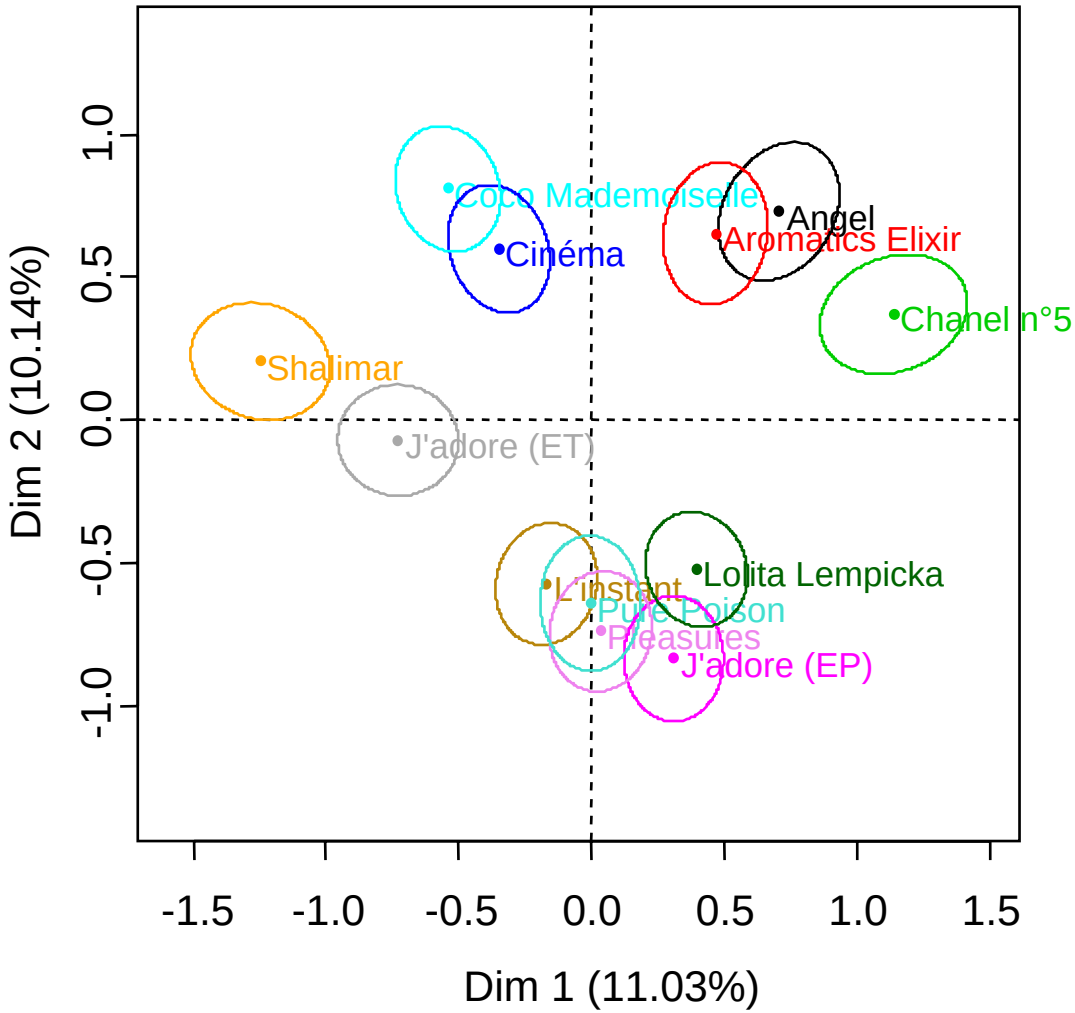
Confidence ellipses around products



Confidence ellipses for perfumes data



Confidence ellipses for random data



Explanation

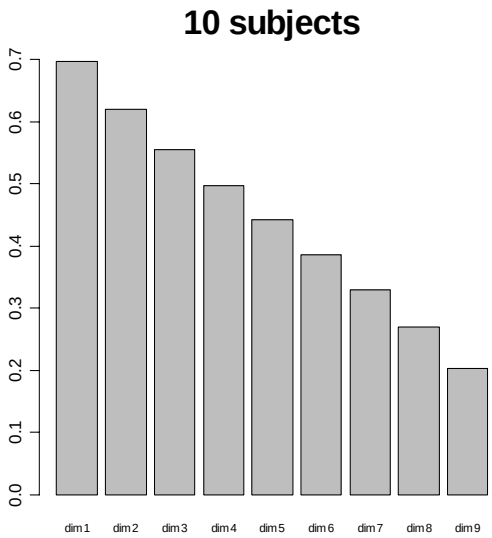
- Number of columns >> number of rows
- Automatic production of common dimensions
- Looking for an indicator of consensus between subjects

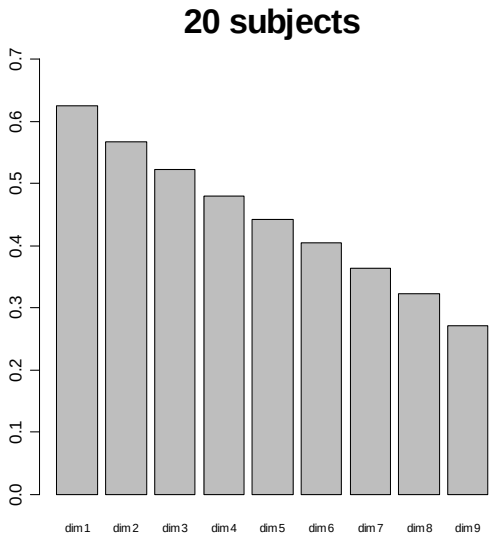
Significance of the results

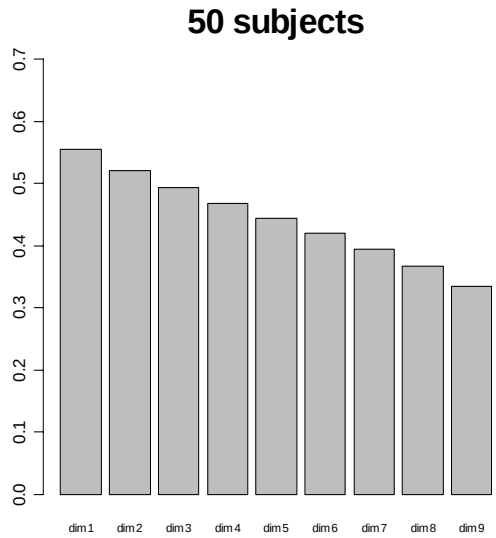
- H_0: absence of consensus
- Indicator: first eigenvalue
- Evolution of the indicator under H_0:

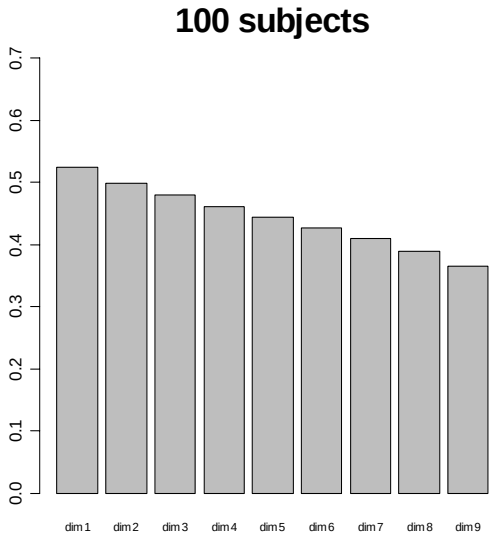
		Number of subjects			
		10	**20**	**50**	**100**
Number of products	**10**	0,69566	0,624637	0,557816	0,524184
	20	0,471051	0,389234	0,319656	0,286282
	50	0,306195	0,228844	0,167827	0,14006
	100	0,236413	0,16401	0,109948	0,086396

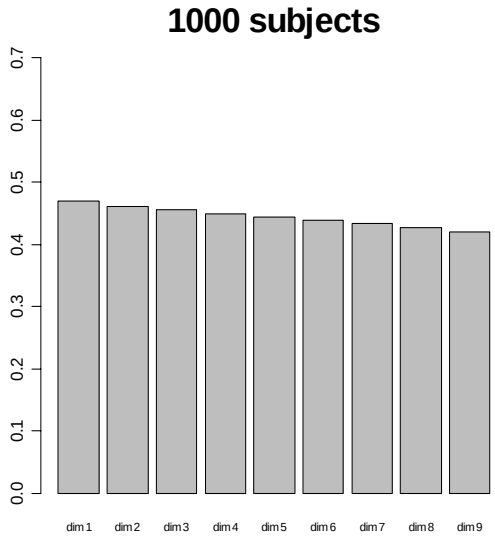
Bar plots of the eigenvalues (10 products)







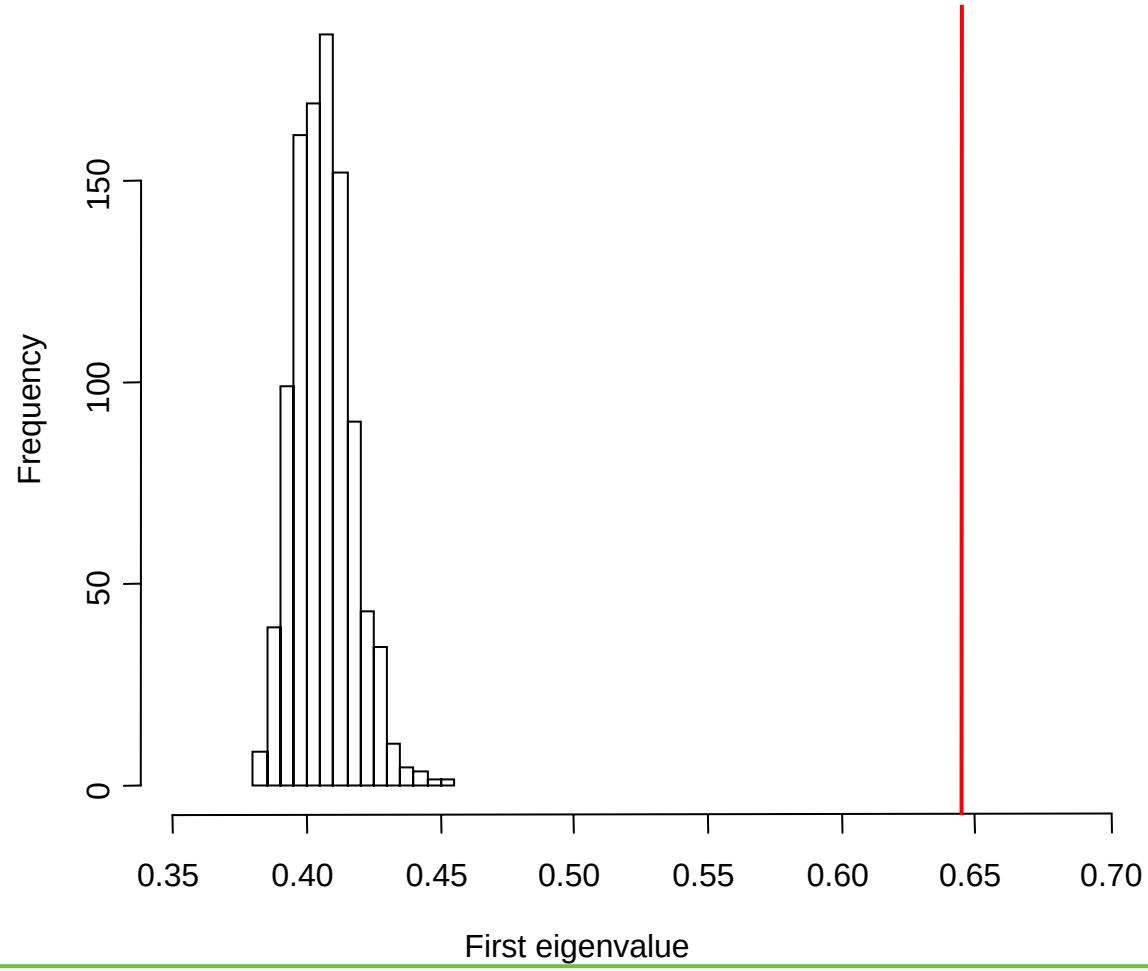




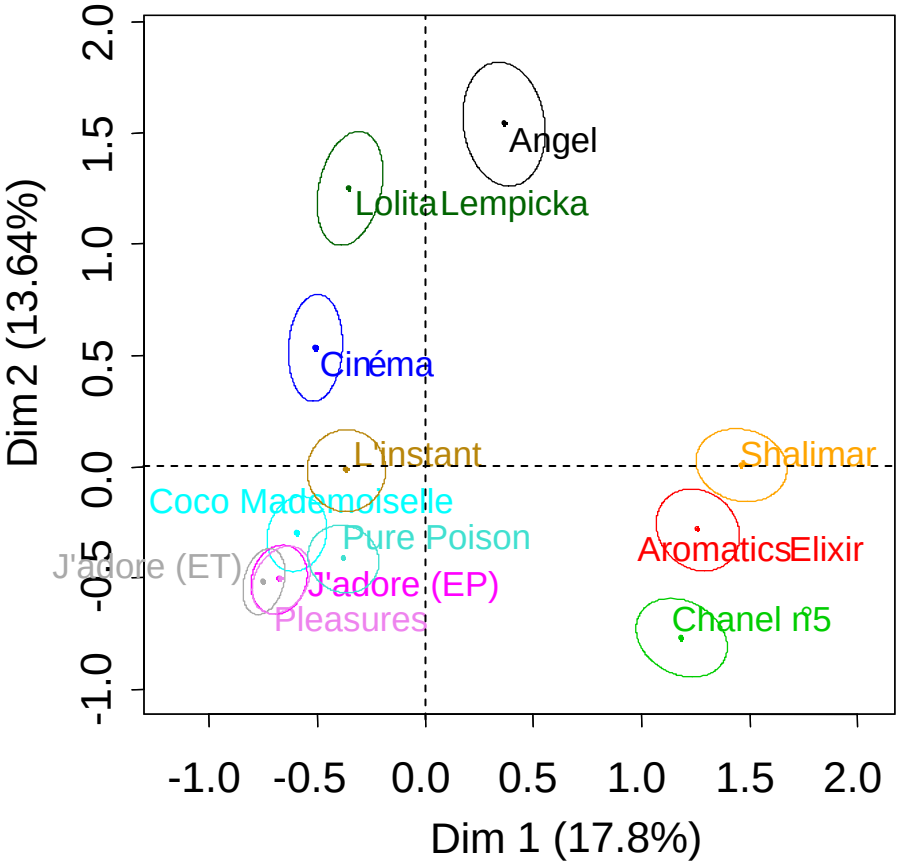
Significance of the indicator for a given data table

- Calculate the p-value associated to the first eigenvalue of the MCA:
  1. Repeat a great number of times:
     1. Independent row permutations within each column
     2. Calculate the first eigenvalue associated to the permutated table
  2. Distribution of the eigenvalues under H_0
  3. Identify the observed eigenvalue in this distribution to get the p-value

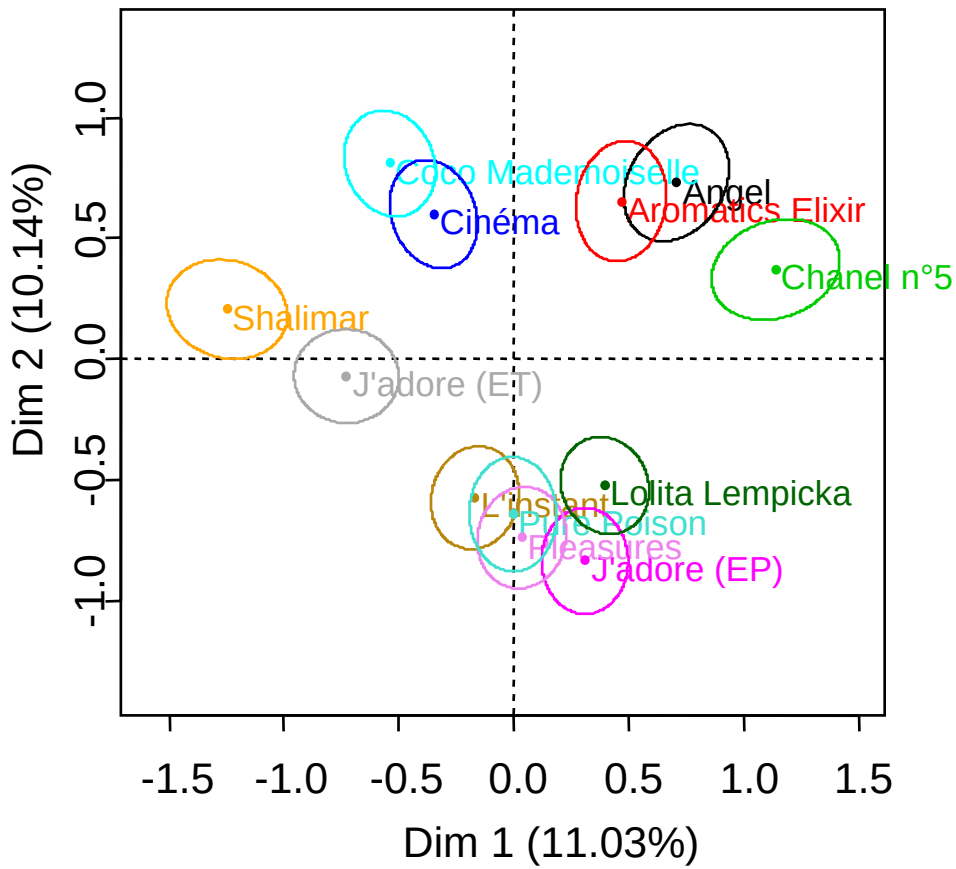
Significance of the first dimension for perfumes data



Confidence ellipses



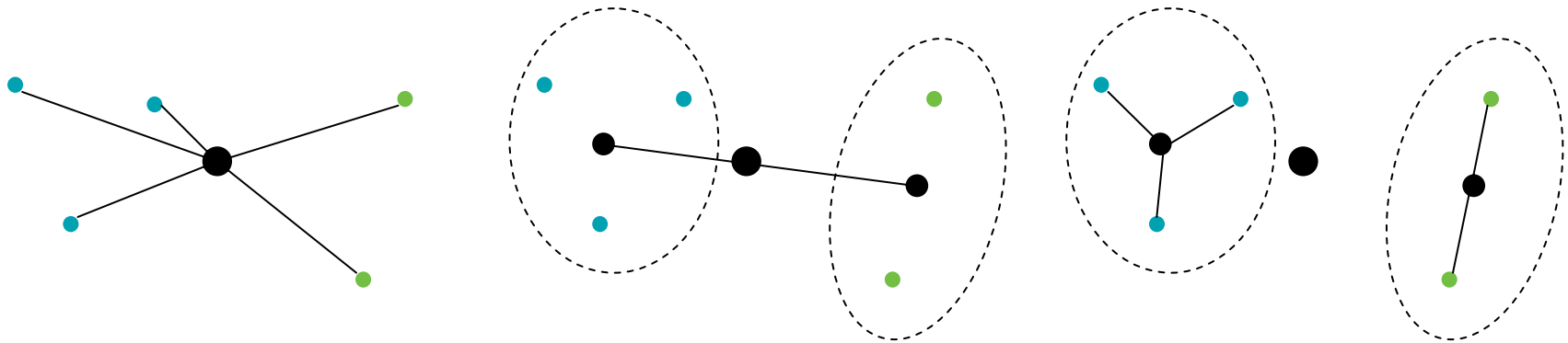
Perfume data



Random data

Second empirical indicator

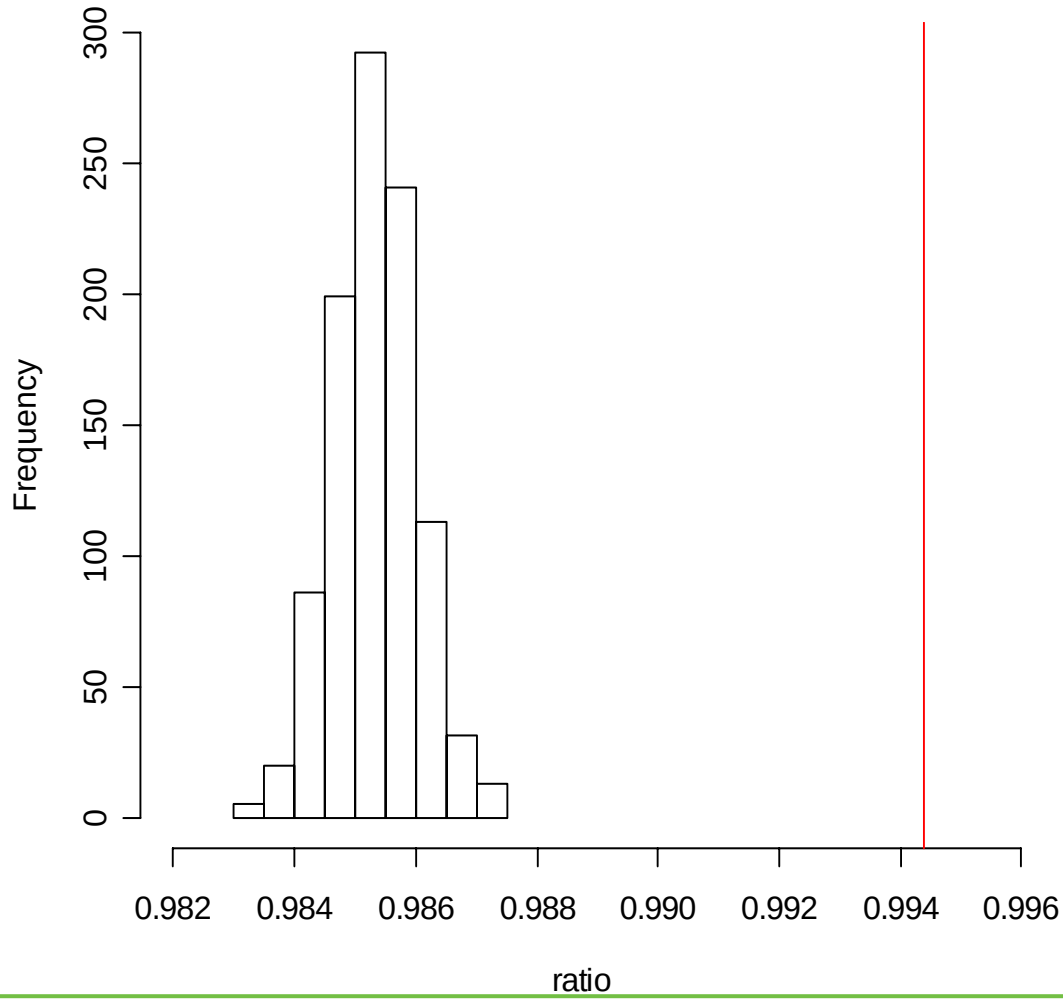
- Ellipses overlapping



Total inertia = Between inertia + Within inertia

- Calculate the inertia ratio: $0 \leq \frac{\text{between inertia}}{\text{total inertia}} \leq 1$

Significance of the inertia ratio for perfumes data



Conclusion

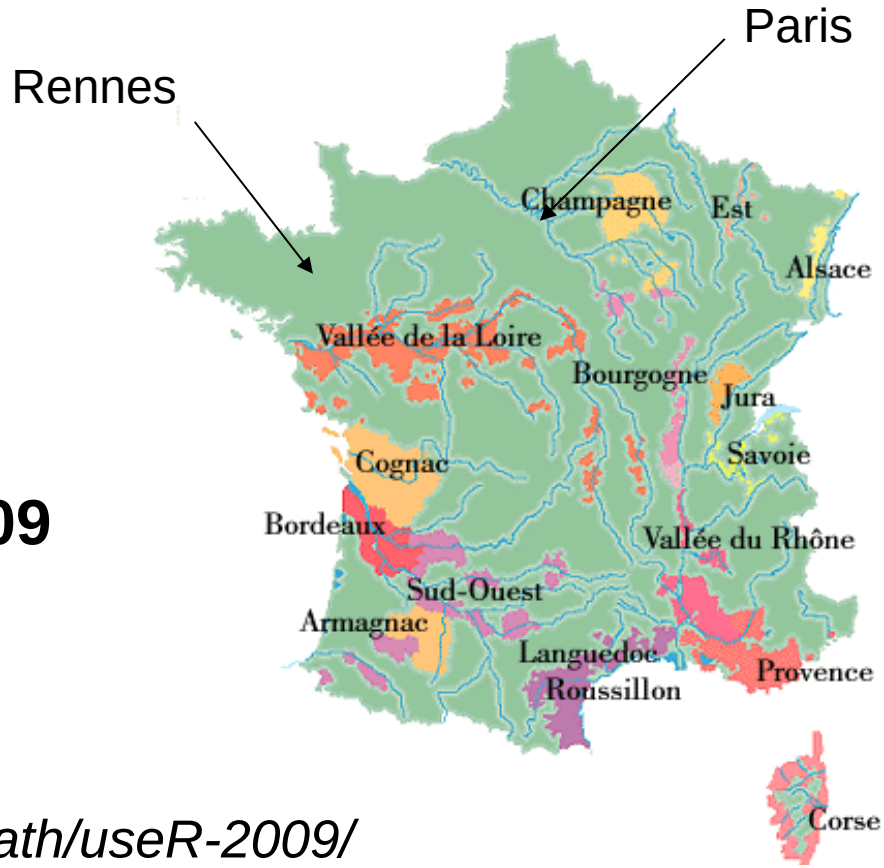
- MCA: suitable method for categorization data
- Sensory data: few rows and many columns
- By construction, many relationships
- External validation

SensoMineR a package for sensory data analysis
Journal of sensory studies (2008)

FactoMineR: an R package for multivariate analysis
Journal of statistical software (2008)

http://www.agrocampus-rennes.fr/math/

The applied mathematics department of Agrocampus organizes



The R User Conference 2009

July 8-10 in Rennes, France

http://www.agrocampus-rennes.fr/math/useR-2009/