

CARME-2003, Barcelona
Correspondence Analysis and Related Methods
(Invited lecture)

Validation procedures in Principal Axes Techniques.

Ludovic Lebart,
CNRS – ENST, Paris, France.
lebart@enst.fr

Validation procedures in Principal Axes Techniques

1. External validation

- Illustrative information
- Multiple comparisons

2. Internal validation

2.1 Analytical validation

2.2 Bootstrap techniques

- Partial bootstrap
- Global bootstrap
- Customized bootstrap

I. External validation:

- 1. Main pragmatic validation for clustering
 - Dissection *vs* discovery of clusters
- 2. The problem of multiple comparisons
 - B to B From Bonferroni to Bootstrap

- External information could coincide, in some contexts, with the so-called **supplementary (or illustrative) elements** (extra rows or columns of the data table that are projected afterwards onto the principal visualization planes) (Benzécri, Cazes, 1967; Gower, 1966).
- In some other contexts, external information takes the form of **instrumental variables** whose effects on the data must be eliminated beforehand (leading, for instance, to analyse partial correlations instead of correlations).
- In an exploratory approach, numerous supplementary elements could be projected, leading to as many test-values or p-values (statistical parameters expressing the significance of these projections).
- Such procedure must then be protected against repeated measurements or **multiple comparisons effects** (see: Saville, 1990; Westfall and Young, 1993).

II.1 Internal validation : The quagmire of analytical validation!

Example: ASSESSMENT OF RESULTS IN « Principal component analysis »

Distribution of eigenvalues

matrix $\mathbf{S} = \mathbf{X}'\mathbf{X}$ ($p(p+1)/2$ distinct elements)

Wishart, $W(p, n, \Sigma)$ whose density $f(\mathbf{S})$ is :

$$f(\mathbf{S}) = C(n, p, \Sigma) |\mathbf{S}|^{-\frac{n-p-1}{2}} \exp \left\{ -\frac{1}{2} \text{trace}(\Sigma^{-1} \mathbf{S}) \right\}$$

$$C(n, p, \Sigma) = 2^{-\frac{np}{2}} |\Sigma|^{-\frac{n}{2}} \pi^{-\frac{p(p-1)}{4}} \prod_{k=1}^p \Gamma\left(\frac{1}{2}(n+1-k)\right)$$

Distribution of eigenvalues (continuation)

Distribution of Eigenvalues from Wishart :

Fisher (1939), Girshick (1939), Hsu (1939) and Roy (1939), then Mood (1951). Anderson (1958), Muirhead (1982).

$$f(\mathbf{S}) = C(n, p, \mathbf{I}) \left(\prod_{k=1}^p \lambda_k \right)^{-\frac{n-p-1}{2}} \exp \left\{ -\frac{1}{2} \sum_{k=1}^p \lambda_k \right\} \quad (\text{If } \Sigma = \mathbf{I})$$

$$g(\Lambda) = D(n, p) \left(\prod_{k=1}^p \lambda_k \right)^{-\frac{n-p-1}{2}} \exp \left\{ -\frac{1}{2} \sum_{k=1}^p \lambda_k \right\} \prod_{k < j}^p (\lambda_k - \lambda_j)$$

Case of largest eigenvalues:

Pillai (1965), Krishnaiah et Chang (1971), Mehta (1960, 1967)

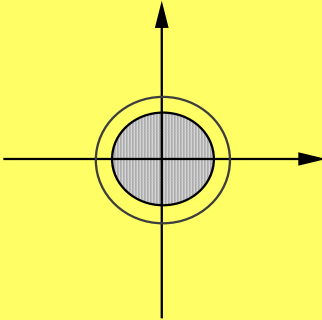
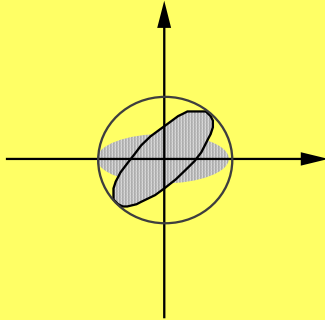
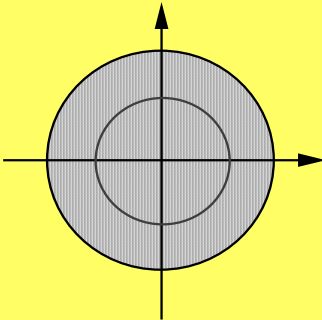
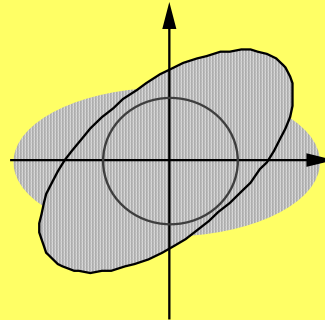
In practice, all these results are both unrealistic and unpractical

In Correspondence analysis, for a contingency table (n, p) , the eigenvalues are those obtained from a Wishart matrix : $W (n-1, p-1)$ (L. L., 1974)

As a consequence, under the hypothesis of independence, the percentage of variance are independent from the trace, which is the usual chi-square with $(n-1, p-1)$ degrees of freedom.

However, in the case of Multiple Correspondence Analysis, or in the case of binary data, the trace has not the same meaning, and the percentages of variance are misleading measure of information.

First eigenvalue

Cloud		Spherical	Non spherical"
Inertia	small inertia	 <p>1- INDEPENDENCE</p>	 <p>2- DEPENDENCE</p>
	Large inertia	 <p>3- DEPENDENCE</p>	 <p>4- DEPENDENCE</p>

Chi-squared

Quality of the structural compression of data

Approximation formula

$$\mathbf{X}^* = \sum_{\alpha=1}^q \sqrt{\lambda_{\alpha}} \mathbf{v}_{\alpha} \mathbf{u}_{\alpha}' \quad \text{con } q < p$$

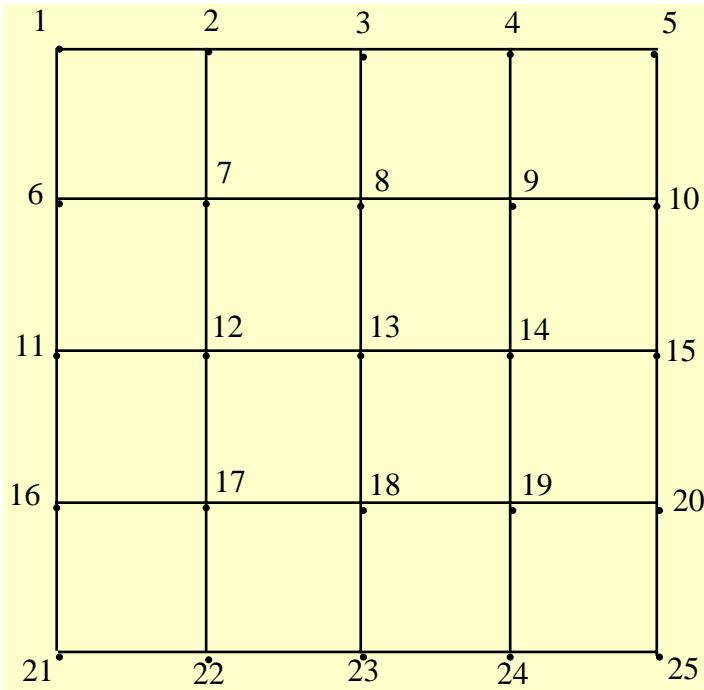
(Compression formula)

Measurement of the quality of the approximation

$$\tau_q = \frac{\sum_{\alpha=1}^q \lambda_{\alpha}}{\sum_{\alpha=1}^p \lambda_{\alpha}} = \frac{\text{tr}\{\mathbf{X}^{*'} \mathbf{X}^*\}}{\text{tr}\{\mathbf{X}' \mathbf{X}\}} = \frac{\sum_{i,j=1}^p (\mathbf{x}_{ij}^*)^2}{\sum_{i,j=1}^p (\mathbf{x}_{ij})^2}$$

CORRESPONDENCE ANALYSIS APPLIED TO SOME PARTICULAR BINARY TABLES

Example of a graph **G** ($n = 25$) associated with a squared lattice



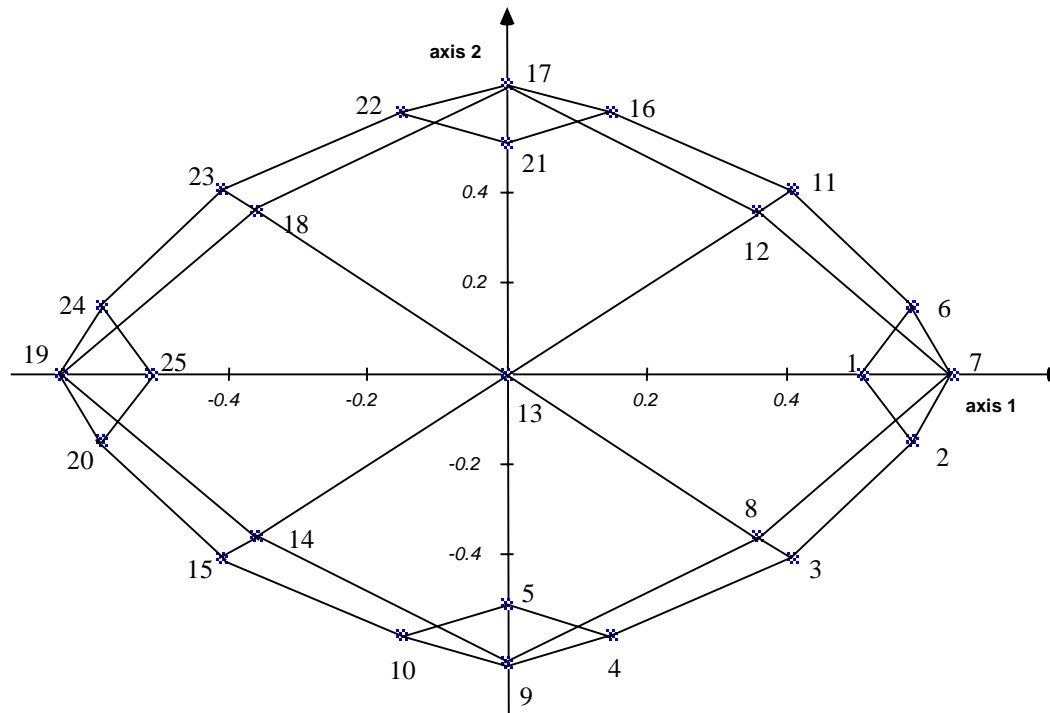
... and its associated matrix **M**

matrix:

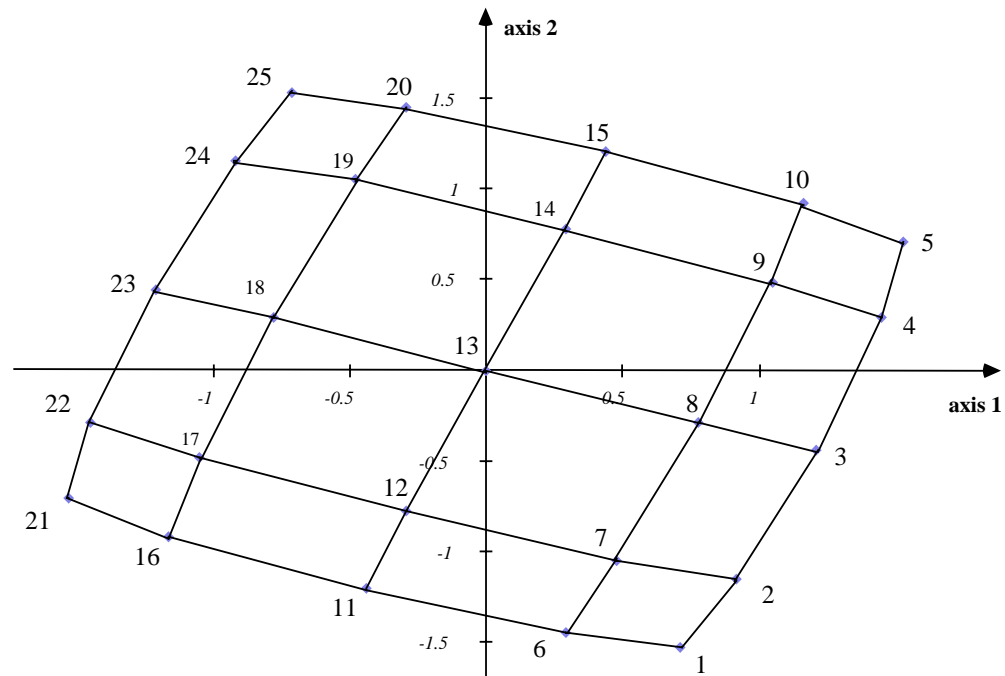
M

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
r01	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
r02	1	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
r03	0	1	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
r04	0	0	1	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
r05	0	0	0	1	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
r06	1	0	0	0	0	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
r07	0	1	0	0	0	1	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
r08	0	0	1	0	0	0	1	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
r09	0	0	0	1	0	0	0	1	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
r10	0	0	0	0	1	0	0	0	1	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
r11	0	0	0	0	0	1	0	0	0	0	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0
r12	0	0	0	0	0	0	1	0	0	0	1	1	1	0	0	0	1	0	0	0	0	0	0	0	0
r13	0	0	0	0	0	0	0	1	0	0	0	1	1	1	0	0	0	1	0	0	0	0	0	0	0
r14	0	0	0	0	0	0	0	0	1	0	0	0	1	1	1	0	0	0	1	0	0	0	0	0	0
r15	0	0	0	0	0	0	0	0	0	1	0	0	0	1	1	0	0	0	0	1	0	0	0	0	0
r16	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	1	0	0	0	1	0	0	0	0
r17	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	1	1	0	0	0	1	0	0	0
r18	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	1	1	0	0	0	1	0	0
r19	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	1	1	0	0	0	1	0
r20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	1	0	0	0	0	1
r21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	1	0	0	0
r22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	1	1	0	0
r23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	1	1	0
r24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	1	1
r25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	1

Description of **G** through
Principal Component Analysis of **M**



Description of **G** through
Correspondence Analysis of M



Explanation :

$$\text{Local variance} = \mathbf{y}'(\mathbf{I} - \mathbf{N}^{-1}\mathbf{M}) \mathbf{y}$$

$$\text{Global variance} = \mathbf{y}'\mathbf{y}$$

Bounds for $c(\mathbf{y}) = \text{contiguity coefficient}$.

$$c(\mathbf{y}) = \mathbf{y}'(\mathbf{I} - \mathbf{N}^{-1}\mathbf{M}) \mathbf{y} / \mathbf{y}' \mathbf{y}$$

minimum of $c(\mathbf{y})$, μ , is the smallest eigenvalue of:

$$(\mathbf{I} - \mathbf{N}^{-1}\mathbf{M}) \boldsymbol{\psi} = \mu \boldsymbol{\psi}$$

Equivalently:

$$\mathbf{N}^{-1}\mathbf{M} \boldsymbol{\psi} = (1 - \mu) \boldsymbol{\psi}$$

transition formulae, CA of \mathbf{M} : $\mathbf{N}^{-1}\mathbf{M} \boldsymbol{\phi} = \varepsilon \sqrt{\lambda} \boldsymbol{\phi}$

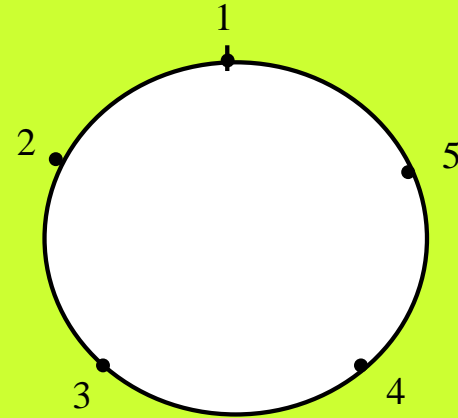
if $\varepsilon = +1$, direct factor, if $\varepsilon = -1$, inverse factor.

Min $\mu = \text{Max } \lambda$, λ_{\max} if ($\varepsilon = +1$).

Thus: $\text{Min } [c(\mathbf{y})] = 1 - \sqrt{\lambda_{\max}}$

Misleading measures of information : Case of a cycle

$$\mathbf{M} = \begin{bmatrix} 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 \end{bmatrix}$$



$$\varphi_{\alpha}(j) = \cos\left(\frac{2j\alpha\pi}{n}\right) \quad \text{and} \quad \psi_{\alpha}(j) = \sin\left(\frac{2j\alpha\pi}{n}\right)$$

$$\lambda_{\alpha} = \cos^2\left(\frac{2\alpha\pi}{n}\right)$$

$$\tau_{\alpha} = \frac{2}{n} \cos^2\left(\frac{2\alpha\pi}{n}\right)$$

Other tools for internal validation

Stability (Escofier and Leroux, 1972)

Sensitivity (Tanaka, 1984)

Confidence zones using Delta method (Gifi, 1990)

.

II.2. Internal validation: Bootstrap, opportunity of the method

- In order to compute estimates precision, many reasons lead to the Bootstrap method :
 - highly complex computation in the analytical approach
 - to get free from beforehand assumptions
 - possibility to master every statistical computation for each sample replication
 - no assumption about the underlying distributions
 - availability of cumulative frequency functions, which offers various possibilities

Reminder about Bootstrap Method

An example : Confidence areas in statistical mappings.

- The mappings used to visualise multidimensional data in Marketing Research (through *Multidimensional Scaling*, *Principal Component Analysis* or *Correspondence Analysis*) involve complex computation.
- In particular, variances of the locations of points on mappings cannot be easily computed.
- The seminal paper by [Diaconis](#) and [Efron](#) in *Scientific American* (1983) *Computer intensive methods in statistics* precisely dealt with a similar problem in the framework of *Principal Component Analysis*.

PCA case

In PCA case, variants of bootstrap (**partial** and **total** bootstrap) are proposed for active variables, supplementary variables, and supplementary nominal variables as well.

In the case of numerous homogeneous variables, a *bootstrap on variables* is also proposed.

Various examples of application of **partial**, **total** bootstrap, and of **bootstrap on variables** are presented in the context of *semiometric* data (Lebart, Piron and Steiner, 2003).

Ludovic Lebart
Marie Piron
Jean-François Steiner

LA SÉMIOMÉTRIE

Essai de statistique structurale

Au-delà de leur signification, les mots, par les souvenirs qu'ils mobilisent, ont le pouvoir de provoquer en nous des sensations agréables ou désagréables.

De cette observation simple est née une méthode, la Sémiométrie, largement utilisée en marketing et dans les études psychosociologiques. Mais, grâce à la puissance des outils statistiques actuels, son pouvoir d'investigation va, bien au-delà de ces applications pratiques, jusqu'aux confins de la psychanalyse et de la linguistique.

Il semblait indispensable qu'un ouvrage fasse le point sur les principes de cette méthode, les travaux réalisés et les applications potentielles. Le lecteur peut maintenant découvrir, au fil des chapitres, l'étendu du travail d'expérimentation, la sévérité des épreuves de validation, la profondeur et la finesse des résultats obtenus, enfin les promesses de cet outil transdisciplinaire.

Plusieurs niveaux de lecture sont possibles selon les connaissances mathématiques et statistiques du lecteur ; les développements plus techniques sont en effet regroupés dans une annexe.

Ce livre s'adresse aux spécialistes du marketing et de la communication, aux socio-économistes, aux statisticiens, aux psychosociologues, aux linguistes. Il intéresse un large public, allant des sociétés d'études et des instituts de sondage aux étudiants, professeurs et chercheurs des universités ou des grandes écoles.



ISBN 2 10 008105 5

<http://www.dunod.com>



L. LEBART
M. PIRON
J.-F. STEINER

LA SÉMIOMÉTRIE



La Sémiométrie

LUDOVIC LEBART
MARIE PIRON
JEAN-FRANÇOIS STEINER

LUDOVIC LEBART
est directeur de recherches
au CNRS et professeur à
l'ENST (École nationale
supérieure des
télécommunications)

MARIE PIRON
est chargée de recherches
à l'IRD (Institut de
recherche pour le
développement)

JEAN-FRANÇOIS STEINER
est écrivain.

DUNOD

Partial bootstrap making use of projections of replicated elements on the reference subspace provided by SVD of the **observed covariance matrix** has several advantages.

From a descriptive standpoint, this initial subspace is better than any subspace undergoing a perturbation by a random noise.

In fact, this subspace is the expectation of all the perturbed subspaces (replicates).

In this context, one may project the q replicates of variable-points in the common reference subspace, and compute confidence areas (ellipses or convex hulls) for the locations of these replicates.

Example 1 : Validation in Semiometry

The basic idea is to insert in the questionnaire a series of questions consisting uniquely of words (a list of 210 words is currently used, but some abbreviated lists containing a subset of 80 words could be used as well).

The interviewees must rate these words according to a seven levels scale, the lowest level (mark = 1) concerning a "most disagreeable (or unpleasant) feeling about the word", the highest level (mark = 7) concerning a "most agreeable (or pleasant) feeling" about the word.

Questionnaires in 5 languages

FRENCH

l'absolu

l'acharnement

acheter

admirer

adorer

l'ambition

l'âme

l'amitié

l'angoisse

un animal

un arbre

l'argent

une armure

l'art

ENGLISH

absolute

persistence

to buy

to admire

to love

ambition

soul

friendship

anguish

animal

tree

silver

armour

art

GERMAN

absolut

hartnaeckig

kaufen

bewundern

anbeten

der ehrgeiz

die seele

die freundschaft

die angst

ein tier

ein baum

das geld

die ruestung

die kunst

SPANISH

el absoluto

el empeno

comprar

admirar

adorar

la ambicion

el alma

la amistad

la angustia

un animal

un arbol

el dinero

una armadura

el arte

ITALIAN

l'assoluto

l'accanimento

comprare

ammirare

adorare

l'ambizione

l'anima

l'amicizia

l'angoscia

un animale

un albero

il denaro

un'armatura

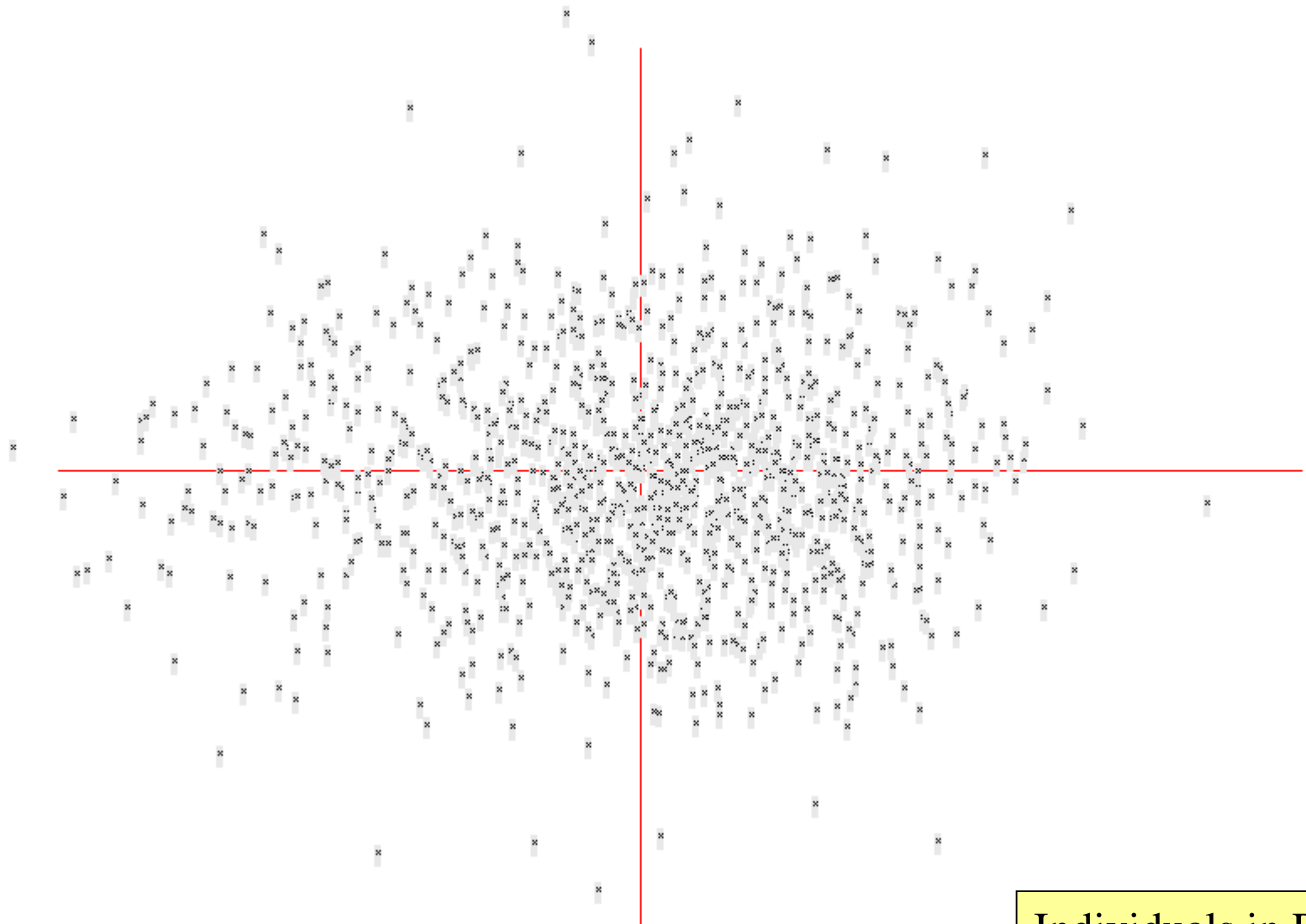
l'arte

122	La modestie	-3	-2	X	0	+1	+2	+3
133	Mcelleux	-3	-2	-1	X	+1	+2	+3
124	La mort	-3	X	-1	0	+1	+2	+3
100	Une muraille	-3	-2	-1	0	+1	+2	+3
085	Un mystère	-3	-2	-1	0	+1	+2	+3
105	Nager	-3	-2	-1	0	+1	+2	+3
043	Une naissance	-3	-2	-1	0	+1	+2	+3
025	Un nid	-3	-2	-1	0	+1	+2	+3
106	La nudité	-3	-2	-1	0	+1	+2	+3
071	Obéir	-3	-2	-1	0	+1	+2	+3
173	L'océan	-3	-2	-1	0	+1	+2	+3
086	Un orage	-3	-2	-1	0	+1	+2	+3

Facsimile of a questionnaire

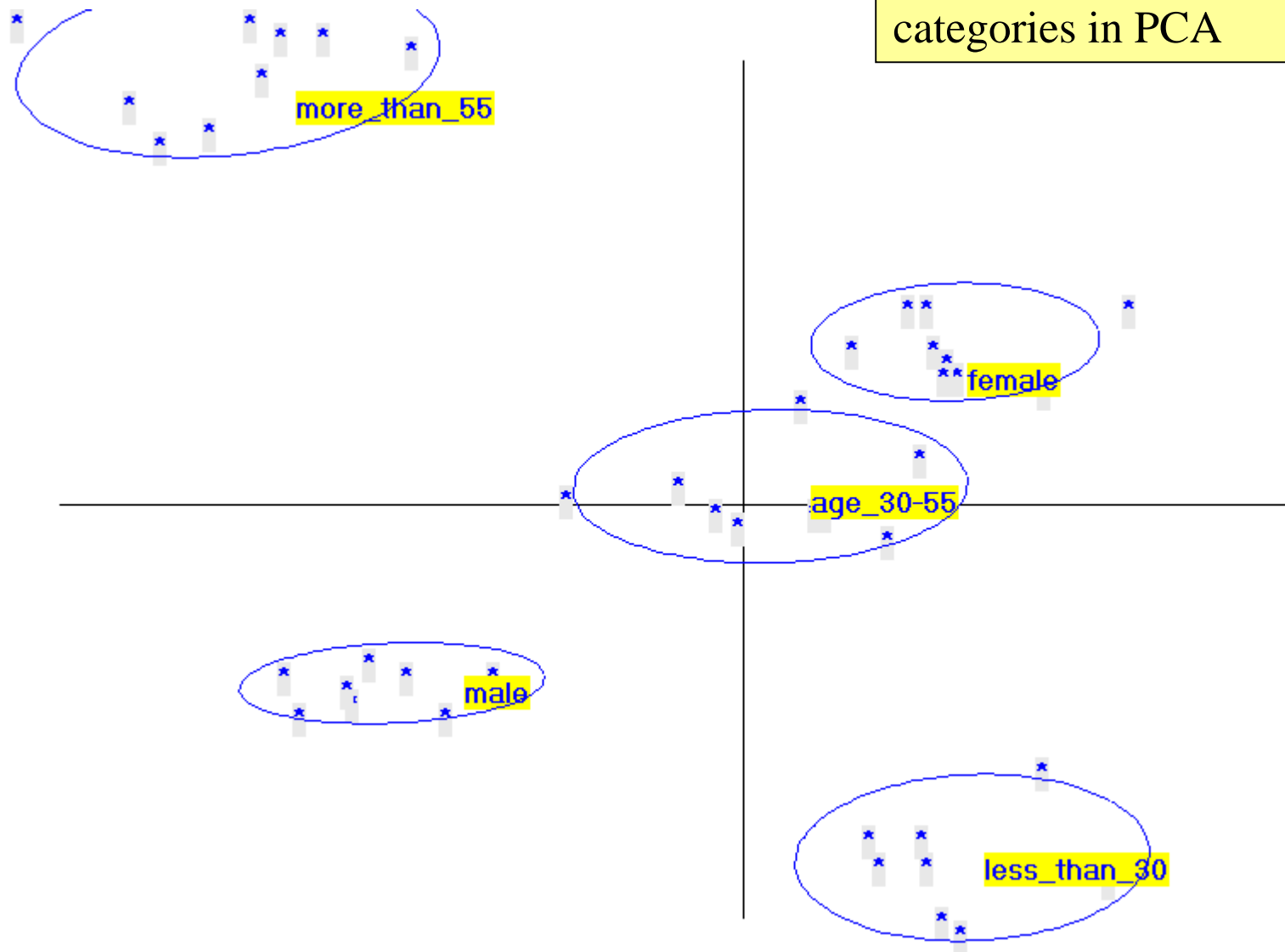
The processing of the filled questionnaires (*mainly through Principal Components Analysis*) produces a stable pattern (*up to 8 stable principal axes*).

Very similar patterns are obtained in ten different countries, despite the problems posed by the translation of the list of words.



Individuals in PCA

Bootstrapping supplementary
categories in PCA



CA and MCA cases

Gifi (1981), Meulman (1982), Greenacre (1984) did pioneering work in addressing the problem in the context of two-way and multiple correspondence analysis.

It is easier to assess eigenvectors than eigenvalues that are much more sensitive to data coding, the replicated eigenvalues being biased replicates of the theoretical ones.

Reminder about Bootstrap Method

Example : Confidence areas in statistical mappings.

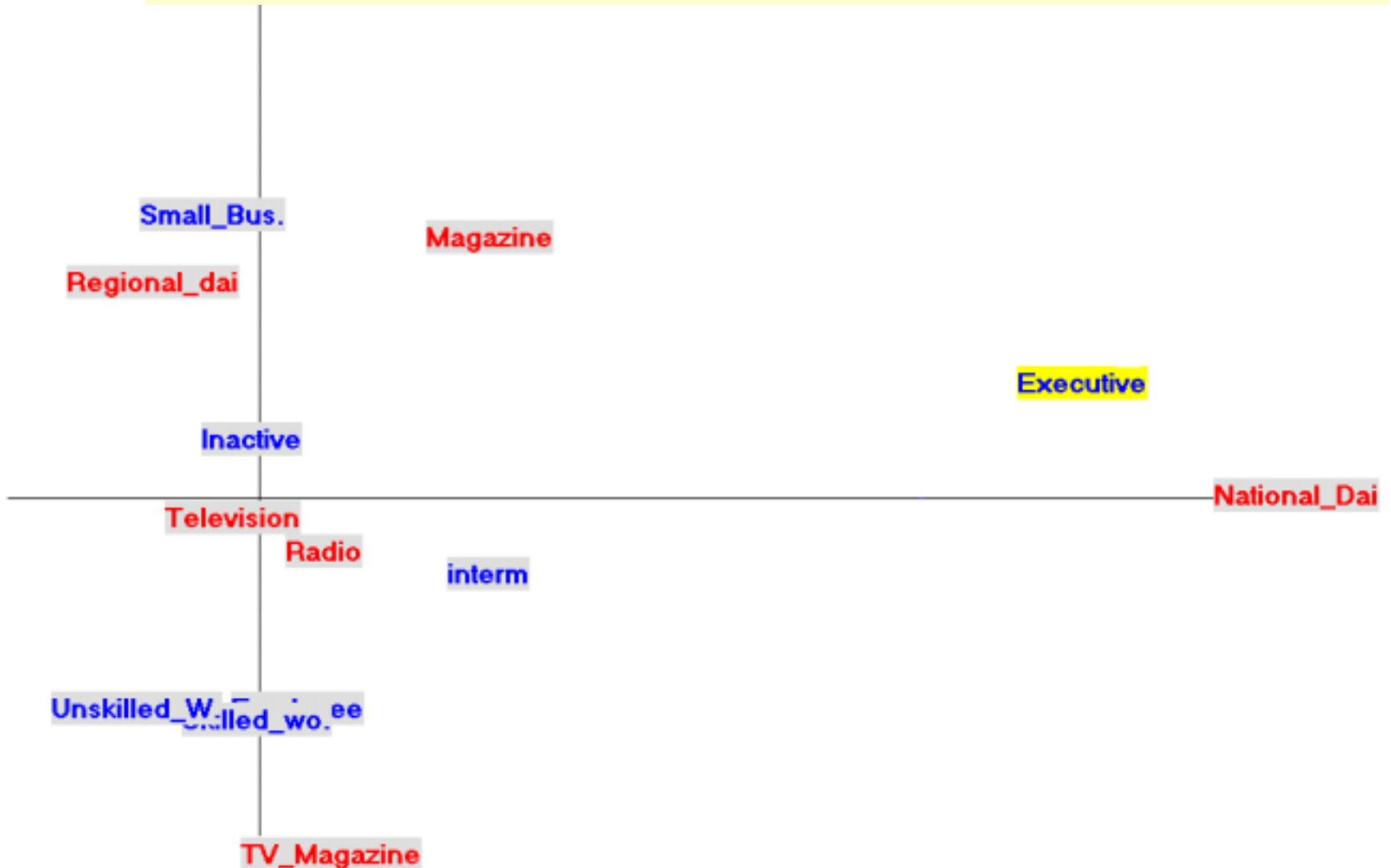
- “Contingency table” (Cross-tabulation) (CESP Multi-Media Survey, 1993).
- In each cell: number of media contacts the day before.
- **Columns : Media** [Radio, TV, National & Regional Daily N., Magazines].
- **Rows : Occupation groups.**

	Radio	Tele	Nat.	Reg.	Maga	TV_Mag
Farmer	96.	118.	2.	71.	50.	17.
Small Business	122.	136.	11.	76.	49.	41.
Executive	193.	184.	74.	63.	103.	79.
Intermediate	360.	365.	63.	145.	141.	184.
Employee	511.	593.	57.	217.	172.	306.
Skilled worker	385.	457.	42.	174.	104.	220.
Unskilled worker	156.	185.	8.	69.	42.	85.
Housewives, Ret.	1474.	1931.	181.	852.	642.	782.

Farmer

Visualisation of associations between occupation and media contacts

[Example of C.A. mapping]

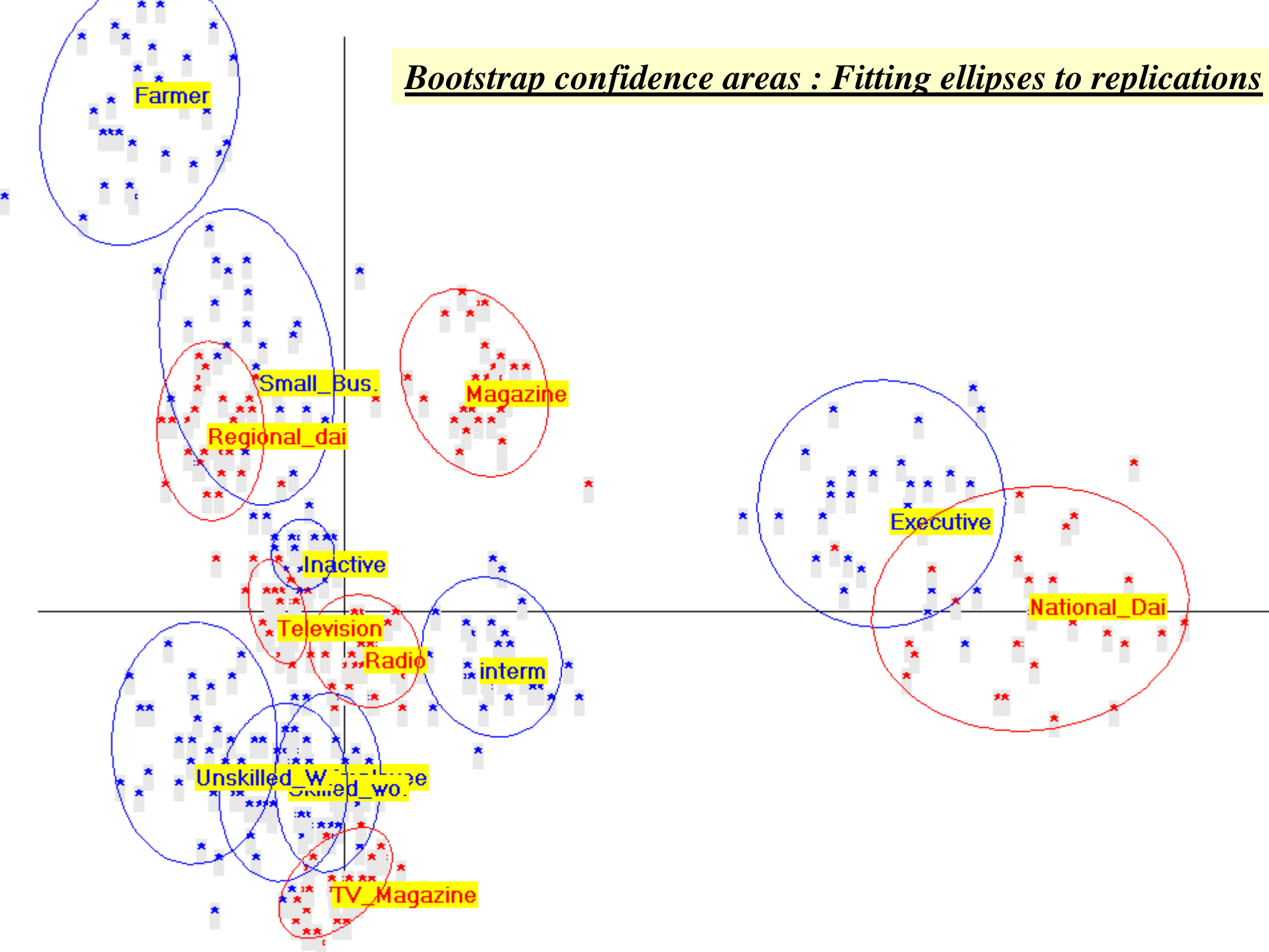


(Reminder about Bootstrap Method)

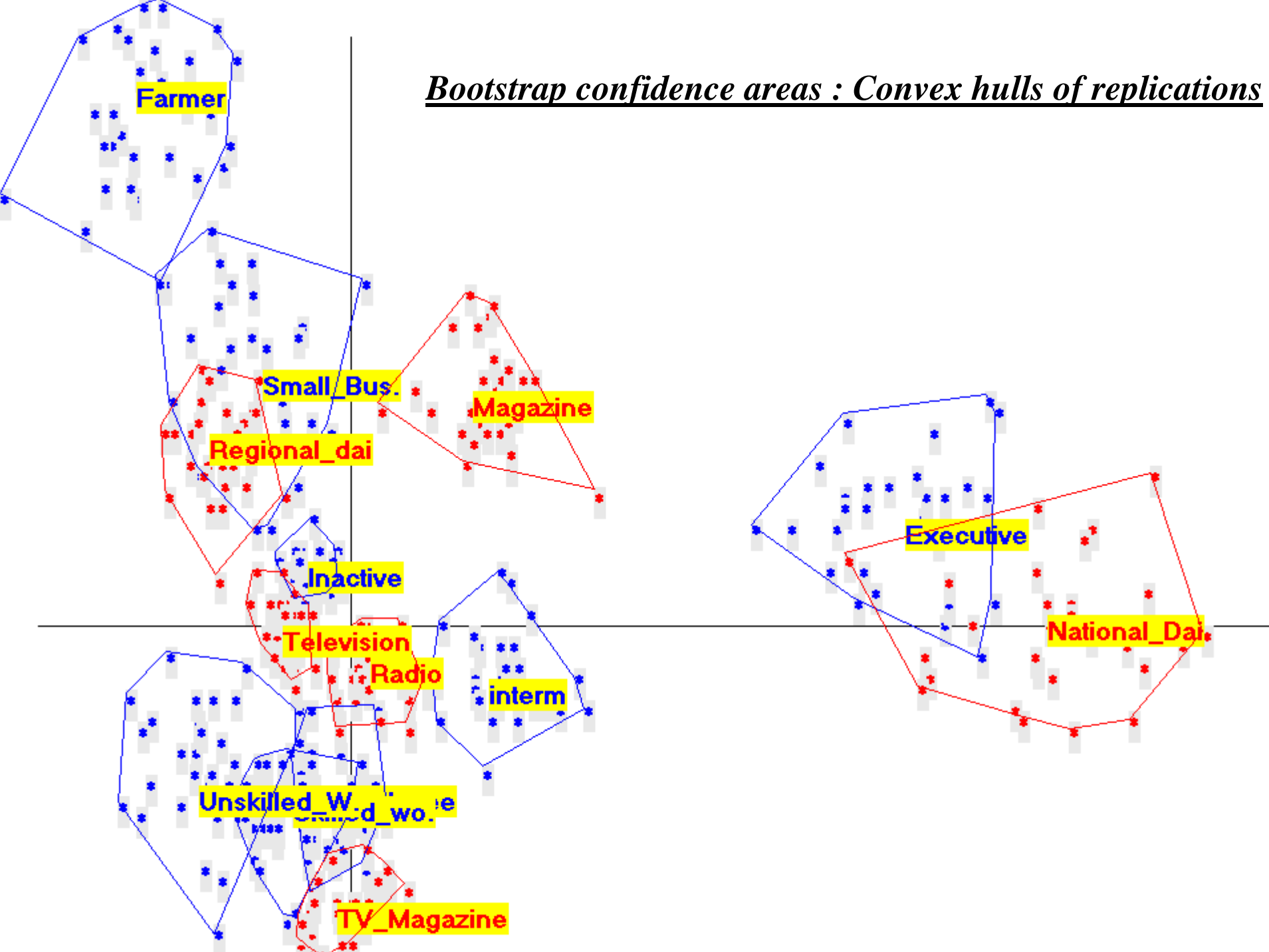
- Example of a replicated table

	Radio	Tele	Nat.	Reg.	Maga	TV_M
• Farmer	109.	120.	1.	78.	48.	20.
• Small Business	126.	142.	8.	76.	53.	30.
• Executive	196.	181.	80.	77.	109.	72.
• Intermediate	384.	365.	60.	133.	138.	203.
• Employee	514.	596.	59.	228.	172.	316.
• Skilled worker	378.	467.	33.	171.	100.	223.
• Unskilled worker	169.	188.	8.	79.	38.	81.
• Housewives, Ret.	1519.	1961.	158.	893.	632.	764.

Bootstrap confidence areas : Fitting ellipses to replications



Bootstrap confidence areas : Convex hulls of replications



Example: Open question in a sample surveys

The following open-ended question was asked :

"What is the single most important thing in life for you?"

It was followed by the probe:

"What other things are very important to you?".

This question was included in a multinational survey conducted in seven countries (Japan, France, Germany, Italy, Nederland, United Kingdom, USA) in the late nineteen eighties (Hayashi *et al.*, 1992).

Our illustrative example is limited to the British sample (Sample size: 1043).

Examples of responses to “Life” question

<i>Gender</i>	<i>Educ.</i>	<i>Age</i>	<i>Responses</i>
1	1	4	happiness in people around me, contented family, would make me happy
1	2	2	my own time, not dictated by other people
1	2	2	freedom of choice as to what I do in my leisure time
1	3	2	I suppose work
1	2	1	firm, my work, which is my dad's firm
2	1	6	just the memory of my last husband
2	2	6	well-being of my handicapped son
1	1	5	my wife, she gave me courage to carry on even in the bad times
2	2	3	my sons, my kids are very important to me, being on my own, I am responsible for their education
1	3	3	job, being a teacher I love my job, for the well-being of the children

Example, *continuation*

The counts for the first phase of numeric coding are as follows:

Out of **1043** responses, there are **13 669** occurrences (*tokens*),
with **1 413** distinct words (*types*).

When the words appearing at least **16** times are selected,
there remain **10 357** occurrences of these words (*tokens*),
with **135** distinct words (*types*).

Example, *continuation*

The same questionnaire also had a number of closed-end questions (among them, the socio-demographic characteristics of the respondents, which play a major role).

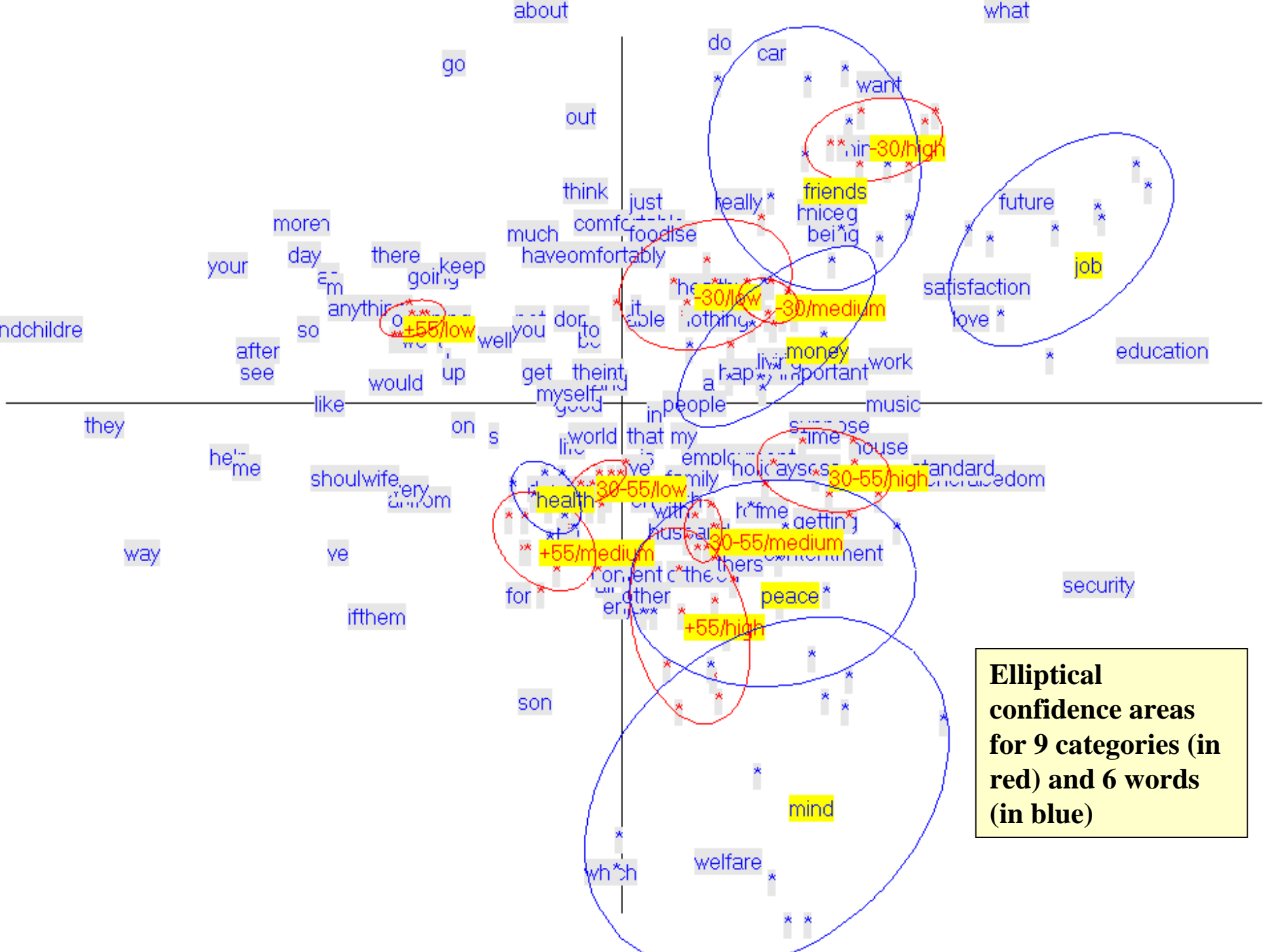
In this example we focus on a partitioning of the sample into *nine* categories, obtained by cross-tabulating **age** (*three* categories) with **educational level** (*three* categories).

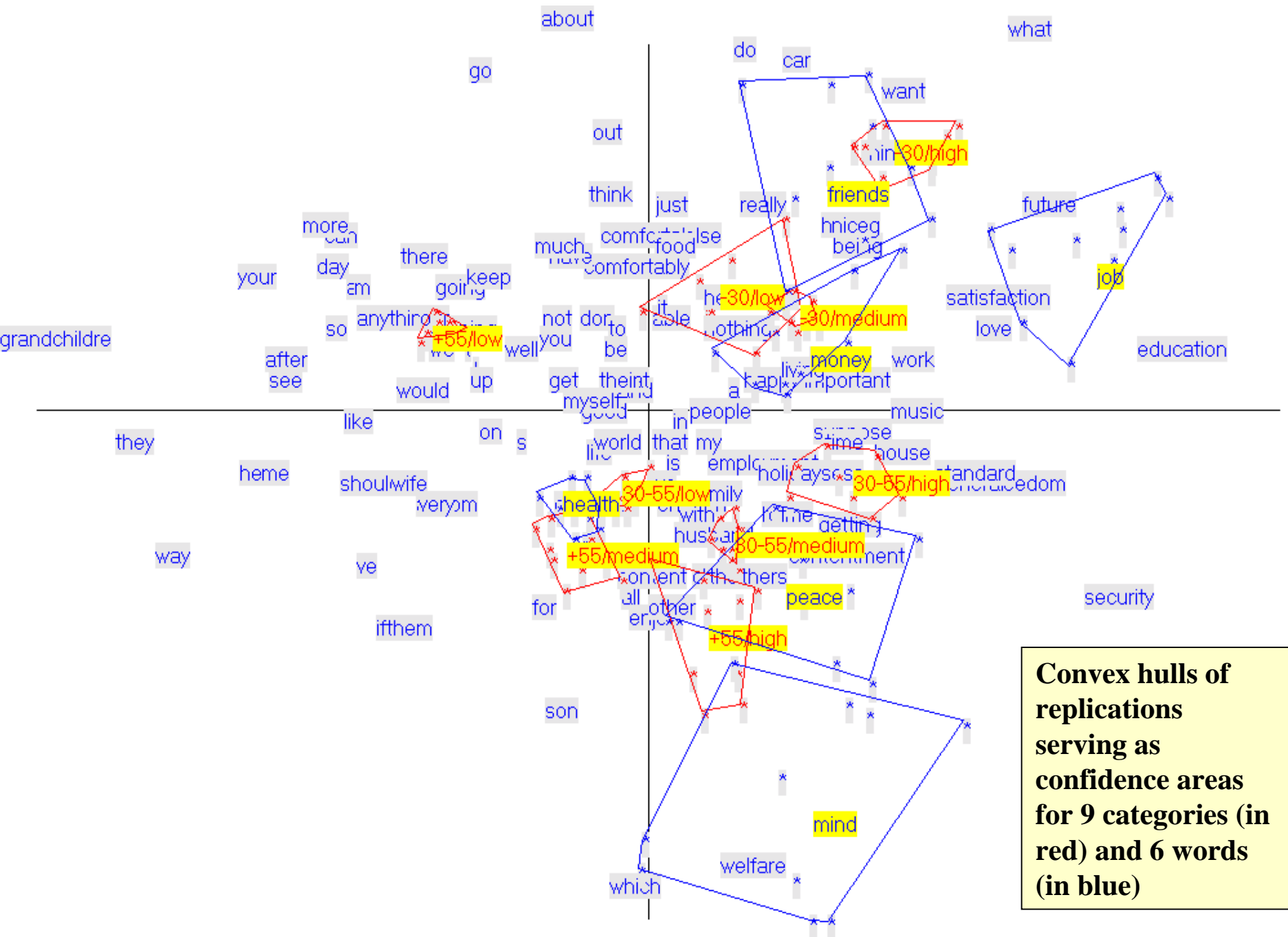
Example of a lexical contingency table

Partial listing of lexical table cross-tabulating 135 words of frequency greater than or equal to 16 with 9 age-education categories

	L-30	L-55	L+55	M-30	M-55	M+55	H-30	H-55	H+55
I	2	46	92	30	25	19	11	21	2
I'm	2	5	9	3	2	1	0	0	0
a	10	56	66	54	44	19	20	22	7
able	1	9	16	9	7	4	4	5	0
about	0	3	13	7	1	2	4	1	0
after	1	8	11	3	1	2	0	0	0
all	1	24	19	8	18	6	3	5	2
and	8	89	148	86	73	30	25	32	13
anything	0	4	9	1	3	0	1	1	0

- The two forthcoming diapositives show the principal plane produced by a correspondence analysis of the previous lexical contingency table.
- Proximity between 2 category-points (columns) means similarity of lexical profiles of the 2 categories.
- Proximity between 2 word-points (rows) means similarity of lexical profiles of these words.
- Both ellipses and convex hulls describe the uncertainty of the location of the points.
- 9 categories points, **in red** (all the categories, in fact)
- 6 selected word-points, **in blue**.

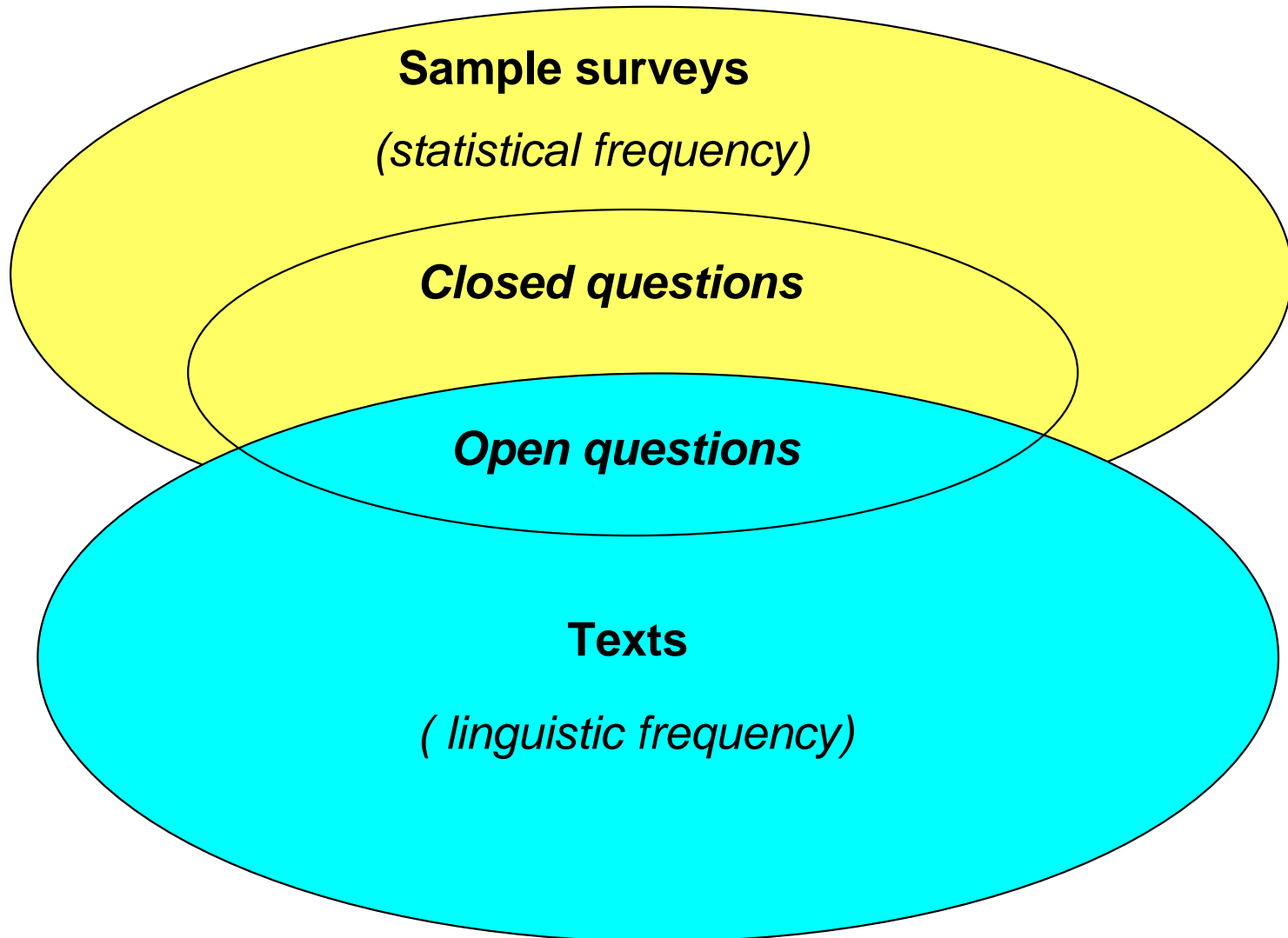




- When dealing with textual data, these resampling techniques can help solving the problem of **plurality** of statistical units.
- In fact, two (or more) levels of statistical units coexist in textual data analysis.
- On the one hand, observations or individuals (with their usual meaning in statistics) could be **respondents** (case of sample surveys), **documents** or **abstracts** (Information retrieval) or **cybernauts**, **web-users** (WebMining).
- On the other hand, within the same textual corpus, the observations or individuals could be **occurrences** (token), **words**, **lemmas**, **phrases**.
- At an intermediate level, it could be also for some other applications: **pages**, **sentences**, **frames**

Ambiguity of frequencies:

statistical frequency versus « linguistic frequency »



Due to the discrepancies of text sizes, a structural pattern could be significant when the statistical unit is the word, and not relevant if the statistical unit is the respondent or the web user.

The replication scheme can be customized to all mentioned levels, leading to conclusions adapted to the expected inference.

[replications built through drawing with replacement of words]

versus

[replications built through drawing with replacements of respondents]

Conclusion

From a statistical perspective, most of available data sets that **require** and **deserve** an exploration involving visualizations (including **textual data**) are:

- ✓ **High dimensional,**
- ✓ **Qualitative,**
- ✓ **Sparse,**
- ✓ ... with a high level of **noise**.

Validation procedures are then badly needed,
... and resampling techniques
(mainly bootstrap for unsupervised approaches)
provide the versatile tools that transform a nice
visualization into a scientific document.

Choukrane

Ευχαριστω πολι

Domo Arigato

Thank You

Grazie

Merci

Gracies

Danke

Obrigado

Gracias

Computers are useless,
they can only give you answers.

Pablo Picasso