LUDOVIC LEBART'S APPROACH: A WAY FOR TEACHING APPLIED MULTIVARIATE STATISTICS IN GRADUATE COURSES WITH A HETEROGENEOUS AUDIENCE

Hebe Goldenhersch
Universidad Nacional de Córdoba, Argentina
hebegold@eco.unc.edu.ar

First of all we describe different kinds of audiences in a graduate multivariate course. We talk about how to start a Multivariate Statistics Course. We show a survey as an Example of Multivariate Data. Factorial Methods and Adjustment Criteria are discussed, within a general analysis .The importance of the "illustrative variables" according to Lebart´s approach are discussed, and finally we discuss the complementary applications of factorial methods with cluster methods to analyze the surveys when the data are quantitative and when they are qualitative.

INTRODUCTION

Statistics is, without any doubt, an invaluable tool which contributes to the understanding of reality. This understanding is essential in order to modify it.

With regard to the role of statistics today, I quote a paragraph, slightly adapted, from the lecture delivered by Estela Bee Dagun on the occasion of receiving a PhD Honoris Causa at the University of Naples. Her words, in my opinion, encapsulate remarkably well the challenges posed to statisticians at the beginning of the XXI century when investigating information about critical societal issues:

> "Good information will not be enough; it is necessary to develop innovative statistical models and methods to explain new realities. These two needs require statisticians with strong theoretical knowledge as well as excellent empirical statisticians who, contrary to current practice, must keep in permanent scientific communication… My conviction is that the success of statistical research and analysis lies in the interaction of the 'watching and wondering' stated by Nicholas Tinbergen in his 1973 Nobel Prize lecture. 'Watching' being supported by informative, accurate and well-timed data while the 'wondering' should lead us to an appropriate use of data coupled with the right methods in order to offer solutions to new problems."

Multivariate methods constitute, undoubtedly, an important tool in the achievement of these objectives. The complexity of reality can only, though sometimes partially, be represented with their help.

MASTERS AND PHD TRACKS. DESCRIPTION OF THE CANDIDATES AND THEIR INTERESTS.

Given an audience with sufficient basic knowledge that needs to delve deeply into the methods and to learn its applications, the presentation of each topic and the explanation of its scope must be followed or preceded by the basic mathematical developments which allow for these applications. We are thinking in terms of groups of students who have a background in mathematics and in basic statistics.

On some occasions, however, when a proper presentation cannot be made on account of the scarce time available, some bibliography can be recommended so that the candidates can have access to the theoretical background. *Análisis de Datos Multivariantes* by Daniel Peña (2002) constitutes a possible option in Spanish, while *Applied Multivariate Statistical Analysis* by Johnson and Wichern (1992) is another good option in English.

When the courses are aimed at a heterogeneous audience, without a proper mathematical background, and the main objective is to understand and interpret the applications of the different methods, it is advisable to emphasize the importance of the selection of the appropriate technique and the interpretation of results. This would be the situation with Masters like the MBA, or those oriented towards some aspect of Economic or Social Sciences. In these cases it is always possible

to teach multivariate applications without the theoretical analysis, although it is always desirable to have at least an intuitive presentation of the mathematical background of each method so that its application should not just be a matter of using the results provided by the computer software.

HOW TO START A MULTIVARIATE STATISTICS COURSE?

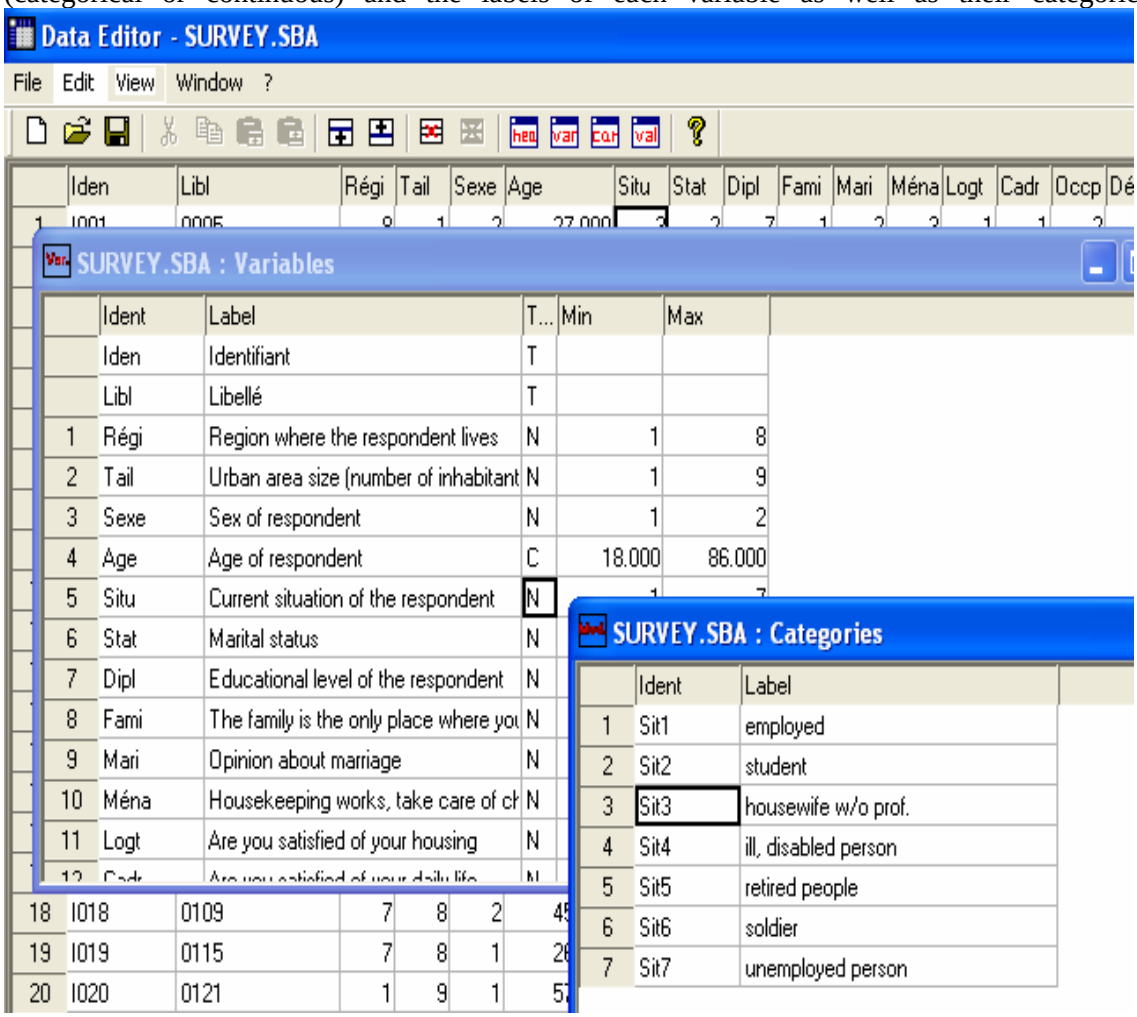
To begin any course on Multivariate Statistics, we consider it convenient to start with the presentation of an example and then continue explaining how data should be presented, what a data matrix is and what the possible strategies for its interpretation and representation are.

A SURVEY AS AN EXAMPLE OF MULTIVARIATE DATA

We consider that the survey carried out periodically by the "Centre de recherche pour l'étude et l'óbsérvation des conditions de vie" in France (Lebart, Morineau and Piron, 1997, Sec.2) provides an excellent example to start with and to develop a course on multivariate statistics whatever the audience might be.

SPAD, a software package specially designed to process these data and to teach the methods, was until very recently only available in French, but its latest version (which appeared mid 2005, in its 6.0 version) is also available in English. (We suggest consulting the website of the firm which commercializes this software: www.decisia.com.) *SPAD* is a commercial package. According to Ludovic Lebart´s internet site http://egsh.enst.fr/lebart/ (he is the author of the original programs), it is possible to download a free version of the software *DTM* (Data and Text Mining), less "friendly" than *SPAD*, but with almost all its possibilities for performing multivariate analysis.

See *SPAD* screens below illustrating the format of the survey, the type of the variables (categorical or continuous) and the labels of each variable as well as their categories.



	Ident	Label	T...	Min	Max
	Iden	Identifiant	T		
	Libl	Libellé	T		
1	Régi	Region where the respondent lives	N	1	8
2	Tail	Urban area size (number of inhabitant	N	1	9
3	Sexe	Sex of respondent	N	1	2
4	Age	Age of respondent	C	18.000	86.000
5	Situ	Current situation of the respondent	N		
6	Stat	Marital status	N		
7	Dipl	Educational level of the respondent	N		
8	Fami	The family is the only place where you	N		
9	Mari	Opinion about marriage	N		
10	Ména	Housekeeping works, take care of ch	N		
11	Logt	Are you satisfied of your housing	N		

	Ident	Label
1	Sit1	employed
2	Sit2	student
3	Sit3	housewife w/o prof.
4	Sit4	ill, disabled person
5	Sit5	retired people
6	Sit6	soldier
7	Sit7	unemployed person

A discussion of the objectives of factorial and cluster procedures in this work leads to a summary of current methods of analysis.

Another example of multivariate data can be a big contingency table (3096 students pre-enrolled to enter the Faculty of Economic Sciences, National University of Cordoba in 1997, classified according their parents educational level).

Crosstab: Father studies * Mother studies												
		Mother studies										
		no studies	inc.prim	comp. prim	inc. sec.	comp. sec	inc. terc	comp. terc	inc. univ.	comp. univ.	postg rad.	Total
Father studies	no studies	9	6	7	2	2	0	2	0	1	0	29
	inc.prim	4	92	45	28	23	0	14	1	3	0	210
	comp.prim	6	45	332	100	74	6	29	13	5	0	610
	inc.sec	1	33	100	200	105	16	70	37	34	3	599
	comp.sec	3	6	84	70	255	9	77	31	33	8	576
	inc.terc	0	0	1	5	8	5	7	5	1	0	32
	comp.terc	0	4	7	9	18	2	32	9	6	0	87
	inc.univ.	0	10	16	49	75	4	51	87	78	10	380
	comp.univ.	0	2	15	31	89	10	66	68	182	17	480
	postgrad.	0	0	1	5	17	2	9	12	25	22	93
Total		23	198	608	499	666	54	357	263	368	60	3096

Generalizing from the examples above, the data matrix can be symbolized as follows

	Var 1	Var 2	. . .	Var p
Obs. 1	x_{11}	x_{12}	. . .	x_{1p}
Obs. 2	x_{21}	x_{22}	. . .	x_{2p}
. . .	. . .	. . .	. . .	. . .
Obs. n	x_{n1}	x_{n2}	. . .	x_{np}

Sometimes, like in the crosstab of the previous example, columns are not variables and rows are not individuals or observations, but both of them are known to be categories of nominal variables.

The following diagrams (Lebart *et al.*, 1997, pp. 9-10) are remarkably adequate for the understanding of the topic.

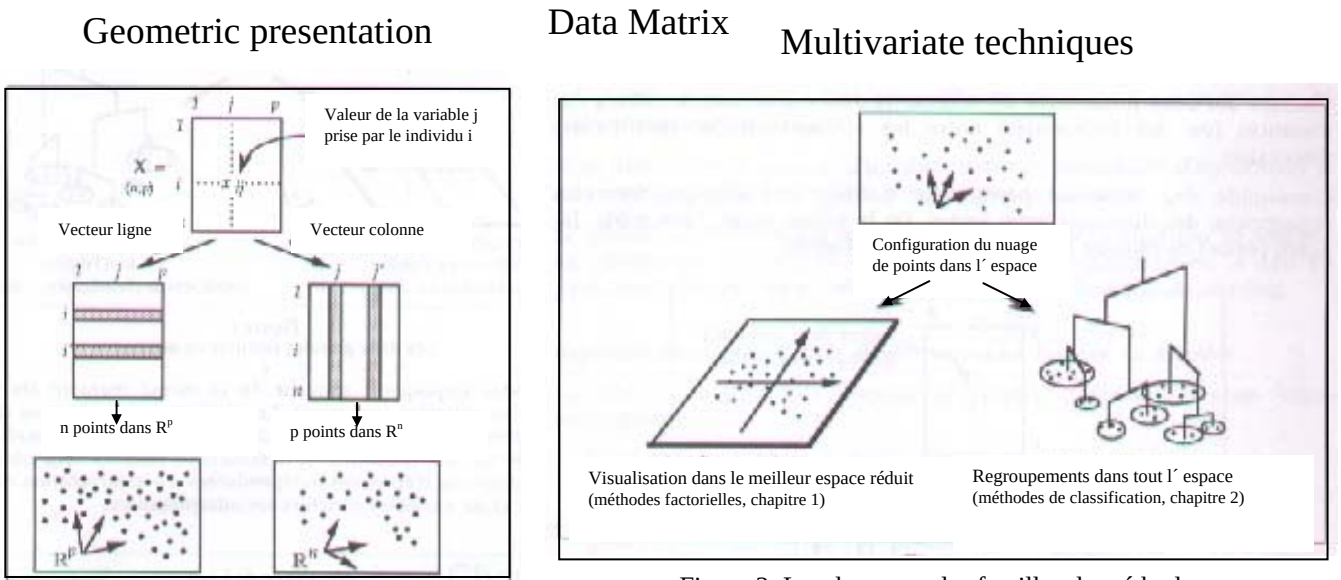


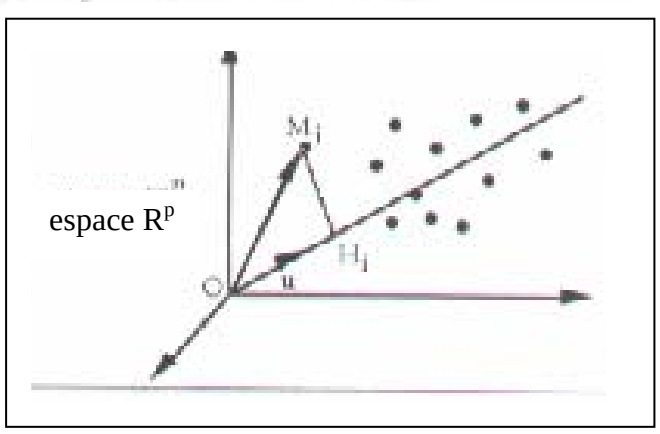
Figure 2: Les deux grandes familles de méthodes

Once the objectives of factorial and cluster methods represented in the diagrams above has been explained, it is possible to summarize the existing methods for the treatment of data, explaining what can be achieved by means of each of them

At this point it is possible to classify using different criteria: exploratory and confirmatory criteria with and without dependent variables, for quantitative and qualitative variables and so on. It is advisable to introduce examples in each case and then to start with the analysis of factorial methods.

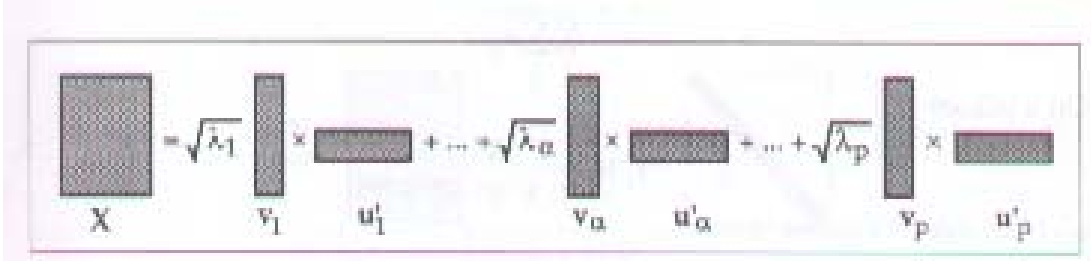
FACTORIAL METHODS: GENERAL ANALYSIS; ADJUSTMENT CRITERIA.

The introduction of the outline of the factorial methods can be made with the mathematical background or according a more intuitive approach, as we have explained above. The following diagram illustrates the construction of the first factor and the better adjustment criterion.



(Lebart *et al.*, 1997, p. 17)

After developing the resolutions from the eigenvalues and eigenvectors of matrixes XX´ or X'X (X being the data matrix), we explain how to obtain the various dimensions on which the "dotted cloud" is projected, ending with an interesting conclusion about the reconstruction of the original data from the eigenvalues and eigenvectors.



(Lebart *et al.*, 1997, p. 23)

This general outline adapts itself to the treatment of various data matrixes: if we deal with quantitative variables, the Principal Components Analysis should be used, from which it is possible to go to the classical Factor Analysis without too much problem. If we deal with Qualitative Variables we use the Correspondence Factor Analysis (in the case of two variables) or the Multiple Correspondence Factor Analysis (in the case of more than two variables).

THE IMPORTANCE OF THE "ILLUSTRATIVE VARIABLES" ACCORDING TO LEBART'S APROACH

The use of the illustrative variables and the complementary use of factor and cluster methods, are perhaps the most distinctive characteristics of the methods developed by the French School, of which the text by Lebart, Morineau and Piron (1997) is an excellent example.

When we explain the factorial methods, we must first determine the axes, then we project the individuals or the variables on them and at last we interpret their meaning.

The application of a particular method requires homogeneous variables, either quantitative of qualitative, but besides these considerations, the variables must be homogeneous in their relationship to the phenomenon under study.

It is important to remark that if the variables used to determine the factor axes are mixed with other variables which are not directly related to the issue being represented, it will not be possible to interpret the results thus obtained. In the case of an opinion survey for example, it is not possible to carry out a Correspondence Analysis if those variables representing opinions are mixed with others which indicate the socio-economic characteristics of the individuals. However, the authors mentioned above as well as the *SPAD* software raise the possibility of including "illustrative" individuals or variables. These are individuals or variables (lines or columns in general) which have not been used to determine the factor axes but which can be used to project on, thus helping the process of interpretation in a remarkable way. Is very simple to apply this procedure: the illustrative variables are incorporated as new columns in the matrix data (or as new rows when we are performing Principal Components Analysis and the illustrative variables are nominal), and they are projected into the factor space; their position related to the axes is very useful in order to understand what each factor means.

THE COMPLEMENTARY APPLICATIONS OF FACTORIAL METHODS WITH CLUSTER METHODS TO ANALYSE THE SURVEYS WHEN THE DATA ARE QUANTITATIVE AND WHEN THEY ARE QUALITATIVE.

When factorial methods are used, it is very interesting to carry out a cluster analysis with the factor coordinates. If we analyze the following output given by *SPAD* related to the mentioned survey, it is clear why we can say that the complementary use of both factorial and cluster methods are the appropriate way to understand multivariate data.

```
DESCRIPTION AND CHARACTERISATION OF PARTITIONS
DESCRIPTION OF: CUT "a" OF THE TREE INTO   3 CLUSTER
CLUSTERS CHARACTERISATION BY CATEGORIES
CLUSTERS CHARACTERISATION BY CONTINUOUS VARIABLES
CHARACTERISATION BY CATEGORIES OF CLUSTERS OR CATEGORIES
OF CUT "a" OF THE TREE INTO   3 CLUSTERS
Cluster  1 /  3
```

T.VALUE	PROB.	GRP/CAT	CAT/GRP	GLOBAL	CHARACTERISTIC CATEGORIES	OF VARIABLES	WEIGHT
				20.32	Cluster 1 / 3		64
10.13	0.000	89.74	54.69	12.38	not at all	Are you worried about the risk of having a serious illness	39
9.92	0.000	89.47	53.13	12.06	not at all	Are you worried about the risk of a road accident	38
7.76	0.000	53.16	65.63	25.08	not at all	Are you worried about the risk of a nuclear plant accident	79
7.73	0.000	51.81	67.19	26.35	not at all	Are you worried about the risk of a mugging	83
6.87	0.000	52.94	56.25	21.59	not at all	Are you worried about the risk of umemployment	68
2.97	0.001	44.83	20.31	9.21	I do not know	Do you think the society needs to change	29
2.77	0.003	33.33	35.94	21.90	no	Do you have children	69
2.72	0.003	38.10	25.00	13.33	single	Marital status	42
2.69	0.004	58.33	10.94	3.81	I do not know	Your opinion on the justice running in 1986	12

```
Cluster  2 /  3
```

T.VALUE	PROB.	GRP/CAT	CAT/GRP	GLOBAL	CHARACTERISTIC CATEGORIES	OF VARIABLES	WEIGHT
				42.86	Cluster 2 / 3		135
7.57	0.000	74.49	54.07	31.11	enough	Are you worried about the risk of a road accident	98
6.96	0.000	77.63	43.70	24.13	enough	Are you worried about the risk of having a serious illness	76
5.03	0.000	64.89	45.19	29.84	a little	Are you worried about the risk of a nuclear plant accident	94
4.37	0.000	66.20	34.81	22.54	enough	Are you worried about the risk of umemployment	71
4.12	0.000	71.74	24.44	14.60	enough	Are you worried about the risk of a mugging	46
3.72	0.000	66.67	26.67	17.14	a little	Are you worried about the risk of having a serious illness	54
3.68							

```
Cluster  3 /  3
-------------------------------------------------------------------------------------------------------------------------
T.VALUE PROB. ---- PERCENTAGES  ----  CHARACTERISTIC                                                              WEIGHT
            GRP/CAT CAT/GRP GLOBAL  CATEGORIES      OF VARIABLES
-------------------------------------------------------------------------------------------------------------------------
                            36.83  Cluster  3 /  3                                                                  116
 11.66  0.000   69.86   87.93  46.35  a lot          Are you worried about the risk of having a serious illness     146
 11.41  0.000   77.39   76.72  36.51  a lot          Are you worried about the risk of a road accident              115
  9.69  0.000   78.26   62.07  29.21  a lot          Are you worried about the risk of a mugging                     92
  9.55  0.000   78.65   60.34  28.25  a lot          Are you worried about the risk of a nuclear plant accident      89
  9.29  0.000   68.00   73.28  39.68  a lot          Are you worried about the risk of unemployment                 125
  5.20  0.000   79.41   23.28  10.79  a little       Are you satisfied of your health                                34
  3


CHARACTERISATION BY CONTINUOUS VARIABLES OF CLUSTERS OR CATEGORIES
OF CUT "a" OF THE TREE INTO   3 CLUSTERS
Cluster  1 /  3
+--------+-------+-------------------+-------------------+-----------------------------------------------+
| T.VALUE| PROB. |       MEANS       |   STD. DEVIATION  |                    CHARACTERISTIC VARIABLES   |
|        |       |   GROUP   OVERALL |   GROUP  OVERALL  | NUM.LABEL                                     |
+--------+-------+-------------------+-------------------+-----------------------------------------------+
|                 Cluster  1 /  3      ( WEIGHT =    64.00        COUNT =   64 )                          |
|                                                                                                         |
|        |       |                   |                   |                                               |
|  -2.60 | 0.005 |     1.37     1.86 |     1.24     1.67 |  28.Number of children                        |
|  -2.85 | 0.002 |     6.31     6.65 |     1.43     1.06 |  41.Family, children : importance given       |
+--------+-------+-------------------+-------------------+-----------------------------------------------+
```

In this kind of output, it is possible to see the characterization of each cluster (the clusters are made with the factor coordinates of the individuals). For each category of the active and illustrative variables we can see the difference between the group percentage and the global one (columns 4 and 5 of the output).The first column shows the *t*-value of that difference (the printout shows only the significant differences). In the case of continuous variables (last Table, columns 3 and 4), the output shows the mean-difference between group and global. By this approach it is possible to describe each cluster.

CONCLUDING REMARKS

The experiences we want to communicate here show that it is possible to teach multivariate statistical methods to students with or without a proper mathematical and statistical background. We think that Ludovic Lebart and the French School point of view, supported by the appropriate software (i.e., *SPAD*) provide a good frame for this kind of graduate courses. Students are able to carry out research and properly apply the methodology.

REFERENCES

Johnson, R. H. and Winchern, D. W. (1992) *Applied Multivariate Statistical Analysis* (3rd edition). New York: Prentice Hall.

Lebart, L., Morineau, A., and Piron, M. (1997) *Statistique Exploratoire Multidimensionnelle* (2nd edition). Paris: Dunod.

Peña, D. (2002). *Análisis de Datos Multivariantes*. Madrid: McGraw Hill.