SYSPAD: FRENCH METHODS OF DATA ANALYSIS

Stephane Corre - Coref

Introduction

Given a set of observed measures between entities (objects, products, persons, etc...) the multidimensional analysis find a representation of these entities as points in an euclidean space such that the interpoint distances in some sense match the observed proximities. A monotonic relationship between the interpoint distances and the observed proximities is assumed. The output from multidimensional analysis is a spatial representation consisting on a geometric configuration of objects. It reflects the 'hidden structure' in the observed proximities. The user can specify the dimensionality in which the representation is desired.
To sum up, the multidimensional analysis works on big multivariate data bases to reduce them in easy-reading set to understand the main effects given by the data.

These methods are very powerful in marketing survey and each time we need to understand results given by a large data matrix.

The French Methods for Data Analysis

In the last 20 years, some particular methods and statistical practices have been developed in France in the fields of data analysis.

- Correspondence analysis on contingency tables and multiple correspondence analysis.
- Hierarchical ascending clustering based on aggregation strategies according to variance and essentially using the chi-square metric or the euclidean metric.
- Techniques of the interpretation of the analysis on the analysis on principal component analysis or the correspondence analysis with helpful graphs and tables.

But the French methods are presently available only in the form of specific programs which are scattered and not very compatible.

- This situation is particularly harmful to the spread of these techniques outside france, and even to their large-scale use in France. How many firms, universities, and French research centres which have only a very incomplete (and sometimes even obsolete) set of programs that prevent them from using these programs with profit.
- While waiting for the possible development of projects that may be able group these techniques into more coherent and generally portable sets, it seemed to us very interesting to integrate the French methods and practices of data analysis into the SAS language with two main reasons:
  1. To give to the statistician, in a simple and homogeneous language, a synthesis of the techniques of data analysis practiced in France.
  2. To facilitate, through a connection with a very widespread product the promotion of the French methods abroad.

SYSPAD Procedures

SYSPAD procedures are SAS procedures. These procedures analyse SAS data sets, compute statistics, print outputs and also can create others SAS data sets.

*** PROC CORRESP ***

CORRESP procedure is based on a factorial correspondences analysis. It computes the eigenvalues and eigenvectors of the analysis, prints the interpretation aids for the observations and the variables (contributions, coordinates, squared cosinus) and saves the factors in a SAS data set.
These factors have to be saved to be visualized with a graph when using the graph procedure (PROC GRAPH) or to be used in a clustering analysis (PROC CAH).
Input data can be standard contingency tables, burt tables or qualitative data coded with binary variables (0 or 1).

Syntax

```
PROC CORRESP   OPTIONS;
          ID   VARIABLE ;
          IDOBS   VARIABLES;
          VAR  VARIABLES;
          FREQ   VARIABLE;
          BY  VARIABLES;
          VSUP  VARIABLES;
```

with the options we can mainly specify;

- the name of the SAS data set entering
- the name of the data which will contain the factors
- the type of the data set to be analysed
- the number of eigenvalues and eigenfactors to be printed on the output listing
- other options are available to help the interpretation (CF SYSPAD User's Guide)

Missing Values

CORRESP does not treat the observations when one variable is missing or negative. An observation filled only with zeros is omitted.

Table 1: CORRESP OUTPUTS

Parameters and Options

DATA....: EX.TELPER
EDITION.: SORTED
VALP....: 3
VECP....: 3
CONTVAR.: 5
CONTOBS.: 3
INTVSUP.: 0
INTOBSUP: 0
ESPACE..: 0
ECVAR...: 0
ECOBS...: 0
MOD.....: CONT
OUT.....: EX.OUTCO
FACREC..: 5
FILL....: ALL

Number of Variables and Observations

ACTIVES VARIABLES.....: 6
SUPPLEMENTARY VARIABLES.....: 0
OBSERVATIONS.....: 6
SUPPLEMENTARY OBSERVATIONS.....: 0
DELETED OBSERVATIONS.....: 0

List of Actives Variables

FARMERFA DEALERFA CAPINDFA LIBPROFA CLERKFAT MANUALFA

TOTAL INERTIA: 0.63665079

TABLE OF EIGENVALUES AND EIGENVECTORS

NUMBER	EIGENVALUE 0	EIGENVALUE 1	EIGENVALUE 2	EIGENVALUE 3
VECTOR	1.00000000	0.36498896	0.19882025	0.07284158
AXIS 1	0.54241837	-.79807750	-.25757406	0.05001782
AXIS 2	0.42257713	0.27944986	-.12019171	-.75932683
AXIS 3	0.34255939	0.36540081	-.34803201	0.30136770
AXIS 4	0.28272243	0.32775674	-.32618575	0.49400637
AXIS 5	0.30304576	0.19496067	-.01262220	-.19089211
AXIS 6	0.48795004	0.07764137	0.83158091	0.22274738

Table 1 (Continued)

EIGENVALUES

NUM	EIGENVALUE	PERCENT	CUM.PCT	PLOT
1	0.36498896	55.627	55.627	**********
2	0.19882025	30.302	85.929	**********
3	0.07284158	11.102	97.030	*********

SORTED CONTRIBUTION

AXIS 1

VAR	1.AXIS	CTR	CO2	CO2	CTR	1.AXIS	VAR
FARMERFA	-8888	6369	9456	5783	1335	6444	CAPINDFA
				4640	1074	7004	LIBPROFA
				3775	781	3995	DELEARFA
				6127	380	3887	CLERKFAT
				153	60	961	MANUALFA

AXIS 2

VAR	2.AXIS	CTR	CO2	CO2	CTR	2.AXIS	VAR
CAPINDFA	-4529	1211	2858	9589	6915	7599	MANUALFA
LIBPROFA	-5143	1064	2504				
FARMERFA	-2116	663	537				
DEALERFA	-1267	144	380				
CLERKFAT	-185	2	14				

**** PROC PRINCIP ****

It carries out the principal component analysis with the same outputs and savings as PROC CORRESP (Correspondence analysis is a generalised principal components analysis with a CHI-SQUARE metric.)

The factors have to be saved in order to be displayed with the PROC GRAPH or the PROC CAH.

Input data are numeric variables or correlation matrix.

Syntax

```
PROC PRINCIP   OPTIONS;
       ID     VARIABLE;
       IDOBS  VARIABLES;
       VAR   VARIABLES;
       FREQ   VARIABLE;
       BY     VARIABLE;
       VSUP   VARIABLES;
```

Missing Values

When one value is missing for one or more variable, PROC PRINCIP does not use these observations for the analysis.
In case of centered analysis (option) constant variables are excluded.

**** PROC GRAPH ****

This procedure gives graphic representations associated to an analysis produced by PROC CORRESP or PROC PRINCIP.

Reading the data set issued from one of these procedures; PROC GRAPH runs simultaneous representations of pairs of factors for observations and variables. The elements are represented by labels (1 to 4 characters) as specified by the user.

The GRAPH procedure options:

- the choice of axis for the map
- the choice of points to be plotted
    - observations of the analysis
    - supplementary observations
    - any subset of observations characterized by the modality of a qualitative variable
    - variables of the analysis

- supplementary variables
- any combination of the points mentioned above on the same map

- the choice of an identification variable for the observations (any alphanumeric variable 1 to 8 characters in the data set) to be printed on the map
- the choice of the kind of representaion
  - density GRAPH
  - GRAPH point by point
- the choice of the centre of the GRAPH and the scales on the axes
- a zoom spotted on the centre of the GRAPH, or at any other point
- representation of the most significant point only
- printing a circle (radius=1) for the map of the principal components analysis variables

Syntax

```
PROC GRAPH   OPTIONS;
        AXIS  M1  M2   OPTIONS;
        BY   VARIABLES;
```

**** PROC CAH ****

The CAH procedure performs a hierarchical ascending clustering minimising at each step the reduction of the between class variance.

This procedure can be performed on three kind of data sets:

- contingency table, burt table of qualitative data logically coded
- numerical variables data set
- factors data set saved from CORRESP or PRINCIP procedures

The CAH procedure prints the nodes tables, the description of hierarchy clusters containing, the GRAPH of clustering tree variables contributions (interpretation aids). The nodes set is saved in a SAS data set which could be used with the PROC AFFECT to add a cluster variable to the initial data set.one, two or three partitions can be made with part tree printing and description of the initial partition clusters.

Syntax

```
PROC CAH    OPTIONS;
       ID  VARIABLE;
       BY  VARIABLES;
       VAR  VARIABLES;
       FREQ VARIABLE;
       WEIGHT  VARIABLE;
```

Missing Values:

When a missing value is found for a variable, the connected observation is excluded from clustering.

Table 2: CAH Outputs (Abstracts)

Number of Variables and Observations

VARIABLES.....	7
OBSERVATIONS.....	16

VARIABLES

MURDER RAPE ROBBERY ASSAULT BURGLARY LARCERY AUTO_TH

CITY CRIME

FROM THE U.S STATISTICAL ABSTRACT (1970)

PLOT FOR THE CLUSTERING TREE