ABSTRACT

The presentation of cluster profiles is a graph where each cluster corresponds to a line (its profile). The variables of the cluster analysis form the x-axis, the cluster means form the y-axis and the means of a cluster are joined by a straight line (Figure 6). The presentation is optimal if the order of the variables is such that the number of intersections of the cluster profiles is minimized.

The optimal presentation yields better interpretation of the clusters' interrelationships. This paper presents the CLPF SAS procedure (Cluster Profile) that produces the optimal presentation of the cluster profiles.

1. INTRODUCTION

In this paper the expressions "presentation of cluster profiles", "cluster profiles" and "profile" have the same meaning. About the optimal presentation of the cluster profiles in general see (6).

1.1 Problem

\[(A_{ij})\]

is the matrix of cluster means, where \(i=1,\ldots,n\) and \(n\) is the number of variables,

\[(D_{ij})\]

is the intersection matrix, where an element \(d_{ij}\) is the number of intersections in the profile between variable \(i\) and \(j\) when they are adjacent. The matrix is \(n \times n\) and symmetrical.

\[p,\ldots,\vec{p}\]

is a permutation of the cluster \(1\) to \(n\) variables.

\[S(p,\ldots,\vec{p}) = d_{12} + d_{23} + d_{3n} + \ldots + d_{n1}\]

is the total number of intersections in the profile, if the variables have the order of \(p,\ldots,\vec{p}\).

To obtain the optimal presentation of the profiles, a permutation of the variables has to be found such that \(S\) is minimum.

Figure 1 shows a profile of three clusters. The variables of the cluster analysis are \(V_1, V_2, \ldots, V_5\). The number of intersections in the profile is 8. Figure 2 shows the optimal profile of the same clusters. Here the number of intersections is only 2.

The problem of finding the optimal profile can be translated into a slightly modified traveling salesman problem. If a complete graph is created whose vertices are the cluster variables, then the intersection matrix defined above is a distance matrix of that graph. Our task is this: starting from a variable, the remaining variables must be visited exactly once creating a path with the minimum length. The path does not return to the starting variable.

Unfortunately the traveling salesman problem belongs to the class of NP-complete problems, thus we can not hope for an exact solution in reasonable time if the number of variables is large. Procedure CLPF has two methods to find the optimum solution.
1.2 Full enumeration

In order to obtain the exact solution, we have to examine every possible permutation of the variables and select that permutation which gives the minimum number of intersections in the profile.

Because the number of permutations increases factorially, this "full enumeration" method can be applied only if the number of variables is \( \leq 10 \).

Since the reverse permutations \( p_1, p_2, \ldots, p_n \) yield the same number of intersections (matrix \( D \) is \( i,j \) symmetrical), we need to examine only \( n!/2 \) permutations.

Procedure CLPF generates the permutations using Robinson's method (2). The second \( n!/2 \) permutations given by this algorithm are the reverse of the first \( n!/2 \) permutations.

Usually the number of variables is not very high or can be reduced by an earlier factor analysis; thus in most cases the "full enumeration" method can be used.

1.3 Approximate method

If the number of variables is greater than 10, CLPF uses an approximation algorithm. Since our distance matrix is symmetrical and satisfies the triangle inequality, the heuristic methods perform on it well. The approximate methods that CLPF is able to apply are three kinds of insertion methods: nearest, farthest and arbitrary (1, 3). These methods are simple, work fast and produce satisfactory results.

1.4 Transformation

The following formula transforms the cluster means into a new, uniform scale:

\[
a \rightarrow \text{scmin} + (\text{scmax} - \text{scmin}) \times \frac{a - \text{min}}{\text{max} - \text{min}}
\]

where \( \text{min} \) is the minimum of variable \( i \), \( \text{max} \) is the maximum of variable \( i \), \( \text{scmin} \), \( \text{scmax} \) defines the two endpoints of the new scale.

The number of intersections in the profile is invariant with respect to this transformation and procedure CLPF executes this transformation if it is requested.

2. PROCEDURE CLPF

2.1 Overview

Procedure CLPF calculates the cluster means, the intersection matrix, determines the optimal order of the variables, draws the graph of the optimal profile and creates an output data set. If the number of variables is greater than 10, the approximation method is used. If the number of variables \( \leq 10 \), the user can specify the method (the default is "full enumeration").

The input is either the output of a cluster procedure or the user's own data set. The input data set must contain one cluster identifying variable and at least three cluster variables.

The procedure is written in FORTRAN and is wholly compatible with the SAS environment. It accepts the SAS system options such as LINESIZE, OVP/NOOVP, CENTER, C48/C60 etc.

2.2 Specification

PROC CLPF options:

VAR variables;
CLUSTER variable;

The options that may appear in the PROC CLPF statement are given below:

- DATA=SASdataset gives the name of the input data set. If it is omitted, the most recently created SAS data set is used.
- OUT=SASdataset requests that CLPF create a new SAS data set containing the optimal order of the variables and the cluster means.
- NOPRINT suppresses all printed output.
- PLOT causes CLPF to produce a plot of the cluster profiles. The cluster variables form the x-axis and the cluster means form the y-axis. The plot normally occupies one page, unless VPOS or HPOS is specified.
- VPOS=n specifies the number of print positions on the vertical axis.
- HPOS=n specifies the number of print positions on the horizontal axis.
- NOTICK suppresses the tickmarks on the vertical axis.
- METHOD=name specifies the method of finding the optimal order of the variables.
METHOD=FULL  "full enumeration" method  
   (if the number of variables is greater than 10
   then this is replaced by METHOD=APPR). 
METHOD=NINS  nearest insertion method. 
METHOD=FINS  farthest insertion method. 
METHOD=AINS  arbitrary insertion method. 
METHOD=APPR  the best of the above
   three insertion methods.
METHOD=NONE  no optimal order is required.
   This option is used when the only
   purpose of using CLPF is to create the cluster
   profile with the order of variables defined in
   the VAR statement.
SCALEMIN=n  gives the minimum and
SCALEMAX=n  maximum endpoints of the
   new scale if the cluster
   means have to be trans­
   formed into a new scale.

Statement "VAR variables;" defines the names of the
   cluster variables. They are all numerical
   variables and their number is at least
   three. If the user omits this statement, all
   numerical variables but the cluster identifying
   variable are taken.
Statement "CLUSTER variable;" or "CL variable;"
   gives the name of the cluster identifying vari­
   able that can be either numerical or charac­
   ter. This statement must appear in a CLPF
   procedure.

2.3 Output data set

CLPF creates an output data set that can be an
   input of a SAS/GRAPH procedure. The variables
   in the data set are these:
   CLUSTER, the cluster identifying variable,
   ORDER, the order of the variables in the
   optimal profile (its values are
   integers between 1 and n),
   VARIABLE, the order of the variables in the
   optimal profile (its values are
   integers between 1 and n followed
   by the names of the variables),
   MEAN  cluster means.
   The number of observations in the data set is
   n*k.

2.4 Using CLPF with other SAS procedures

If we want the optimal cluster profile drawn by
   GPlot, we have to request the output data set
   and use the following PLOT statement in GPlot:
   PLOT MEAN = VARIABLE = CLUSTER;

The input of CLPF can be an output data set of
   a SAS cluster analysis procedure. When the
   cluster procedure is CLUSTER, we have to follow
   the next commands:

   PROC CLUSTER OUTTREE=TREE ... ;
      VAR variables;
   PROC TREE DATA=TREE OUT=CLUS LEVEL=n
      NOPRINT ... ;
      COPY variables;
   PROC CLPF DATA=CLUS ...
      VAR variables;
      CL CLUSTER;

   LEVEL=n defines the number of clusters the user
   wants to get. The variables in the VAR and
   COPY statements are identical.

When the cluster analysis is FASTCLUS, we can
   follow the next two procedures:
   PROC FASTCLUS OUT=CLUS ...;
      VAR variables;
   PROC CLPF DATA=CLUS ...;
      VAR variables;
      CL CLUSTER;
   or
   PROC FASTCLUS MEAN=CLUS ...;
      VAR variables;
   PROC CLPF DATA=CLUS ...;
      VAR variables;
      CL CLUSTER;

2.5 Missing data

If an observation has a missing value for the
   cluster identifying variable or any of the
   variables in the VAR statement, it is
   omitted from the analysis.

3. Example

Figure 3 is a SAS program that processes the
   well known "Fisher (1936) Iris Data". The data
   are taken from the SAS Sample Library and form
   the data set FISHER.IRIS. The first CLPF runs
   with METHOD=NONE and draws the graph of the
   cluster profiles having the original order of
   variables (Figure 4). The points in the graph
   are connected manually. Note the number of
   intersections in the profile is 4. The second
   CLPF uses the default method of "full enumera­
   tion" and gives the optimal order of the vari­
   ables. In this case the number of intersections
   is only 2 (Figure 5). An output data set is
   created and it is the input of the GPlot pro­
   cedure. The plot of the optimal profile is in
   Figure 6.

In this example the number of intersections in
   the profile decreased from 4 to 2 (which is the
   minimum). If the number of clusters and/or the
   number of variables is higher, the reduction
   can be greater.

4. Discussion

This paper describes a SAS procedure called
   CLPF. It is run after a cluster analysis pro­
   cedure and produces the optimal presentation of
   the cluster profiles. This optimal presentation
   gives the user an opportunity for better inter­
   pretation of the interrelationships between the
   clusters.
5. REFERENCE


6. MORE INFORMATION

For more information please contact the author:

Akos Felsovalyi
J.C. Penney, 39th Floor
1301 Avenue of the Americas
New York, NY 10019
(212) 957-6696

SAS and SAS/GRAPH are registered trademarks of
SAS Institute Inc., Cary, NC, USA

```
PROC FASTCLUS DATA=FISHER.IRIS MAXC=3 MAXITER=10 MEAN=CLUS;
  TITLE 'FISHER (1936) IRIS DATA';
PROC CLPFF DATA=CLUS METHOD=NONE PLOT VPOS=30 HPOS=72;
  VAR SEPALLEN SEPALWID PETALLEN PETALWID;
  CL CLUSTER;
PROC CLPFD DATA=CLUS OUTF=PROFILE;
  VAR SEPALLEN SEPALWID PETALLEN PETALWID;
  CL CLUSTER;
DATA PROFILE;
  SET PROFILE;
  LENGTH SPECIES $ 10;
  IF CLUSTER EQ 1 THEN SPECIES='SETOSA';
  ELSE IF CLUSTER EQ 2 THEN SPECIES='VIRGINICA';
  ELSE SPECIES='VERSICOLOR';
PROC GPLOT DATA=PROFILE;
  PLOT MEAN * VARIABLE = SPECIES;
  SYMBOL2 C=RED I=JOIN L=1;
  SYMBOL3 C=GREEN I=JOIN L=2;
  SYMBOL1 C=BLUE I=JOIN L=3;
  TITLE 'F=SIMPLEX H=1 FISHER (1936) IRIS DATA';
  TITLE2 OPTIMUM CLUSTER PROFILE;
  FOOTNOTE 'F=SIMPLEX H=1 Figure 6';
```

*Figure 3*
FISHER (1936) IRIS DATA

CLUSTER PROFILE

<table>
<thead>
<tr>
<th>SEPALLEN</th>
<th>SEPALWID</th>
<th>PETALLEN</th>
<th>PETALWID</th>
</tr>
</thead>
<tbody>
<tr>
<td>50.0800</td>
<td>34.2800</td>
<td>14.6200</td>
<td>2.4600</td>
</tr>
<tr>
<td>68.5000</td>
<td>30.7368</td>
<td>57.4211</td>
<td>20.7105</td>
</tr>
<tr>
<td>59.0161</td>
<td>27.4839</td>
<td>43.9355</td>
<td>14.3387</td>
</tr>
</tbody>
</table>

INTERSECTION MATRIX

<table>
<thead>
<tr>
<th></th>
<th>SEPALLEN</th>
<th>SEPALWID</th>
<th>PETALLEN</th>
<th>PETALWID</th>
</tr>
</thead>
<tbody>
<tr>
<td>SEPALLEN</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>SEPALWID</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>PETALLEN</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>PETALWID</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

VARIABLE ORDER IN THE OPTIMUM CLUSTER PROFILE

1 = PETALWID  2 = PETALLEN  3 = SEPALLEN  4 = SEPALLEN

NUMBER OF INTERSECTIONS IN THE PROFILE = 2

Figure 4

Figure 5
FISHER (1936) IRIS DATA
OPTIMUM CLUSTER PROFILE

LEGEND: SPECIES
--- SETOSA
--- VERSICOLOR
--- VIRGINICA

Figure 6

977