GRAPHICAL SOLUTIONS FOR MARKET INTELLIGENCE

Frederick Pratter, Abt Associates Inc., Cambridge, MA

Perceptual mapping has been in common use as an analytic tool in market research for a number of years, yet there are probably almost as many definitions of the term as researchers who use it. A generally acceptable definition would certainly include the following features. First, a group of respondents (customers) are asked to compare a set of products or brands on a collection of attributes. The ratings can include customers' perceptions as well as preferences, as long as the data are ordinal. These ratings or rankings are then combined into linear combinations (principal components) that each account for some unique portion of the variation in the original data. Finally, these linear functions are plotted in order to display the information graphically.

The SAS® System can be used to produce several kinds of perceptual maps, in particular *multidimensional preference (MDPREF) plots*. The SAS Institute offers a two-day public course called *"Marketing Research: Practical Applications Using the SAS® System"* which covers several graphical techniques, including MDPREF. The original idea for this paper came out of some of the discussions in this course, as well as from several papers by Warren Kuhfeld and others at the SAS Institute.[1]

It should also be noted that the graphical methods presented here are not "cutting edge" technologies. They can all be run using only the SAS/STAT® and SAS/GRAPH® products; SAS/IML® or SAS/AF® is not required. The focus is instead on *why* one would want to use them--how this kind of perceptual mapping can help to discover research solutions to market questions.

The 1980s car preference data set (included as PRQEX1 in the SAS/STAT® sample library) is surely one of the most analyzed response sets in market research. Even for those who do not remember the Ford Pinto or the Plymouth Volare, the original MDPREF analysis is nonetheless a useful starting point. Based on that earlier example, some (hopefully) new ideas will be illustrated using these well worn data.

Twenty five raters were asked their opinion (on a scale of 1 to 10) of 17 automobiles, ranging from the Pinto to the Cadillac Eldorado. They were also asked to judge the reliability, mileage, and ride of the vehicles (on a five point scale). A principal components analysis was carried out on the ratings of each, resulting in the factor solutions illustrated in the *scree* plot shown in Figure 1. The code that produced this plot is simply:

```
proc factor data=CARPREF scree;
   var subj1-subj25;
```

For those unfamiliar with the use of principal components analysis to reduce the number of observations, this must seem like a strange idea. Usually, one uses PROC FACTOR to cluster variables, not respondents.[2] Nevertheless, this transposition of the usual technique is essential to how the SAS System does MDPREF. The first principal component is the linear combination of individual ratings that accounts for the most variation among respondents, the second principal component accounts for the next most variation, and so forth. The scree plot compares the "eigenvalue" (variance) attributable to each factor to the number of factors. As Figure 1 shows, there are five factors with eigenvalues greater than 1.0, but the first three seem to explain a relatively large proportion of the variation in the original data. One would be led to the conclusion that the three factor solution is the most parsimonious.

In practice, PROC PRINQUAL is used instead of PROC FACTOR, and the results are fed into PROC TRANSREG to get the points to be plotted. The following code example illustrates one way these data can be mapped; Figure 2 shows the resulting plot.[3]

```
proc prinqual data=CARPREF2 out=PRESULTS
         (drop=subj1-subj25)
   n=2 standard scores;
   id model mpg reliable ride;
   transform identity(subj1-subj25);
   title 'MDPREF Analysis';

proc transreg data=PRESULTS;
   model identity (mpg reliable ride)=
         identity(prin1 prin2);
   output coefficients replace out=TRESULT;
   id model;
```

The *standard* option in PROC PRINQUAL standardizes the principal component scores to mean 0 and variance 1, while the *scores* option includes the standardized scores in the output data set. The *identity* transformation specifies that the transformed variables are the same as the originals. (PRINQUAL can also find a monotonic transformation that preserves the ordering of the items but optimally scales them.) A two factor solution (n=2) is specified because only two dimensions can be plotted in a single graph (but see below). The **PRESULTS** data set includes the factor scores as well as the mean ratings by the 25 subjects on the three additional items: miles per gallon, ride and reliability.

The **TRANSREG** procedure step separately fits three regression models:

(1) ride = prin1 prin2
(2) reliable = prin1 prin2
(3) mpg = prin1 prin2

where *prin1* and *prin2* are the first two principal components produced in the previous step. Each of these three regression models produces a set of coordinates for the end point of an attribute vector. The options specify that the regression coefficients are to replace the items with the same names on the output dataset. Again, an identity transformation (that is, no transformation) is used.

The **%PLOTIT** macro (available from the SAS Institute) plots the vectors, labels the data points (nicely handling overlapping values), and outputs the plot to a graphic stream file (GSF). The syntax is as follows:

```
%plotit (data=tresult,
   datatype=vector 2.5,
   method=print)
```

%PLOTIT works by saving a printer plot to a file in the current directory using PROC PRINTTO, then displaying it as an Annotate data set using PROC GANNO. The macro supports several types of plots, of which the Preference Mapping Vector Model is only one.[4]

Only biplots are available in the macro, however. As noted in the discussion of the scree plot above, however, there seem to be least three principal components that could be plotted. Fortunately, the SAS® System offers a procedure, PROC G3D, that allows three dimensional surface and scatter plots. Figure 3 illustrates a plot of the first three dimensions, produced by the following code:

```
data LIST;
     set TRESULT;
     length text $16 sval style $8;
     if _type_='SCORE' then size=1.5;
            else size=2;
     text=_name_;
     x=prin1; y=prin2; z=prin3;
  retain color 'black' when 'a'
            xsys ysys '2' hsys '3'
            function 'label';

proc g3grid data=LIST out=GRID;
    grid prin2*prin1=prin3/spline
    axis1=-2 to 2 by .1
    axis2=-2 to 2 by .1;

proc g3d data=GRID anno=LIST;
   plot prin2*prin1=prin3/
            grid min=-3 zmax=2;
```

The PRINQUAL and TRANSREG steps are run as above, substituting *prin1, prin2,* and *prin3* in the regressions. Then an Annotate data set is created with labels for the points. The observations with _TYPE_ equal to 'SCORE' are used to plot the (x,y) coordinates for each the car model. The remaining observations are the coordinates for the three attributes: MPG, ride, and reliability. The raw coordinates are run through PROC G3GRID to smooth the curves, using a spline transformation and setting the x and y axis tick marks. Finally, PROC G3D creates the response surface map shown. (A little fudging with the *position* parameter-- not shown here-- was required to make sure the labels would not overlap, a problem that is handled automatically by the %PLOTIT macro.)

The resulting elegant graphic is sure to impress a client. It certainly is an improvement over the two dimensional biplot, and seems to convey more information. The problem is, what does it all mean? It seems reasonable somehow that the luxury cars by Cadillac and Lincoln are way over in one corner while the mini-compacts Pinto and Chevette are at the opposite side. But why is the Mustang at the top of a spike? Are the Fairmont, Rabbit, and Citation really so close in the judgement of the raters?

Unfortunately, using principal component analysis in this way to produce preference maps has a number of significant drawbacks. First, information is being thrown away by using the mean values of MPG, RIDE, and RELIABLE. There is likely to be an interaction effect between the individual rater's opinion of the car models and each one's judgement of the ride, reliability and mileage attributes. Using the mean values across the twenty five raters in the TRANSREG procedure obscures this effect. Second, the response data must be transposed so that *raters* are the variable dimension and car *models* are the observations. (This is the reverse of the usual questionnaire data file format.) With twenty five observations, PROC FACTOR iterates over twenty five variables; a 1000 responses would mean solving a 1000 variable problem. According to the documentation, the "amount of time that FACTOR takes is roughly proportional to the *cube* of the number of variables. factoring 100 variables therefore takes about 1000 times as long as factoring 10 variables."[5]

Finally, and perhaps most important, as Figure 3 illustrates, the principal components are *arbitrary* constructs. They are expressed in no particular units, and have no clear meaning. A better technique might be to find summary functions that are defined in terms of the actual responses. Fortunately, such a technique does exist, and is available as part of the SAS/STAT® package. "The CANDISC procedure derives *canonical variables* (linear combinations of quantitative variables) that summarize between class variation in much the same way that principal components summarize total variation."[6]

Canonical discriminant function analysis can be used to derive the set of functions that best *discriminates* between the car models, based on the raters' judgement of mileage, reliability, ride, and overall value. The advantage of derived canonical functions over principal components is that the resulting factors usually represent identifiable dimensions. Figure 4 illustrates the scores produced by the following:

```
proc candisc data=CARPREF            outstat=OUTCAN
short;
      class model;
      var rating mpg reliable ride;

proc transpose data=OUTCAN out=OUTRAN;
      by _type_ model notsorted;
      where _type_ in ('STRUCTUR','CANMEAN');
      var rating mpg reliable ride;

proc sort data=OUTRAN nodupkey;
   by _type_ can1 can2 can3 model;

proc print data=OUTRAN noobs;
      title 'Figure 4';
```

This program uses the individual ratings rather than the transposed data set, so the interaction effects are preserved and the number of observations is not problematic. The records on the output listing with a _TYPE_ of 'CANMEAN' are the class means of the canonical variables by model, while the 'STRUCTUR' records are the correlations between the canonical variables and the attributes. From a review of the latter it is evident that CAN1 represents MPG, CAN2 is ride, CAN3 is reliability, and CAN4 the overall rating by model type.

The first three canonical variables can be plotted just like the first three principal components. Figure 5 shows the result. The following code produces a three dimensional scatter plot:

```
data LIST;
   set OUTRAN;
   where (_type_ eq 'CANMEAN');
   select(model);
      when ('Accord')
                     text='Accord/Civic';
      when('Civic') delete;
      when('Fairmont')
                     text='Fairmont/Malibu';
      when('Malibu') delete;
      otherwise text=model;
   end;
   x=can2; y=can1; z=can3;
   retain function 'label'
            hsys '3' xsys ysys '2'
            style 'swissb' size 1.5;
     title h=2 'Figure 5. Plot of
     Canonical Variables';

proc g3d data=LIST;
     scatter can1 * can2 = can3/
     anno=LIST
  shape='POINT'
  xticknum=6 yticknum=6 zticknum=6
  zmin=-3 zmax=3;
run;
```

The Annotate data set is quite simple to create; the only wrinkle is that the two vehicles by Honda as well as the Fairmont/Malibu combination had exactly the same mean ratings of ride, mileage, and reliability. Each vehicle is shown in precise relationship to all of the others on these three dimensions. Alas, there is not all that much to be learned from a display of the relative positions in this trivial example. effective manner.

It is important to note that in practice these graphics are not well suited for presentation. There is not currently any way in SAS/Graph to scale the axes of the plots; the **haxis** and **vaxis** commands are not supported. More seriously, there is no way to prevent label collision on the plots (as there is in the **%plotit** macro). This can mean many tedious hours are necessary to get each plot to be clear and readable. Consequently, final presentation graphics are usually prepared in some other desktop plotting package. For the purposes of exploratory data analysis, however, SAS/Graph is unsurpassed in its ability to produce multiple slices of the data.

In conclusion, Multidimensional Preference plots, whether based on principal components or canonical discrimination function analysis, can be a useful tool for compiling market intelligence. As brand management becomes more and more of an issue in the marketplace, the ability of the SAS System to produce perceptual maps provides an increasingly important capability. The combination of sophisticated statistical tools with an integrated graphical capability allows the visualization of complex product interrelationships that could not be easily distinguished from even the densest tabular presentation.

References

[1] Warren F. Kuhfeld. *Marketing Research Methods in the SAS® System: A Collection of Papers and Handouts.* SAS Institute Inc., Cary, NC. August 9, 1993.
Marketing Research: Practical Applications Using the SAS® System, Course Notes. SAS Institute Inc., Cary, NC. 1996.
SAS® Technical Report P-229, SAS/STAT® Software: Changes and Enhancement, Release 6.07. Cary, NC: SAS Institute Inc.. 1992. cf. "Plotting Points with Labels Using the Plot Procedure", pp.481-498; "The PRINQUAL Procedure", pp.503-510; and "The TRANSREG Procedure", pp. 551-596.
SAS® System for Statistical Graphics, First Edition. Cary, NC: SAS Institute Inc., 1991. cf. "Annotated Three-Dimensional Scatterplots", pp.207-210; "Biplot: Plotting Variables and Observations Together", pp.437-445; and "Canonical Discriminant Analysis and MANOVA" pp.490-498.

[2] *SAS/STAT® User's Guide, Version 6, Fourth Edition, Volume 1.* Cary, NC: SAS Institute Inc., 1989. cf. "Introduction to Clustering Procedures", p.53-55.

[3] *Course Notes,* p. 187.

[4] The %PLOTIT macro is documented internally and in Kuhfeld, 1993

[5] *SAS/STAT® User's Guide, Volume 1,* p.797

[6] *SAS/STAT® User's Guide, Volume I,* p.387.