

Bootstrapping a Multidimensional Preference Analysis

E. Barry Moser and Xiaoming Liang, Department of Experimental Statistics,
Louisiana State University Agricultural Center, Baton Rouge, LA 70803-5606

ABSTRACT

Preference data are frequently collected to help decision makers prioritize items, or are measured on subjects, such as animals, and used to help infer preference for or avoidance of resources. Multidimensional preference analysis may be used to construct a biplot of the items and subjects in the space of a few components as an aid to interpretation and presentation of results. The variability associated with the objects in the plot, however, is frequently not shown. The SAS/STAT® survey sampling procedure SURVEYSELECT is used to select bootstrap samples from a preference data set, then multidimensional preference analysis is performed with the SAS/STAT® PRINQUAL procedure to produce estimates of the variabilities of objects. Next the SAS/GRAPH® Gplot procedure is used to produce a variety of graphics using the bootstrap information that can then show the precision with which objects are estimated in the components space.

INTRODUCTION

Preference data arise in a variety of situations where items are rank-ordered according to some preference (or avoidance) of the items. In food science, a set of food preparations may be rank-ordered by judges based upon flavor characteristics. In marketing, product presentations may be ranked by a sample of potential consumers, while in ecology, habitat types may be "rank-ordered" by animals based upon their frequency of usage of habitats as determined from radio telemetry data. In an agricultural economics project of Professor Kenneth Paxton, Louisiana State University Agricultural Center, cotton farmers were asked to rank-order a dozen research thrusts that the Agricultural Center might pursue with respect to cotton production. Upon consultation with Dr. Paxton, one of the analyses that we suggested was a multidimensional preference analysis.

Multidimensional preference analysis (MDPREF) is one of a number of perceptual mapping techniques for graphically displaying product preferences in a low dimensional space of principal components (Carroll, 1972; see Kuhfeld, 1992). In the data matrix, the columns or variables correspond to the judges, while the rows or observations correspond with the items or products. A "biplot" can be produced where the items are given as points in the display, while the judges can each be plotted as vectors emanating from the origin. The longer the vector, the more of that judge's information that is contained in the display. The items are projected onto the vectors to interpret which items are most and least preferred by a judge. In a trained panel, such as in food science, it is often hoped that the judges are homogeneous and so their vectors would be coincident in a perfect world. In product presentation studies, variations among consumers, as indicated by different directions of their vectors, may be related to demographic or other characteristics of the consumers. In some cases, these characteristics can also be superimposed on the biplot.

Unlike regression analysis, these vector and item points do not come with a standard error or confidence interval formula. Thus although the data in the study may be a sample from a much larger population of interest, the MDPREF results are presented as if the sample is the population of interest. Since we are dealing with rank-ordered responses that have special mathematical properties, we decided to use a Monte Carlo method to derive measures of variability and confidence regions for the MDPREF analysis. In particular, bootstrap samples (see Manly, 1997) were constructed and the replications used to build confidence regions. Monteleone et al. (1998) used a similar approach to form preference maps for starchy food consumption.

JOURNAL PREFERENCES EXAMPLE

E. Roskam (as taken from Giffi, 1990) reported on journal preferences in a psychological research area as determined by 39 scientists. Each scientist rank-ordered the list of 10 journals as to their importance to the field. Note that for this type of data, not only do we have positive integer values, but also the sum of the ranks for each scientist is a constant. Thus multivariate normality is not likely plausible, and the dimensionality of the problem is at least 1 less than the number of items being ranked. A SAS® program was used to construct the MDPREF analysis of these data, then to bootstrap the data set, and to construct indicators as to the variability of the original display.

First, the original data set is constructed. In most settings, the observations will consist of the responses from each judge, while the columns will correspond with the items being ranked. Notice that this data organization is the transpose of the way in which we will need to have the data for analysis.

```
Data Roskam;
  Input Judge JEXP JAPP JPSP MVBR JCLP
         JEDP PMET HURE BUU HUDE;
Datalines;
1. 7 4 1 8 10 9 5 2 3 6
2. 7 6 2 9 3 8 10 1 4 5
3. 10 5 1 7 4 6 8 2 3 9
---- Data Omitted ----
38. 2 3 6 5 7 8 4 9 1 10
39. 2 6 7 3 10 8 4 9 1 5
;
```

Next the data matrix is transposed to get it into the proper form (PROC TRANSPOSE), the items are reverse-ranked so that the most important items have the highest rank (PROC RANK), and then the MDPREF analysis is performed using the PRINQUAL procedure. The %PLOTIT macro of SAS/STAT® is used to display the results from the PRINQUAL procedure. The basic code used is:

```
Proc Transpose Data=Roskam Out=Roskam2
  Prefix=Judge Name=Journal;
  Var JEXP--HUDE;
  Id Judge;
Run;

Proc Rank Data=Roskam2 Out=Roskam3
  Descending;
  Var Judge1-Judge39;
Run;

Proc Prinqual Data=Roskam3 Out=PQResults
  N=2 Replace MDPREF MaxIter=150;
  Id Journal;
  Transform Monotone(Judge1-Judge39);
Run;

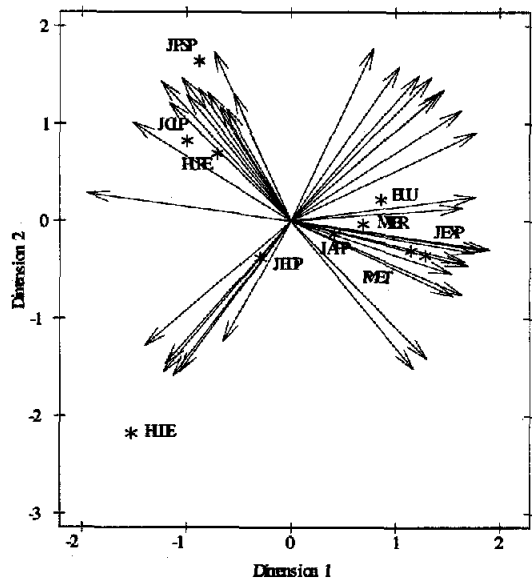
Proc Print Data=PQResults;
Run;

%Plotit(Data=PQResults,
  Datatype=MDPREF 2);
```

The results from this analysis are shown in the following figure. There, it is observed that there appear to be different

POSTERS

groups of scientists in terms of what journals they prefer. For example, the journal HUDE (*Human Development*) is highly preferred by about 6 judges, 5 of which happen to be from Developmental and Educational Psychology departments.



BOOTSTRAP ANALYSIS

Let n be the size of the original data set. The basic bootstrap process takes a simple random sample with replacement of size n from the original data set. The SURVEYSELECT procedure using the METHOD=URS with $n=39$ provides such a sample. In fact, by setting REP=M, where M is some positive integer, we can generate M such bootstrap samples in a single call to SURVEYSELECT. Note that the original data set (though of reverse-ordered scores) is bootstrapped, as we believe that the scientists (judges) are the random sampling units, not the journals.

```
/*
 * Rearrange like original data but with
 * reverse-ranked values
 */
Proc Transpose Data=Roskam3
    Out=Roskam4 (Drop=_LABEL_) Name=Judge;
Var Judge1-Judge39;
Id Journal;
Run;

Proc SurveySelect Data=Roskam4 Method=URS
    n=39 Rep=1000 Out=Boots;
Run;
```

Once the bootstrap samples are selected, the data are again transposed, but BY REPLICATE, and the PRINQUAL procedure is again called to perform the analysis BY REPLICATE.

```
Proc Transpose Data=Boots Out=Trans
    Prefix=Judge Name=Journal;
Var JEXP--HUDE;
By Replicate;
Run;

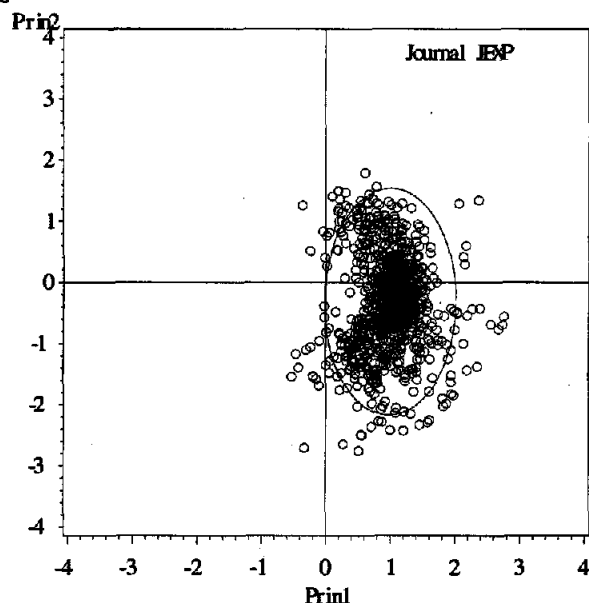
/*
 * Perform separate MDPREF analyses on each rep
 */
Proc Prinqual Data=Trans Out=PQResults2 N=2
    Replace MDPREF NoPrint MaxIter=150;
Id Journal;
Transform Monotone (Judge1-Judge39);
By Replicate;
Run;
```

It is important to note that the direction of a principal component is unimportant. Thus, it is possible that one result is a reflection of one or more axes of the original solution, and so each solution from each replicate must be registered relative to the original solution. We did this by measuring the sum of squared errors from each item point in the bootstrap solution to the item points in the original solution and selecting appropriate reflections to minimize this error. Thus we assumed that the appropriate registration for a solution is the one that "looks" most like the original solution. The code for this involved a couple of data steps to first compute the errors and find the best solution, then a second step to make the necessary conversions to the components in the solutions.

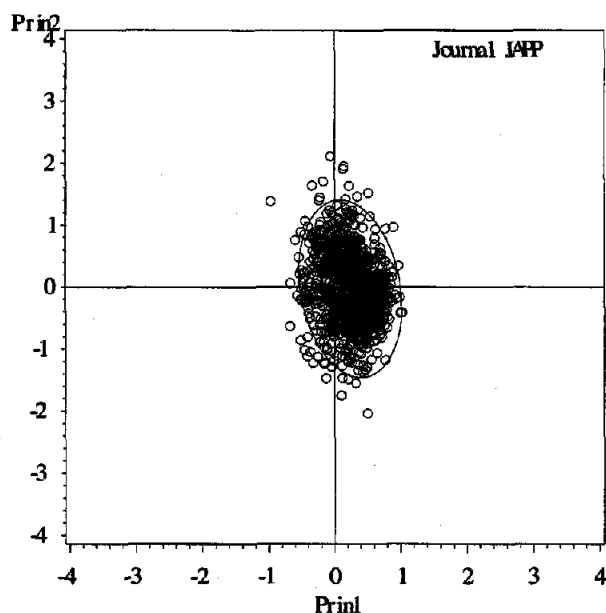
Next the %CONELIP macro obtained from the SAS Institute support web site was used to construct 95% confidence ellipses for each journal using the bootstrap replicates. A simple SAS macro was used to process the results for each journal separately, then to build a composite data set containing all of the ellipses.

```
%Macro Ellipses;
GOptions Reset=Symbol Reset=Axis;
Proc Datasets Library=Work NoList NoWarn;
Delete All;
Run;
Quit;
%Let Journals=JEXP JAPP JPSP MVBR JCLP
JEDP PMET HURE BUU HUDE;
%Do J=1 %To 10;
%Let Journal=%Scan(&Journals,&J);
%Conelip(Data=Scores(Where=(_NAME_=
    "&Journal")), Out=&Journal,
    x1=Prin1, x2=Prin2, mean=no, conf=.95);
Data &Journal;
Length Journal $4.;
Retain Journal "&Journal";
Set &Journal;
Prin1=X1; Prin2=X2;
Run;
Proc Append Data=&Journal Base=All;
Run;
Title1 "Journal &Journal";
Proc GPlot Data=&Journal;
Plot Prin2*Prin1=Id / NoLegend
    VAxis=Axis1 HAxis=Axis1 VRef=0 HRef=0;
Axis1 Length=4in Order=(-4 To 4 By 1);
Symbol1 C=Black V=CIRCLE H=1 I=None;
Symbol2 C=Black V=None I=Join L=1;
Run;
Quit;
%End;
%Mend Ellipses;
%Ellipses;
```

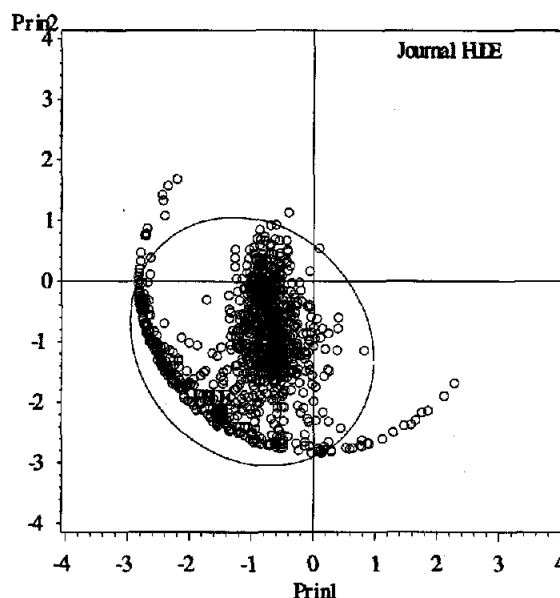
The replicates and superimposed ellipse for the journal JEXP are given below.



Notice that there is quite a bit of variability in the points and that the ellipse doesn't capture the data very well. On the other hand, the points tend to be in the southeast to east direction from the origin.

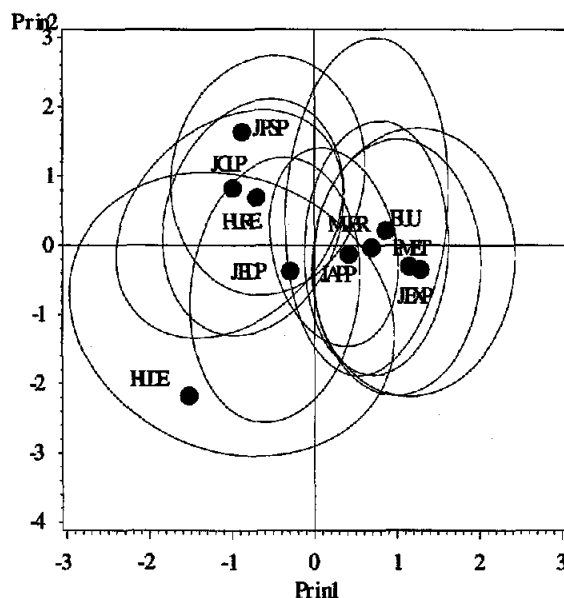


For the JAPP journal shown above, the points congestate a lot near the origin suggesting that this journal is not well represented in this presentation, or is generally preferred in the middle ranks by most everyone. A third journal, HUDE, demonstrates the wide variability that an item might exhibit. For this journal, a number of scientists ranked it as their number 1 journal, while several others ranked it at the very bottom of their lists.



We constructed these plots for each journal, and then the set of ellipses was plotted in a single figure using the SAS/GRAPH code:

```
GOptions Reset=Symbol Reset=Axis;
Proc GPlot Data=ALL (Where=(Id=2))
  Annotate=Centers;
  Plot Prin2*Prin1=Journal / VAxis=Axis1
    HAxis=Axis1 NoLegend VRef=0 HRef=0;
  Axis1 Length=4in;
  Symbol1 C=Black V=None I=Join L=1 R=10;
Run;
Quit;
```



The magnitudes of the ellipses are related to the variability that they exhibit through the selection of sampling units from the population.

CONCLUSION

The bootstrapping technique provides considerable insight into the variability in results that could occur simply as a consequence of the individuals selected into the sample assuming that the original sample is a simple random sample. For example, the journal Human Development (HUDE), shows a considerable range over which its placement could occur and is likely a result of the groups of individuals that ranked the journal either as 1 or 10, and so as these group sizes change as a consequence of sampling, so does the placement of the point. Note that here the sample size of 39 is small and so a few observations can have a profound impact on the results. For larger data sets, this technique should produce much more stable results.

Though confidence ellipses based upon the bivariate normal distribution were used to approximate the region occupied by the journal vectors, the bivariate normal model does not appear to be a good model for several of the journal points. Some nonparametric method such as hull peels (Green, 1981) may be more useful in these situations.

Finally there are a few points corresponding with a few bootstrap replications that appear to be miss-registered as viewed from the plots. Thus a more elegant algorithm for insuring correct registration may be required.

REFERENCES

- Carroll, J.D. 1972. Individual differences and multidimensional scaling. Pages 105-155 in R.N. Shephard, A.K. Romney, and S.B. Nerlove, eds. *Multidimensional Scaling: Theory and Applications in the Behavioral Sciences*, Vol. 1., Seminar Press, NY.
- Gifi, A. 1990. *Nonlinear Multivariate Analysis*. John Wiley & Sons, New York, p. 183.
- Green, P.J. 1981. Peeling bivariate data. Pages 3-19 in Barnett, V., ed. *Interpreting Multivariate Data*. John Wiley & Sons, Inc., NY.
- Kuhfeld, W.F. 1992. Marketing research: uncovering competitive advantages. *Proceedings of the SAS Users Group International Conference* 17, 1304-1312.
- Manly, B.F.J. 1997. *Randomization, Bootstrap, and Monte Carlo Methods in Biology*. Chapman and Hall, Inc., London, 399pp.
- Monteleone, E., L. Frewer, I. Wakeling, and D.J. Mela. 1998. Individual differences in starchy food consumption: the application of preference mapping. *Food Quality and Preference* 9 (4), 211-219.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:

E. Barry Moser
Louisiana State University Agricultural Center
Department of Experimental Statistics
Baton Rouge, LA 70803-5606
Work Phone: 225-578-8303
Fax: 225-578-8344
Email: bmoser@lsu.edu
<http://www.stat.lsu.edu/faculty/moser>