

Paper 90-28

Determining the Dimensionality of Data: A SAS[®] Macro for Parallel AnalysisRobert I. Kabacoff, Ph.D., Management Research Group[®], Portland, ME**ABSTRACT**

While parallel analysis has received considerable empirical support as a method for uncovering the dimensionality of multivariate data, it is rarely used in real world research. One reason for its underutilization is the analyst's lack of access to the methodology in existing statistical packages. This paper provides a SAS macro that allows researchers to easily employ parallel analysis as a method of determining the dimensionality of their data.

INTRODUCTION

A fundamental step in the analysis of multivariate data is the determination of its dimensionality, the so called 'how many factors?' question. Usually the analyst will examine the eigenvalues of the sample correlation matrix in an attempt to answer this question. Numerous rules of thumb exist to aid the analyst, including Kaiser's eigenvalues greater than one rule (Kaiser, 1960), Cattell's scree test (Cattell, 1966), the Maximum Likelihood test (Joreskog, 1967), and Horn's parallel analysis (Horn, 1965). Each has advantages and disadvantages.

While numerous studies have supported the relative merits of the parallel analysis approach (Haikstian, Rogers, & Cattell, 1982; Glorfeld, 1995; Zwick & Velicer, 1986; Hubbard and Allen, 1987; Buja & Eyuboglu, 1992), the technique is rarely used in real world data analysis applications. One reason that the approach is underused is that the computations involved are complex and time consuming and the methodology is not provided as an option in standard statistical packages. This paper presents a simple to use SAS macro that should aid in making the technique more readily available to both researchers and analysts.

METHODOLOGY

Let N represent the number of observations in the dataset, and p represent the number of variables. In general, parallel analysis is completed as follows:

1. Calculate the $p \times p$ sample correlation matrix from the $N \times p$ sample dataset. Create a scree plot by plotting the eigenvalues of the sample correlation matrix against their position from largest to smallest (1, 2, ..., p) and connecting the points with straight lines
2. Generate a simulated dataset with N observations randomly sampled from p independent normal variates. Calculate the $p \times p$ correlation matrix for this simulated data and extract the p eigenvalues and order them from largest (position 1) to smallest (position p).
3. Repeat step 2, k times (e.g., k = 1000).
4. Calculate the median of the of the k eigenvalues at position 1 from steps 2-3, the median of the k largest eigenvalues at position 2, ..., up to the median of the k eigenvalues at position p.
5. Overlay these medians from step 4 on the scree plot created in step 1, connecting the points.
6. The intersection of the two lines is the cutoff for determining the number of dimensions present in the data.

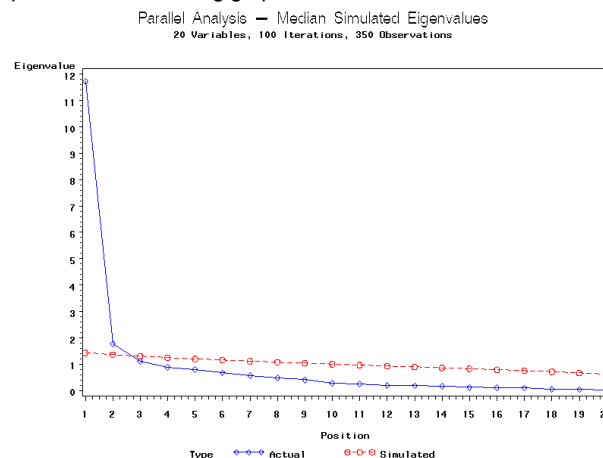
Note: other summary statistics can be used in place of the median at step 4 (e.g., 99th percentile, 95th percentile, etc.).

EXAMPLE

Three hundred fifty mid-level executives were rated by their bosses on 20 measures of leadership effectiveness. The analyst was interested in identifying the number of dimensions underlying this 20 variable dataset to both understand the ratings better and to hopefully create more fundamental and reliable summary measures. The call to the `%parallel` macro given below

```
%parallel(data=Leaders, var=Rating1-Rating20, niter=1000,
statistic=Median);
```

produced the following graph:



Examination of the graph suggests the presence of two underlying dimensions. An examination of the factor loadings suggested that one dimension consisted of items pertaining to business skills, while the other dimension consisted of items pertaining to interpersonal skills. The 20 behaviors were then combined to produce two summary scales (business and people skills) that could be used in subsequent research.

CONCLUSION

Parallel analysis can be a valuable addition to the toolbox of the researcher analyzing multivariate data. The `%parallel` macro can be used to generate Monte Carlo simulations useful for identifying the number of dimensions underlying a set of data.

REFERENCES

- Buja, A. & Eyuboglu, N. (1992). Remarks on parallel analysis. *Multivariate Behavioral Research*, 27, 509-540.
- Cattell, R. B. ((1966). The scree test for the number of factors. *Multivariate Behavioral Research*, 1, 245-276.
- Glorfeld, L. W. (1995). An improvement on Horn's parallel analysis methodology for selecting the correct number of factors to retain. *Educational and Psychological Measurement*, 55, 377-393.
- Hakstian, A. R., Rogers, W. T., & Cattell, R. B. (1982). The behavior of number of factors rules with simulated data. *Multivariate Behavioral Research*, 17, 193-210.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30, 179-185.

Hubbard, R. & Allen, S. J. (1987). An empirical comparison of alternative methods for principal components extraction. *Journal of Business Research*, 15, 173-190.

Joreskog, K. G. (1967). Some contributions to maximum likelihood factor analysis. *Psychometrika*, 32, 433-482.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, 20, 141-151.

Zwick, W. R. & Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological Bulletin*, 99, 432-442.

PARELLEL MACRO

```
%macro parallel(data=_LAST_, var=_NUMERIC_,
niter=1000, statistic=Median);
/*-----*
| Macro Parallel
| Parameters
| data = dataset to be analyzed
|         (default: _LAST_)
| var = variables to be analyzed
|         (default: _NUMERIC_)
| niter= number of simulated datasets
|         to create (default: 1000)
| statistic = statistic used to
|             summarized eigenvalues
|             (default: Median. Other
|             possible values: P90,
|             P95, P99)
| Output
| Graph of actual vs. simulated
| eigenvalues
|-----*/
data _temp;
set &data;
keep &var;

run;
/* obtain number of observations and
variables in dataset */
ods output Attributes=Params;
ods listing close;
proc contents data=_temp ;
run;
ods listing;
data _NULL_;
set Params;
if Label2 eq 'Observations' then
call
symput('Nobs',Trim(Left(nValue2)));
else if Label2 eq 'Variables' then
call
symput('NVar',Trim(Left(nValue2)));
run;
/* obtain eigenvalues for actual data */
proc factor data=_temp nfact=&nvar noprint
outstat=E1(where=( _TYPE_ = 'EIGENVAL'));
var &var;
run;
data E1;
set E1;
array A1{&nvar} &var;
array A2{&nvar} X1-X&nvar;
do J = 1 to &nvar;
A2{J} = A1{J};
end;
```

```
keep X1-X&nvar;
run;
/* generate simulated datasets and obtain
eigenvalues */
%DO K = 1 %TO &niter;
data raw;
array X {&nvar} X1-X&nvar;
keep X1-X&nvar;
do N = 1 to &nobs;
do I = 1 to &nvar;
X{I} = rannor(-1);
end;
output;
end;
run;
proc factor data=raw nfact=&nvar noprint
outstat=E(where=( _TYPE_ =
'EIGENVAL'));
var X1-X&nvar;
proc append base=Eigen
data=E(keep=X1-X&nvar);
run;
%END;
/* summarize eigenvalues for simulated
datasets */
proc means data=Eigen noprint;
var X1-X&nvar;
output out=Simulated(keep=X1-X&nvar)
&statistic=;
proc datasets nolist;
delete Eigen;
run;
proc transpose data=E1 out=E1;
run;
proc transpose data=Simulated out=Simulated;
run;
/* plot actual vs. simulated eigenvalues */
data plotdata;
length Type $ 9;
Position+1;
if Position eq (&nvar + 1)
then Position = 1;
set E1(IN=A)
Simulated(IN=B);
if A then Type = 'Actual';
if B then Type = 'Simulated';
rename Coll = Eigenvalue;
run;
title height=1.5 "Parallel Analysis -
&statistic Simulated Eigenvalues";
title2 height=1 "&nvar Variables, &niter
Iterations, &nobs Observations";
proc print data = plotdata;
run;
symbol1
interpol = join
value=diamond
height=1
line=1
color=blue
;
symbol2
interpol = join
value=circle
height=1
line=3
color=red
;
proc gplot data = plotdata;
plot Eigenvalue * Position = Type;
run;
quit;
%mend parallel;
```

CONTACT INFORMATION

Questions and comments are welcome. The author can be contacted at:

Robert I. Kabacoff, Ph.D.
Management Research Group
14 York Street, Suite 301
Portland, Maine 04101
Work Phone: (207) 775-2173
Fax: (207) 775-2173
Email: rob.Kabacoff@mrg.com
Web: www.mrg.com

A copy of the source code can be downloaded from
<http://www.mrg.com/articles/parallel.sas>.

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.