Behavior Research Methods, Instruments, & Computers
2000, 32 (3), 396-402

SPSS and SAS programs for determining the number of components using parallel analysis and Velicer's MAP test

BRIAN P. O'CONNOR
Lakehead University, Thunder Bay, Ontario, Canada

Popular statistical software packages do not have the proper procedures for determining the number of components in factor and principal components analyses. Parallel analysis and Velicer's minimum average partial (MAP) test are validated procedures, recommended widely by statisticians. However, many researchers continue to use alternative, simpler, but flawed procedures, such as the eigenvalues-greater-than-one rule. Use of the proper procedures might be increased if these procedures could be conducted within familiar software environments. This paper describes brief and efficient programs for using SPSS and SAS to conduct parallel analyses and the MAP test.

Users of factor and principal components analyses are required to make decisions on a number of technical issues, including the number factors to retain, extraction and rotation techniques, and the procedure for computing factor scores. The choices and controversies involved in each step have probably led many to shy away from the procedure or to be suspicious of its results. It seems only logical to assume that the many possible routes through the decision tree result in differing results for the same data. However, the crucial decision is that of determining how many factors to retain. Assorted decisions on the other issues generally produce similar results when the optimal number of factors is specified (Zwick & Velicer, 1986). In addition to conflicting findings, other problems also emerge when nonoptimal numbers of factors are extracted. Under-extraction compresses variables into a small factor space, resulting in a loss of important information, a neglect of potentially important factors, a distorted fusing of two or more factors, and an increase in error in the loadings. Over-extraction diffuses variables across a large factor space, potentially resulting in factor splitting, in factors with few high loadings, and in researchers' attributing excessive substantive importance to trivial factors (see Wood, Tataryn, & Gorsuch, 1996; Zwick & Velicer, 1986, for reviews).

Users who are concerned with extracting the optimal number of factors are nevertheless confronted with a variety of decision rules that have been described in the literature (see Coovert & McNelis, 1988; Floyd & Widaman, 1995; Gorsuch, 1997; Merenda, 1997; Tinsley & Tinsley, 1987; Turner, 1998; and Zwick & Velicer, 1986, for reviews). The discussions are sometimes technical, and many users simply trust the default decision rule implemented in their statistical software packages (typically the eigenvalues-greater-than-one rule). Other users examine scree plots of eigenvalues, which are also available in popular statistical packages (such as SPSS and SAS), before making their decisions. Unfortunately, these two highly popular decision rules are problematic. The eigenvalues-greater-than-one rule typically overestimates, and sometimes underestimates, the number of components (Zwick & Velicer, 1986). This overly mechanical and somewhat arbitrary rule also does not always result in components that are reliable, as was originally believed (Cliff, 1988). The scree test has been a strongly promoted alternative rule of thumb (Cattell & Vogelmann, 1977). But it involves eyeball searches of plots for sharp demarcations between the eigenvalues for major and trivial factors. In practice, such demarcations do not always exist or there may be more than one demarcation point. Not surprisingly, the reliability of scree plot interpretations is low, even among experts (Crawford & Koopman, 1979; Streiner, 1998).

Fortunately, there is increasing consensus among statisticians that two less well-known procedures, parallel analysis and Velicer's minimum average partial (MAP) test, are superior to other procedures and typically yield optimal solutions to the number of components problem (Wood et al., 1996; Zwick & Velicer, 1982, 1986). These procedures are statistically based, rather than being mechanical rules of thumb. In parallel analysis, the focus is on the number of components that account for more variance than the components derived from random data. In the MAP test, the focus is on the relative amounts of systematic and unsystematic variance remaining in a correlation matrix after extractions of increasing numbers of components. The popular SPSS and SAS statistical software packages do not permit users to perform these recommended tests. However, the packages do permit users to write their own programs. The present paper describes how parallel analyses and the MAP test can be readily con-

This work was supported by a grant from the Social Sciences and Humanities Research Council of Canada. Address correspondence to B. P. O'Connor, Department of Psychology, Lakehead University, 955 Oliver Road, Thunder Bay, ON P7B 5E1, Canada (e-mail: brian.oconnor@lakeheadu.ca).

ducted within these familiar computing environments. The two procedures have been used in both principal components and common factor analyses, and the computational procedures described in the literature (and in this paper) are the same in both cases. This is because researchers must determine how many components or factors to extract before they begin their factor extractions. The computations for the present programs are performed within the matrix processing environments that are provided by SPSS (Matrix–End Matrix) and SAS (Proc IML). The Matrix–End Matrix procedure is a standard part of SPSS packages, whereas Proc IML is typically a separate module in SAS.

The MAP Test

Velicer's (1976) MAP test involves a complete principal components analysis followed by the examination of a series of matrices of partial correlations. Specifically, on the first step, the first principal component is partialed out of the correlations between the variables of interest, and the average squared coefficient in the off-diagonals of the resulting partial correlation matrix is computed. On the second step, the first *two* principal components are partialed out of the original correlation matrix and the average squared partial correlation is again computed. These computations are conducted for k (the number of variables) minus one steps. The average squared partial correlations from these steps are then lined up, and the number of components is determined by the step number in the analyses that resulted in the lowest average squared partial correlation. The average squared coefficient in the original correlation matrix is also computed, and if this coefficient happens to be lower than the lowest average squared partial correlation, then no components should be extracted from the correlation matrix. Statistically, components are retained as long as the variance in the correlation matrix represents systematic variance. Components are no longer retained when there is proportionately more unsystematic variance than systematic variance.

SPSS commands for the MAP test appear in Appendix A, and SAS commands appear in Appendix B. Users simply read in their data as they normally do, request a matrix of correlations or principal components analysis of the variables of interest, and specify that the correlation matrix be saved in a matrix file. The programs then read the saved matrix file, conduct the necessary analyses, and print the results. Sample output from using the SPSS program in Appendix A on data provided by Harman (1967, p. 80) appears in Appendix E. Harman's data were used in this example because they were also analyzed by Velicer (1976). The data were eight physical measurement variables (e.g., height, weight) obtained from 305 children. The squared correlation for "Step 0" in the output is the average squared off-diagonal correlation for the unpartialed correlation matrix. It is even possible to run the MAP program with a single command. For example, if the SPSS Matrix–End Matrix statements were saved in a file called C:\velicer.map, then the following commands would compute the correlation matrix and run the MAP program:

```
corr var1 to var10 / matrix out ('C:\datafile') / missing = listwise.
include file = 'C:\velicer.map'.
```

In SAS, this would be accomplished by saving the program as a module and running it by using a CALL MODULE statement. It is also possible (and very simple) to enter a correlation matrix into the matrix processing program directly, instead of having the program read a saved matrix of correlations. For example, in SPSS one would use the COMPUTE statement from the Matrix–End Matrix procedure instead of the MGET statement that appears in Appendix A.

Parallel Analysis

The second recommended procedure for deciding on the number of components involves extracting eigenvalues from random data sets that parallel the actual data set with regard to the number of cases and variables. For example, if the original data set consists of 305 observations for each of 8 variables, then a series of random data matrices of this size (305 × 8) would be generated, and eigenvalues would be computed for the correlation matrices for the original data and for each of the random data sets. The eigenvalues derived from the actual data are then compared to the eigenvalues derived from the random data. In Horn's (1965) original description of this procedure, the mean eigenvalues from the random data served as the comparison baseline, whereas a currently recommended practice is to use the eigenvalues that correspond to the desired percentile (typically the 95th) of the distribution of random data eigenvalues (Cota, Longman, Holden, Fekken, & Xinaris, 1993; Glorfeld, 1995; although see Cota, Longman, Holden, & Fekken, 1993, and Turner, 1998). Factors or components are retained as long as the ith eigenvalue from the actual data is greater than the ith eigenvalue from the random data.

This computationally intensive procedure for determining the number of components can be performed surprisingly quickly on modern personal computers. SPSS commands for parallel analysis appear in Appendix C, and SAS commands appear in Appendix D. The user simply specifies the number of cases, variables, data sets, and the desired percentile for the analysis at the start of the program. Unlike the MAP program, the commands in Appendices C and D do not read in the user's correlation matrix (although it would be a simple matter to have the programs do so). The user's correlation matrix is left out of the analyses to permit greater flexibility in use of the programs.

Sample output from using the SPSS program in Appendix C for random data that parallel Harman's (1967, p. 80) data—that is, for 305 cases and 8 variables—is provided in Appendix E. Parallel analysis involves comparing the actual eigenvalues with the random data eigenvalues. The eigenvalues derived from Harman's actual data were listed by the MAP program (see Appendix E) and can otherwise be obtained from regular factor or principal

Table 1
Processing Time Required by SPSS Parallel Analysis Program for Data Specifications

Cases	Variables	Data Sets	Time
250	25	100	00:10
250	25	1,000	01:26
250	50	100	00:25
250	50	1,000	04:02
500	25	100	00:15
500	25	1,000	02:21
500	50	100	00:37
500	50	1,000	06:07

components analysis procedures. In the sample output, it is clear that the first two eigenvalues from the actual data are larger than the corresponding first two 95th percentile (and mean) random data eigenvalues. However, the third eigenvalue from the actual data is less than the third 95th percentile (and mean) random data eigenvalue. This indicates that two components should be retained.

In all parallel analyses focusing on a chosen percentile of the distributions of random data eigenvalues, consideration must be given to the relationship between the chosen percentile value and the number of random data sets generated. The multiplication of the chosen percentile (e.g., 95) by the number of data sets, divided by 100, should result in an integer [e.g., (95 * 1000) / 100 = 950]. This is because the program searches the distribution of eigenvalues (for a given root) for the eigenvalue whose rank order corresponds to the specified percentile, based on the number of random data sets generated. (In the example above, the program searches for the 950th largest eigenvalue in the set of 1,000.) If the user's specifications translate into rank orders that are not integers, the program rounds the computed rank order to the closest integer.

The processing time required by the SPSS parallel analysis program, running on a 233-MHz personal computer, was recorded for a number of data specifications, with the results shown in Table 1 (in minutes and seconds).

All parallel analysis programs use random number generators, and different programs or even different runs of the same program may produce slight differences in the results (e.g., a .04 difference in the 95th percentile eigenvalues from one run to another). This is due to differences in the random number generators and/or differences in the seeds that are used by the random number generators. The variation in the results becomes increasingly small and essentially disappears as the number of random data sets increases. In cases where a random data eigenvalue is very similar in magnitude to the eigenvalue for actual data, it is recommended that the parallel analysis be run again using a large number of data sets for more precise and reliable results.

The SPSS and SAS random number generators have been deemed "safe" on the basis of tests conducted by Onghena (1993). The present SAS parallel analysis program samples from normal parent distributions. The SPSS program samples from uniform parent distributions because a normal random deviate facility is presently not available in the SPSS Matrix–End Matrix environment. However, the uniform deviates generated in the SPSS program are converted to standard normal deviates using the Box–Muller algorithm (see Brysbaert, 1991, and the COMPUTE X = statement in Appendix C).

Parallel analyses and the MAP test typically result in the same decision regarding the number of components to retain, as they did for Harman's data (1967, p. 80; see Appendix E). However, researchers have been encouraged to run both tests because identical results do not always emerge (Zwick & Velicer, 1986). When differences do emerge, the number of random data sets in the parallel analysis should be increased, and the average squared correlations from the MAP test should be scrutinized for close calls. The two procedures complement each other nicely, in that the MAP tends to err (when it does err) in the direction of underextraction, whereas parallel analysis tends to err (when it does err) in the direction of overextraction. Optimal decisions are thus likely to be made after the results of both analytic procedures have been considered.

FORTRAN programs for conducting parallel analysis and the MAP test have been reported in the literature (Longman, Cota, Holden, & Fekken, 1989; Reddon, 1985). A parallel analysis program for personal computers is also described in a recent article in this journal (Kaufman & Dunlap, 2000). The present programs give results identical to the results from previous programs. Their primary benefit is their likely convenience to many researchers. The programs are brief and simple to use, and they run efficiently in the familiar computing environments of popular statistical packages. It is to be hoped that these features will facilitate the use of valid procedures for determining the number of components. It was previously much more practical for factor analysts to continue using problematic procedures, but this is no longer the case.

Program Availability

The programs may be downloaded from the following internet address: http://flash.lakeheadu.ca/~boconno2/nfactors.html. The programs may also be obtained by e-mail from the author at: brian.oconnor@lakeheadu.ca, or by sending a stamped, self-addressed disk mailer to Department of Psychology, Lakehead University, 955 Oliver Road, Thunder Bay, ON P7B 5E1, Canada.

REFERENCES

Brysbaert, M. (1991). Algorithms for randomness in the behavioral sciences: A tutorial. *Behavior Research Methods, Instruments, & Computers*, **23**, 45-60.

Cattell, R. B., & Vogelmann, S. (1977). A comprehensive trial of the scree and KG criteria for determining the number of factors. *Multivariate Behavioral Research*, **12**, 289-325.

Cliff, N. (1988). The eigenvalues-greater-than-one rule and the reliability of components. *Psychological Bulletin*, **103**, 276-279.

Coovert, M. D., & McNelis, K. (1988). Determining the number of common factors in factor analysis: A review and program. *Educational & Psychological Measurement*, **48**, 687-692.

Cota, A. A., Longman, R. S., Holden, R. R., & Fekken, G. C. (1993). Comparing different methods for implementing parallel analysis: A practical index of accuracy. *Educational & Psychological Measurement*, **53**, 865-875.

Cota, A. A., Longman, R. S., Holden, R. R., Fekken, G. C., & Xinaris, S. (1993). Interpolating 95th percentile eigenvalues from random data: An empirical example. *Educational & Psychological Measurement*, **53**, 585-596.

Crawford, C. B., & Koopman, P. (1979). Inter-rater reliability of scree test and mean square ratio test of number of factors. *Perceptual & Motor Skills*, **49**, 223-226.

Floyd, F. J., & Widaman, K. F. (1995). Factor analysis in the development and refinement of clinical assessment instruments. *Psychological Assessment*, **7**, 286-299.

Glorfeld, L. W. (1995). An improvement on Horn’s parallel analysis methodology for selecting the correct number of factors to retain. *Educational & Psychological Measurement*, **55**, 377-393.

Gorsuch, R. L. (1997). Exploratory factor analysis: Its role in item analysis. *Journal of Personality Assessment*, **68**, 532-560.

Harman, H. H. (1967). *Modern factor analysis* (2nd ed.). Chicago: University of Chicago Press.

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

Kaufman, J. D., & Dunlap, W. P. (2000). Determining the number of factors to retain: A Windows-based FORTRAN-IMSL program for parallel analysis. *Behavior Research Methods, Instruments, & Computers*, **32**, 389-395.

Longman, R. S., Cota, A. A., Holden, R. R., & Fekken, G. C. (1989). PAM: A double-precision FORTRAN routine for the parallel analysis method in principal components analysis. *Behavior Research Methods, Instruments, & Computers*, **21**, 477-480.

Merenda, P. F. (1997). A guide to the proper use of factor analysis in the conduct and reporting of research: Pitfalls to avoid. *Measurement & Evaluation in Counseling & Development*, **30**, 156-164.

Onghena, P. (1993). A theoretical and empirical comparison of mainframe, microcomputer, and pocket calculator pseudorandom number generators. *Behavior Research Methods, Instruments, & Computers*, **25**, 384-395.

Reddon, J. R. (1985). MAPF and MAPS: Subroutines for the number of principal components. *Applied Psychological Measurement*, **9**, 97.

Streiner, D. L. (1998). Factors affecting reliability of interpretations of scree plots. *Psychological Reports*, **83**, 687-694.

Tinsley, H. E. A., & Tinsley, D. J. (1987). Uses of factor analysis in counseling psychology research. *Journal of Counseling Psychology*, **34**, 414-424.

Turner, N. E. (1998). The effect of common variance and structure on random data eigenvalues: Implications for the accuracy of parallel analysis. *Educational & Psychological Measurement*, **58**, 541-568.

Velicer, W. F. (1976). Determining the number of components from the matrix of partial correlations. *Psychometrika*, **41**, 321-327.

Wood, J. M., Tataryn, D. J., & Gorsuch, R. L. (1996). Effects of under- and overextraction on principal axis factor analysis with varimax rotation. *Psychological Methods*, **1**, 354-365.

Zwick, W. R., & Velicer, W. F. (1982). Factors influencing four rules for determining the number of components to retain. *Multivariate Behavioral Research*, **17**, 253-269.

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

APPENDIX A
SPSS Syntax for Velicer’s Minimum Average Partial (MAP) Test

```
correlation var1 to var25 / matrix out ('C:\data.cor') / missing = listwise.
factor var= var1 to var25 / matrix out (cor = 'C:\data.cor').


matrix.

mget /type= corr /file='C:\data.cor' .

call eigen (cr,eigvect,eigval).
compute loadings = eigvect * sqrt(mdiag(eigval)).
compute fm = make(nrow(cr),2,-9999).
compute fm(1,2) = (mssq(cr) - ncol(cr)) / (ncol(cr)*(ncol(cr)-1))).
loop #m = 1 to ncol(cr) - 1.
compute a = loadings(:,1:#m).
compute partcov = cr - (a * t(a)).
compute d = mdiag( 1 / (sqrt(diag(partcov))) ).
compute pr = d * partcov * d.
compute fm(#m+1,2) = (mssq(pr) - ncol(cr)) / (ncol(cr)*(ncol(cr)-1))).
end loop.

* identifying the smallest fm value & its location (= the # of factors).
compute minfm = fm(1,2).
compute nfactors = 0.
loop #s = 1 to nrow(fm).
compute fm(#s,1) = #s -1.
do if ( fm(#s,2) < minfm ).
compute minfm = fm(#s,2).
compute nfactors = #s - 1.
end if.
end loop.

print eigval /title="Eigenvalues".
print fm /title="Velicer’s Average Squared Correlations".
print minfm /title="The smallest average squared correlation is".
print nfactors /title="The number of components is".

end matrix.
```

APPENDIX B
SAS Syntax for Velicer's Minimum Average Partial (MAP) Test

```
proc corr data = rawdata outp = cormatrix;
run;


options nocenter nodate nonumber linesize=90; title;
proc iml;

use cormatrix;
read all var _num_ into whole;
cr = whole[4:nrow(whole),];

call eigen (eigval,eigvect,cr);
loadings = eigvect * sqrt(diag(eigval));
fm = j(nrow(cr),2,-9999);
fm[1,2] = (ssq(cr) - ncol(cr))/(ncol(cr)*(ncol(cr)-1));
do m = 1 to ncol(cr) - 1;
a = loadings[,1:m];
partcov = cr - (a * t(a));
d = diag( 1 / (sqrt(vecdiag(partcov))) );
pr = d * partcov * d;
fm[m+1,2] = (ssq(pr)-ncol(cr)) / (ncol(cr)*(ncol(cr)-1));
end;

/* identifying the smallest fm value & its location (= the of factors) */
minfm = fm[1,2];
nfactors = 0;
do s = 1 to nrow(fm);
fm[s,1] = s - 1;
if ( fm[s,2] < minfm ) then do;
minfm = fm[s,2];
nfactors = s - 1;
end;

end;
print, "Eigenvalues", eigval;
print, "Velicer's Average Squared Correlations", fm[format=15.9];
print, "The smallest average squared correlation is", minfm;
print, "The number of components is", nfactors;

quit;
```

APPENDIX C
SPSS Syntax for Parallel Analysis

```
set mxloops=9000 length=none printback=none width=80 seed = 1953125.
matrix.

* enter your specifications here.
compute Ncases = 500.
compute Nvars = 50.
compute Ndatsets = 1000.
compute percent = 95.

* computing random data correlation matrices & eigenvalues.
compute evals = make(nvars,ndatsets,-9999).
compute nm1 = 1 / (ncases-1).
loop #nds = 1 to ndatsets.
compute x = sqrt(2 * (ln(uniform(ncases,nvars)) * -1) ) &*
            cos(6.283185 * uniform(ncases,nvars) ).
compute vcv = nm1 * (sscp(x) - ((t(csum(x))*csum(x))/ncases)).
compute d = inv(mdiag(sqrt(diag(vcv)))).
compute evals(:,#nds) = eval(d * vcv * d).
end loop.

* identifying the eigenvalues corresponding to the desired percentile.
compute num = rnd((percent*ndatsets)/100).
compute results = { t(1:nvars), t(1:nvars), t(1:nvars) }.
loop #root = 1 to nvars.
```

APPENDIX C (Continued)

```
compute ranks = rnkorder(evals(#root,:)).
loop #col = 1 to ndatsets.
do if (ranks(1,#col) = num).
compute results(#root,3) = evals(#root,#col).
break.
end if.
end loop.
end loop.
compute results(:,2) = rsum(evals) / ndatsets.

compute specifs = {ncases; nvars; ndatsets; percent}.
print specifs /title="Specifications for this Run:"
 /rlabels="Ncases" "Nvars" "Ndatsets" "Percent".

print results /title="Random Data Eigenvalues"
 /clabels="Root" "Means" "Prcntyle".

end matrix.
```

APPENDIX D
SAS Syntax for Parallel Analysis

```
options nocenter nodate nonumber linesize=90; title;

proc iml;
seed = 1953125;

/* enter your specifications here */
Ncases = 305;
Nvars = 8;
Ndatsets = 1000;
percent = 95;

/* computing random data correlation matrices & eigenvalues */
evals = j(nvars,ndatsets,-9999);
nm1 = 1 / (ncases-1);
do nds = 1 to ndatsets;
x = normal( j(ncases,nvars)) ;
vcv = nm1 * (t(x)*x - ((t(x[+,])*x[+,])/ncases));
d = inv(diag(sqrt(vecdiag(vcv))));
evals[,nds] = eigval(d * vcv * d);
end;

/* identifying the eigenvalues corresponding to the desired percentile */
num = round((percent*ndatsets)/100);
results = j(nvars,3,-9999);
s = 1:nvars;
results[,1] = t(s);
do root = 1 to nvars;
ranks = rank(evals[root,] );
do col = 1 to ndatsets;
if (ranks[1,col] = num) then do;
results[root,3] = evals[root,col];
col = ndatsets;
end;
end;
end;
results[,2] = evals[,+] / ndatsets;

specifs = (ncases // nvars // ndatsets // percent);
rlabels = {"Ncases" "Nvars" "Ndatsets" "Percent"};
print, "Specifications for this Run:", specifs[rowname=rlabels];

clabels={"Root" "Means" "Prcntyle"};
print, "Random Data Eigenvalues", results[colname=clabels format=15.9];
quit;
```

APPENDIX E
Sample Output

SPSS Output from Velicer's MAP Test

```
Eigenvalues
    4.672880
    1.770983
     .481035
     .421441
     .233221
     .186674
     .137304
     .096463

Velicer's Average Squared Correlations
     .000000     .312475
    1.000000     .245121
    2.000000     .066445
    3.000000     .127594
    4.000000     .204203
    5.000000     .271829
    6.000000     .434591
    7.000000    1.000000

The smallest average squared correlation is
     .066445

The number of components is
 2
```

SPSS Output from a Parallel Analysis

```
Specifications for this Run:
Ncases    305
Nvars       8
Ndatsets 1000
Percent    95

Random Data Eigenvalues
       Root       Means     Prcntyle
    1.000000    1.245463    1.325851
    2.000000    1.154223    1.212952
    3.000000    1.083692    1.128706
    4.000000    1.022316    1.063478
    5.000000     .965652    1.004431
    6.000000     .908654     .950213
    7.000000     .846994     .895851
    8.000000     .773006     .831101
```

(Manuscript received June 23, 1999;
revision accepted for publication March 11, 2000.)