SAS Statistics Research and Applications
Paper #2023-01

Determining the Number of Latent Factors Using PROC FACTOR: A Case Study of the Big Five Personality Traits Model

Yiu-Fai Yung and Clay Thompson, SAS Institute Inc., Cary, NC

This paper reviews several methods for determining the number of factors in exploratory factor analysis. It uses the FACTOR procedure in SAS/STAT® software, version 14.3 (SAS Institute Inc. 2017) or later, including all SAS Viya versions, to illustrate the application of these methods. Four methods, which are based on the minimum eigenvalue, proportion of variance explained, minimum average partial correlation, and parallel analysis, respectively, are the focus of the discussion. After explaining the logic of these four methods and their corresponding features in PROC FACTOR, the paper demonstrates an application to a large data set, which records the responses to a questionnaire that contains 50 items for measuring the Big Five personality traits. Together, the four methods suggest a plausible range for the number of factors to explain the data and provide an objective ground for validating the questionnaire. We conclude that these methods, when used appropriately and judiciously in practical research, can provide valuable insights about the latent dimensionality and thus should be routinely used together with substantive theory to determine the number of factors.

INTRODUCTION

Common factor analysis was developed in the field of psychology to measure latent (unobserved or unobservable) mental abilities or personality traits (Spearman 1904). Individuals respond to the many items (or variables) on a questionnaire or a test that is designed to measure a set of latent attributes called common factors. The following equation describes a common factor model for the random variables involved:

$$y_j = \lambda_{j1} f_1 + \lambda_{j2} f_2 + \cdots + \lambda_{jm} f_m + e_j \qquad (j = 1, 2, \ldots, p)$$

where $y_1, y_2, \ldots, y_p$ represent p observed variable; $f_1, f_2, \ldots, f_m$ represent m latent factors; e_j represents the error of the jth variable; and $\lambda_{j1}, \lambda_{j2}, \ldots, \lambda_{jm}$ are parameters called factor loadings, which relate the observed variables to the latent factors. Without loss of generality, the observed variables y_j in the model equation are assumed to have been standardized with a mean of 0 and a variance of 1, and the variances of the common factors are all set to 1. In addition, the error terms e_j are independent of each other and also independent of the latent factors.

The term *common factors* refers to the latent factors that are defined in the preceding common factor model equation. This terminology is to be distinguished from the general usage of *factors*, which in other contexts could refer to principal components, matrix factors, and so on. Because the scope of this paper is common factor analysis only, hereafter we can use terms such as *factors* and *factor analysis* without worrying about potential confusion.

Given the factor model equation with uncorrelated latent factors, the variance of the jth observed variable is prescribed as

$$\text{Var}(y_j) = \lambda_{j1}^2 + \lambda_{j2}^2 + \cdots + \lambda_{jm}^2 + \text{Var}(e_j)$$

In this equation, the variance of y_j is composed of two portions. The first portion, $\sum_k^m \lambda_{jk}^2$, is called the common variance, or the *communality*, of the variable. This is the portion of variance that is explained by the factors and is usually assumed to be less than 1. The remaining portion is the error variance, Var(e_j).

Under the same factor model, the correlation between observed variables y_i and y_j is prescribed as

$$\text{Corr}(y_i, y_j) = \sum_k^m \lambda_{ik}\lambda_{jk}$$

This equation highlights one of the main goals of factor analysis: to use a small number of factors (m) to explain the correlations among a much larger number of variables (p).

For an elementary discussion of the common factor model, see Gorsuch (1974) and Kim and Mueller (1978a, b). Harman (1976) provides a more in-depth treatment of the topic. Morrison (1976) and Mardia, Kent, and Bibby (1979) introduce the factor analysis model from a more statistics-based perspective.

Exploratory Factor Analysis and Determining the Number of Factors

Factor analysis can be done in a confirmatory way or an exploratory way. In a confirmatory factor analysis (CFA), factors are hypothetical constructs that have a *known* pattern of relationships with observed variables. Such a known pattern of relationships is characterized by a highly *structured* $p \times m$ matrix of factor loadings. First, most loadings in the matrix are fixed to zero. Second, each variable is associated with only a *small* number of factors (usually 1). Third, each factor is associated with a distinct (or nearly distinct) cluster of observed variables, as indicated by the nonzero loading parameters. A CFA is usually conducted within the framework of structural equation modeling, which is not covered here. (See, for example, Loehlin (2004), Bollen (1989), Everitt (1984), or Long (1983) for an introduction to the topic.)

This paper is entirely about exploratory factor analysis (EFA), in which each observed variable might be associated with any factor in the model. The factor loading matrix is "unstructured," in the sense that all elements in the matrix are estimated so that the strengths of factor-variable relationships would be determined empirically from analyzing the sample. In fact, even the number of factors m is not precisely known before the analysis. Consequently, the factors in an EFA must be interpreted on the basis of the number of factors and the estimated factor loading matrix.

To conduct an EFA, the correlation matrix of the observed variables is computed and then *factor-analyzed*. This process entails the following three steps:

1. Determine the number of factors.
2. Extract the factors to obtain an initial factor solution.
3. Transform the initial factor solution to get a final-rotated factor solution for interpretation.

Because of the *exploratory* nature of an EFA, these three steps might not be entirely independent of each other in actual practice. That is, the determination of the number of factors is often interwoven with the interpretability of the factor solutions that are obtained in the third step. Despite this complication, the paper focuses on the first step and initially treats the determination of the number of factors as an isolated topic for the exposition of basic ideas.

Four numerical methods for determining the number of factors are described and illustrated by applying the FACTOR procedure along with specific options:

- Minimum eigenvalue (MINEIGEN= option)
- Proportion (PROPORTION= option)
- Minimum average partial correlation (MAP option)
- Parallel analysis (PARALLEL option)

To demonstrate the basic uses of these four methods, we analyze a small data set. Emphasis is on how to use these options in PROC FACTOR and how to appropriately interpret the results.

Then, to demonstrate how to synthesize the information obtained from these methods and the interpretability of the factor solutions, we analyze a large real data set. Emphasis is on how to use these methods in a real research context so that an informed decision about the number of factors can be made.

Big Five Personality Traits: A Five-Factor Model

To motivate the practical utility of EFA, this section describes the well-known and widely accepted Big Five personality theory in psychology. The Big Five theory proposes the following five main personality traits and descriptions of individuals who have those traits:

- Extroversion—Tendency to seek interaction with people and environment: assertive, sociable, outgoing, and so on.
- Agreeableness—Interaction with others: cooperative, trusting, sympathetic, and so on.
- Conscientiousness—Ability to control impulses and concentrate on directed goals: competent, organized, self-disciplined, and so on.
- Emotional stability—Maintaining a good mood: relaxed, not bothered by things, seldom upset, and so on.
- Intellect, imagination, or openness—Mental and intellectual strengths: imaginative, creative, full of ideas, and so on.

The Big Five factor theory is supported widely by empirical studies. See Goldberg (1990) for a review of the supporting evidence. The main analytical tool in all these supporting studies is exploratory factor analysis. In the factor model, personality traits are treated as the latent factors of the items (or variables) on a personality questionnaire. For example, an item on a questionnaire like this can be a behavioral description such as the following:

“I do not talk a lot.”

Respondents rate themselves on a five- or seven-point scale to indicate whether the item accurately describes their behavior, with “Very Inaccurate” at the lowest end and “Very Accurate” at the highest end of the scale. All items on the questionnaire are rated in a similar fashion. The correlation matrix of the items is then computed and factor-analyzed.

The Big Five factor theory is considered to demonstrate generality because repeated studies using factor analysis all indicate that five factors are sufficient to account for the sample correlations among the items. In addition, these five factors were all identified to be the Big Five personality traits in these repeated studies.[1]

But how did the researchers determine the number of factors to retain in their studies? They used a variety of methods. To provide some insights, the next section demonstrates the use of a visual tool called a *scree plot* (Cattell 1966) for determining the number of factors to retain in the factor model for a large data set that contains 50 items for measuring personality traits.

Using Scree Plots to Determine the Number of Factors

Figure 1 shows the scree plot (left graph) and the plot of proportions and cumulative proportions (right graph) that were produced by the following statements:[2]

```
ods graphics on;
proc factor data=big5cor plot=scree priors=smc;
run;
```

The DATA= option inputs the data set big5cor, which contains the correlations of 50 personality items (variables) that are computed from a raw data set that has 874,434 complete observations. These 50 items were designed to measure the Big Five personality traits that were described in the previous section. See the EXAMPLE section for a more detailed description of the data.

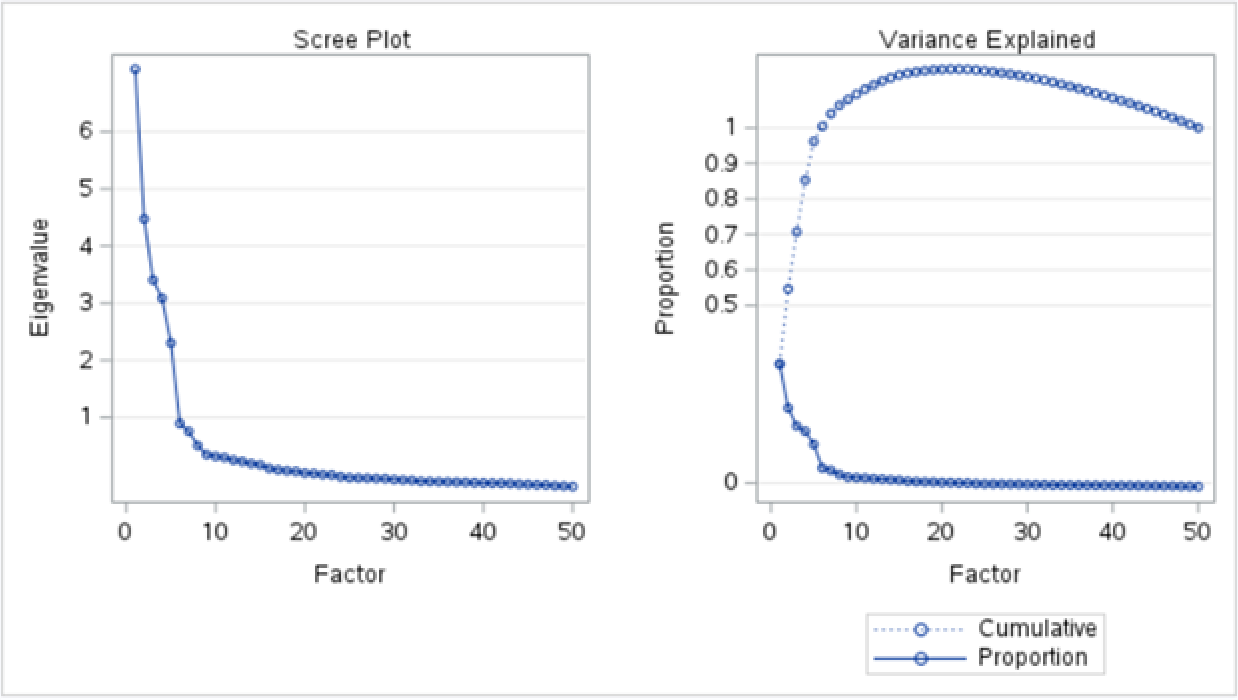
To explore the latent dimensionality of the data (and hence the number of factors), you use the PLOT=SCREE option to produce a scree plot of the eigenvalues of the correlation matrix or the reduced correlation matrix. Because the current analysis uses the PRIORS=SMC option, the eigenvalues of the reduced correlation matrix are plotted in Figure 1. The reason for using the reduced correlation matrix is explained in the next section.

In a scree plot, the eigenvalues are plotted in descending order. Each eigenvalue represents the part of the total variances of the observed variables that a factor explains. The larger the eigenvalue, the more salient the corresponding factor. Therefore, essentially, the number of salient factors corresponds to the number of “nontrivial” eigenvalues. A scree plot helps you visually identify these nontrivial eigenvalues.

[1] After you use the rotated factor loading or pattern matrix to distinguish factors, identifying factors in a factor analysis result might involve subjective interpretations and labeling.

[2] For illustration purposes, a modified ODS graphical template was used to produce the plots in Figure 1. The code for the modified template is available from the authors upon request.

Figure 1: Scree Plot with Reduced Correlation Matrix



To use a scree plot, you retain those factors before the “elbow” point, where the curve starts to level off. For example, the left graph in Figure 1 shows that the number of factors could be 8 to 10. However, because a perfectly leveled curve for eigenvalues almost never occurs in real-data applications, where the elbow point is located in a scree plot could be somewhat ambiguous because of the scaling of the plot. If you look at the bottom curve of the right graph in Figure 1, the rescaled scree plot might suggest only 5 to 8 factors.

Closely related to the scree plot is the plot of cumulative proportion of variance explained by factors. This is the top curve of the right graph in Figure 1. In applications, researchers set a specific level for the proportion of variance that must be explained by the included factors. Conventional levels can be set between 80% and 100%. For this analysis, these criterion levels translate to the inclusion of 4 and 6 factors, respectively.

Correlation Matrix or Reduced Correlation Matrix?

You might wonder why the PRIORS=SMC option is specified in the preceding PROC FACTOR statement. The reason is that the factor model does not assume that the factors can explain 100% of the variance of the observed variables. Instead, the variance of a standardized variable that is accounted for by the (common) factors, or the so-called communality, is in general less than 1 in the model. If the original correlation matrix (with ones on the diagonal) were used directly for factoring, the initial communalities would be set to 1, violating the basic idea of using a (common) factor model. In addition, to avoid overfactoring, factor analysts in the psychometric field tend to recommend that values less than 1 be used as initial communality estimates in exploratory factor analysis.

When you use the PRIORS=SMC option, PROC FACTOR replaces the diagonal elements of the correlation matrix with the squared multiple correlations (SMCs) to yield the so-called reduced correlation matrix. The SMC of a variable is its proportion of variance that can be predicted from all other observed variables in the analysis. Theoretically, the SMC of an observed variable provides the lower-bound estimate of its communality. It is usually less than 1 unless the variable is linearly dependent on other variables.

Consequently, when you use the PRIORS=SMC option, the eigenvalues and the proportion of (common) explained variance in the output results (such as those in Figure 1) correspond to the reduced correlation matrix, but not to the original correlation matrix. Specifically, the proportion of variance explained by the factors refers to the common variance (which is the sum of eigenvalues of the reduced correlation matrix) and not to the total variance of the original variables. As a result, some cumulative proportions can be greater than 1 because the eigenvalues of the reduced correlation matrix can be negative.

Note that the default prior communality estimates option for PROC FACTOR is PRIORS=ONE, which sets the initial communalities to ones, as in the original correlation matrix. This option is most relevant when you are trying to determine the number of principal components. In contrast, the PRIORS=SMC option is a more reasonable (and the most popular) way to set the initial communality estimates for factor analysis. Other choices of initial communality estimates are also available in PROC FACTOR.

Limitations of Scree Plots

The idea of the scree plot is simple and intuitive: you pick those strong factors (which have large eigenvalues) until the curve levels off. However, scree plots can sometimes be ambiguous, depending on the scale that you use to plot the eigenvalues. Therefore, other methods that are based on numerical results could provide more objective means to suggest the number of factors. The next section describes numerical methods of this type that are available in PROC FACTOR.

FOUR METHODS FOR DETERMINING THE NUMBER OF FACTORS

This section describes the logic of four numerical methods for determining the number of factors. To simplify the presentation of the data and the output results, we factor-analyze a smaller data set. To this end, the correlation matrix of 13 job rating variables for 103 police officers is specified by the following DATA step:

```
data JobRating(type=corr);
   input CommunicationSkills ProblemSolving  LearningAbility
         JudgmentUnderPressure ObservationalSkills WillingnessConfrontProblems
         InterestInPeople InterpersonalSensitivity DesireForSelfImprovement
         Appearance Dependability PhysicalAbility Integrity;
   datalines;
1.000 0.628 0.555 0.554 0.538 0.527 0.439 0.503 0.564 0.491 0.547 0.219 0.508
0.628 1.000 0.569 0.620 0.428 0.501 0.397 0.440 0.409 0.387 0.455 0.320 0.385
```

```
0.555 0.569 1.000 0.489 0.623 0.525 0.274 0.185 0.574 0.399 0.511 0.227 0.314
0.554 0.620 0.489 1.000 0.373 0.400 0.623 0.613 0.483 0.227 0.547 0.348 0.588
0.538 0.428 0.623 0.373 1.000 0.730 0.262 0.165 0.598 0.418 0.563 0.427 0.391
0.527 0.501 0.525 0.400 0.730 1.000 0.223 0.129 0.531 0.482 0.487 0.487 0.326
0.439 0.397 0.274 0.623 0.262 0.223 1.000 0.805 0.486 0.268 0.607 0.377 0.745
0.503 0.440 0.185 0.613 0.165 0.129 0.805 1.000 0.371 0.260 0.541 0.218 0.692
0.564 0.409 0.574 0.483 0.598 0.531 0.486 0.371 1.000 0.447 0.598 0.375 0.566
0.491 0.387 0.399 0.227 0.418 0.482 0.268 0.260 0.447 1.000 0.509 0.382 0.414
0.547 0.455 0.511 0.547 0.563 0.487 0.607 0.541 0.598 0.509 1.000 0.446 0.654
0.219 0.320 0.227 0.348 0.427 0.487 0.377 0.218 0.375 0.382 0.446 1.000 0.381
0.508 0.385 0.314 0.588 0.391 0.326 0.745 0.692 0.566 0.414 0.654 0.381 1.000
;
```

Minimum Eigenvalue Method (MINEIGEN= Option)

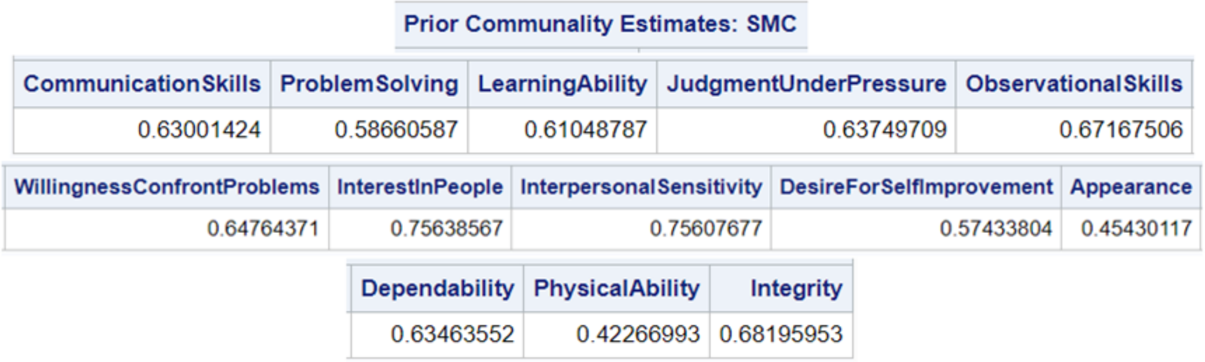
When there is no common factor that explains correlations among the observed variables, the correlation matrix would be an identity matrix in which all off-diagonal elements are zeros. The eigenvalues of such a matrix would be all ones. Therefore, it is reasonable to require an eigenvalue to be greater than 1 as an indication of a factor. This logic leads to the use of the eigen-1 criterion for determining the number of factors—that is, the number of factors is the number of eigenvalues that are greater than 1 (Kaiser 1960).

The following PROC FACTOR statement uses the MINEIGEN=1 option to specify the eigen-1 criterion to determine the number of factors:

```
proc factor data=JobRating(type=corr) nobs=103 priors=smc mineigen=1;
run;
```

The PRIORS=SMC option specifies the use of the squared multiple correlations of the variables as the prior (or initial) communality estimates, which are shown in Figure 2.

Figure 2: Prior Communality Estimates Using Squared Multiple Correlations



Prior Communality Estimates: SMC

CommunicationSkills	ProblemSolving	LearningAbility	JudgmentUnderPressure	ObservationalSkills
0.63001424	0.58660587	0.61048787	0.63749709	0.67167506

WillingnessConfrontProblems	InterestInPeople	InterpersonalSensitivity	DesireForSelfImprovement	Appearance
0.64764371	0.75638567	0.75607677	0.57433804	0.45430117

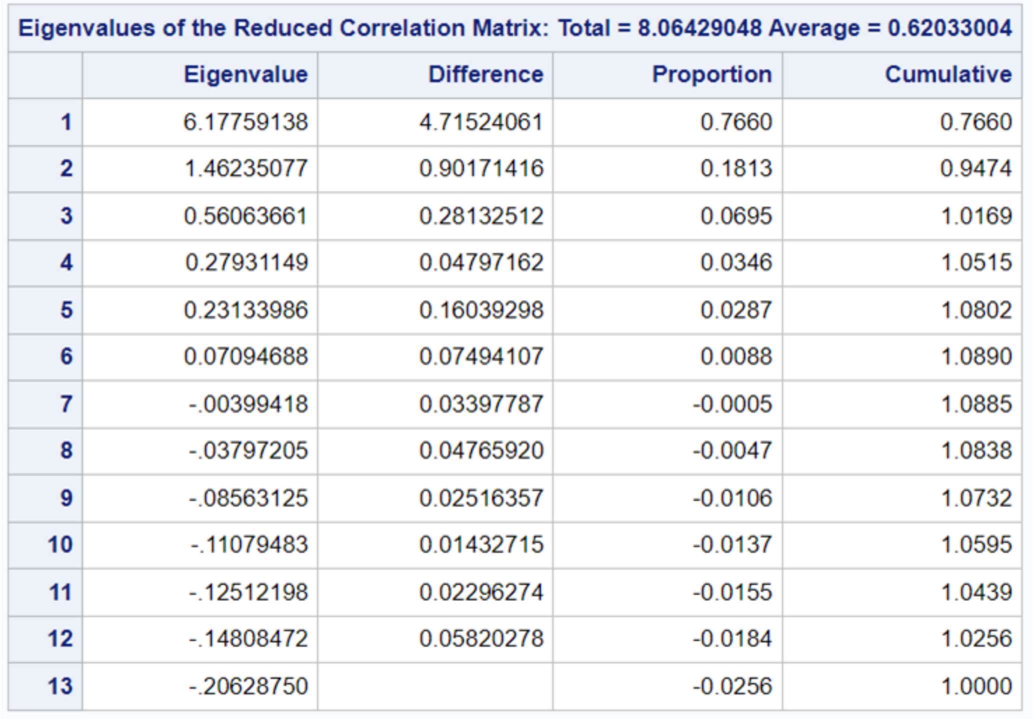
Dependability	PhysicalAbility	Integrity
0.63463552	0.42266993	0.68195953

These prior communality estimates replace the diagonal elements of the original correlation matrix to form the reduced correlation matrix, of which the eigenvalues and their cumulative proportions are computed and shown in Figure 3.

Because only the first two eigenvalues in Figure 3 are greater than 1, the eigen-1 criterion suggests two factors in an output message:

```
NOTE: 2 factors will be retained by the MINEIGEN criterion.
```

Figure 3: Eigenvalues of the Reduced Correlation Matrix



Eigenvalues of the Reduced Correlation Matrix: Total = 8.06429048 Average = 0.62033004

	Eigenvalue	Difference	Proportion	Cumulative
1	6.17759138	4.71524061	0.7660	0.7660
2	1.46235077	0.90171416	0.1813	0.9474
3	0.56063661	0.28132512	0.0695	1.0169
4	0.27931149	0.04797162	0.0346	1.0515
5	0.23133986	0.16039298	0.0287	1.0802
6	0.07094688	0.07494107	0.0088	1.0890
7	-.00399418	0.03397787	-0.0005	1.0885
8	-.03797205	0.04765920	-0.0047	1.0838
9	-.08563125	0.02516357	-0.0106	1.0732
10	-.11079483	0.01432715	-0.0137	1.0595
11	-.12512198	0.02296274	-0.0155	1.0439
12	-.14808472	0.05820278	-0.0184	1.0256
13	-.20628750		-0.0256	1.0000

The eigen-1 criterion generalizes to the minimum eigenvalue method, which you can use to specify any required minimum level of eigenvalue. In PROC FACTOR, you can specify this minimum value as any positive integer in the MINEIGEN= option.

Proportion Method (PROPORTION= Option)

The proportion method retains the minimum number of factors that can exceed a required cumulative proportion of eigenvalues. The required proportion is usually set at a high level, such as a value between 0.8 and 1.

The following PROC FACTOR statement uses the PROPORTION=0.9 option to specify the proportion method to determine the number of factors:

```
proc factor data=JobRating(type=corr) nobs=103 priors=smc proportion=0.9;
run;
```

The output table that summarizes the eigenvalues and the proportion of common variance explained for the current analysis is the same as the table shown previously in Figure 3. This table shows the cumulative proportions in the last column. With 2 factors, the cumulative proportion is 0.9474, which just exceeds the 0.9 criterion value. Therefore, the proportion method suggests 2 factors in the following output message:

```
NOTE: 2 factors will be retained by the PROPORTION criterion.
```

As explained previously, because the PRIORS=SMC option is used, some of the eigenvalues of the reduced correlation matrix can be negative (see the “Proportion” column in Figure 3).

Consequently, the cumulative proportion can reach beyond 1 at some point before it drops to 1 at the last factor.

In PROC FACTOR, you can specify any value between 0 and 1 in the PROPORTION= option. For example, if you specified PROPORTION=1 for the current analysis, 3 factors (with a cumulative proportion of 1.0169) would have been suggested.

Minimum Average Partial Correlations (MAP Option)

Velicer (1976) proposed the criterion of minimum average partial (MAP) correlations to determine the number of factors. This criterion selects the number of factors that corresponds to the number of principal components (PCs) that yields the smallest average residual (or partial) squared correlations among the observed variables after they are regressed on (or partialed out from) their PCs. An extension was proposed by Velicer, Eaton, and Fava (2000) that uses the fourth-powered partial correlations instead of the squared counterparts to compute the average. Through simulation studies, the MAP method was proven to be superior to the eigen-1 and scree plot methods (Zwick and Velicer 1986), but it might underestimate the number of factors in some situations.[3]

The following PROC FACTOR statement uses the NFACTORS=MAP option to request that the MAP method be used to determine the number of factors:

```
ods graphics on;
proc factor data=JobRating(type=corr) nobs=103 nfactors=map plots=map;
run;
```

In addition, the PLOTS=MAP option produces a plot of average partial correlations at each number of PCs that are partialed out. Figure 4 shows the numerical values of the average partial correlations, squared and fourth-powered, and the corresponding plot.

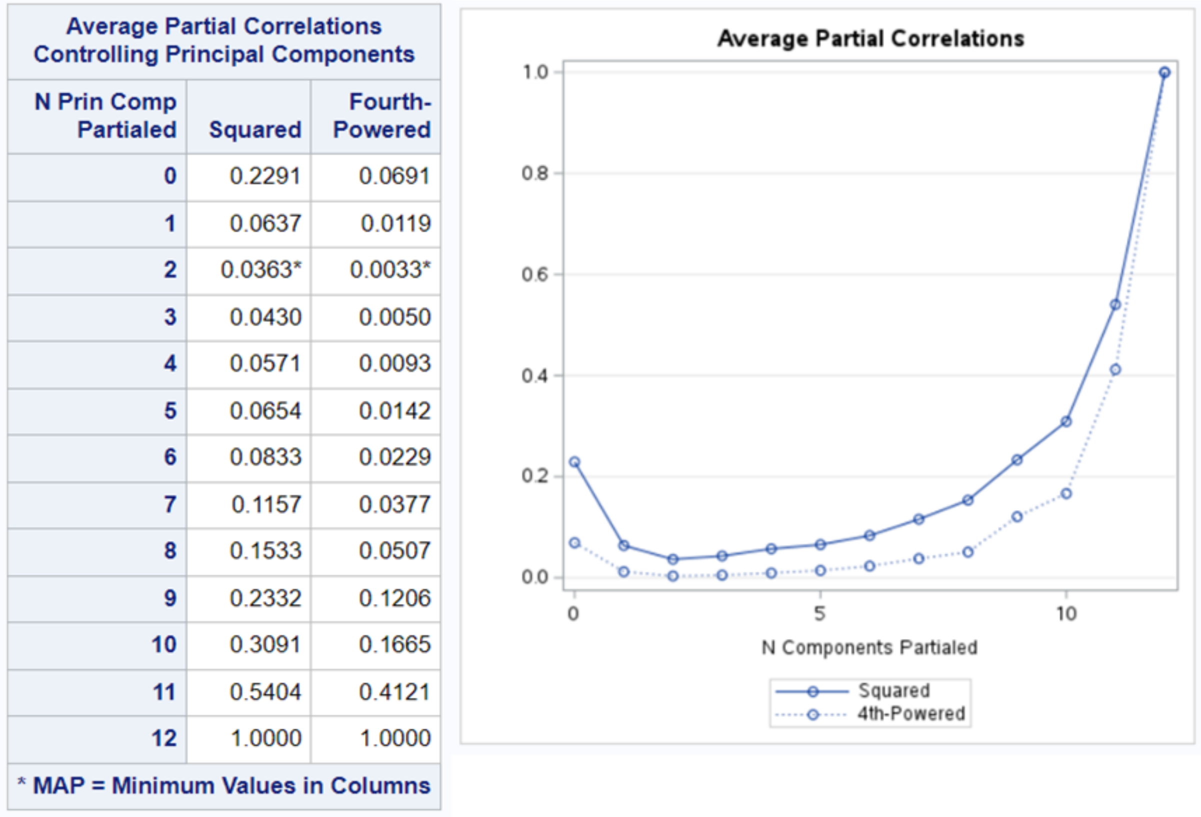
The MAP method picks the number of factors that corresponds to the minimum average partial correlation. The table in Figure 4 shows that the minimum is attained at 2 factors, using the average of either squared or fourth-powered correlations. You reach the same conclusion by finding the minimum points of the two curves in the plot of the same figure. The following output message confirms this result of the MAP method:

```
NOTE: 2 factors will be retained by the method of minimum average
      squared partial correlation (MAP2).
```

This message has assumed the default use of the *squared* partial correlations for computing the averages. If you want to use the fourth-powered partial correlations instead, you should specify the NFACTORS=MAP4 option.

[3] Some researchers reject the MAP method as a legitimate method for determining the number of common factors on the “logical” ground that it is based on the analysis of principal components. The authors of this paper do not side with this argument because it appears that perhaps except for Rao’s significance test of number of factors, none of the currently popular psychometric methods for determining the number of factors have been rigorously derived in a statistical-inferential setup. Most of the time, a recommended method in the field is the one that has received supporting results from simulation studies. In this regard, the MAP method should be recommended, because it certainly has received supporting simulation results.

Figure 4: MAP Analysis of the Job Rating Data



Average Partial Correlations Controlling Principal Components

N Prin Comp Partialed	Squared	Fourth-Powered
0	0.2291	0.0691
1	0.0637	0.0119
2	0.0363*	0.0033*
3	0.0430	0.0050
4	0.0571	0.0093
5	0.0654	0.0142
6	0.0833	0.0229
7	0.1157	0.0377
8	0.1533	0.0507
9	0.2332	0.1206
10	0.3091	0.1665
11	0.5404	0.4121
12	1.0000	1.0000

* MAP = Minimum Values in Columns

Note that the MAP method is based on computing residual correlations, which are produced by partialing out the PCs from the *full* (original) correlation matrix. Therefore, unlike the minimum eigenvalue or proportion method, the MAP method is *not* affected by the choice of prior communality estimates that you specify in the PRIORS= option.

Parallel Analysis (PARALLEL Option)

Recognizing that the eigen-1 criterion is essentially based on comparing the observed *sample* eigenvalues with those obtained from the identity matrix at the *population*, Horn (1965) proposed a simulation method, now widely known as “parallel analysis,” to take the sampling fluctuations into account in evaluating the sample eigenvalues. Parallel analysis is considered to be the most accurate method of determining the number of factors (see, for example, Zwick and Velicer (1986) and Dinno (2009)). In addition, although parallel analysis is based on simulating random normal data, Glorfeld (1995) and Dinno (2009) showed that it is not sensitive to the distribution form of the data and therefore is widely applicable.

In a parallel analysis, random data sets (each with the same numbers of observations and variables as in the sample) are generated from a hypothetical population where the variables are uncorrelated. In each random data set, the eigenvalues are computed and put in descending order. Averages of the ordered eigenvalues from the random data sets are then computed. The parallel analysis compares the ordered eigenvalues of the sample correlation matrix to the corresponding simulated average eigenvalues. The number of factors is determined to be m, which is the number of first m sample eigenvalues that are greater than the simulated average eigenvalues.

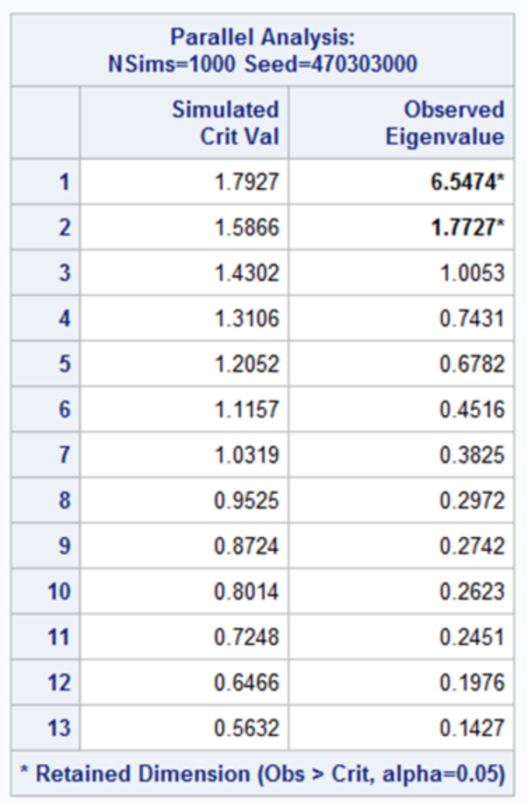
Glorfeld (1995) proposed an extension that treats parallel analysis like the hypothesis testing of significance of the sample eigenvalues. The simulated eigenvalues at each rank order form the null distribution under the zero-factor hypothesis. Adopting the hypothesis testing framework, you would then conclude that a factor is “real” when its corresponding observed eigenvalue is greater than the simulated *critical* value at a certain prespecified α-level. An example will make this logic clear.

The following PROC FACTOR statement uses the NFACTORS=PARALLEL option to request that parallel analysis be used to determine the number of factors:

```
ods graphics on;
proc factor data=JobRating(type=corr) nobs=103 nfactors=parallel
            plots=parallel;
run;
```

In addition, the PLOTS=PARALLEL option produces a plot of observed and simulated eigenvalues. Figure 5 shows the numerical and graphical results of the parallel analysis.[4] The observed and simulated critical eigenvalues are ordered and compared. A factor is retained if the corresponding observed eigenvalue value is greater than the simulated critical value.

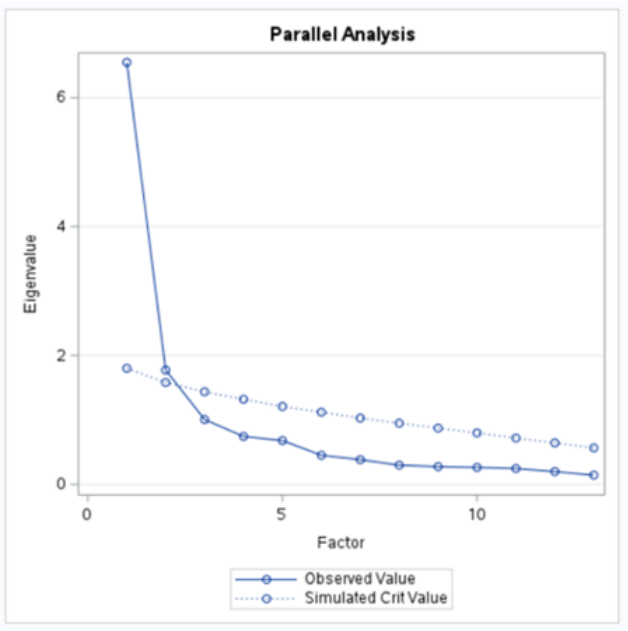
Figure 5: Parallel Analysis of the Job Rating Data



Parallel Analysis:
NSims=1000 Seed=470303000

	Simulated Crit Val	Observed Eigenvalue
1	1.7927	**6.5474***
2	1.5866	**1.7727***
3	1.4302	1.0053
4	1.3106	0.7431
5	1.2052	0.6782
6	1.1157	0.4516
7	1.0319	0.3825
8	0.9525	0.2972
9	0.8724	0.2742
10	0.8014	0.2623
11	0.7248	0.2451
12	0.6466	0.1976
13	0.5632	0.1427

* Retained Dimension (Obs > Crit, alpha=0.05)



[4] The paper uses a recently updated ODS output template for the parallel analysis table. The updated template is available from the authors upon request.

In the table shown in Figure 5, the first two factors are marked to be retained. The corresponding plot in Figure 5 shows the same result: the number of factors to retain is at the point just before the two curves cross. Here, the parallel analysis suggests 2 factors in the following output message:

```
NOTE: 2 factors will be retained by the method of parallel analysis
      (with alpha=0.05).
```

Three important suboptions for parallel analysis PROC FACTOR supports are described as follows:

- ALPHA= suboption

  You can use the ALPHA= suboption (with a value between 0 and 1) to control the criterion level of the parallel analysis. The smaller the specified value, the more stringent the criterion for accepting a factor. For example, the NFACTORS=PARALLEL(ALPHA=0.01) option generates higher critical eigenvalues than the default value 0.05 does.

  Note that PROC FACTOR implements Glorfeld's more general procedure rather than Horn's original parallel analysis, which computes critical eigenvalues by averaging the ordered eigenvalues from random data sets. Assuming that the distributions of the simulated eigenvalues are all (approximately) symmetrical, specifying ALPHA=0.5 would lead to Horn's original parallel analysis.

- NSIMS= suboption

  You can use the NSIMS= suboption to specify the number of simulation samples (or replications) for parallel analysis. To ensure that the simulated critical values are trustworthy, this number should not be small. PROC FACTOR uses a default of 1,000 simulations, which should be a reasonable number for most applications. But you can change the default. For example, the NFACTORS=PARALLEL(NSIMS=2000) option generates 2,000 random samples to compute the critical values.

- SEED= suboption

  You can use the SEED= suboption to specify a fixed seed number to maintain the replicability of simulation results of a parallel analysis. For example, using the NFACTORS=PARALLEL(SEED=12479) option generates the same parallel analysis results each time you input the same correlation matrix. In contrast, if you omit the SEED= suboption, PROC FACTOR sets the seed by using the computer's clock time at code execution. Hence, the numerical results of parallel analysis might fluctuate in different runs at different times.

Note that, as originally proposed by Horn, the parallel analysis method is based on analyzing the eigenvalues of the full correlation matrix (not the reduced correlation matrix). PROC FACTOR adopts the same basis to determine the number of factors.[5] Therefore, unlike the minimum eigenvalue or proportion method, this method is *not* affected by the choice of prior communality estimates (that is, the PRIORS= option).

[5] The authors of this paper recognize that some researchers have proposed that eigenvalues of the reduced correlation matrix, instead of the full correlation matrix, be analyzed in parallel analysis. This proposal is certainly logical and reasonable. However, our understanding is that previous supporting simulation results for parallel analysis were based on analyzing the full correlation matrix. Thus, supporting evidence for using the reduced correlation matrix in parallel analysis is needed for the authors to become advocates of the proposal.

Exploring the Number of Factors in Practice

If you omit the NFACTORS= option, in most situations PROC FACTOR determines the number of factors as the minimum value among those that are determined by the following criteria:

- MINEIGEN=1
- PROPORTION=1
- NFACTORS=*number of observed variables*

Neither the MAP method nor parallel analysis is conducted by default.

However, in most situations where exploratory factor analysis is appropriate, researchers want to look at all these analytical results before making their decision about the number of factors. PROC FACTOR enables you to explore all these analytical results from various methods and criteria without actually extracting the factors. For example, you can specify the following statements for the JobRating data set to request all the previously mentioned methods:

```
ods graphics on;
proc factor data=JobRating(type=corr) nobs=103 nfactors=0 priors=smc
            map parallel plots=(scree map parallel);
run;
```

When you specify the NFACTORS=0 option, no factors are extracted, although the eigenvalues and the cumulative proportions, such as those displayed in Figure 3, would still be available for exploration. The MAP and PARALLEL options produce results regarding minimum average partial correlations and parallel analysis. The PLOTS= option produces graphical plots. Hence, the specification here can be used as a code template for exploring the number of factors in any application.

EXAMPLE: NUMBER OF FACTORS IN THE BIG5COR DATA

To provide more insights into how to determine the number of factors in real applications, the four methods discussed in the preceding section are now applied to the big5cor correlation data set, which is described briefly in the INTRODUCTION section. The current section provides more background details about the data and demonstrates a more complete analysis of the number-of-factors problem.

The data were collected online between 2016 and 2018. Individuals responded to 50 items that were constructed using the Big Five factor markers from the International Personality Item Pool (IPIP) (Goldberg 1992). The 50 items were constructed so that each of the Big Five personality traits was represented primarily by exactly 10 items in the questionnaire used in the research. Participants rated items on a five-point scale to indicate whether each of the 50 items was an accurate description of their behavior. For a detailed description of the 50 items, see the IPIP website (International Personality Item Pool 2023).

The data were downloaded from the Kaggle website (Kaggle Inc. 2020). The total number of observations in the data set is 1,015,341. However, because of missing values, only 874,434 complete observations were used to compute the correlation matrix of the 50 items. Those items that were constructed as negative measures of the Big Five personality traits were reverse-scored before the computation of the correlation matrix, which was saved as a CORR-type SAS data set named big5cor.

An interesting question here would be whether the four methods could provide consistent suggestions about the number of factors. Another interesting question would be whether the analytic results could provide a validation of the questionnaire, which was supposed to measure only the Big Five factors.

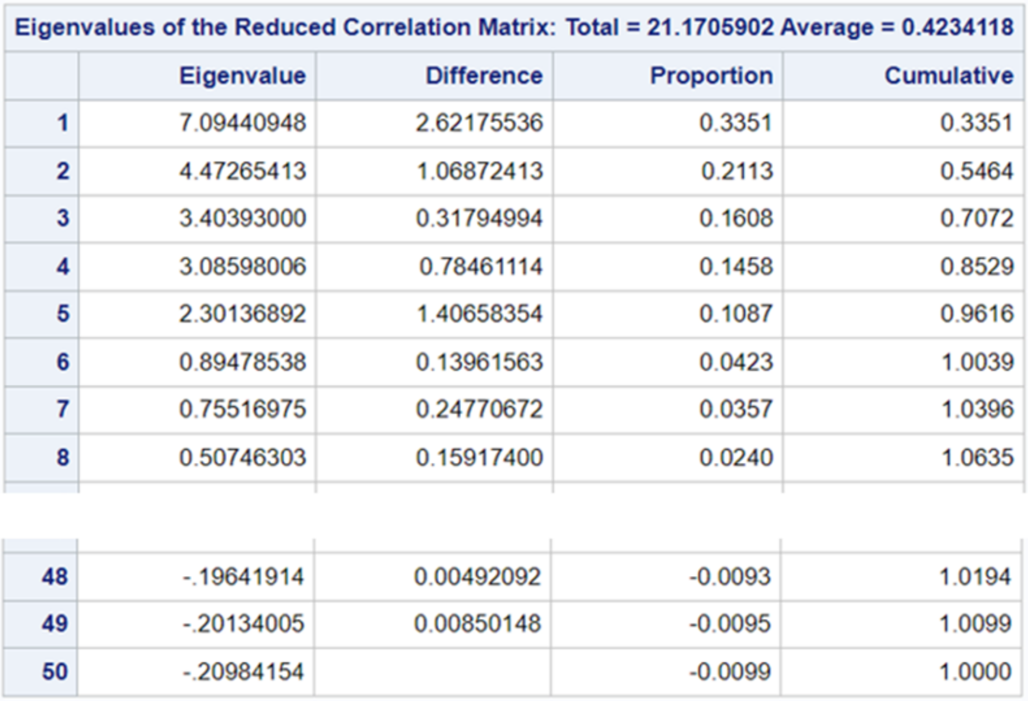
The following PROC FACTOR statement specifies various methods for exploring the number of factors in the data:

```
ods graphics on;
proc factor data=big5cor nfactor=0 map parallel(seed=12345)
     plots=all priors=smc;
run;
```

Note that the big5cor data set includes the information about the number of observations for computing the correlation matrix, and therefore the specification of the NOBS= option in the PROC FACTOR statement is not necessary.

Figure 6 displays a portion of results about the eigenvalues of the reduced correlation matrix. In the "Eigenvalue" column, the first five eigenvalues are greater than 1, and the eigenvalue of the sixth factor drops below 1 drastically. Looking at the last column, you can see that the 100% cumulative proportion of common variance is first attained at the sixth factor. Therefore, 5 factors would have been suggested with the MINEIGEN=1 option and 6 factors with the PROPORTION=1 option.

Figure 6: Eigenvalues

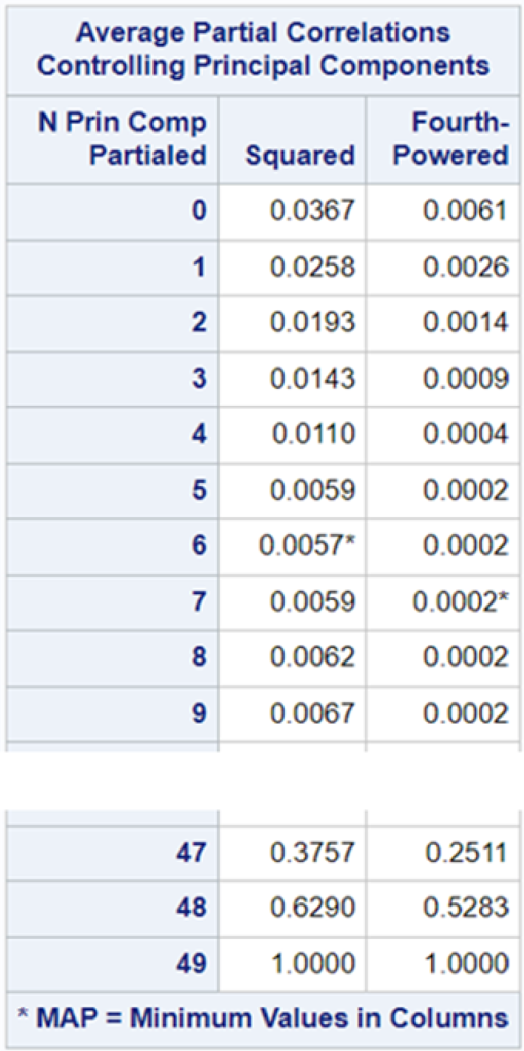


Eigenvalues of the Reduced Correlation Matrix: Total = 21.1705902 Average = 0.4234118

	Eigenvalue	Difference	Proportion	Cumulative
1	7.09440948	2.62175536	0.3351	0.3351
2	4.47265413	1.06872413	0.2113	0.5464
3	3.40393000	0.31794994	0.1608	0.7072
4	3.08598006	0.78461114	0.1458	0.8529
5	2.30136892	1.40658354	0.1087	0.9616
6	0.89478538	0.13961563	0.0423	1.0039
7	0.75516975	0.24770672	0.0357	1.0396
8	0.50746303	0.15917400	0.0240	1.0635
48	-.19641914	0.00492092	-0.0093	1.0194
49	-.20134005	0.00850148	-0.0095	1.0099
50	-.20984154		-0.0099	1.0000

Figure 7 and Figure 8 show the main results of the MAP and parallel analyses, respectively. The MAP2 criterion suggests 6 factors, whereas the MAP4 criterion suggests 7 factors. The parallel analysis suggests 8 factors at the default 0.05 α-level.

Figure 7: Average Partial Correlations

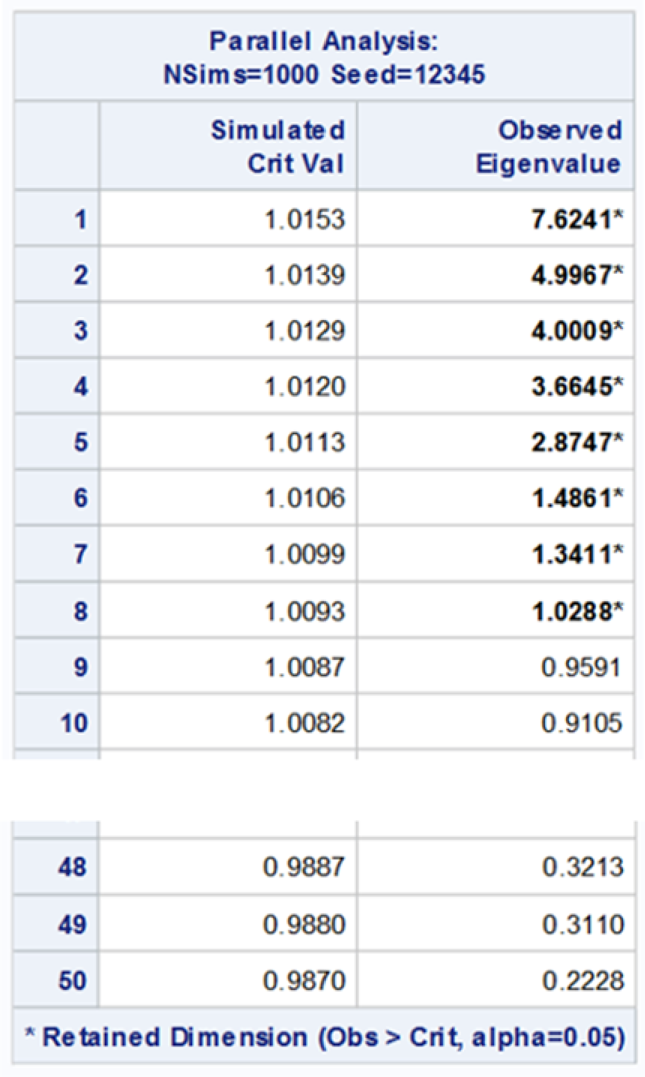


Average Partial Correlations Controlling Principal Components

N Prin Comp Partialed	Squared	Fourth-Powered
0	0.0367	0.0061
1	0.0258	0.0026
2	0.0193	0.0014
3	0.0143	0.0009
4	0.0110	0.0004
5	0.0059	0.0002
6	0.0057*	0.0002
7	0.0059	0.0002*
8	0.0062	0.0002
9	0.0067	0.0002
47	0.3757	0.2511
48	0.6290	0.5283
49	1.0000	1.0000

* MAP = Minimum Values in Columns

Figure 8: Parallel Analysis



Parallel Analysis: NSims=1000 Seed=12345

	Simulated Crit Val	Observed Eigenvalue
1	1.0153	**7.6241***
2	1.0139	**4.9967***
3	1.0129	**4.0009***
4	1.0120	**3.6645***
5	1.0113	**2.8747***
6	1.0106	**1.4861***
7	1.0099	**1.3411***
8	1.0093	**1.0288***
9	1.0087	0.9591
10	1.0082	0.9105
48	0.9887	0.3213
49	0.9880	0.3110
50	0.9870	0.2228

* Retained Dimension (Obs > Crit, alpha=0.05)

If you combine all these analytical results, the suggested number of factors ranges from 5 to 8. Which is the most plausible number for the current analysis? If you believe that the parallel analysis is *always* the best method among the four investigated here, then you should probably conclude that the number of factors is 8. However, because the sample size is extremely large in the current analysis (relative to most factor analysis applications), it is quite possible that the power to detect a factor by parallel analysis is so high that even trivial factors could have been included.

In fact, a more convincing statistical reason for not rushing to accept the results of the parallel analysis is that the null hypothesis of parallel analysis is simply that there are no factors due to uncorrelated variables. When you detect significant eigenvalues in practice, the logical conclusion is simply to reject the null hypothesis and conclude that there are *some* factors—no more and no less. To the best of our knowledge, there is actually no established logical and statistical ground for estimating the number of factors by using the number of observed significant (greater than 1) eigenvalues. Supporting evidence of parallel analysis has been based mainly on simulation results alone.

In addition, although the eigen-1 (MINEIGEN=1 option) criterion has been frequently criticized for its inability to take sampling fluctuations into account, it does present a compelling and clear picture about the plausible number of factors here. Again, as shown in Figure 6, the first five eigenvalues are very strong, ranging from 2.3 to 7.09, whereas the sixth eigenvalue (and all eigenvalues after it) suddenly drops below 1. Because the sample size in the current analysis is quite large, the unambiguous suggestion of 5 factors by the eigen-1 criterion cannot be dismissed simply as a haphazard result.

Finally, both the PROPORTION=1 and MAP2 methods suggest 6 factors, which is fewer than the 8 factors suggested by the parallel analysis. Therefore, we should be more cautious not to jump on the parallel analysis bandwagon too soon and conclude that the number of factors must be 8 for the current analysis. The next section explores several rotated solutions that use different numbers of factors and attempts to use the interpretability of the rotated solutions to make a more informed decision.

Rotated Solution with 5 Factors

When you use the NFACTORS=5 option, the following PROC FACTOR statement specifies a 5-factor solution for the big5cor data:

```
proc factor data=big5cor prior=smc method=prinit
            rotate=quartimin fuzz=0.3 nfactors=5;
run;
```

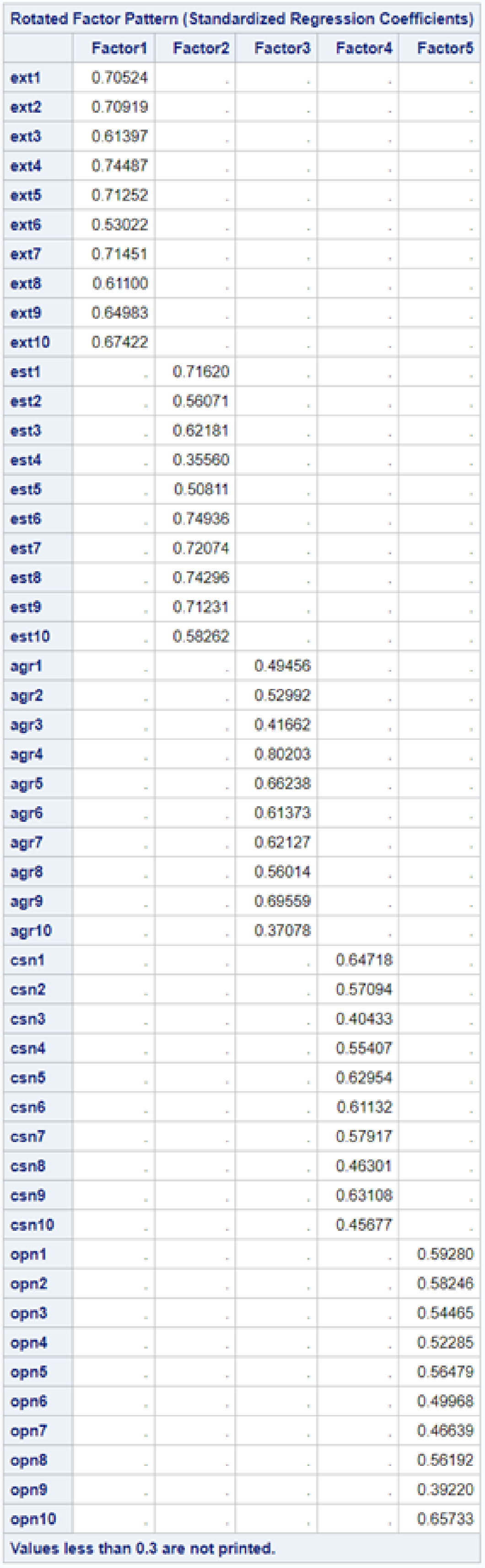
The PRIORS=SMC option uses the squared multiple correlations as the initial or prior communality estimates so that an initial principal factor solution is obtained. The METHOD=PRINIT option iterates the principal factor solutions until the communality estimates do not change. This iterative process boosts the communalities and compensates for the downward bias of using SMCs as initial estimates.

The ROTATE=QUARTIMIN option rotates the factor solution by the popular quartimin rotation, which transforms the initial orthogonal factors into oblique factors so that the rotated factor pattern or loading matrix attains a simple structure. With a simple factor pattern, the factors are easier to interpret than the orthogonal counterparts.

To provide an overall picture of the factor loading pattern, the FUZZ=0.3 option is used to hide the factor loadings whose magnitudes are less than 0.3. These small loadings are represented by dots (.) in Figure 9, which shows that the rotated factor pattern is consistent with the Big Five factor theory—that is, each of the 50 variables loads nontrivially (specifically, loadings are 0.3 or greater in magnitude) on exactly one factor, and each of the 5 factors is reflected exactly by the 10 items (observed variables) that were designed to measure it. With this 5-factor solution, it seems as if you could not ask for a better validation of the questionnaire that was designed for measuring the Big Five personality traits.

Figure 10 shows the correlation among the rotated factors. Figure 11 shows the common variance explained by these 5 factors.

Figure 9: Rotated Loadings with 5 Factors

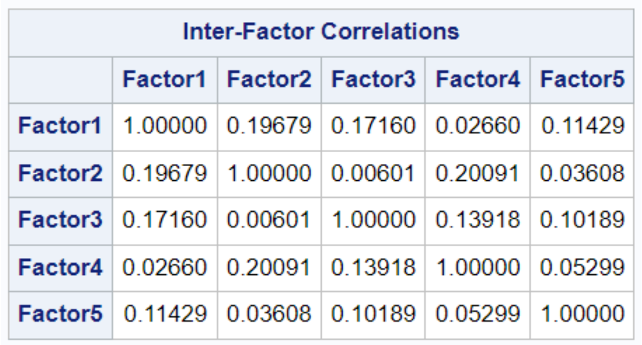


Rotated Factor Pattern (Standardized Regression Coefficients)

	Factor1	Factor2	Factor3	Factor4	Factor5
ext1	0.70524	.	.	.	.
ext2	0.70919	.	.	.	.
ext3	0.61397	.	.	.	.
ext4	0.74487	.	.	.	.
ext5	0.71252	.	.	.	.
ext6	0.53022	.	.	.	.
ext7	0.71451	.	.	.	.
ext8	0.61100	.	.	.	.
ext9	0.64983	.	.	.	.
ext10	0.67422	.	.	.	.
est1	.	0.71620	.	.	.
est2	.	0.56071	.	.	.
est3	.	0.62181	.	.	.
est4	.	0.35560	.	.	.
est5	.	0.50811	.	.	.
est6	.	0.74936	.	.	.
est7	.	0.72074	.	.	.
est8	.	0.74296	.	.	.
est9	.	0.71231	.	.	.
est10	.	0.58262	.	.	.
agr1	.	.	0.49456	.	.
agr2	.	.	0.52992	.	.
agr3	.	.	0.41662	.	.
agr4	.	.	0.80203	.	.
agr5	.	.	0.66238	.	.
agr6	.	.	0.61373	.	.
agr7	.	.	0.62127	.	.
agr8	.	.	0.56014	.	.
agr9	.	.	0.69559	.	.
agr10	.	.	0.37078	.	.
csn1	.	.	.	0.64718	.
csn2	.	.	.	0.57094	.
csn3	.	.	.	0.40433	.
csn4	.	.	.	0.55407	.
csn5	.	.	.	0.62954	.
csn6	.	.	.	0.61132	.
csn7	.	.	.	0.57917	.
csn8	.	.	.	0.46301	.
csn9	.	.	.	0.63108	.
csn10	.	.	.	0.45677	.
opn1	.	.	.	.	0.59280
opn2	.	.	.	.	0.58246
opn3	.	.	.	.	0.54465
opn4	.	.	.	.	0.52285
opn5	.	.	.	.	0.56479
opn6	.	.	.	.	0.49968
opn7	.	.	.	.	0.46639
opn8	.	.	.	.	0.56192
opn9	.	.	.	.	0.39220
opn10	.	.	.	.	0.65733

Values less than 0.3 are not printed.

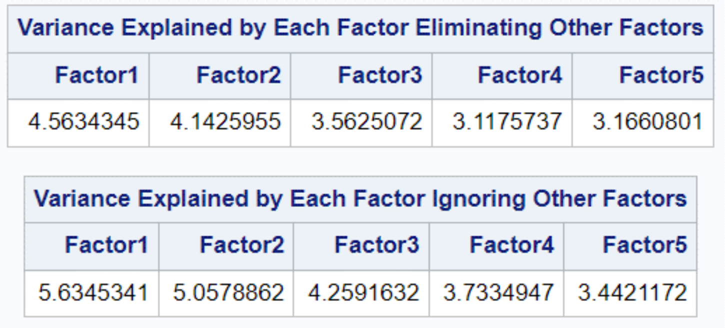
Figure 10: Factor Correlations with 5 Factors



Inter-Factor Correlations

	Factor1	Factor2	Factor3	Factor4	Factor5
Factor1	1.00000	0.19679	0.17160	0.02660	0.11429
Factor2	0.19679	1.00000	0.00601	0.20091	0.03608
Factor3	0.17160	0.00601	1.00000	0.13918	0.10189
Factor4	0.02660	0.20091	0.13918	1.00000	0.05299
Factor5	0.11429	0.03608	0.10189	0.05299	1.00000

Figure 11: Variance Explained by 5 Factors



Variance Explained by Each Factor Eliminating Other Factors

Factor1	Factor2	Factor3	Factor4	Factor5
4.5634345	4.1425955	3.5625072	3.1175737	3.1660801

Variance Explained by Each Factor Ignoring Other Factors

Factor1	Factor2	Factor3	Factor4	Factor5
5.6345341	5.0578862	4.2591632	3.7334947	3.4421172

Because the largest correlation in Figure 10 is barely above 0.2, you conclude that the 5 rotated factors are nonoverlapping. As indicated by the two tables shown in Figure 11, you can compute the common variance explained by the rotated factor in two ways. The first table computes the variances while eliminating the contribution from other factors. It shows that the 5 factors explain the total common variance quite evenly, ranging from 3.1 to 4.6. The second table computes the variances while ignoring the contribution from other factors. Thus, these variances are naturally larger than their counterparts in the first table. Again, the distribution of the variances explained by the factors is quite even in this table.

Rotated Solutions with 6 and 7 Factors

The 5-factor solution for the big5cor data set is almost perfect in that it matches the Big Five personality theory quite well. However, you should not stop exploring other feasible solutions that use different numbers of factors, because the MAP and parallel analyses do suggest that 6–8 factors are plausible. This section investigates the 6- and 7-factor solutions to see whether they can provide alternative or better interpretations.

Similar specifications to that of the 5-factor solution are now applied in order to obtain the 6- and 7-factor solutions, as shown in the following statements:

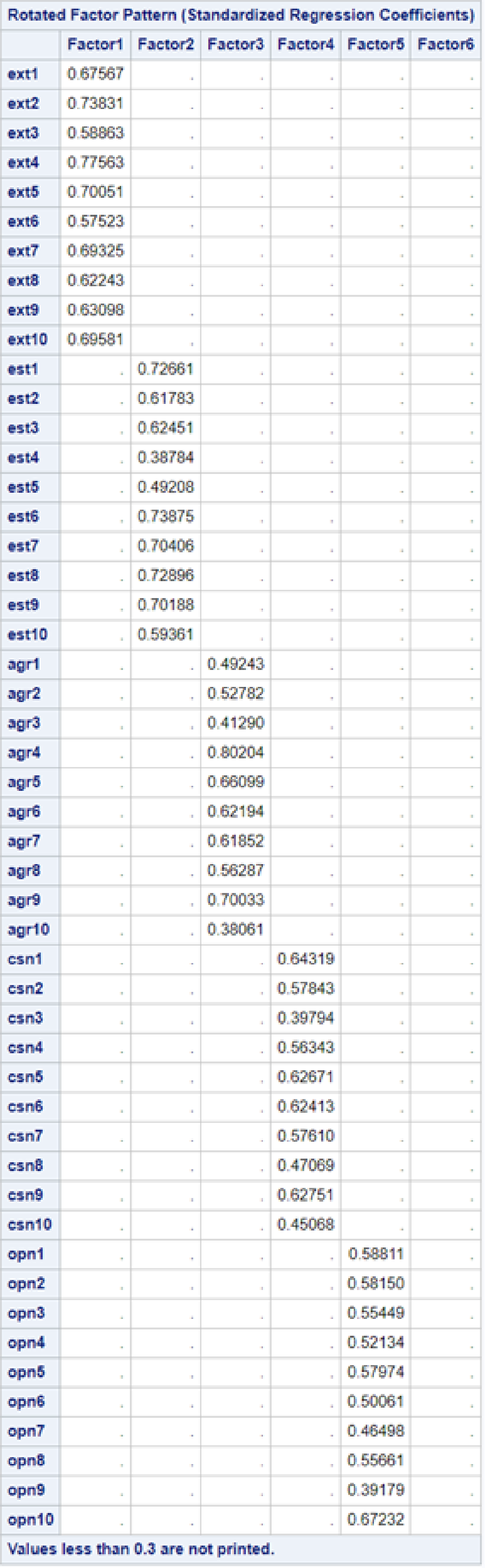
```
proc factor data=big5cor prior=smc method=prinit
            rotate=quartimin fuzz=0.3 nfactors=6;
run;

proc factor data=big5cor prior=smc method=prinit
            rotate=quartimin fuzz=0.3 nfactors=7;
run;
```

Figure 12 shows that the last (sixth) factor column in the 6-factor solution is filled in with only weak (or small) loadings—that is, they are all less than 0.3 in magnitude. Because the pattern of the first 5 factors in the 6-factor solution mimics that of the 5-factor solution, these 5 factors are identified with the Big Five personality traits. As a result, the last factor in the 6-factor solution is neither theoretically based nor strong and unique enough to be interpreted meaningfully.

Figure 13 shows that the last (seventh) factor column in the 7-factor solution, again, is filled in with only small loadings. Thus, the interpretability of this factor is in doubt. The fifth and sixth factors in the 7-factor solution seem as if they are “artificially” split from Factor 5 of the 5-factor solution. This makes neither the fifth nor sixth factor easier to interpret in the 7-factor solution. An additional problem is that the variable opn9 no longer has a salient loading on any factor.

Figure 12: Rotated Loadings with 6 Factors

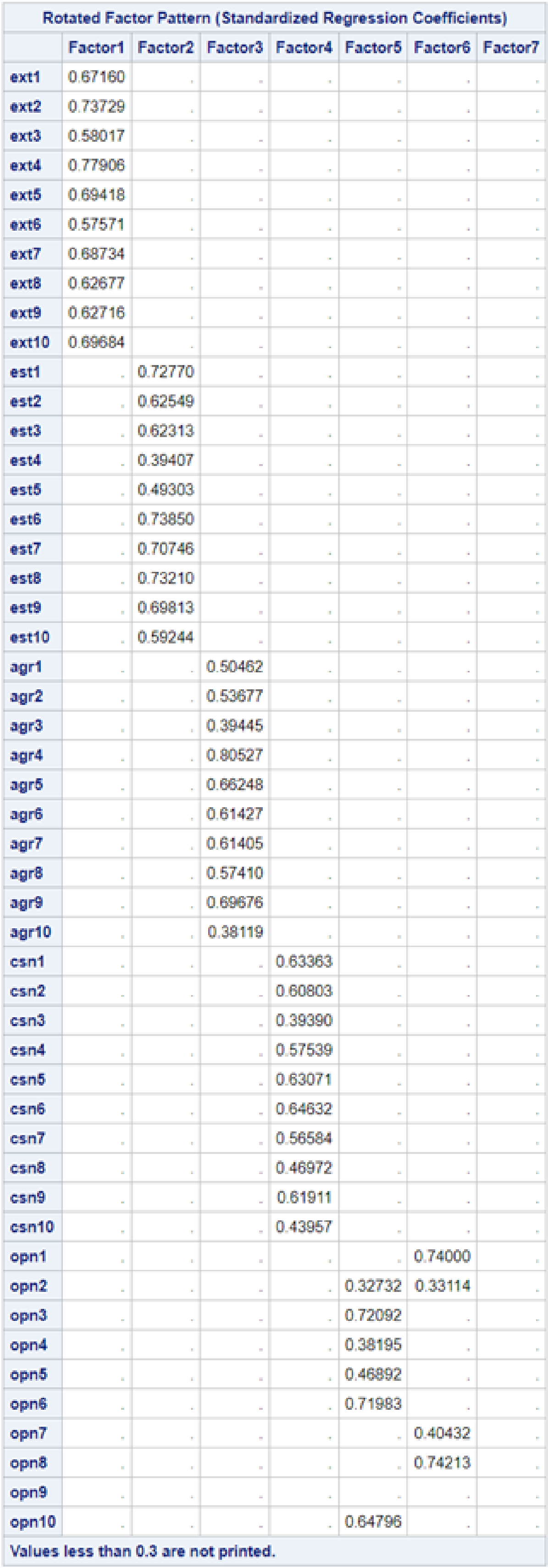


Rotated Factor Pattern (Standardized Regression Coefficients)

	Factor1	Factor2	Factor3	Factor4	Factor5	Factor6
ext1	0.67567	.	.	.	.	.
ext2	0.73831	.	.	.	.	.
ext3	0.58863	.	.	.	.	.
ext4	0.77563	.	.	.	.	.
ext5	0.70051	.	.	.	.	.
ext6	0.57523	.	.	.	.	.
ext7	0.69325	.	.	.	.	.
ext8	0.62243	.	.	.	.	.
ext9	0.63098	.	.	.	.	.
ext10	0.69581	.	.	.	.	.
est1	.	0.72661	.	.	.	.
est2	.	0.61783	.	.	.	.
est3	.	0.62451	.	.	.	.
est4	.	0.38784	.	.	.	.
est5	.	0.49208	.	.	.	.
est6	.	0.73875	.	.	.	.
est7	.	0.70406	.	.	.	.
est8	.	0.72896	.	.	.	.
est9	.	0.70188	.	.	.	.
est10	.	0.59361	.	.	.	.
agr1	.	.	0.49243	.	.	.
agr2	.	.	0.52782	.	.	.
agr3	.	.	0.41290	.	.	.
agr4	.	.	0.80204	.	.	.
agr5	.	.	0.66099	.	.	.
agr6	.	.	0.62194	.	.	.
agr7	.	.	0.61852	.	.	.
agr8	.	.	0.56287	.	.	.
agr9	.	.	0.70033	.	.	.
agr10	.	.	0.38061	.	.	.
csn1	.	.	.	0.64319	.	.
csn2	.	.	.	0.57843	.	.
csn3	.	.	.	0.39794	.	.
csn4	.	.	.	0.56343	.	.
csn5	.	.	.	0.62671	.	.
csn6	.	.	.	0.62413	.	.
csn7	.	.	.	0.57610	.	.
csn8	.	.	.	0.47069	.	.
csn9	.	.	.	0.62751	.	.
csn10	.	.	.	0.45068	.	.
opn1	.	.	.	.	0.58811	.
opn2	.	.	.	.	0.58150	.
opn3	.	.	.	.	0.55449	.
opn4	.	.	.	.	0.52134	.
opn5	.	.	.	.	0.57974	.
opn6	.	.	.	.	0.50061	.
opn7	.	.	.	.	0.46498	.
opn8	.	.	.	.	0.55661	.
opn9	.	.	.	.	0.39179	.
opn10	.	.	.	.	0.67232	.

Values less than 0.3 are not printed.

Figure 13: Rotated Loadings with 7 Factors



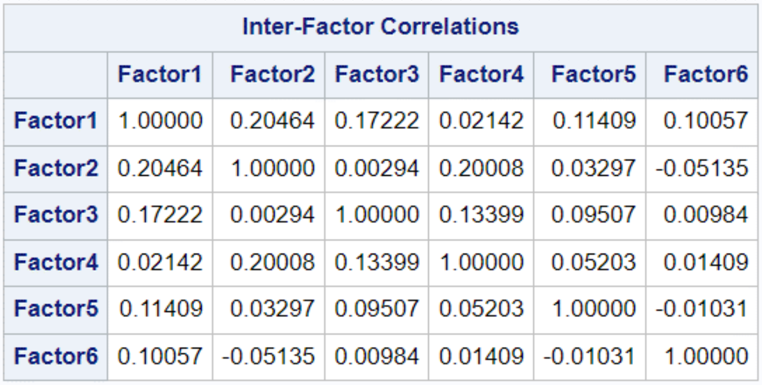
Rotated Factor Pattern (Standardized Regression Coefficients)

	Factor1	Factor2	Factor3	Factor4	Factor5	Factor6	Factor7
ext1	0.67160	.	.	.	.	.	.
ext2	0.73729	.	.	.	.	.	.
ext3	0.58017	.	.	.	.	.	.
ext4	0.77906	.	.	.	.	.	.
ext5	0.69418	.	.	.	.	.	.
ext6	0.57571	.	.	.	.	.	.
ext7	0.68734	.	.	.	.	.	.
ext8	0.62677	.	.	.	.	.	.
ext9	0.62716	.	.	.	.	.	.
ext10	0.69684	.	.	.	.	.	.
est1	.	0.72770	.	.	.	.	.
est2	.	0.62549	.	.	.	.	.
est3	.	0.62313	.	.	.	.	.
est4	.	0.39407	.	.	.	.	.
est5	.	0.49303	.	.	.	.	.
est6	.	0.73850	.	.	.	.	.
est7	.	0.70746	.	.	.	.	.
est8	.	0.73210	.	.	.	.	.
est9	.	0.69813	.	.	.	.	.
est10	.	0.59244	.	.	.	.	.
agr1	.	.	0.50462	.	.	.	.
agr2	.	.	0.53677	.	.	.	.
agr3	.	.	0.39445	.	.	.	.
agr4	.	.	0.80527	.	.	.	.
agr5	.	.	0.66248	.	.	.	.
agr6	.	.	0.61427	.	.	.	.
agr7	.	.	0.61405	.	.	.	.
agr8	.	.	0.57410	.	.	.	.
agr9	.	.	0.69676	.	.	.	.
agr10	.	.	0.38119	.	.	.	.
csn1	.	.	.	0.63363	.	.	.
csn2	.	.	.	0.60803	.	.	.
csn3	.	.	.	0.39390	.	.	.
csn4	.	.	.	0.57539	.	.	.
csn5	.	.	.	0.63071	.	.	.
csn6	.	.	.	0.64632	.	.	.
csn7	.	.	.	0.56584	.	.	.
csn8	.	.	.	0.46972	.	.	.
csn9	.	.	.	0.61911	.	.	.
csn10	.	.	.	0.43957	.	.	.
opn1	.	.	.	.	.	0.74000	.
opn2	.	.	.	.	0.32732	0.33114	.
opn3	.	.	.	.	0.72092	.	.
opn4	.	.	.	.	0.38195	.	.
opn5	.	.	.	.	0.46892	.	.
opn6	.	.	.	.	0.71983	.	.
opn7	.	.	.	.	.	0.40432	.
opn8	.	.	.	.	.	0.74213	.
opn9	.	.	.	.	.	.	.
opn10	.	.	.	.	0.64796	.	.

Values less than 0.3 are not printed.

Figure 14 and Figure 15 display the factor correlations, respectively, for the 6- and 7-factor solutions. Figure 14 shows that Factor 6 in the 6-factor solution is very mildly correlated with the first 5 factors, which have been identified as the Big Five personality traits from the factor pattern results. The factor correlation pattern here does not raise any concerns—but still, the interpretability of the sixth factor in the 6-factor solution is the main issue.

Figure 14: Factor Correlations with 6 Factors

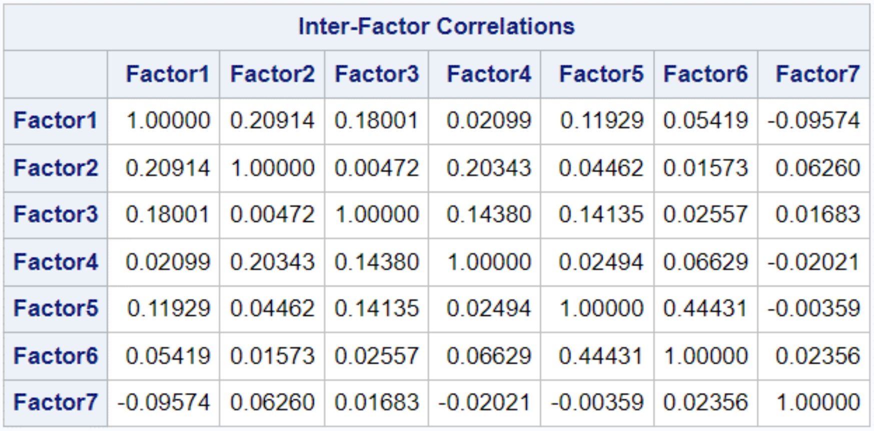


Inter-Factor Correlations

	Factor1	Factor2	Factor3	Factor4	Factor5	Factor6
Factor1	1.00000	0.20464	0.17222	0.02142	0.11409	0.10057
Factor2	0.20464	1.00000	0.00294	0.20008	0.03297	-0.05135
Factor3	0.17222	0.00294	1.00000	0.13399	0.09507	0.00984
Factor4	0.02142	0.20008	0.13399	1.00000	0.05203	0.01409
Factor5	0.11409	0.03297	0.09507	0.05203	1.00000	-0.01031
Factor6	0.10057	-0.05135	0.00984	0.01409	-0.01031	1.00000

Figure 15 shows that Factor 7 in the 7-factor solution is very mildly correlated with all other factors. Although such a correlation pattern of the seventh factor is not a concern, its interpretability is still an issue. Moreover, Factors 5 and 6 in this solution are moderately correlated at 0.44 here. This correlation is an indication that these two factors could have been combined to form a single factor much like the fifth factor in the 5-factor solution.

Figure 15: Factor Correlations with 7 Factors

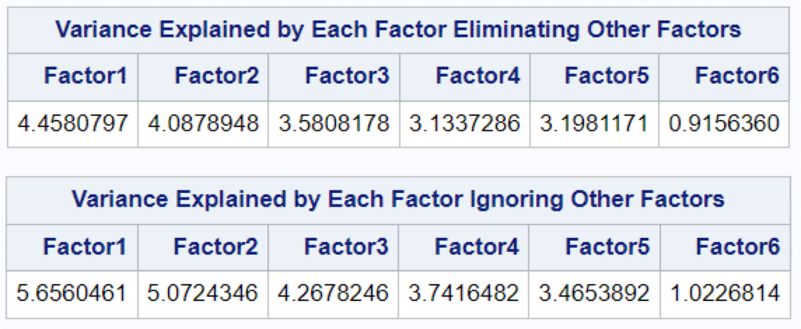


Inter-Factor Correlations

	Factor1	Factor2	Factor3	Factor4	Factor5	Factor6	Factor7
Factor1	1.00000	0.20914	0.18001	0.02099	0.11929	0.05419	-0.09574
Factor2	0.20914	1.00000	0.00472	0.20343	0.04462	0.01573	0.06260
Factor3	0.18001	0.00472	1.00000	0.14380	0.14135	0.02557	0.01683
Factor4	0.02099	0.20343	0.14380	1.00000	0.02494	0.06629	-0.02021
Factor5	0.11929	0.04462	0.14135	0.02494	1.00000	0.44431	-0.00359
Factor6	0.05419	0.01573	0.02557	0.06629	0.44431	1.00000	0.02356
Factor7	-0.09574	0.06260	0.01683	-0.02021	-0.00359	0.02356	1.00000

Figure 16 and Figure 17 display the variances explained by the factors in the 6- and 7-factor solutions. In Figure 16, Factor 6 seems to be explaining much less common variance relative to all other factors. It confirms that Factor 6 is a relatively weak factor.

Figure 16: Variance Explained by 6 Factors

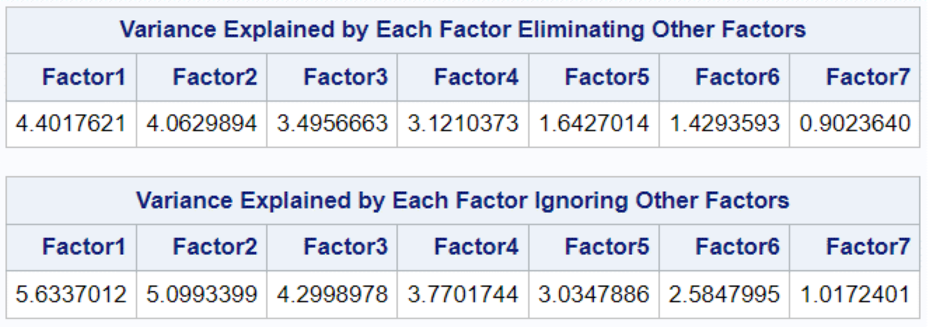


Variance Explained by Each Factor Eliminating Other Factors					
Factor1	Factor2	Factor3	Factor4	Factor5	Factor6
4.4580797	4.0878948	3.5808178	3.1337286	3.1981171	0.9156360

Variance Explained by Each Factor Ignoring Other Factors					
Factor1	Factor2	Factor3	Factor4	Factor5	Factor6
5.6560461	5.0724346	4.2678246	3.7416482	3.4653892	1.0226814

In Figure 17, Factors 5, 6, and 7 explain relatively small portions of common variance, eliminating the contributions from other factors. This pattern itself is not an issue if these factors are interpretable. But when you compare this pattern to that of the 5-factor solution in Figure 11, it confirms that Factors 6 and 7 dilute the common variance explained by the theoretical Big Five personality traits. This dilution especially compromises the interpretation of the fifth factor (intellect, imagination, or openness) in the Big Five factor theory.

Figure 17: Variance Explained by 7 Factors



Variance Explained by Each Factor Eliminating Other Factors						
Factor1	Factor2	Factor3	Factor4	Factor5	Factor6	Factor7
4.4017621	4.0629894	3.4956663	3.1210373	1.6427014	1.4293593	0.9023640

Variance Explained by Each Factor Ignoring Other Factors						
Factor1	Factor2	Factor3	Factor4	Factor5	Factor6	Factor7
5.6337012	5.0993399	4.2998978	3.7701744	3.0347886	2.5847995	1.0172401

The 8-factor solution exhibits issues similar to those found in the 6- and 7-factor solutions. To save space, no further explanations are presented for the 8-factor solution. In summary, the 6- and 7-factor solutions are not as plausible as the 5-factor solution for the following reasons:

- The 6- and 7-factor solutions both have interpretability issues with the additional weak factors: their corresponding loadings are too small to be identified with meaningful constructs.
- When compared to the 5-factor solution, the 7-factor solution has two extra factors that seem to have been yielded by an artificial split from the Big Five personality traits. This especially weakens the interpretability of the intellect (imagination or openness) factor in the Big Five personality theory.

SUMMARY AND CONCLUSION

This paper explains and demonstrates several methods for determining the number of factors to retain in exploratory factor analysis. You can use scree plots, minimum eigenvalue criteria, proportion of common variance criteria, minimum average partial correlations, and parallel analysis to tackle the problem. All these methods are supported by the FACTOR procedure in SAS/STAT® software.

To summarize, it is recommended that you use multiple methods (or criteria) to tackle the number-of-factors problem in practical exploratory factor analysis. A plausible range of numbers of factors can be established by applying these methods. Then, possibly with the help of substantive theory, you should carefully examine different scenarios for different numbers of factors and identify the most interpretable factor solutions.

REFERENCES

Bollen, K. A. (1989). *Structural Equations with Latent Variables*. New York: John Wiley & Sons.

Cattell, R. B. (1966). "The Scree Test for the Number of Factors." *Multivariate Behavioral Research* 1:245–276.

Dinno, A. (2009). "Exploring the Sensitivity of Horn's Parallel Analysis to the Distributional Form of Random Data." *Multivariate Behavioral Research* 44:362–388.

Everitt, B. S. (1984). *An Introduction to Latent Variable Methods*. London: Chapman & Hall.

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.

Goldberg, L. R. (1990). "An Alternative 'Description of Personality': The Big-Five Factor Structure." *Journal of Personality and Social Psychology* 59:1216–1229.

Goldberg, L. R. (1992). "The Development of Markers for the Big-Five Factor Structure." *Psychological Assessment* 4:26–42.

Gorsuch, R. L. (1974). *Factor Analysis*. Philadelphia: W. B. Saunders.

Harman, H. H. (1976). *Modern Factor Analysis*. 3rd 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.

International Personality Item Pool (2023). "Big-Five Factor Markers." Accessed November 22, 2023. https://ipip.ori.org/newBigFive5broadKey.htm.

Kaggle Inc. (2020). "Big 5 Personality Scores." Accessed November 22, 2023. https://www.kaggle.com/datasets/lucasgreenwell/big5personalitydataset.

Kaiser, H. F. (1960). "The Application of Electronic Computers to Factor Analysis." *Educational and Psychological Measurement* 20:141–151.

Kim, J. O., and Mueller, C. W. (1978a). *Factor Analysis: Statistical Methods and Practical Issues.* Vol. 07-014 of Sage University Paper Series on Quantitative Applications in the Social Sciences. Beverly Hills, CA: Sage Publications.

Kim, J. O., and Mueller, C. W. (1978b). *Introduction to Factor Analysis: What It Is and How to Do It.* Vol. 07-013 of Sage University Paper Series on Quantitative Applications in the Social Sciences. Beverly Hills, CA: Sage Publications.

Loehlin, J. C. (2004). *Latent Variable Models: An Introduction to Factor, Path, and Structural Analysis*. 4th ed. Mahwah, NJ: Lawrence Erlbaum Associates.

Long, J. S. (1983). *Covariance Structure Models: An Introduction to LISREL*. Beverly Hills, CA: Sage Publications.

Mardia, K. V., Kent, J. T., and Bibby, J. M. (1979). *Multivariate Analysis*. London: Academic Press.

Morrison, D. F. (1976). *Multivariate Statistical Methods*. 2nd ed. New York: McGraw-Hill.

SAS Institute Inc. (2017). *SAS/STAT 14.3 User's Guide*. Cary, NC: SAS Institute Inc. http://go.documentation.sas.com/?docsetId=statug&docsetTarget=titlepage.htm&docsetVersion=14.3&locale=en.

Spearman, C. (1904). "General Intelligence Objectively Determined and Measured." *American Journal of Psychology* 15:201–293.

Velicer, W. F. (1976). "Determining the Number of Components from the Matrix of Partial Correlations." *Psychometrika* 41:321–327.

Velicer, W. F., Eaton, C. A., and Fava, J. L. (2000). "Construct Explication through Factor or Component Analysis: A Review and Evaluation of Alternative Procedures for Determining the Number of Factors or Components." In *Problems and Solutions in Human Assessment: Honoring Douglas N. Jackson at Seventy*, edited by R. D. Goffin and E. Helmes, 41–71. Boston: Kluwer.

Zwick, W. R., and Velicer, W. F. (1986). "Comparison of Five Rules for Determining the Number of Components to Retain." *Psychological Bulletin* 99:432–442.

ACKNOWLEDGMENT

The authors would like to thank Ed Huddleston for editing the paper. Any mistakes are ours.