1 Determining the Number of Components in PCA

1.1 Introduction

(Continuing from 7 question questionnaire example.)

- The preceding section may have created the impression that, if a principal component analysis were performed on data from the 7-item job satisfaction questionnaire, only two components would be created. However, such an impression would not be entirely correct.
- **Important:** Correctly, the actual number of components extracted in a principal component analysis ***is equal*** to the number of observed variables being analyzed. This means that an analysis of the 7-item questionnaire would actually result in seven components, not two.
- However, in most analyses, only the first few components account for meaningful amounts of variance, so only these first few components are retained, interpreted, and used in subsequent analyses (such as in multiple regression analyses).
- For example, in your analysis of the 7-item job satisfaction questionnaire, it is likely that only the first two components would account for a meaningful amount of variance; therefore only these would be retained for interpretation. You would assume that the remaining five components accounted for only trivial amounts of variance. These latter components would therefore be discarded.

1.2 Methods of Determining the Number of Meaningful Components to Retain

- **Important:** As you carry out the analysis, you must decide just how many of these components are truly meaningful and worthy of being retained for further analysis.
- In general, you expect that only the first few components will account for meaningful amounts of variance, and that the later components will tend to account for only trivial variance.
- The next step of the analysis, therefore , is to determine how many meaningful components should be retained for interpretation.
- The followings section will describe four criteria that may be used in making this decision:

  1. the scree test,
  2. the eigenvalue-one criterion,
  3. the proportion of variance accounted for,
  4. the interpretability criterion.

  In the next section, we will go through them one by one.

2 Selection Techniques

2.1 Method 1: The Scree Test

- With the **Scree Test** *(Cattell, 1966)*, eigenvalues associated with each component are plotted in descending order. We look for a break between the components with relatively large eigenvalues and those with small eigenvalues.
- The components that appear before the break are assumed to be meaningful and are retained for analysis; those apppearing after the break are assumed to be unimportant and are not retained.
- **Remark:** The word "scree" refers to the loose rubble that lies at the base of a cliff. When performing a scree test, you normally hope that the scree plot will take the form of a cliff: At the top will be the eigenvalues for the few meaningful components, followed by a break (the edge of the cliff). At the bottom of the cliff will lie the scree: eigenvalues for the trivial components.

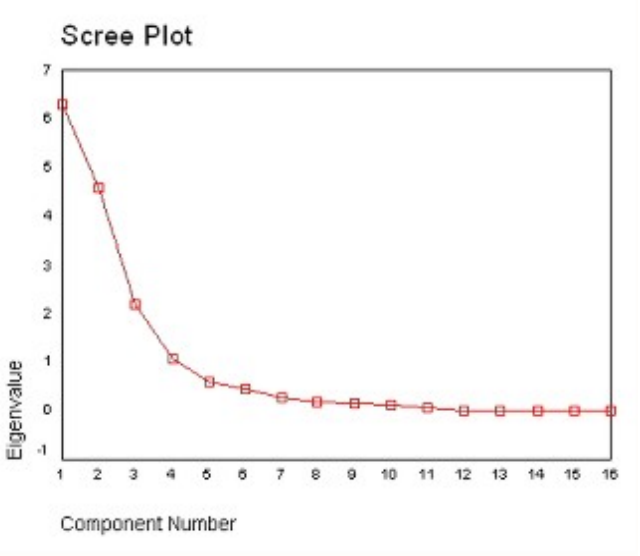

Figure 1: Example of a Scree Plot

- Sometimes a scree plot will display several large breaks. When this is the case, you could look for the last big break before the eigenvalues begin to level off. Only the components that appear before this last large break should be retained.

Advantages and Disadvantages

- The scree test can be expected to provide reasonably accurate results, provided the sample is large (over 200) and most of the variable communalities are large *(Stevens, 1986)*.

- However, this criterion has its own weaknesses as well, most notably the ambiguity that is often displayed by scree plots under typical research conditions: Very often, it is difficult to determine exactly where in the scree plot a break exists, or even if a break exists at all.

2.2 Method 2: The Eigenvalue-One Criterion

- In principal component analysis, one of the most commonly used criteria for solving the number-of-components problem is the eigenvalue-one criterion, also known as the Kaiser criterion *(Kaiser, 1960)*.
- **Important:** With this approach, you retain and interpret any component with an eigenvalue greater than 1.00. (hence the name)
- The rationale for this criterion is simple:
  - Each observed variable contributes one unit of variance to the total variance in the data set. Any component that displays an eigenvalue greater than 1.00 is accounting for a greater amount of variance than had been contributed by one variable. Such a component is therefore accounting for a meaningful amount of variance, and is worthy of being retained.
  - On the other hand, a component with an eigenvalue less than 1.00 is accounting for less variance than had been contributed by one variable.
- The purpose of principal component analysis is to reduce a number of observed variables into a relatively smaller number of components; this cannot be effectively achieved if you retain components that account for less variance than had been contributed by individual variables. For this reason, components with eigenvalues less than 1.00 are viewed as trivial, and are not retained.

Advantages and Disadvantages

- The eigenvalue-one criterion has a number of positive features that have contributed to its popularity. Perhaps the most important reason for its widespread use is its simplicity: You do not make any subjective decisions, but merely retain components with eigenvalues greater than one.
- On the positive side, it has been shown that this criterion very often results in retaining the correct number of components, particularly when a small to moderate number of variables are being analyzed and the variable communalities are high.
  Stevens (1986) reviews studies that have investigated the accuracy of the eigenvalue-one criterion, and recommends its use when less than 30 variables are being analyzed and communalities are greater than 0.70, or when the analysis is based on over 250 observations and the mean communality is greater than or equal to 0.60.
- There are a number of problems associated with the eigenvalue-one criterion, however. As was suggested in the preceding paragraph, it can lead to retaining the wrong number of components under circumstances that are often encountered in research (e.g., when many variables are analyzed, when communalities are small).

Total Variance Explained

Factor	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	3.550	23.668	23.668	2.917	19.445	19.445
2	1.298	8.624	32.293	.671	4.474	23.919
3	1.204	8.029	40.317	.552	3.680	27.600
4	1.152	7.680	47.997	.516	3.440	31.039
5	1.055	7.035	55.033	.175	1.169	32.208
6	.964	6.427	61.459			
7	.751	5.004	66.463			
8	.702	4.680	71.143			
9	.686	4.562	75.705			
10	.671	4.476	80.181			
11	.651	4.337	84.518			
12	.615	4.103	88.621			
13	.503	3.356	92.977			
14	.587	3.912	96.889			
15	.527	3.511	100.000			

Extraction Method: Maximum Likelihood.

- Also, the mindless application of this criterion can lead to retaining a certain number of components when the actual difference in the eigenvalues of successive components is only trivial.
    - For example, if component 2 displays an eigenvalue of 1.002 and component 3 displays an eigenvalue of 0.998, then component 2 will be retained but component 3 will not; this may mislead you into believing that the third component was meaningless when, in fact, it accounted for almost exactly the same amount of variance as the second component.
- In short, the eigenvalue-one criterion can be helpful when used judiciously, but the thoughtless application of this approach can lead to serious errors of interpretation.

2.3 (For Next Method) Total Variance in the context of PCA

- To understand the meaning of **total variance** as it is used in a principal component analysis, remember that the observed variables are standardized in the course of the analysis. This means that each variable is transformed so that it has a mean of zero and a variance of one.
- **Important** In the SPSS output they are equivalent to the eigen-values for each component.
- The total variance in the data set is simply the sum of the variances of these observed variables. Because they have been standardized to have a variance of one, each observed variable contributes one unit of variance to the total variance in the data set. Because of this, the total variance in a principal component analysis ***will always be equal*** to the number of observed variables being analyzed.

- For example, if seven variables are being analyzed, the total variance will equal seven. The components that are extracted in the analysis will partition this variance: perhaps the first component will account for 3.2 units of total variance; perhaps the second component will account for 2.1 units. The analysis continues in this way until all of the variance in the data set (i.e. the remaining 1.7 units). has been accounted for.

2.4 Method 3: Proportion of Variance Accounted For

- A third criterion in solving the number of factors problem involves retaining a component if it accounts for a specified proportion (or percentage) of variance in the data set.
- For example, you may decide to retain any component that accounts for at least 5% or 10% of the total variance. This proportion can be calculated with a simple formula:

$$\text{Proportion} = \frac{\text{Eigenvalue for the component of interest}}{\text{Total eigenvalues of the correlation matrix}}$$

- In principal component analysis, the total eigenvalues of the correlation matrix is equal to the total number of variables being analyzed (because each variable contributes one unit of variance to the analysis).
- An alternative criterion is to retain enough components so that the cumulative percent of variance accounted for is equal to some minimal value. Suppose that,in a PCA procedure, that components 1, 2, 3, and 4 accounted for approximately 37%, 33%, 13%, and 10% of the total variance, respectively.
- Suppose that it was required to account for 90% of the variance. Adding these percentages together results in a sum of 93%. This means that the cumulative percent of variance accounted for by components 1, 2, 3, and 4 is 93%.

Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	3.88155846	2.24699126	0.4313	0.4313
2	1.63456720	0.57436630	0.1816	0.6129
3	1.06020090	0.10480537	0.1178	0.7307
4	0.95539554	0.42415968	0.1062	0.8369
5	0.53123586	0.10477119	0.0590	0.8959
6	0.42646467	0.13882496	0.0474	0.9433
7	0.28763971	0.16930381	0.0320	0.9752
8	0.11833590	0.01373414	0.0131	0.9884
9	0.10460176		0.0116	1.0000

Figure 2: Eigenvalue Table

Advantages and Disadvantages

- The proportion of variance criterion has a number of positive features. For example, in most cases, you would not want to retain a group of components that, combined, account for only a minority of the variance in the data set (for example, 30%).
- Nonetheless, many critical values discussed earlier are obviously arbitrary. Because of these and related problems, this approach has sometimes been criticized for its subjectivity *(Kim and Mueller, 1978)*.

2.5 Method 4: The Interpretability Criteria

Perhaps the most important criterion for solving the ***number of-components*** problem is the interpretability criterion: interpreting the substantive meaning of the retained components and verifying that this interpretation makes sense in terms of what is known about the data under investigation, a Post-Hoc analysis in other words.

Remark: There is a lot more to this in practice, but we will not have time to consider it fully.

3 Rotations

- Ideally, you would like to review the correlations between the variables and the components and use this information to interpret the components; that is, to determine what construct seems to be measured by component 1, what construct seems to be measured by component 2, and so forth.
- **Important:** Unfortunately, when more than one component has been retained in an analysis, the interpretation of an unrotated factor pattern is usually quite difficult. To make interpretation easier, you will normally perform an operation called a **rotation**.
- **Definition:** A rotation is a linear transformation that is performed on the factor solution for the purpose of making the solution easier to interpret.