Principal Component and Factor Analysis

8

Learning Objectives

After reading this chapter, you should understand:

– The basics of principal component and factor analysis.
– The principles of exploratory and confirmatory factor analysis.
– Key terms, such as communality, eigenvalues, factor loadings, factor scores, and uniqueness.
– What rotation is.
– The principles of exploratory and confirmatory factor analysis.
– How to determine whether data are suitable for carrying out an exploratory factor analysis.
– How to interpret Stata principal component and factor analysis output.

– The principles of reliability analysis and its execution in Stata.
– The concept of structural equation modeling.

8.1   Introduction

Principal component analysis (**PCA**) and **factor analysis** (also called **principal factor analysis** or **principal axis factoring**) are two methods for identifying structure within a set of variables. Many analyses involve large numbers of variables that are difficult to interpret. Using PCA or factor analysis helps find interrelationships between variables (usually called items) to find a smaller number of unifying variables called **factors**. Consider the example of a soccer club whose management wants to measure the satisfaction of the fans. The management could, for instance, measure fan satisfaction by asking how satisfied the fans are with the (1) assortment of merchandise, (2) quality of merchandise, and (3) prices of merchandise. It is likely that these three items together measure satisfaction with the merchandise. Through the application of PCA or factor analysis, we can determine whether a single factor represents the three satisfaction items well. Practically, PCA and factor analysis are applied to understand much larger sets of variables, tens or even hundreds, when just reading the variables' descriptions does not determine an obvious or immediate number of factors.

PCA and factor analysis both explain patterns of correlations within a set of observed variables. That is, they identify sets of highly correlated variables and infer an underlying factor structure. While PCA and factor analysis are very similar in the way they arrive at a solution, they differ fundamentally in their assumptions of the variables' nature and their treatment in the analysis. Due to these differences, the methods follow different research objectives, which dictate their areas of application. While the PCA's objective is to *reproduce* a data structure, as well as possible only using a few factors, factor analysis aims to *explain* the variables' correlations by means of factors (e.g., Hair et al. 2013; Matsunaga 2010; Mulaik 2009).[1] We will discuss these differences and their implications in this chapter.

Both PCA and factor analysis can be used for exploratory or confirmatory purposes. What are exploratory and confirmatory factor analyses? Comparing the left and right panels of Fig. 8.1 shows us the difference. **Exploratory factor analysis**, often simply referred to as EFA, does not rely on previous ideas on the factor structure we may find. That is, there may be relationships (indicated by the arrows) between each factor and each item. While some of these relationships may be weak (indicated by the dotted arrows), others are more pronounced, suggesting that these items represent an underlying factor well. The left panel of Fig. 8.1 illustrates this point. Thus, an exploratory factor analysis reveals the number of factors and the items belonging to a specific factor. In a **confirmatory factor**

---

[1]Other methods for carrying out factor analyses include, for example, unweighted least squares, generalized least squares, or maximum likelihood. However, these are statistically complex and inexperienced users should not consider them.
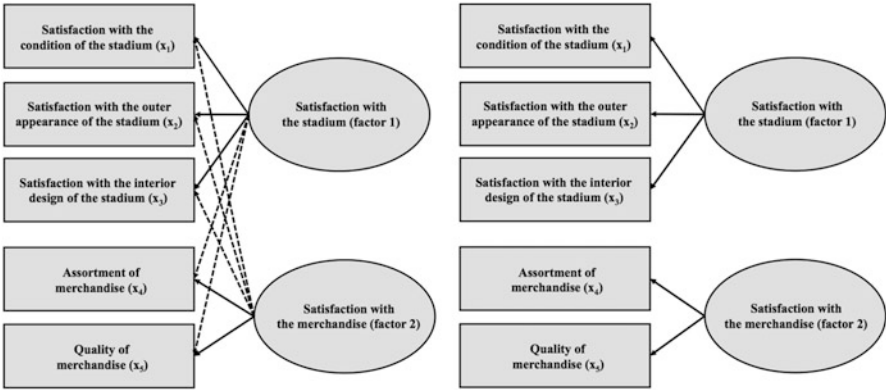
Fig. 8.1 Exploratory factor analysis (*left*) and confirmatory factor analysis (*right*)

analysis, usually simply referred to as CFA, there may only be relationships between a factor and specific items. In the right panel of Fig. 8.1, the first three items relate to factor 1, whereas the last two items relate to factor 2. Different from the exploratory factor analysis, in a confirmatory factor analysis, we have clear expectations of the factor structure (e.g., because researchers have proposed a scale that we want to adapt for our study) and we want to test for the expected structure.

In this chapter, we primarily deal with exploratory factor analysis, as it conveys the principles that underlie all factor analytic procedures and because the two techniques are (almost) identical from a statistical point of view. Nevertheless, we will also discuss an important aspect of confirmatory factor analysis, namely **reliability analysis**, which tests the consistency of a measurement scale (see Chap. 3). We will also briefly introduce a specific confirmatory factor analysis approach called **structural equation modeling** (often simply referred to as SEM). Structural equation modeling differs statistically and practically from PCA and factor analysis. It is not only used to evaluate how well observed variables relate to factors but also to analyze hypothesized relationships between factors that the researcher specifies prior to the analysis based on theory and logic.

8.2 Understanding Principal Component and Factor Analysis

8.2.1 Why Use Principal Component and Factor Analysis?

Researchers often face the problem of large questionnaires comprising many *items*. For example, in a survey of a major German soccer club, the management was particularly interested in identifying and evaluating performance features that relate to soccer fans' satisfaction (Sarstedt et al. 2014). Examples of relevant features include the stadium, the team composition and their success, the trainer, and the

Table 8.1 Items in the soccer fan satisfaction study

Satisfaction with...	
Condition of the stadium	Public appearances of the players
Interior design of the stadium	Number of stars in the team
Outer appearance of the stadium	Interaction of players with fans
Signposting outside the stadium	Volume of the loudspeakers in the stadium
Signposting inside the stadium	Choice of music in the stadium
Roofing inside the stadium	Entertainment program in the stadium
Comfort of the seats	Stadium speaker
Video score boards in the stadium	Newsmagazine of the stadium
Condition of the restrooms	Price of annual season ticket
Tidiness within the stadium	Entry fees
Size of the stadium	Offers of reduced tickets
View onto the playing field	Design of the home jersey
Number of restrooms	Design of the away jersey
Sponsors' advertisements in the stadium	Assortment of merchandise
Location of the stadium	Quality of merchandise
Name of the stadium	Prices of merchandise
Determination and commitment of the players	Pre-sale of tickets
Current success regarding matches	Online-shop
Identification of the players with the club	Opening times of the fan-shops
Quality of the team composition	Accessibility of the fan-shops
Presence of a player with whom fans can identify	Behavior of the sales persons in the fan shops

management. The club therefore commissioned a questionnaire comprising 99 previously identified items by means of literature databases and focus groups of fans. All the items were measured on scales ranging from 1 ("very dissatisfied") to 7 ("very satisfied"). Table 8.1 shows an overview of some items considered in the study.

As you can imagine, tackling such a large set of items is problematic, because it provides quite complex data. Given the task of identifying and evaluating performance features that relate to soccer fans' satisfaction (measured by "Overall, how satisfied are you with your soccer club"), we cannot simply compare the items on a pairwise basis. It is far more reasonable to consider the factor structure first. For example, satisfaction with the condition of the stadium (x_1), outer appearance of the stadium (x_2), and interior design of the stadium (x_3) cover similar aspects that relate to the respondents' satisfaction with the stadium. If a soccer fan is generally very satisfied with the stadium, he/she will most likely answer all three items positively. Conversely, if a respondent is generally dissatisfied with the stadium, he/she is most likely to be rather dissatisfied with all the performance aspects of the stadium, such as the outer appearance and interior design. Consequently, these three items are likely to be highly correlated—they cover related aspects of the respondents' overall satisfaction with the stadium. More precisely, these items can be interpreted

as manifestations of the factor capturing the "joint meaning" of the items related to it. The arrows pointing from the factor to the items in Fig. 8.1 indicate this point. In our example, the "joint meaning" of the three items could be described as *satisfaction with the stadium*, since the items represent somewhat different, yet related, aspects of the stadium. Likewise, there is a second factor that relates to the two items x_4 and x_5, which, like the first factor, shares a common meaning, namely *satisfaction with the merchandise*.

PCA and factor analysis are two statistical procedures that draw on item correlations in order to find a small number of factors. Having conducted the analysis, we can make use of few (uncorrelated) factors instead of many variables, thus significantly reducing the analysis's complexity. For example, if we find six factors, we only need to consider six correlations between the factors and overall satisfaction, which means that the recommendations will rely on six factors.

8.2.2   Analysis Steps

Like any multivariate analysis method, PCA and factor analysis are subject to certain requirements, which need to be met for the analysis to be meaningful. A crucial requirement is that the variables need to exhibit a certain degree of correlation. In our example in Fig. 8.1, this is probably the case, as we expect increased correlations between x_1, x_2, and x_3, on the one hand, and between x_4 and x_5 on the other. Other items, such as x_1 and x_4, are probably somewhat correlated, but to a lesser degree than the group of items x_1, x_2, and x_3 and the pair x_4 and x_5. Several methods allow for testing whether the item correlations are sufficiently high.

Both PCA and factor analysis strive to reduce the overall item set to a smaller set of factors. More precisely, PCA extracts factors such that they account for variables' variance, whereas factor analysis attempts to explain the correlations between the variables. Whichever approach you apply, using only a few factors instead of many items reduces its precision, because the factors cannot represent all the information included in the items. Consequently, there is a trade-off between simplicity and accuracy. In order to make the analysis as simple as possible, we want to extract only a few factors. At the same time, we do not want to lose too much information by having too few factors. This trade-off has to be addressed in any PCA and factor analysis when deciding how many factors to extract from the data.

Once the number of factors to retain from the data has been identified, we can proceed with the interpretation of the factor solution. This step requires us to produce a label for each factor that best characterizes the joint meaning of all the variables associated with it. This step is often challenging, but there are ways of facilitating the interpretation of the factor solution. Finally, we have to assess how well the factors reproduce the data. This is done by examining the solution's goodness-of-fit, which completes the standard analysis. However, if we wish to continue using the results in further analyses, we need to calculate the factor scores.
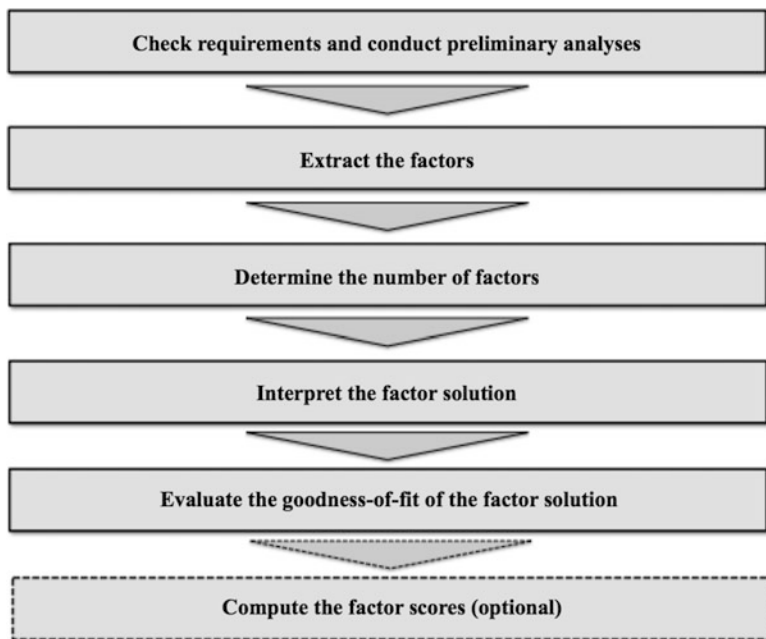
Fig. 8.2 Steps involved in a PCA

Factor scores are linear combinations of the items and can be used as variables in follow-up analyses.

Figure 8.2 illustrates the steps involved in the analysis; we will discuss these in more detail in the following sections. In doing so, our theoretical descriptions will focus on the PCA, as this method is easier to grasp. However, most of our descriptions also apply to factor analysis. Our illustration at the end of the chapter also follows a PCA approach but uses a Stata command (`factor, pcf`), which blends the PCA and factor analysis. This blending has several advantages, which we will discuss later in this chapter.

8.3   Principal Component Analysis

8.3.1   Check Requirements and Conduct Preliminary Analyses

Before carrying out a PCA, we have to consider several requirements, which we can test by answering the following questions:

– Are the measurement scales appropriate?
– Is the sample size sufficiently large?
– Are the observations independent?
– Are the variables sufficiently correlated?

Are the measurement scales appropriate?
For a PCA, it is best to have data measured on an interval or ratio scale. In practical applications, items measured on an ordinal scale level have become common. Ordinal scales can be used if:

– the scale points are equidistant, which means that the difference in the wording between scale steps is the same (see Chap. 3), *and*
– there are five or more response categories.

Is the sample size sufficiently large?
Another point of concern is the sample size. As a rule of thumb, the number of (valid) observations should be at least ten times the number of items used for analysis. This only provides a rough indication of the necessary sample size. Fortunately, researchers have conducted studies to determine minimum sample size requirements, which depend on other aspects of the study. MacCallum et al. (1999) suggest the following:

– When all communalities (we will discuss this term in Sect. 8.3.2.4) are above 0.60, small sample sizes of below 100 are adequate.
– With communalities around 0.50, sample sizes between 100 and 200 are sufficient.
– When communalities are consistently low, with many or all under 0.50, a sample size between 100 and 200 is adequate if the number of factors is small and each of these is measured with six or more indicators.
– When communalities are consistently low and the factors numbers are high or are measured with only few indicators (i.e., 3 or less), 300 observations are recommended.

Are the observations independent?
We have to ensure that the observations are independent. This means that the observations need to be completely unrelated (see Chap. 3). If we use dependent observations, we would introduce "artificial" correlations, which are not due to an underlying factor structure, but simply to the same respondents answered the same questions multiple times.

Are the variables sufficiently correlated?
As indicated before, PCA is based on correlations between items. Consequently, conducting a PCA only makes sense if the items correlate sufficiently. The problem is deciding what "sufficient" actually means.

An obvious step is to examine the correlation matrix (Chap. 5). Naturally, we want the correlations between different items to be as high as possible, but they will not always be. In our previous example, we expect high correlations between x_1, x_2, and x_3, on the one hand, and x_4 and x_5 on the other. Conversely, we might

expect lower correlations between, for example, x_1 and x_4 and between x_3 and x_5. Thus, not all of the correlation matrix's elements need to have high values. The PCA depends on the *relative* size of the correlations. Therefore, if single correlations are very low, this is not necessarily problematic! Only when all the correlations are around zero is PCA no longer useful. In addition, the statistical significance of each correlation coefficient helps decide whether it differs significantly from zero.

There are additional measures to determine whether the items correlate sufficiently. One is the **anti-image**. The anti-image describes the portion of an item's variance that is independent of another item in the analysis. Obviously, we want all items to be highly correlated, so that the anti-images of an item set are as small as possible. Initially, we do not interpret the anti-image values directly, but use a measure based on the anti-image concept: **The Kaiser–Meyer–Olkin** (**KMO**) statistic. The KMO statistic, also called the **measure of sampling adequacy** (**MSA**), indicates whether the other variables in the dataset can explain the correlations between variables. Kaiser (1974), who introduced the statistic, recommends a set of distinctively labeled threshold values for KMO and MSA, which Table 8.2 presents.

To summarize, the correlation matrix with the associated significance levels provides a first insight into the correlation structures. However, the final decision of whether the data are appropriate for PCA should be primarily based on the KMO statistic. If this measure indicates sufficiently correlated variables, we can continue the analysis of the results. If not, we should try to identify items that correlate only weakly with the remaining items and remove them. In Box 8.1, we discuss how to do this.

Table 8.2 Threshold values for KMO and MSA

KMO/MSA value	Adequacy of the correlations
Below 0.50	Unacceptable
0.50–0.59	Miserable
0.60–0.69	Mediocre
0.70–0.79	Middling
0.80–0.89	Meritorious
0.90 and higher	Marvelous

Box 8.1 Identifying Problematic Items
Examining the correlation matrix and the significance levels of correlations allows identifying items that correlate only weakly with the remaining items. An even better approach is examining the variable-specific MSA values, which are interpreted like the overall KMO statistic (see Table 8.2). In fact, the KMO statistic is simply the overall mean of all variable-specific MSA values. Consequently, all the MSA values should also lie above the threshold

> **Box 8.1** (continued)
> level of 0.50. If this is not the case, consider removing this item from the analysis. An item's communality or uniqueness (see next section) can also serve as a useful indicators of how well the factors extracted represent an item. However, communalities and uniqueness are mostly considered when evaluating the solution's goodness-of-fit.

8.3.2   Extract the Factors

8.3.2.1 Principal Component Analysis vs. Factor Analysis

Factor analysis assumes that each variable's variance can be divided into common variance (i.e., variance shared with all the other variables in the analysis) and unique variance (Fig. 8.3), the latter of which can be further broken down into specific variance (i.e., variance associated with only one specific variable) and error variance (i.e., variance due to measurement error). The method, however, can only reproduce common variance. Thereby factor analysis explicitly recognizes the presence of error. Conversely, PCA assumes that all of each variable's variance is common variance, which factor extraction can explain fully (e.g., Preacher and MacCallum 2003). These differences entail different interpretations of the analysis's outcomes. PCA asks:

Which umbrella term can we use to summarize a set of variables that loads highly on a specific factor?

Conversely, factor analysis asks:

What is the common reason for the strong correlations between a set of variables?

From a theoretical perspective, the assumption that there is a unique variance for which the factors cannot fully account, is generally more realistic, but simultaneously more restrictive. Although theoretically sound, this restriction can sometimes lead to complications in the analysis, which have contributed to the widespread use of PCA, especially in market research practice.

Researchers usually suggest using PCA when data reduction is the primary concern; that is, when the focus is to extract a minimum number of factors that account for a maximum proportion of the variables' total variance. In contrast, if the primary concern is to identify latent dimensions represented in the variables, factor analysis should be applied. However, prior research has shown that both approaches arrive at essentially the same result when:

– more than 30 variables are used, *or*
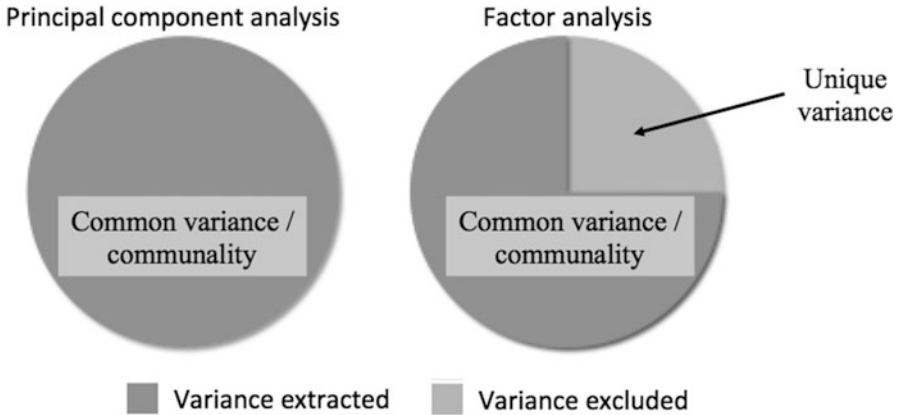– most of the variables' communalities exceed 0.60.

Fig. 8.3 Principal component analysis vs. factor analysis

With 20 or fewer variables and communalities below 0.40—which are clearly undesirable in empirical research—the differences are probably pronounced (Stevens 2009).

Apart from these conceptual differences in the variables' nature, PCA and factor analysis differ in the aim of their analysis. Whereas the goal of factor analysis is to explain the correlations between the variables, PCA focuses on explaining the variables' variances. That is, the PCA's objective is to determine the linear combinations of the variables that retain as much information from the original variables as possible. Strictly speaking, PCA does not extract factors, but **components**, which are labeled as such in Stata.

Despite these differences, which have very little relevance in many common research settings in practice, PCA and factor analysis have many points in common. For example, the methods follow very similar ways to arrive at a solution and their interpretations of statistical measures, such as KMO, eigenvalues, or factor loadings, are (almost) identical. In fact, Stata blends these two procedures in its `factor, pcf` command, which runs a factor analysis but rescales the estimates such that they conform to a PCA. That way, the analysis assumes that the entire variance is common but produces (rotated) loadings (we will discuss factor rotation later in this chapter), which facilitate the interpretation of the factors. In contrast, if we would run a standard PCA, Stata would only offer us eigenvectors whose (unrotated) weights would not allow for a concluding interpretation of the factors. In fact, in many PCA analyses, researchers are not interested in the interpretation of the extracted factors but merely use the method for data reduction. For example, in sensory marketing research, researchers routinely use PCA to summarize a large set of *sensory variables* (e.g., haptic, smell, taste) to derive a set of factors whose scores are then used as input for cluster analyses (Chap. 9). This approach allows for identifying distinct groups of products from which one or more representative products can then be chosen for a more detailed comparison using qualitative

research or further assessment in field experiments (e.g., Carbonell et al. 2008; Vigneau and Qannari 2002).

Despite the small differences of PCA and factor analysis in most research settings, researchers have strong feelings about the choice of PCA or factor analysis. Cliff (1987, p. 349) summarizes this issue well, by noting that proponents of factor analysis "insist that components analysis is at best a common factor analysis with some error added and at worst an unrecognizable hodgepodge of things from which nothing can be determined." For further discussions on this topic, see also Velicer and Jackson (1990) and Widaman (1993).[2]

8.3.2.2 How Does Factor Extraction Work?

PCA's objective is to reproduce a data structure with only a few factors. PCA does this by generating a new set of factors as linear composites of the original variables, which reproduces the original variables' variance as best as possible. These linear composites are called **principal components**, but, for simplicity's sake, we refer to them as factors. More precisely, PCA computes **eigenvectors**. These eigenvectors include so called **factor weights**, which extract the maximum possible variance of all the variables, with successive factoring continuing until a significant share of the variance is explained.

Operationally, the first factor is extracted in such a way that it maximizes the variance accounted for in the variables. We can visualize this easily by examining the vector space illustrated in Fig. 8.4. In this example, we have five variables (x_1–x_5) represented by five vectors starting at the zero point, with each vector's length standardized to one. To maximize the variance accounted for, the first factor F_1 is fitted into this vector space in such a way that the sum of all the angles between this factor and the five variables in the vector space is minimized. We do this to interpret the angle between two vectors as correlations. For example, if the factor's vector and a variable's vector are congruent, the angle between these two is zero,

Fig. 8.4 Factor extraction (Note that Fig. 8.4 describes a special case, as the five variables are scaled down into a two-dimensional space. In this set-up, it would be possible for the two factors to explain all five items. However, in real-life, the five items span a five-dimensional vector space.)

---
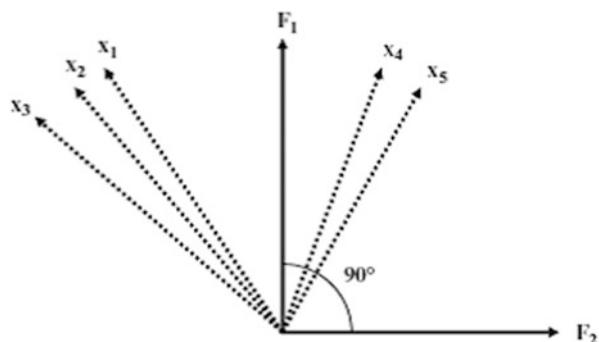
[2]Related discussions have been raised in structural equation modeling, where researchers have heatedly discussed the strengths and limitations of factor-based and component-based approaches (e.g., Sarstedt et al. 2016, Hair et al. 2017a, b).

indicating that the factor and the variable correlate perfectly. On the other hand, if the factor and the variable are uncorrelated, the angle between these two is 90°. This correlation between a (unit-scaled) factor and a variable is called the **factor loading**. Note that factor weights and factor loadings essentially express the same thing—the relationships between variables and factors—but they are based on different scales.

After extracting F_1, a second factor (F_2) is extracted, which maximizes the remaining variance accounted for. The second factor is fitted at a 90° angle into the vector space (Fig. 8.4) and is therefore uncorrelated with the first factor.[3] If we extract a third factor, it will explain the maximum amount of variance for which factors 1 and 2 have hitherto not accounted. This factor will also be fitted at a 90° angle to the first two factors, making it independent from the first two factors (we don't illustrate this third factor in Fig. 8.4, as this is a three-dimensional space). The fact that the factors are uncorrelated is an important feature, as we can use them to replace many highly correlated variables in follow-up analyses. For example, using uncorrelated factors as independent variables in a regression analysis helps solve potential collinearity issues (Chap. 7).

> The Explained Visually webpage offers an excellent illustration of two- and three-dimensional factor extraction, see http://setosa.io/ev/principal-component-analysis/

An important PCA feature is that it works with standardized variables (see Chap. 5 for an explanation of what standardized variables are). Standardizing variables has important implications for our analysis in two respects. First, we can assess each factor's eigenvalue, which indicates how much a specific factor extracts all of the variables' variance (see next section). Second, the standardization of variables allows for assessing each variable's communality, which describes how much the factors extracted capture or reproduce each variable's variance. A related concept is the uniqueness, which is 1–communality (see Sect. 8.3.2.4).

8.3.2.3 What Are Eigenvalues?

To understand the concept of **eigenvalues**, think of the soccer fan satisfaction study (Fig. 8.1). In this example, there are five variables. As all the variables are standardized prior to the analysis, each has a variance of 1. In a simplified way, we could say that the overall information (i.e., variance) that we want to reproduce by means of factor extraction is 5 units. Let's assume that we extract the two factors presented above.

The first factor's eigenvalue indicates how much variance of the total variance (i.e., 5 units) this factor accounts for. If this factor has an eigenvalue of, let's say

---

[3]Note that this changes when oblique rotation is used. We will discuss factor rotation later in this chapter.
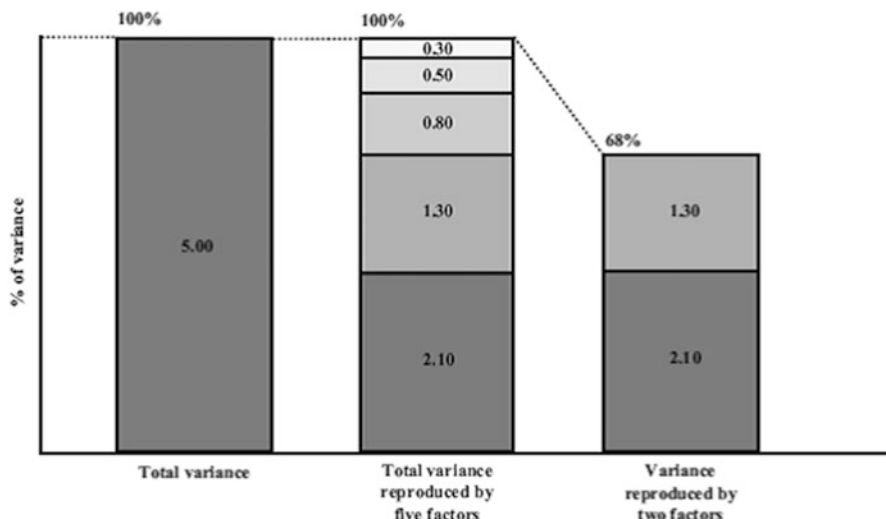
Fig. 8.5 Total variance explained by variables and factors

2.10, it covers the information of 2.10 variables or, put differently, accounts for 2.10/5.00 = 42% of the overall variance (Fig. 8.5).

Extracting a second factor will allow us to explain another part of the remaining variance (i.e., 5.00 − 2.10 = 2.90 units, Fig. 8.5). However, the eigenvalue of the second factor will always be smaller than that of the first factor. Assume that the second factor has an eigenvalue of 1.30 units. The second factor then accounts for 1.30/5.00 = 26% of the overall variance. Together, these two factors explain (2.10 + 1.30)/5.00 = 68% of the overall variance. Every additional factor extracted increases the variance accounted for until we have extracted as many factors as there are variables. In this case, the factors account for 100% of the overall variance, which means that the factors reproduce the complete variance.

Following the PCA approach, we assume that factor extraction can reproduce each variable's entire variance. In other words, we assume that each variable's variance is common; that is, the variance is shared with other variables. This differs in factor analysis, in which each variable can also have a unique variance.

8.3.2.4 What Are Communality and Uniqueness?

Whereas the eigenvalue tells us how much variance each factor is accounts for, the **communality** indicates how much variance of each variable, factor extraction can reproduce. There is no commonly agreed threshold for a variable's communality, as this depends strongly on the complexity of the analysis at hand. However, generally, the extracted factors should account for at least 50% of a variable's variance. Thus, the communalities should be above 0.50. Note that Stata does not indicate each variable's communality but its **uniqueness**, which is 1–communality. Hence, uniqueness gives the proportion of a variable's variance that the factors do *not* capture. For uniqueness the same threshold as for communality applies. Thus, the

uniqueness values should be below 0.50. Every additional factor extracted will increase the explained variance, and if we extract as many factors as there are items (in our example five), each variable's communality would be 1.00 and its uniqueness equal to 0. The factors extracted would then fully explain each variable; that is, the first factor will explain a certain amount of each variable's variance, the second factor another part, and so on.

However, since our overall objective is to reduce the number of variables through factor extraction, we should extract only a few factors that account for a high degree of overall variance. This raises the question of how to decide on the number of factors to extract from the data, which we discuss in the following section.

8.3.3   Determine the Number of Factors

Determining the number of factors to extract from the data is a crucial and challenging step in any PCA. Several approaches offer guidance in this respect, but most researchers do not pick just one method, but determine the number of factors resulting from the application of multiple methods. If multiple methods suggest the same number of factors, this leads to greater confidence in the results.

8.3.3.1 The Kaiser Criterion
An intuitive way to decide on the number of factors is to extract all the factors with an eigenvalue greater than 1. The reason for this is that each factor with an eigenvalue greater than 1 accounts for more variance than a single variable (remember, we are looking at standardized variables, which is why each variable's variance is exactly 1). As the objective of PCA is to reduce the overall number of variables, each factor should of course account for more variance than a single variable can. If this occurs, then this factor is useful for reducing the set of variables. Extracting all the factors with an eigenvalue greater than 1 is frequently called the **Kaiser criterion** or **latent root criterion** and is commonly used to determine the number of factors. However, the Kaiser criterion is well known for overspecifying the number of factors; that is, the criterion suggests more factors than it should (e.g., Russell 2002; Zwick and Velicer 1986).

8.3.3.2 The Scree Plot
Another popular way to decide on the number of factors to extract is to plot each factor's eigenvalue (*y*-axis) against the factor with which it is associated (*x*-axis). This results in a **scree plot**, which typically has a distinct break in it, thereby showing the "correct" number of factors (Cattell 1966). This distinct break is called the "elbow." It is generally recommended that all factors should be retained *above* this break, as they contribute most to the explanation of the variance in the dataset. Thus, we select one factor less than indicated by the elbow. In Box 8.2, we introduce a variant of the scree plot. This variant is however only available when using the `pca` command instead of the `factor, pcf` command, which serves as the basis for our case study illustration.

> **Box 8.2 Confidence Intervals in the Scree Plot**
> A variant of the scree plot includes the confidence interval (Chap. 6, Sect. 6.6.7) of each eigenvalue. These confidence intervals allow for identifying factors with an eigenvalue significantly greater than 1. If a confidence interval's lower bound is above the 1 threshold, this suggests that the factor should be extracted. Conversely, if an eigenvalue's confidence interval overlaps with the 1 threshold or falls completely below, this factor should not be extracted.

8.3.3.3 Parallel Analysis

A large body of review papers and simulation studies has produced a prescriptive consensus that Horn's (1965) **parallel analysis** is the best method for deciding how many factors to extract (e.g., Dinno 2009; Hayton et al. 2004; Henson and Roberts 2006; Zwick and Velicer 1986). The rationale underlying parallel analysis is that factors from real data with a valid underlying factor structure should have larger eigenvalues than those derived from randomly generated data (actually pseudorandom deviates) with the same sample size and number of variables.

Parallel analysis involves several steps. First, a large number of datasets are randomly generated; they have the same number of observations and variables as the original dataset. Parallel PCAs are then run on each of the datasets (hence, parallel analysis), resulting in many slightly different sets of eigenvalues. Using these results as input, parallel analysis derives two relevant cues.

First, parallel analysis adjusts the original eigenvalues for sampling error-induced collinearity among the variables to arrive at adjusted eigenvalues (Horn 1965). Analogous to the Kaiser criterion, only factors with adjusted eigenvalues larger than 1 should be retained.

Second, we can compare the randomly generated eigenvalues with those from the original analysis. Only factors whose original eigenvalues are larger than the 95th percentile of the randomly generated eigenvalues should be retained (Longman et al. 1989).

8.3.3.4 Expectations

When, for example, replicating a previous market research study, we might have a priori information on the number of factors we want to find. For example, if a previous study suggests that a certain item set comprises five factors, we should extract the same number of factors, even if statistical criteria, such as the scree plot, suggest a different number. Similarly, theory might suggest that a certain number of factors should be extracted from the data.

Strictly speaking, these are confirmatory approaches to factor analysis, which blur the distinction between these two factor analysis types. Ultimately however, we should not fully rely on the data, but keep in mind that the research results should be interpretable and actionable for market research practice.

When using factor analysis, Stata allows for estimating two further criteria called the **Akaike Information Criterion** (**AIC**) and the **Bayes Information Criterion** (**BIC**). These criteria are relative measures of goodness-of-fit and are used to compare the adequacy of solutions with different numbers of factors. "Relative" means that these criteria are not scaled on a range of, for example, 0 to 1, but can generally take any value. Compared to an alternative solution with a different number of factors, smaller AIC or BIC values indicate a better fit. Stata computes solutions for different numbers of factors (up to the maximum number of factors specified before). We therefore need to choose the appropriate solution by looking for the smallest value in each criterion. When using these criteria, you should note that AIC is well known for overestimating the "correct" number of factors, while BIC has a slight tendency to underestimate this number.

Whatever combination of approaches we use to determine the number of factors, the factors extracted should account for at least 50% of the total variance explained (75% or more is recommended). Once we have decided on the number of factors to retain from the data, we can start interpreting the factor solution.

8.3.4   Interpret the Factor Solution

8.3.4.1 Rotate the Factors

To interpret the solution, we have to determine which variables relate to each of the factors extracted. We do this by examining the *factor loadings*, which represent the correlations between the factors and the variables and can take values ranging from −1 to +1. A high factor loading indicates that a certain factor represents a variable well. Subsequently, we look for high *absolute* values, because the correlation between a variable and a factor can also be negative. Using the highest absolute factor loadings, we "assign" each variable to a certain factor and then produce a label for each factor that best characterizes the joint meaning of all the variables associated with it. This labeling is subjective, but a key PCA step. An example of a label is the respondents' satisfaction with the stadium, which represents the items referring to its condition, outer appearance, and interior design (Fig. 8.1).

We can make use of **factor rotation** to facilitate the factors' interpretation.[4] We do not have to rotate the factor solution, but it will facilitate interpreting the findings, particularly if we have a reasonably large number of items (say six or more). To understand what factor rotation is all about, once again consider the factor structure described in Fig. 8.4. Here, we see that both factors relate to the

---

[4]Note that factor rotation primarily applies to factor analysis rather than PCA—see Preacher and MacCallum (2003) for details. However, our illustration draws on the `factor, pcf` command, which uses the factor analysis algorithm to compute PCA results for which rotation applies.
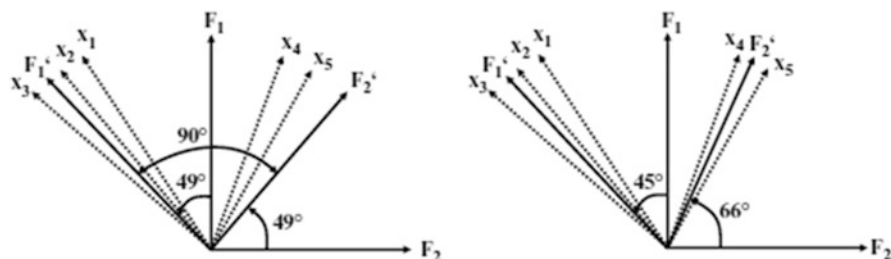
Fig. 8.6 Orthogonal and oblique factor rotation

variables in the set. However, the first factor appears to generally correlate more strongly with the variables, whereas the second factor only correlates weakly with the variables (to clarify, we look for small angles between the factors and variables). This implies that we "assign" all variables to the first factor without taking the second into consideration. This does not appear to be very meaningful, as we want both factors to represent certain facets of the variable set. Factor rotation can resolve this problem. By rotating the factor axes, we can create a situation in which a set of variables loads highly on only one specific factor, whereas another set loads highly on another. Figure 8.6 illustrates the factor rotation graphically.

On the left side of the figure, we see that both factors are orthogonally rotated 49°, meaning that a 90° angle is maintained between the factors during the rotation procedure. Consequently, the factors remain uncorrelated, which is in line with the PCA's initial objective. By rotating the first factor from F_1 to F_1', it is now strongly related to variables x_1, x_2, and x_3, but weakly related to x_4 and x_5. Conversely, by rotating the second factor from F_2 to F_2' it is now strongly related to x_4 and x_5, but weakly related to the remaining variables. The assignment of the variables is now much clearer, which facilitates the interpretation of the factors significantly.

Various **orthogonal rotation** methods exist, all of which differ with regard to their treatment of the loading structure. The **varimax rotation** (the default option for orthogonal rotation in Stata) is the best-known one; this procedure aims at maximizing the dispersion of loadings within factors, which means a few variables will have high loadings, while the remaining variables' loadings will be considerably smaller (Kaiser 1958).

Alternatively, we can choose between several **oblique rotation** techniques. In oblique rotation, the 90° angle between the factors is not maintained during rotation, and the resulting factors are therefore correlated. Figure 8.6 (right side) illustrates an example of an oblique factor rotation. **Promax rotation** (the default option for oblique rotation in Stata) is a commonly used oblique rotation technique. The Promax rotation allows for setting an exponent (referred to as *Promax power* in Stata) that needs to be greater than 1. Higher values make the loadings even more extreme (i.e., high loadings are amplified and weak loadings are reduced even further), which is at the cost of stronger correlations between the factors and less total variance explained (Hamilton 2013). The default value of 3 works well for most applications. **Oblimin rotation** is a popular alternative oblique rotation type.

Oblimin is a class of rotation procedures whereby the degree of obliqueness can be specified. This degree is the *gamma*, which determines the level of the correlation allowed between the factors. A gamma of zero (the default) ensures that the factors are—if at all—only moderately correlated, which is acceptable for most analyses.[5]

Oblique rotation is used when factors are possibly related. It is, for example, very likely that the respondents' satisfaction with the stadium is related to their satisfaction with other aspects of the soccer club, such as the number of stars in the team or the quality of the merchandise. However, relinquishing the initial objective of extracting uncorrelated factors can diminish the factors' interpretability. We therefore recommend using the varimax rotation to enhance the interpretability of the results. Only if the results are difficult to interpret, should an oblique rotation be applied. Among the oblique rotation methods, researchers generally recommend the promax (Gorsuch 1983) or oblimin (Kim and Mueller 1978) methods but differences between the rotation types are typically marginal (Brown 2009).

8.3.4.2 Assign the Variables to the Factors

After rotating the factors, we need to interpret them and give each factor a name, which has some descriptive value. Interpreting factors involves assigning each variable to a specific factor based on the highest *absolute* (!) loading. For example, if a variable has a 0.60 loading with the first factor and a 0.20 loading with the second, we would assign this variable to the first factor. Loadings may nevertheless be very similar (e.g., 0.50 for the first factor and 0.55 for the second one), making the assignment ambiguous. In such a situation, we could assign the variable to another factor, even though it does not have the highest loading on this specific factor. While this step can help increase the results' face validity (see Chap. 3), we should make sure that the variable's factor loading with the designated factor is above an acceptable level. If very few factors have been extracted, the loading should be at least 0.50, but with a high number of factors, lower loadings of above 0.30 are acceptable. Alternatively, some simply ignore a certain variable if it does not fit with the factor structure. In such a situation, we should re-run the analysis without variables that do not load highly on a specific factor. In the end, the results should be interpretable and actionable, but keep in mind that this technique is, first and foremost, exploratory!

8.3.5   Evaluate the Goodness-of-Fit of the Factor Solution

8.3.5.1 Check the Congruence of the Initial and Reproduced
Correlations

While PCA focuses on explaining the variables' variances, checking how well the method approximates the correlation matrix allows for assessing the quality of the

---

[5]When the gamma is set to 1, this is a special case, because the value of 1 represents orthogonality. The result of setting gamma to 1 is effectively a varimax rotation.

solution (i.e., the goodness-of-fit) (Graffelman 2013). More precisely, to assess the solution's goodness-of-fit, we can make use of the differences between the correlations in the data and those that the factors imply. These differences are also called **correlation residuals** and should be as small as possible.

In practice, we check the proportion of correlation residuals with an absolute value higher than 0.05. Even though there is no strict rule of thumb regarding the maximum proportion, a proportion of more than 50% should raise concern. However, high residuals usually go hand in hand with an unsatisfactory KMO measure; consequently, this problem already surfaces when testing the assumptions.

8.3.5.2 Check How Much of Each Variable's Variance Is Reproduced by Means of Factor Extraction

Another way to check the solution's goodness-of-fit is by evaluating how much of each variable's variance the factors reproduce (i.e., the communality). If several communalities exhibit low values, we should consider removing these variables. Considering the variable-specific MSA measures could help us make this decision. If there are more variables in the dataset, communalities usually become smaller; however, if the factor solution accounts for less than 50% of a variable's variance (i.e., the variable's communality is less than 0.50), it is worthwhile reconsidering the set-up. Since Stata does not provide communality but uniqueness values, we have to make this decision in terms of the variance that the factors do *not* reproduce. That is, if several variables exhibit uniqueness values larger than 0.50, we should reconsider the analysis.

8.3.6   Compute the Factor Scores

After the rotation and interpretation of the factors, we can compute the **factor scores**, another element of the analysis. Factor scores are linear combinations of the items and can be used as separate variables in subsequent analyses. For example, instead of using many highly correlated independent variables in a regression analysis, we can use few uncorrelated factors to overcome collinearity problems.

The simplest ways to compute factor scores for each observation is to sum all the scores of the items assigned to a factor. While easy to compute, this approach neglects the potential differences in each item's contribution to each factor (Sarstedt et al. 2016).

Drawing on the eigenvectors that the PCA produces, which include the factor weights, is a more elaborate way of computing factor scores (Hershberger 2005). These weights indicate each item's relative contribution to forming the factor; we simply multiply the standardized variables' values with the weights to get the factor scores. Factor scores computed on the basis of eigenvectors have a zero mean. This means that if a respondent has a value greater than zero for a certain factor, he/she scores above the above average in terms of the characteristic that this factor describes. Conversely, if a factor score is below zero, then this respondent exhibits the characteristic below average.

Different from the PCA, a factor analysis does not produce determinate factor scores. In other words, the factor is indeterminate, which means that part of it remains an arbitrary quantity, capable of taking on an infinite range of values (e.g., Grice 2001; Steiger 1979). Thus, we have to rely on other approaches to computing factor scores such as the **regression method**, which features prominently among factor analysis users. This method takes into account (1) the correlation between the factors and variables (via the item loadings), (2) the correlation between the variables, and (3) the correlation between the factors if oblique rotation has been used (DiStefano et al. 2009). The regression method z-standardizes each factor to zero mean and unit standard deviation.[6] We can therefore interpret an observation's score in relation to the mean and in terms of the units of standard deviation from this mean. For example, an observation's factor score of 0.79 implies that this observation is 0.79 standard deviations above the average with regard to the corresponding factor.

Another popular approach is the **Bartlett method**, which is similar to the regression method. The method produces factor scores with zero mean and standard deviations larger than one. Owing to the way they are estimated, the factor scores that the Bartlett method produces are considered are considered more accurate (Hershberger 2005). However, in practical applications, both methods yield highly similar results. Because of the z-standardization of the scores, which facilitates the comparison of scores across factors, we recommend using the regression method.

In Table 8.3 we summarize the main steps that need to be taken when conducting a PCA in Stata. Our descriptions draw on Stata's `factor, pcf` command, which carries out a factor analysis but rescales the resulting factors such that the results conform to a standard PCA. This approach has the advantage that it follows the fundamentals of PCA, while allowing for analyses that are restricted to factor analysis (e.g., factor rotation, use of AIC and BIC).

8.4   Confirmatory Factor Analysis and Reliability Analysis

Many researchers and practitioners acknowledge the prominent role that exploratory factor analysis plays in exploring data structures. Data can be analyzed without preconceived ideas of the number of factors or how these relate to the variables under consideration. Whereas this approach is, as its name implies, exploratory in nature, the *confirmatory factor analysis* allows for testing hypothesized structures underlying a set of variables.

In a confirmatory factor analysis, the researcher needs to first specify the constructs and their associations with variables, which should be based on previous measurements or theoretical considerations.

---

[6]Note that this is not the case when using factor analysis if the standard deviations are different from one (DiStefano et al. 2009).

Table 8.3 Steps involved in carrying out a PCA in Stata

Theory	Stata
Check Assumptions and Carry Out Preliminary Analyses	
Select variables that should be reduced to a set of underlying factors (PCA) or should be used to identify underlying dimensions (factor analysis)	► Statistics ► Multivariate analysis ► Factor and principal component analysis ► Factor analysis. Enter the variables in the **Variables** box.
Are the variables interval or ratio scaled?	Determine the measurement level of your variables (see Chap. 3). If ordinal variables are used, make sure that the scale steps are equidistant.
Is the sample size sufficiently large?	Check MacCallum et al.'s (1999) guidelines for minimum sample size requirements, dependent on the variables' communality. For example, if all the communalities are above 0.60, small sample sizes of below 100 are adequate. With communalities around 0.50, sample sizes between 100 and 200 are sufficient.
Are the observations independent?	Determine whether the observations are dependent or independent (see Chap. 3).
Are the variables sufficiently correlated?	Check whether at least some of the variable correlations are significant. ► Statistics ► Summaries, tables, and tests ► Summary and descriptive statistics ► Pairwise correlations. Check **Print number of observations for each entry** and **Print significance level for each entry**. Select **Use Bonferroni-adjusted significance level** to maintain the familywise error rate (see Chap. 6).
	`pwcorr s1 s2 s3 s4 s5 s6 s7 s8, obs sig bonferroni`
	Is the KMO ≥ 0.50? ► Statistics ► Postestimation ► Factor analysis reports and graphs ► Kaiser-Meyer-Olkin measure of sample adequacy. Then click on **Launch** and **OK**. Note that this analysis can only be run after the PCA has been conducted.
	`estat kmo`
Extract the factors	
Choose the method of factor analysis	► Statistics ► Multivariate analysis ► Factor and principal component analysis ► Factor analysis. Click on the **Model 2** tab and select **Principal component factor**.
	`factor s1 s2 s3 s4 s5 s6 s7 s8, pcf`
Determine the number of factors	
Determine the number of factors	Kaiser criterion: ► Statistics ► Multivariate analysis ► Factor and principal component analysis ► Factor analysis. Click on the **Model 2** tab and enter **1** under **Minimum value of eigenvalues to be retained**.

Table 8.3 (continued)

Theory	Stata
	```factor s1 s2 s3 s4 s5 s6 s7 s8 pcf mineigen (1)```
	Parallel analysis: Download and install *paran* (help paran) and enter ```paran s1 s2 s3 s4 s5 s6 s7 s8, centile (95) q all graph```
	Extract factors (1) with adjusted eigenvalues greater than 1, and (2) whose adjusted eigenvalues are greater than the random eigenvalues.
	Scree plot: ► Statistics ► Postestimation ► Factor analysis reports and graphs ► Scree plot of eigenvalues. Then click on **Launch** and **OK**.
	```screeplot```
	Pre-specify the number of factors based on a priori information: ► Statistics ► Multivariate analysis ► Factor and principal component analysis ► Factor analysis. Under the **Model 2** tab, tick **Maximum number of factors to be retained** and specify a value in the box below (e.g., **2**).
	```factor s1 s2 s3 s4 s5 s6 s7 s8, factors(2)```
	Make sure that the factors extracted account for at least 50% of the total variance explained (75% or more is recommended): Check the Cumulative column in the PCA output.
Interpret the Factor Solution	
Rotate the factors	Use the varimax procedure or, if necessary, the promax procedure with gamma set to 3 (both with Kaiser normalization): ► Statistics ► Postestimation ► Principal component analysis reports and graphs ► Rotate factor loadings. Select the corresponding option in the menu.
	Varimax: ```rotate, kaiser```
	Promax: ```rotate, promax(3) oblique Kaiser```
Assign variables to factors	Check the Factor loadings (pattern matrix) table in the output of the rotated solution. Assign each variable to a certain factor based on the highest absolute loading. To facilitate interpretation, you may also assign a variable to a different factor, but check that the loading is at an acceptable level (0.50 if only a few factors are extracted, 0.30 if many factors are extracted).
	Consider making a loadings plot: ► Statistics ► Postestimation ► Factor analysis reports

(continued)

Table 8.3 (continued)

Theory	Stata
	and graphs ▶ Plot of factor loadings. Under **Plot all combinations of the following**, indicate the number of factors for which you want to plot. Check which items load highly on which factor.
Compute factor scores	Save factor scores as new variables: ▶ Statistics ▶ Postestimation ▶ Predictions ▶ Regression and Bartlett scores. Under **New variable names or variable stub*** enter *factor**. Select **Factors scored by the regression scoring method**
	`predict score*, regression`
Evaluate the Goodness-of-fit of the Factor Solution	
Check the congruence of the initial and reproduced correlations	Create a reproduced correlation matrix: ▶ Statistics ▶ Postestimation ▶ Factor analysis reports and graphs ▶ Matrix of correlation residuals. Is the proportion of residuals greater than $0.05 \leq 50\%$?
	`estat residuals`
Check how much of each variable's variance is reproduced by means of factor extraction	Check the **Uniqueness** column in the PCA output. Are all the values lower than 0.50?

Instead of allowing the procedure to determine the number of factors, as is done in an exploratory factor analysis, a confirmatory factor analysis tells us how well the actual data fit the pre-specified structure. Reverting to our introductory example, we could, for example, assume that the construct *satisfaction with the stadium* can be measured by means of the three items x_1 (condition of the stadium), x_2 (appearance of the stadium), and x_3 (interior design of the stadium). Likewise, we could hypothesize that *satisfaction with the merchandise* can be adequately measured using the items x_4 and x_5. In a confirmatory factor analysis, we set up a theoretical model linking the items with the respective construct (note that in confirmatory factor analysis, researchers generally use the term construct rather than factor). This process is also called operationalization (see Chap. 3) and usually involves drawing a visual representation (called a **path diagram**) indicating the expected relationships.

Figure 8.7 shows a path diagram—you will notice the similarity to the diagram in Fig. 8.1. Circles or ovals represent the constructs (e.g., Y_1, satisfaction with the stadium) and boxes represent the items (x_1 to x_5). Other elements include the relationships between the constructs and respective items (the loadings l_1 to l_5), the error terms (e_1 to e_5) that capture the extent to which a construct does not explain a specific item, and the correlations between the constructs of interest (r_{12}).

Having defined the individual constructs and developed the path diagram, we can estimate the model. The relationships between the constructs and items (the loadings l_1 to l_5) and the item correlations (not shown in Fig. 8.7) are of particular
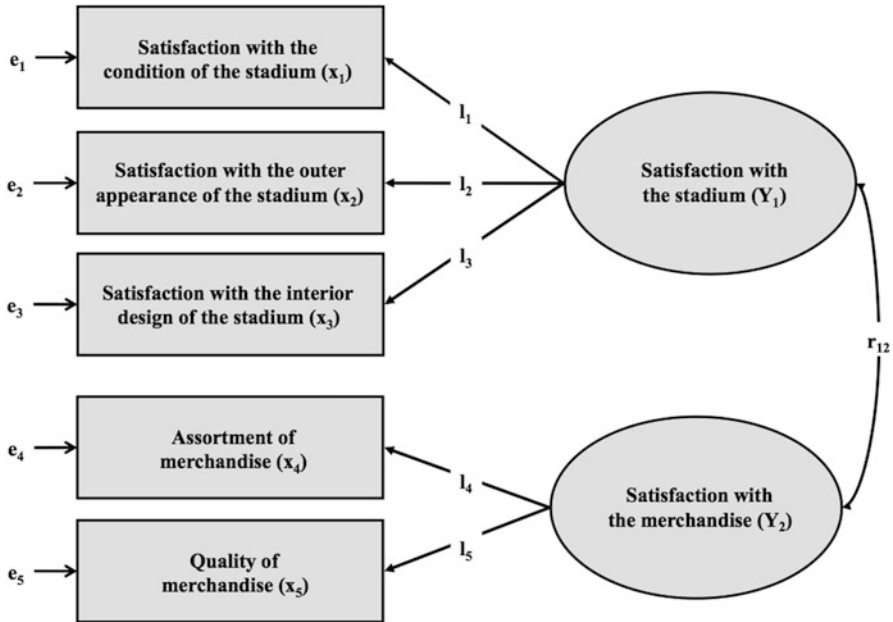
Fig. 8.7 Path diagram (confirmatory factor analysis)

interest, as they indicate whether the construct has been reliably and validly measured.

Reliability analysis is an important element of a confirmatory factor analysis and essential when working with measurement scales. The preferred way to evaluate reliability is by taking two independent measurements (using the same subjects) and comparing these by means of correlations. This is also called **test-retest reliability** (see Chap. 3). However, practicalities often prevent researchers from surveying their subjects a second time.

An alternative is to estimate the **split-half reliability**. In the split-half reliability, scale items are divided into halves and the scores of the halves are correlated to obtain an estimate of reliability. Since all items should be consistent regarding what they indicate about the construct, the halves can be considered approximations of alternative forms of the same scale. Consequently, instead of looking at the scale's test-retest reliability, researchers consider the scale's equivalence, thus showing the extent to which two measures of the same general trait agree. We call this type of reliability the **internal consistency reliability**.

In the example of *satisfaction with the stadium*, we compute this scale's split-half reliability manually by, for example, splitting up the scale into x_1 on the one side and x_2 and x_3 on the other. We then compute the sum of x_2 and x_3 (or calculate the items' average) to form a total score and correlate this score with x_1. A high correlation indicates that the two subsets of items measure related aspects of the same underlying construct and, thus, suggests a high degree of internal consistency.

However, with many indicators, there are many different ways to split the variables into two groups.

Cronbach (1951) proposed calculating the average of all possible split-half coefficients resulting from different ways of splitting the scale items. The **Cronbach's Alpha** coefficient has become by far the most popular measure of internal consistency. In the example above, this would comprise calculating the average of the correlations between (1) x_1 and $x_2 + x_3$, (2) x_2 and $x_1 + x_3$, as well as (3) x_3 and $x_1 + x_2$. The Cronbach's Alpha coefficient generally varies from 0 to 1, whereas a generally agreed lower limit for the coefficient is 0.70. However, in exploratory studies, a value of 0.60 is acceptable, while values of 0.80 or higher are regarded as satisfactory in the more advanced stages of research (Hair et al. 2011). In Box 8.3, we provide more advice on the use of Cronbach's Alpha. We illustrate a reliability analysis using the standard Stata module in the example at the end of this chapter.

Box 8.3 Things to Consider When Calculating Cronbach's Alpha
When calculating Cronbach's Alpha, ensure that all items are formulated in the same direction (positively or negatively worded). For example, in psychological measurement, it is common to use both negatively and positively worded items in a questionnaire. These need to be reversed prior to the reliability analysis. In Stata, this is done automatically when an item is negatively correlated with the other items. It is possible to add the option `, reverse (variable name)` to force Stata to reverse an item or you can stop Stata from reversing items automatically by using `, asis`. Furthermore, we have to be aware of potential subscales in our item set. Some multi-item scales comprise subsets of items that measure different facets of a multidimensional construct. For example, soccer fan satisfaction is a multidimensional construct that includes aspects such as satisfaction with the stadium, the merchandise (as described above), the team, and the coach, each measured with a different item set. It would be inappropriate to calculate one Cronbach's Alpha value for all 99 items. Cronbach's Alpha is always calculated over the items belonging to one construct and not all the items in the dataset!

8.5   Structural Equation Modeling

Whereas a confirmatory factor analysis involves testing if and how items relate to specific constructs, *structural equation modeling* involves the estimation of relations between these constructs. It has become one of the most important methods in social sciences, including marketing research.

There are broadly two approaches to structural equation modeling: **Covariance-based structural equation modeling** (e.g., Jöreskog 1971) and **partial least**
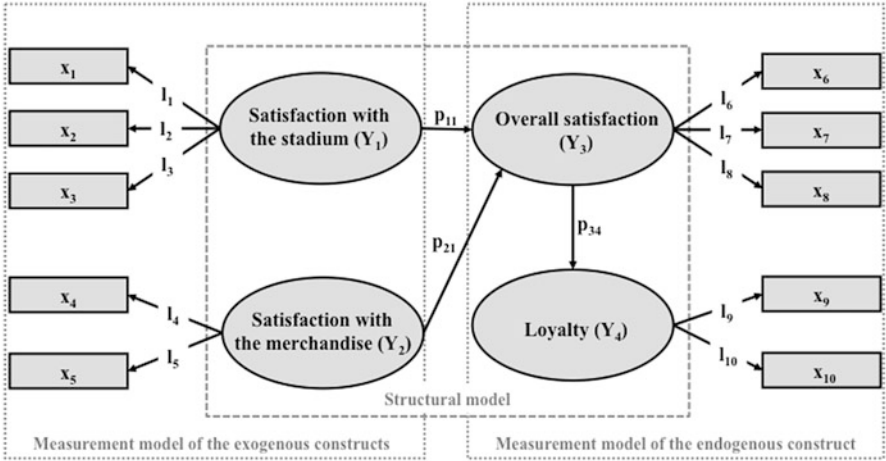
Fig. 8.8 Path diagram (structural equation modeling)

squares structural equation modeling (e.g., Wold 1982), simply referred to as **CB-SEM** and **PLS-SEM**. Both estimation methods are based on the idea of an underlying model that allows the researcher to test relationships between multiple items and constructs.

Figure 8.8 shows an example path diagram with four constructs (represented by circles or ovals) and their respective items (represented by boxes).[7] A path model incorporates two types of constructs: (1) exogenous constructs (here, satisfaction with the stadium (Y_1) and satisfaction with the merchandise (Y_2)) that do not depend on other constructs, and (2) endogenous constructs (here, overall satisfaction (Y_3) and loyalty (Y_4)) that depend on one or more exogenous (or other endogenous) constructs. The relations between the constructs (indicated with p) are called path coefficients, while the relations between the constructs and their respective items (indicated with l) are the indicator loadings. One can distinguish between the structural model that incorporates the relations between the constructs and the (exogenous and endogenous) measurement models that represent the relationships between the constructs and their related items. Items that measure constructs are labeled x.

In the model in Fig. 8.8, we assume that the two exogenous constructs *satisfaction with the stadium* and *satisfaction with the merchandise* relate to the endogenous construct *overall satisfaction* and that *overall satisfaction* relates to *loyalty*. Depending on the research question, we could of course incorporate additional exogenous and endogenous constructs. Using empirical data, we could then test this model and, thus, evaluate the relationships between all the constructs and between each construct and its indicators. We could, for example, assess which of the two

---

[7]Note that we omitted the error terms for clarity's sake.

constructs, Y_1 or Y_2, exerts the greater influence on Y_3. The result would guide us when developing marketing plans in order to increase overall satisfaction and, ultimately, loyalty by answering the research question whether we should rather concentrate on increasing the fans' satisfaction with the stadium or with the merchandise.

The evaluation of a path model analysis requires several steps that include the assessment of both measurement models and the structural model. Diamantopoulos and Siguaw (2000) and Hair et al. (2013) provide a thorough description of the covariance-based structural equation modeling approach and its application. Acock (2013) provides a detailed explanation of how to conduct covariance-based structural equation modeling analyses in Stata. Hair et al. (2017a, b, 2018) provide a step-by-step introduction on how to set up and test path models using partial least squares structural equation modeling.

8.6 Example

In this example, we take a closer look at some of the items from the Oddjob Airways dataset (⤓ Web Appendix → Downloads). This dataset contains eight items that relate to the customers' experience when flying with Oddjob Airways. For each of the following items, the respondents had to rate their degree of agreement from 1 ("completely disagree") to 100 ("completely agree"). The variable names are included below:

– with Oddjob Airways you will arrive on time (*s1*),
– the entire journey with Oddjob Airways will occur as booked (*s2*),
– in case something does not work out as planned, Oddjob Airways will find a good solution (*s3*),
– the flight schedules of Oddjob Airways are reliable (*s4*),
– Oddjob Airways provides you with a very pleasant travel experience (*s5*),
– Oddjob Airways's on board facilities are of high quality (*s6*),
– Oddjob Airways's cabin seats are comfortable (*s7*), and
– Oddjob Airways offers a comfortable on-board experience (*s8*).

Our aim is to reduce the complexity of this item set by extracting several factors. Hence, we use these items to run a PCA using the `factor, pcf` procedure in Stata.

8.6.1 Principal Component Analysis

8.6.1.1 Check Requirements and Conduct Preliminary Analyses
All eight variables are interval scaled from 1 ("very unsatisfied") to 100 ("very satisfied"), therefore meeting the requirements in terms of the measurement scale.
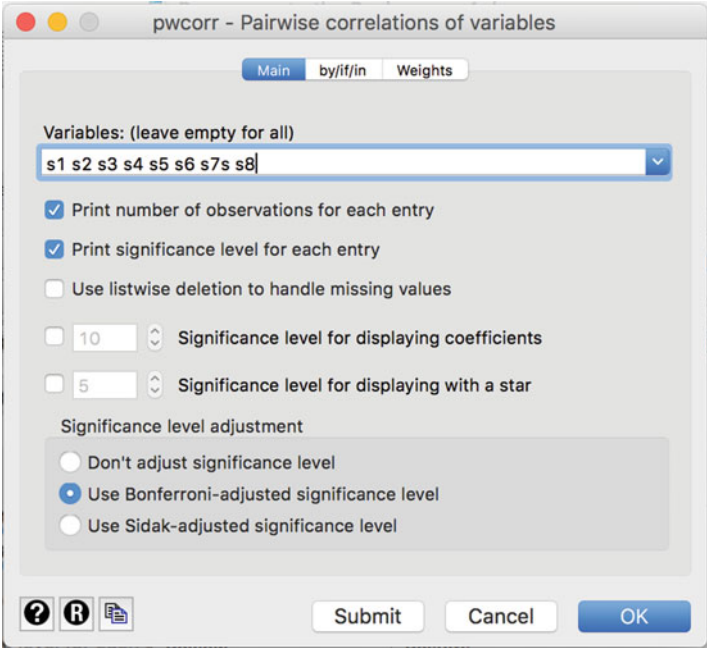
Fig. 8.9 Pairwise correlations of variables

With 1,065 independent observations, the sample size requirements are clearly met, even if the analysis yields very low communality values.

Determining if the variables are sufficiently correlated is easy if we go to ▶ Statistics ▶ Summaries, tables, and tests ▶ Summary and descriptive statistics ▶ Pairwise correlations. In the dialog box shown in Fig. 8.9, either enter each variable separately (i.e., *s1 s2 s3* etc.) or simply write *s1-s8* as the variables appear in this order in the dataset. Then also tick **Print number of observations for each entry**, **Print significance levels for each entry**, as well as **Use Bonferroni-adjusted significance level**. The latter option corrects for the many tests we execute at the same time and is similar to what we discussed in Chap. 6.

Table 8.4 shows the resulting output. The values in the diagonal are all **1.000**, which is logical, as this is the correlation between a variable and itself! The off-diagonal cells correspond to the pairwise correlations. For example, the pairwise correlation between *s1* and *s2* is **0.7392**. The value under it denotes the *p*-value (**0.000**), indicating that the correlation is significant. To determine an absolute minimum standard, check if at least one correlation in all the off-diagonal cells is significant. The last value of **1037** indicates the sample size for the correlation between *s1* and *s2*.

The correlation matrix in Table 8.4 indicates that there are several pairs of highly correlated variables. For example, not only *s1* is highly correlated with *s2* (correlation = **0.7392**), but also *s3* is highly correlated with *s1* (correlation = **0.6189**), just

Table 8.4 Pairwise correlation matrix

```
pwcorr s1-s8, obs sig bonferroni

             |       s1       s2       s3       s4       s5       s6       s7
-------------+---------------------------------------------------------------------
         s1  |   1.0000
             |
             |     1038
             |
         s2  |   0.7392   1.0000
             |   0.0000
             |     1037     1040
             |
         s3  |   0.6189   0.6945   1.0000
             |   0.0000   0.0000
             |      952      952      954
             |
         s4  |   0.7171   0.7655   0.6447   1.0000
             |   0.0000   0.0000   0.0000
             |     1033     1034      951     1035
             |
         s5  |   0.5111   0.5394   0.5593   0.4901   1.0000
             |   0.0000   0.0000   0.0000   0.0000
             |     1026     1027      945     1022     1041
             |
         s6  |   0.4898   0.4984   0.4972   0.4321   0.8212   1.0000
             |   0.0000   0.0000   0.0000   0.0000   0.0000
             |     1025     1027      943     1022     1032     1041
             |
         s7  |   0.4530   0.4555   0.4598   0.3728   0.7873   0.8331   1.0000
             |   0.0000   0.0000   0.0000   0.0000   0.0000   0.0000
             |     1030     1032      947     1027     1038     1037     1048
             |
         s8  |   0.5326   0.5329   0.5544   0.4822   0.8072   0.8401   0.7773
             |   0.0000   0.0000   0.0000   0.0000   0.0000   0.0000   0.0000
             |     1018     1019      937     1014     1029     1025     1032
             |

             |       s8
-------------+---------
         s8  |   1.0000
             |
             |     1034
             |
```

like *s4* (correlation $= \mathbf{0.7171}$). As these variables' correlations with the remaining ones are less pronounced, we suspect that these four variables constitute one factor. As you can see by just looking at the correlation matrix, we can already identify the factor structure that might result.

However, at this point of the analysis, we are more interested in checking whether the variables are sufficiently correlated to conduct a PCA. When we examine Table 8.4, we see that all correlations have *p*-values below 0.05. This result indicates that the variables are sufficiently correlated. However, for a concluding evaluation, we need to take the anti-image and related statistical measures into account. Most importantly, we should also check if the KMO values are at least 0.50. As we can only do this *after* the actual PCA, we will discuss this point later.
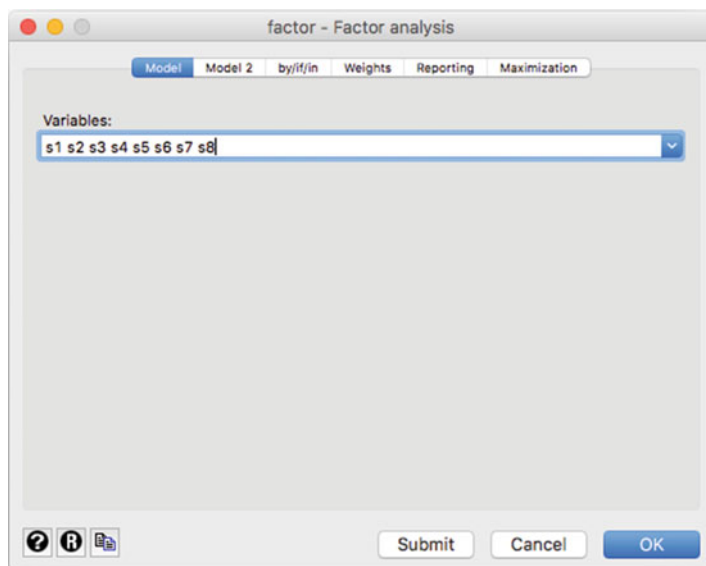
Fig. 8.10 Factor analysis dialog box

8.6.1.2 Extract the Factors and Determine the Number of Factors

To run the PCA, click on ▶ Statistics ▶ Multivariate analysis ▶ Factor and
principal component analysis ▶ Factor analysis, which will open a dialog box
similar to Fig. 8.10. Next, enter *s1 s2 s3 s4 s5 s6 s7 s8* in the **Variables** box.
Alternatively, you can also simply enter *s1-s8* in the box as the dataset includes
these eight variables in consecutive order.

Under the **Model 2** tab (see Fig. 8.11), we can choose the method for factor
extraction, the maximum factors to be retained, as well as the minimum value for
the eigenvalues to be retained. As the aim of our analysis is to reproduce the data
structure, we choose **Principal-component factor**, which initiates the PCA based
on Stata's `factor, pcf` procedure. By clicking on **Minimum value of
eigenvalues to be retained** and entering **1** in the box below, we specify the Kaiser
criterion. If we have a priori information on the factor structure, we can specify the
number of factors manually by clicking on **Maximum number of factors to be
retained**. Click on **OK** and Stata will display the results (Table 8.5).

Before we move to the further interpretation of the PCA results in Table 8.5, let's
first take a look at the KMO. Although we need to know if the KMO is larger than 0.50
to interpret the PCA results with confidence, we can only do this after having run the
PCA. We can calculate the KMO values by going to ▶ Statistics ▶ Postestimation
▶ Principal component analysis reports and graphs ▶ Kaiser-Meyer-Olkin measure
of sample adequacy (Fig. 8.12). In the dialog box that opens, simply click on **OK** to
initiate the analysis.

The analysis result at the bottom of Table 8.6 reveals that the KMO value is
0.9073, which is "marvelous" (see Table 8.2). Likewise, the variable-specific MSA
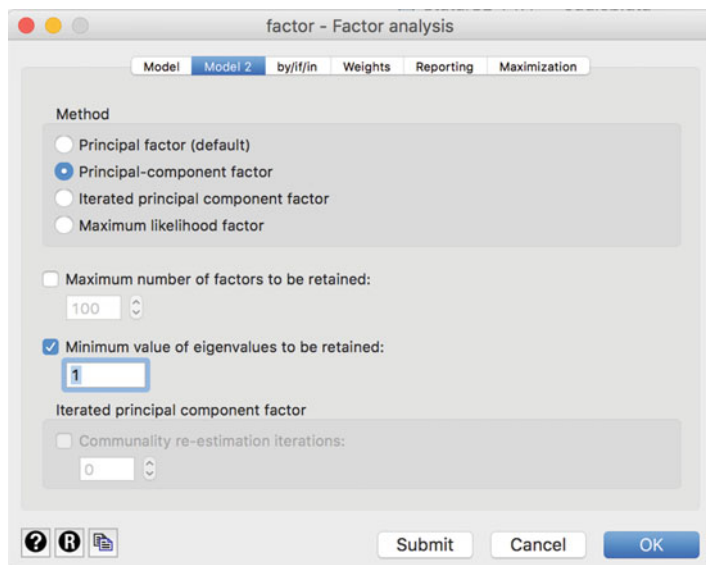
Fig. 8.11 Factor analysis dialog box (options)

values in the table are all above the threshold value of 0.50. For example, *s1* has an MSA value of **0.9166**.

The output in Table 8.5 shows three blocks. On the top right of the first block, Stata shows the number of observations used in the analysis (**Number of obs** = **921**) and indicates that the analysis yielded two factors (**Retained factors** = **2**). In the second block, Stata indicates the eigenvalues for each factor. With an eigenvalue of **5.24886**, the first factor extracts a large amount of variance, which accounts for 5.24886/8 = 65.61% of the total variance (see column: **Proportion**). With an eigenvalue of **1.32834**, factor two extracts less variance (16.60%). Using the **Kaiser criterion** (i.e., eigenvalue >1), we settle on two factors, because the third factor's eigenvalue is clearly lower than 1 (**0.397267**). The **Cumulative** column indicates the cumulative variance extracted. The two factors extract **0.8222** or 82.22% of the variance, which is highly satisfactory. The next block labeled **Factor loadings (pattern matrix) and unique variances** shows the factor loadings along with the **Uniqueness**, which indicates the amount of each variable's variance that the factors cannot reproduce (i.e., 1-communality) and is therefore lost in the process. All uniqueness values are very low, indicating that the factors reproduce the variables' variance well. Specifically, with a value of **0.3086**, *s3* exhibits the highest uniqueness value, which suggests a communality of 1–0.3086 = 0.6914 and is clearly above the 0.50 threshold.

Beside the Kaiser criterion, the scree plot helps determine the number of factors. To create a scree plot, go to ▶ Statistics ▶ Postestimation ▶ Principal component analysis reports and graphs ▶ Scree plot of eigenvalues. Then click on **Launch** and

Table 8.5 PCA output

```
factor s1 s2 s3 s4 s5 s6 s7 s8, pcf mineigen(1)

(obs=921)

Factor analysis/correlation                      Number of obs    =       921
    Method: principal-component factors          Retained factors =         2
    Rotation: (unrotated)                        Number of params =        15

    --------------------------------------------------------------------------
         Factor  |   Eigenvalue   Difference         Proportion   Cumulative
    -------------+------------------------------------------------------------
        Factor1  |     5.24886      3.92053             0.6561       0.6561
        Factor2  |     1.32834      0.93107             0.1660       0.8222
        Factor3  |     0.39727      0.13406             0.0497       0.8718
        Factor4  |     0.26321      0.03202             0.0329       0.9047
        Factor5  |     0.23119      0.03484             0.0289       0.9336
        Factor6  |     0.19634      0.00360             0.0245       0.9582
        Factor7  |     0.19274      0.05067             0.0241       0.9822
        Factor8  |     0.14206          .               0.0178       1.0000
    --------------------------------------------------------------------------
    LR test: independent vs. saturated:  chi2(28) = 6428.36 Prob>chi2 = 0.0000

Factor loadings (pattern matrix) and unique variances

    ------------------------------------------------
       Variable |  Factor1    Factor2 |  Uniqueness
    -------------+--------------------+-------------
            s1  |  0.7855     0.3995  |    0.2235
            s2  |  0.8017     0.4420  |    0.1619
            s3  |  0.7672     0.3206  |    0.3086
            s4  |  0.7505     0.5090  |    0.1777
            s5  |  0.8587    -0.3307  |    0.1533
            s6  |  0.8444    -0.4203  |    0.1104
            s7  |  0.7993    -0.4662  |    0.1439
            s8  |  0.8649    -0.3291  |    0.1436
    ------------------------------------------------
```

OK. Stata will produce a graph as shown in Fig. 8.13. There is an "elbow" in the line at three factors. As the number of factors that the scree plot suggests is one factor less than the elbow indicates, we conclude that two factors are appropriate. This finding supports the conclusion based on the Kaiser criterion.

Stata also allows for plotting each eigenvalue's confidence interval. However, to display such a scree plot requires running a PCA with a different command in combination with a postestimation command. The following syntax produces a scree plot for our example with a 95% confidence interval (heteroskedastic), as well as a horizontal reference line at the 1 threshold.

```
pca s1 s2 s3 s4 s5 s6 s7 s8, mineigen(1)
screeplot, recast(line) ci(heteroskedastic) yline(1)
```
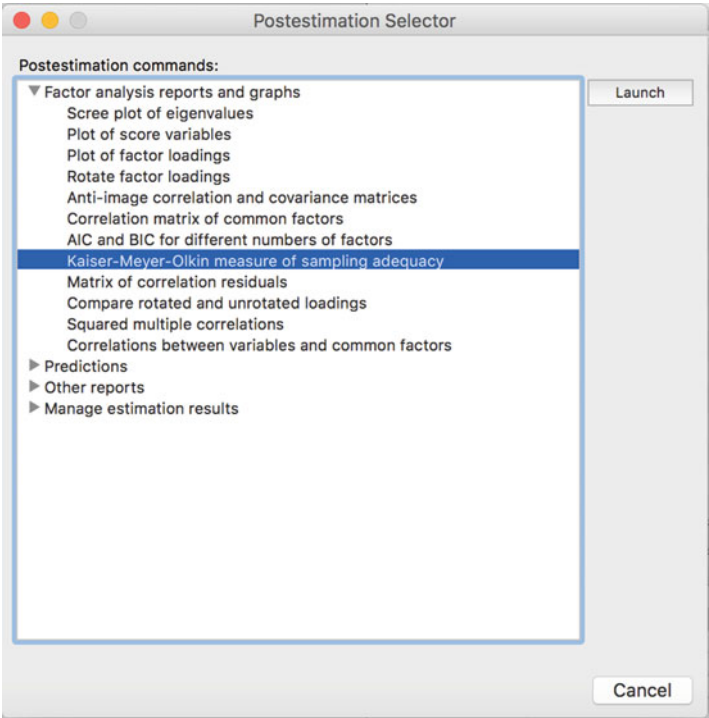
Fig. 8.12 Postestimation dialog box

Table 8.6 The KMO statistic

```
estat kmo

Kaiser-Meyer-Olkin measure of sampling adequacy

      -----------------------
         Variable |     kmo
      ------------+----------
               s1 |  0.9166
               s2 |  0.8839
               s3 |  0.9427
               s4 |  0.8834
               s5 |  0.9308
               s6 |  0.8831
               s7 |  0.9036
               s8 |  0.9180
      ------------+----------
          Overall |  0.9073
```

Fig. 8.13 Scree plot of eigenvalues

While the Kaiser criterion and the scree plot are helpful for determining the number of factors to extract, parallel analysis is a more robust criterion. Parallel analysis can only be accessed through the free add-on package paran. To install the package, type in `help paran` in the command window and follow the instructions to install the package. Having installed the package, type in `paran s1 s2 s3 s4 s5 s6 s7 s8, centile(95) q all graph` in the command window and Stata will produce output similar to Table 8.7 and Fig. 8.14.

Table 8.7 contains two rows of eigenvalues, with the first column (**Adjusted Eigenvalue**) indicating the sampling error-adjusted eigenvalues obtained by parallel analysis. Note that your results will look slightly different, as parallel analysis draws on randomly generated datasets. The second column (**Unadjusted Eigenvalue**) contains the eigenvalues as reported in the PCA output (Table 8.5). Analogous to the original analysis, the two factors exhibit adjusted eigenvalues larger than 1, indicating a two-factor solution. The scree plot in Fig. 8.14 also supports this result, as the first two factors exhibit adjusted eigenvalues larger than the randomly generated eigenvalues. Conversely, the random eigenvalue of the third factor is clearly larger than the adjusted one.

Finally, we can also request the model selection statistics AIC and BIC for different numbers of factors (see Fig. 8.15). To do so, go to ▶ Statistics ▶ Postestimation ▶ AIC and BIC for different numbers of factors. As our analysis draws on eight variables, we restrict the number of factors to consider to 4 (**Specify the maximum number of factors to include in summary table: 4**). Table 8.8 shows the results of our analysis. As can be seen, AIC has the smallest value (**53.69378**) for a four-factor solution, whereas BIC's minimum value occurs for a

Table 8.7 Parallel analysis output

```
paran s1 s2 s3 s4 s5 s6 s7 s8, centile (95) q all graph

Computing: 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%


Results of Horn's Parallel Analysis for principal components
240 iterations, using the p95 estimate

--------------------------------------------------
Component   Adjusted     Unadjusted    Estimated
or Factor   Eigenvalue   Eigenvalue    Bias
--------------------------------------------------
  1         5.1571075    5.2488644     .09175694
  2         1.2542191    1.3283371     .07411802
  3         .3409911     .39726683     .05627573
  4         .23833934    .26320842     .02486908
  5         .22882924    .23118517     .00235593
  6         .23964682    .19634077     -.04330605
  7         .2775334     .19273589     -.0847975
  8         .26333341    .14206139     -.12127203
--------------------------------------------------
Criterion: retain adjusted components > 1
```
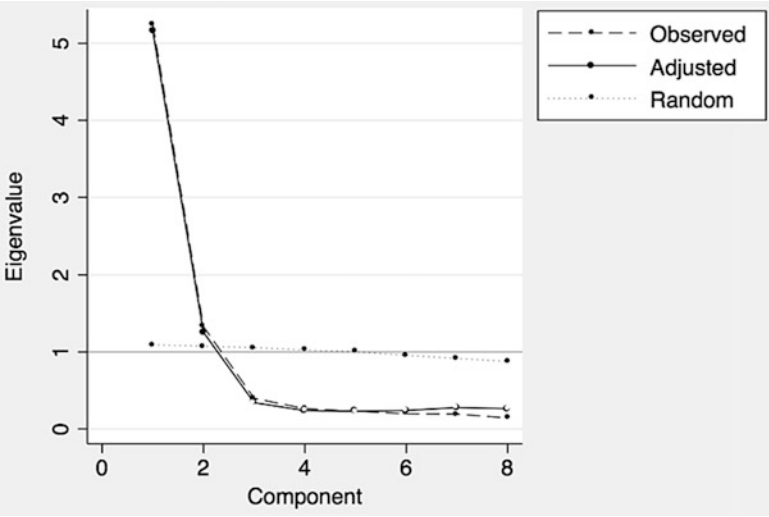

Fig. 8.14 Scree plot of parallel analysis

two-factor solution (**141.8393**). However, as AIC is well known to overspecify the number of factors, this result gives confidence that the two-factor solution as indicated by the BIC is appropriate. Note that Table 8.8 says no Heywood cases encountered.
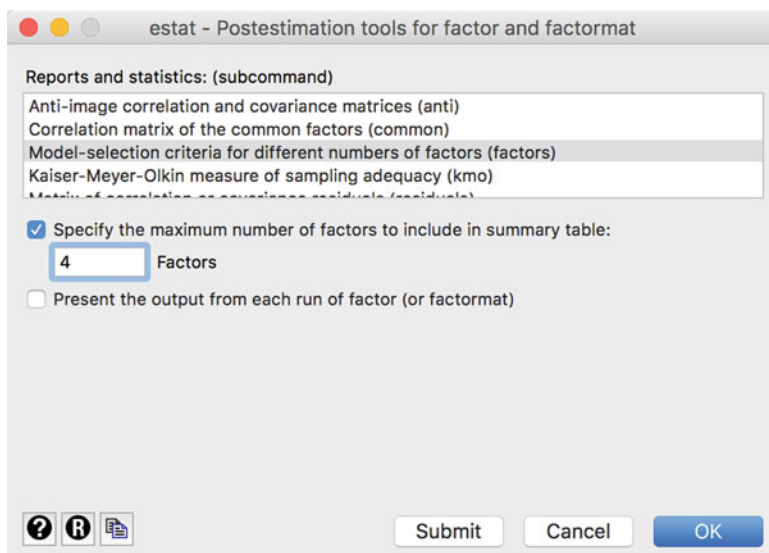
Fig. 8.15   AIC and BIC

Table 8.8   Factor rotation output

```
estat factors, factors(4)

Factor analysis with different numbers of factors (maximum likelihood)

    ----------------------------------------------------------
    #factors |     loglik   df_m   df_r          AIC          BIC
    ---------+------------------------------------------------
         1 |  -771.1381      8     20     1558.276      1596.88
         2 |  -19.72869     15     13     69.45738     141.8393
         3 |  -10.09471     21      7     62.18943     163.5241
         4 |  -.8468887     26      2     53.69378     179.1557
    ----------------------------------------------------------
    no Heywood cases encountered
```

Heywood cases indicate negative estimates of variances or correlation estimates greater than one in absolute value. In the Stata output, they are typically noted as `Beware: solution is a Heywood case`.

8.6.1.3 Interpret the Factor Solution

To facilitate the interpretation of the factor solution, we need to rotate the factors. To initiate this analysis, go to ► Statistics ► Postestimation ► Factor analysis reports and graphs ► Rotate factor loadings. In the dialog box that opens, select **Varimax (default)** under **Orthogonal rotation** and click on **Apply the Kaiser normalization**, followed by **OK**. Table 8.9 shows the resulting output.

Table 8.9 Factor rotation output

```
rotate, kaiser

Factor analysis/correlation                    Number of obs    =        921
    Method: principal-component factors        Retained factors =          2
    Rotation: orthogonal varimax (Kaiser on)   Number of params =         15

    --------------------------------------------------------------------------
        Factor |    Variance   Difference         Proportion   Cumulative
    -------------+------------------------------------------------------------
       Factor1 |    3.41063     0.24405             0.4263       0.4263
       Factor2 |    3.16657        .                0.3958       0.8222
    --------------------------------------------------------------------------
    LR test: independent vs. saturated:  chi2(28) = 6428.36 Prob>chi2 = 0.0000

Rotated factor loadings (pattern matrix) and unique variances

    --------------------------------------------------
       Variable |  Factor1   Factor2 |  Uniqueness
    -------------+--------------------+---------------
           s1 |   0.2989    0.8290 |     0.2235
           s2 |   0.2817    0.8711 |     0.1619
           s3 |   0.3396    0.7590 |     0.3086
           s4 |   0.1984    0.8848 |     0.1777
           s5 |   0.8522    0.3470 |     0.1533
           s6 |   0.9031    0.2719 |     0.1104
           s7 |   0.9017    0.2076 |     0.1439
           s8 |   0.8557    0.3524 |     0.1436
    --------------------------------------------------

Factor rotation matrix

    -------------------------------
            | Factor1   Factor2
    -------------+------------------
       Factor1 |  0.7288    0.6847
       Factor2 | -0.6847    0.7288
    -------------------------------
```

The upper part of Table 8.9 is the same as the standard PCA output (Table 8.5), showing that the analysis draws on **921** observations and extracts two factors, which jointly capture 82.22% of the variance. As its name implies, the **Rotated factor loadings** block shows the factor loadings after rotation. Recall that rotation is carried out to facilitate the interpretation of the factor solution. To interpret the factors, we first "assign" each variable to a certain factor based on its maximum absolute factor loading. That is, if the highest absolute loading is negative, higher values of a particular variable relate negatively to the assigned factor. After that, we should find an umbrella term for each factor that best describes the set of variables associated with that factor. Looking at Table 8.9, we see that *s1–s4* load highly on the second factor, whereas *s5-s8* load on the first factor. For example, *s1* has a **0.2989** loading on the first factor, while its loading is much stronger on the second factor (**0.8290**). Finally, note that the uniqueness and, hence, the communality values are unaffected by the rotation (see Table 8.5).
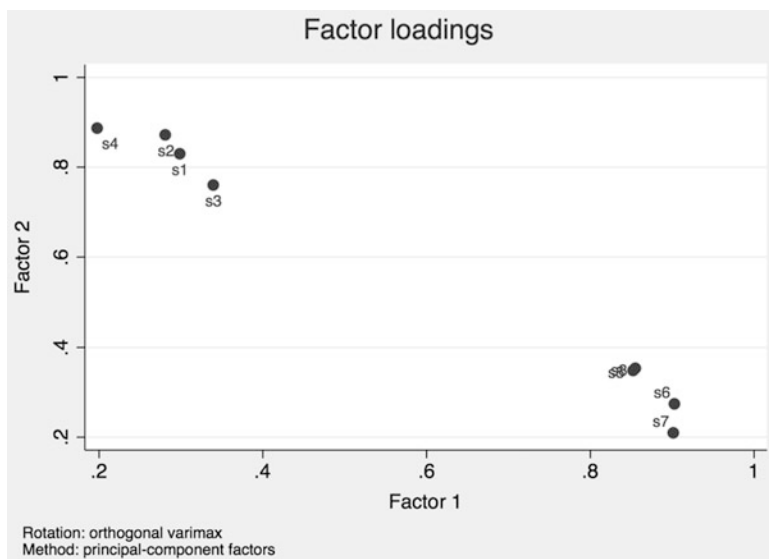
Fig. 8.16 Factor loadings plot

To facilitate the identification of labels, we can plot each item's loading against each factor. To request a factor loadings plot, go to ▶ Statistics ▶ Postestimation ▶ Factor analysis reports and graphs ▶ Plot of factor loadings and click on **Launch**. In the dialog box that follows, retain the default settings and click on **OK**. The resulting plot (Fig. 8.16) shows two cluster of variables, which strongly load on one factor while having low loadings on the other factor. This result supports our previous conclusion in terms of the variable assignments.

Having identified which variables load highly on which factor in the rotated solution, we now need to identify labels for each factor. Looking at the variable labels, we learn that the first set of variables (*s1–s4*) relate to reliability aspects of the journey and related processes, such as the booking. We could therefore label this factor (i.e., factor 2) *reliability*. The second set of variables (*s5–s8*) relate to different aspects of the onboard facilities and the travel experience. Hence, we could label this factor (i.e., factor 1) *onboard experience*. The labeling of factors is of course subjective and you could provide different descriptions.

8.6.1.4 Evaluate the Goodness-of-fit of the Factor Solution

The last step involves assessing the analysis's goodness-of-fit. To do so, we first look at the residuals (i.e., the differences between observed and reproduced correlations) in the reproduced correlation matrix. To create this matrix, go to ▶ Statistics ▶ Postestimation ▶ Factor analysis reports and graphs ▶ Matrix of correlation residuals. In the dialog box that opens, select **Matrix of correlation of covariance residuals (residuals)** and click on **OK**. Table 8.10 shows the resulting output.

Table 8.10 Correlation residual matrix

```
estat residuals

Raw residuals of correlations (observed-fitted)

     -------------------------------------------------------------------------------
       Variable |      s1        s2        s3        s4        s5        s6        s7        s8
     ------------+------------------------------------------------------------------
            s1 |   0.0000
            s2 |  -0.0525    0.0000
            s3 |  -0.1089   -0.0602    0.0000
            s4 |  -0.0598   -0.0557   -0.0907    0.0000
            s5 |  -0.0174   -0.0063   -0.0029    0.0076    0.0000
            s6 |   0.0046    0.0023   -0.0216    0.0159   -0.0464    0.0000
            s7 |   0.0136    0.0182   -0.0118    0.0042   -0.0529   -0.0393    0.0000
            s8 |  -0.0056   -0.0135   -0.0056    0.0052   -0.0404   -0.0265   -0.0645    0.0000
     -------------------------------------------------------------------------------
```

When examining the lower part of Table 8.10, we see that there are several residuals with absolute values larger than 0.05. A quick count reveals that 8 out of 29 (i.e., 27.59%) residuals are larger than 0.05. As the percentage of increased residuals is well below 50%, we can presume a good model fit.

Similarly, our previous analysis showed that the two factors reproduce a sufficient amount of each variable's variance. The uniqueness values are clearly below 0.50 (i.e., the communalities are larger than 0.50), indicating that the factors account for more than 50% of the variables' variance (Table 8.5).

8.6.1.5 Compute the Factor Scores

The evaluation of the factor solution's goodness-of-fit completes the standard PCA analysis. However, if we wish to continue using the results for further analyses, we should calculate the factor scores. Go to ▶ Statistics ▶ Postestimation ▶ Predictions ▶ Regression and Bartlett scores and Stata will open a dialog box similar to Fig. 8.17. Under **New variable names or variable stub*** you can enter a prefix name, which Stata uses to name the saved factor scores. For example, specifying *factor**, as shown in Fig. 8.17, will create two variables called *factor1* and *factor2*. Next, select **Factors scored by the regression scoring method** and click on **OK**. Stata will produce an output table showing the scoring coefficients, which are the weights used to compute the factor scores from the standardized data. However, we are more interested in the actual factor scores, which we can access by clicking on the **Data Browser** button in Stata's menu bar. Figure 8.18 shows the scores of *factor1* and *factor2* for the first ten observations.

Being *z*-standardized, the newly generated variables *factor1* and *factor2* have mean values (approximately) equal to zero and standard deviations equal to 1. Thus, the factor scores are estimated in units of standard deviations from their means. For example, the first observation is about **1.91** standard deviations below average on the *onboard experience* factor (i.e., factor 1) and about **0.91** standard deviations above average on the *reliability* factor (i.e., factor 2). In contrast, the second observation is clearly above average in terms of *reliability* and *onboard experience*. Note that if the original variables include a missing value, the factor score will also be missing (i.e., only a "**.**" (dot) will be recorded).
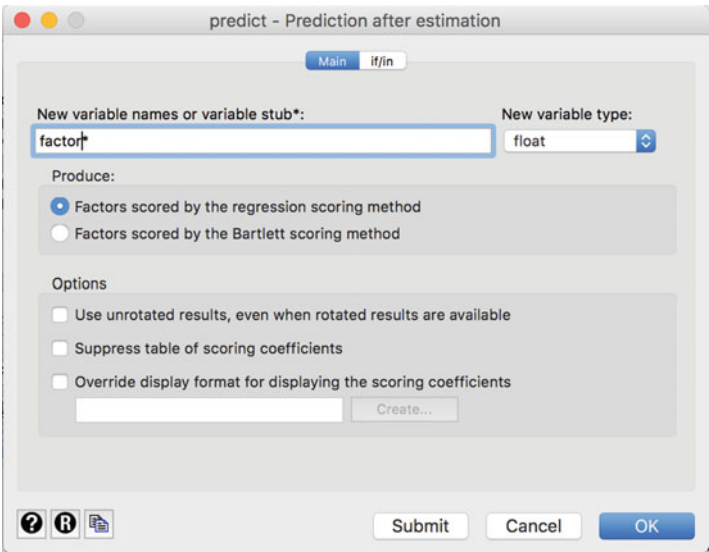
Fig. 8.17 Saving factor scores as new variables

Fig. 8.18 Overview of factor
scores for the first ten
observations

factor1	factor2
−1.911223	.9144877
1.684422	1.297118
−2.432783	.4436803
.1665416	.8420736
−.1204526	−.85245
−.4704992	.3993561
.8271663	−1.990248
.2047913	.9054558
.4886046	1.043682
.5272735	−.7065509

8.6.2 Reliability Analysis

To illustrate its usage, let's carry out a reliability analysis of the first factor *onboard experience* by calculating Cronbach's Alpha as a function of variables *s5* to *s8*. To run the reliability analysis, click on ▶ Statistics ▶ Multivariate analysis ▶ Cronbach's Alpha. A window similar to Fig. 8.19 will appear. Next, enter variables *s5–s8* into the **Variables** box.
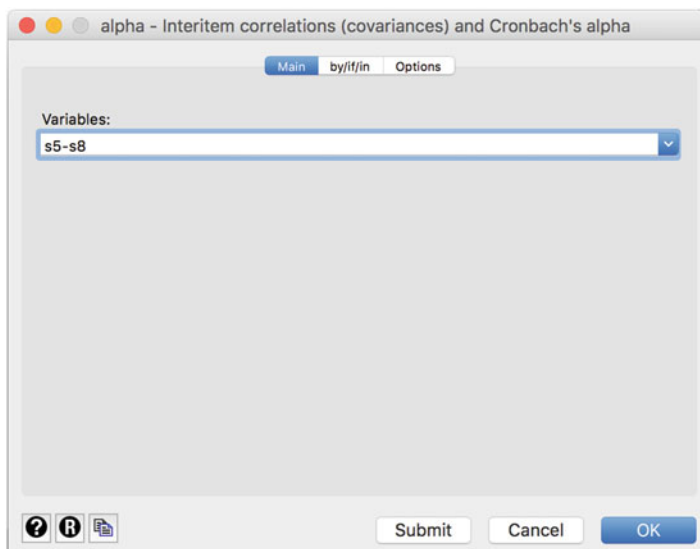
Fig. 8.19   Reliability analysis dialog box

The **Options** tab (Fig. 8.20) provides options for dealing with missing values and requesting descriptive statistics for each item and the entire scale or item correlations. Check **Display item-test and item-rest correlations** and click on **OK**.

The results in Table 8.11 show that the scale exhibits a high degree of internal consistency reliability. With a value of **0.9439** (see row **Test scale**), the Cronbach's Alpha coefficient lies well above the commonly suggested threshold of 0.70. This result is not surprising, since we are simply testing a scale previously established by means of item correlations. Keep in mind that we usually carry out a reliability analysis to test a scale using a different sample—this example is only for illustration purposes! The rightmost column of Table 8.11 indicates what the Cronbach's Alpha would be if we deleted the item indicated in that row. When we compare each of the values with the overall Cronbach's Alpha value, we can see that any change in the scale's set-up would reduce the Cronbach's Alpha value. For example, by removing *s5* from the scale, the Cronbach's Alpha of the new scale comprising only *s6*, *s7*, and *s8* would be reduced to **0.9284**. Therefore, deleting this item (or any others) makes little sense. In the leftmost column of Table 8.11, Stata indicates the number of observations (**Obs**), as well as whether that particular item correlates positively or negatively with the sum of the other items (**Sign**). This information is useful for determining whether reverse-coded items were also identified as such. Reverse-coded items should have a minus sign. The columns **item-test**, **item-rest**, and **average interitem covariance** are not needed for a basic interpretation.
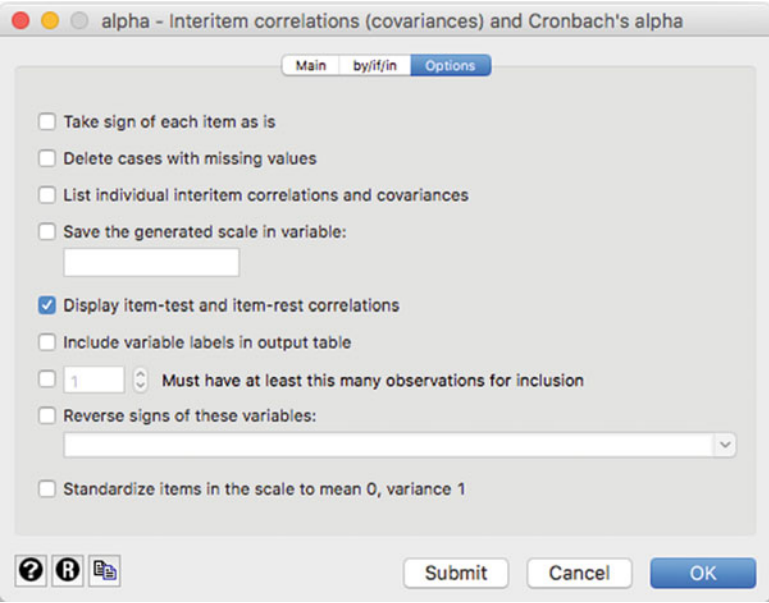
Fig. 8.20 Options tab for Cronbach's alpha

Table 8.11 Reliability statistics

```
alpha s5 s6 s7 s8, item

Test scale = mean(unstandardized items)

                                                        average
                              item-test      item-rest   interitem
Item          |  Obs  Sign  correlation    correlation  covariance    alpha
------------+----------------------------------------------------------------
s5          | 1041   +       0.9211          0.8578      421.5723     0.9284
s6          | 1041   +       0.9422          0.8964      412.3225     0.9168
s7          | 1048   +       0.9222          0.8506      399.2565     0.9330
s8          | 1034   +       0.9203          0.8620      434.3355     0.9282
------------+----------------------------------------------------------------
Test scale  |                                            416.8996     0.9439
------------------------------------------------------------------------------
```

8.7   Customer Satisfaction at Haver and Boecker (Case Study)

Haver and Boecker (http://www.haverboecker.com) is one of the world's leading
and most renowned machine producers in the fields of mineral processing, as well
as the storing, conveying, packing, and loading of bulk material. The family-owned
group operates through its global network of facilities, with manufacturing units,
among others, in Germany, the UK, Belgium, US, Canada, Brazil, China, and India.

The company's relationships with its customers are usually long-term oriented and complex. Since the company's philosophy is to help customers and business partners solve their challenges or problems, they often customize their products and services to meet the buyers' needs. Therefore, the customer is no longer a passive buyer, but an active partner. Given this background, the customers' satisfaction plays an important role in establishing, developing, and maintaining successful customer relationships.

Very early on, the company's management realized the importance of customer satisfaction and decided to commission a market research project in order to identify marketing activities that can positively contribute to the business's overall success. Based on a thorough literature review, as well as interviews with experts, the company developed a short survey to explore their customers' satisfaction with specific performance features and their overall satisfaction. All the items were measured on 7-point scales, with higher scores denoting higher levels of satisfaction. A standardized survey was mailed to customers in 12 countries worldwide, which yielded 281 fully completed questionnaires. The following items (names in parentheses) were listed in the survey:

– Reliability of the machines and systems. (s_1)
– Life-time of the machines and systems. (s_2)
– Functionality and user-friendliness operation of the machines and systems. (s_3)
– Appearance of the machines and systems. (s_4)
– Accuracy of the machines and systems. (s_5)

– Timely availability of the after-sales service. (s_6)
– Local availability of the after-sales service. (s_7)
– Fast processing of complaints. (s_8)
– Composition of quotations. (s_9)
– Transparency of quotations. (s_{10})
– Fixed product prize for the machines and systems. (s_{11})
– Cost/performance ratio of the machines and systems. (s_{12})
– Overall, how satisfied are you with the supplier (*overall*)?

Your task is to analyze the dataset to provide the management of Haver and Boecker with advice for effective customer satisfaction management. The dataset is labeled *haver_and_boecker.dta* (↓ Web Appendix → Downloads).

1. Determine the factors that characterize the respondents by means a factor analysis. Use items s_1–s_{12} for this. Run a PCA with varimax rotation to facilitate interpretation. Consider the following aspects:
   (a) Are all assumptions for carrying out a PCA met? Pay special attention to the question whether the data are sufficiently correlated.
   (b) How many factors would you extract? Base your decision on the Kaiser criterion, the scree plot, parallel analysis, and the model selection statistics AIC and BIC.
   (c) Find suitable labels for the extracted factors.
   (d) Evaluate the factor solution's goodness-of-fit.
2. Use the factor scores and regress the customers' overall satisfaction (*overall*) on these. Evaluate the strength of the model and compare it with the initial regression. What should Haver and Boecker's management do to increase their customers' satisfaction?
3. Calculate the Cronbach's Alpha over items s_1–s_5 and interpret the results.

For further information on the dataset and the study, see Festge and Schwaiger (2007), as well as Sarstedt et al. (2009).

8.8   Review Questions

1. What is factor analysis? Try to explain what factor analysis is in your own words.
2. What is the difference between exploratory factor analysis and confirmatory factor analysis?
3. What is the difference between PCA and factor analysis?
4. Describe the terms communality, eigenvalue, factor loading, and uniqueness. How do these concepts relate to one another?
5. Describe three approaches used to determine the number of factors.
6. What are the purpose and the characteristic of a varimax rotation? Does a rotation alter eigenvalues or factor loadings?

7. Re-run the Oddjob Airways case study by carrying out a factor analysis and compare the results with the example carried out using PCA. Are there any significant differences?
8. What is reliability analysis and why is it important?
9. Explain the basic principle of structural equation modeling.

8.9   Further Readings

Nunnally, J. C., & Bernstein, I. H. (1993). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

Psychometric theory is a classic text and the most comprehensive introduction to the fundamental principles of measurement. Chapter 7 provides an in-depth discussion of the nature of reliability and its assessment.

Sarstedt, M., Hair, J. F., Ringle, C. M., Thiele, K. O., & Gudergan, S. P. (2016). Estimation issues with PLS and CBSEM: where the bias lies! *Journal of Business Research*, *69*(10), 3998–4010.

This paper discusses the differences between covariance-based and partial least squares structural equation modeling from a measurement perspective. This discussion relates to the differentiation between factor analysis and PCA and the assumptions underlying each approach to measure unobservable phenomena.

Stewart, D. W., (1981). The application and misapplication of factor analysis in marketing research. *Journal of Marketing Research*, *18*(1), 51–62.

David Stewart discusses procedures for determining when data are appropriate for factor analysis, as well as guidelines for determining the number of factors to extract, and for rotation.

References

Acock, A. C. (2013). *Discovering structural equation modeling using Stata* (Revised ed.). College Station: Stata Press.

Brown, J. D. (2009). Choosing the right type of rotation in PCA and EFA. *JALT Testing & Evaluation SIG Newsletter, 13*(3), 20–25.

Carbonell, L., Izquierdo, L., Carbonell, I., & Costell, E. (2008). Segmentation of food consumers according to their correlations with sensory attributes projected on preference spaces. *Food Quality and Preference, 19*(1), 71–78.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research, 1*(2), 245–276.

Cliff, N. (1987). *Analyzing multivariate data*. New York: Harcourt Brace Jovanovich.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297–334.

Diamantopoulos, A., & Siguaw, J. A. (2000). *Introducing LISREL: A guide for the uninitiated*. London: Sage.

Dinno, A. (2009). Exploring the sensitivity of Horn's parallel analysis to the distributional form of random data. *Multivariate Behavioral Research, 44*(3), 362–388.

DiStefano, C., Zhu, M., & Mîndriă, D. (2009). Understanding and using factor scores: Considerations fort he applied researcher. *Practical Assessment, Research & Evaluation, 14*(20), 1–11.

Festge, F., & Schwaiger, M. (2007). The drivers of customer satisfaction with industrial goods: An international study. *Advances in International Marketing, 18*, 179–207.

Gorsuch, R. L. (1983). *Factor analysis* (2nd ed.). Hillsdale: Lawrence Erlbaum Associates.

Graffelman, J. (2013). Linear-angle correlation plots: New graphs for revealing correlation structure. *Journal of Computational and Graphical Statistics, 22*(1), 92–106.

Grice, J. W. (2001). Computing and evaluating factor scores. *Psychological Methods, 6*(4), 430–450.

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2013). *Multivariate data analysis. A global perspective* (7th ed.). Upper Saddle River: Pearson Prentice Hall.

Hair, J. F., Ringle, C. M., & Sarstedt, M. (2011). PLS-SEM: Indeed a silver bullet. *Journal of Marketing Theory and Practice, 19*(2), 139–151.

Hair, J. F., Hult, G. T. M., Ringle, C. M., & Sarstedt, M. (2017a). *A primer on partial least squares structural equation modeling (PLS-SEM)* (2nd ed.). Thousand Oaks: Sage.

Hair, J. F., Hult, G. T. M., Ringle, C. M., Sarstedt, M., & Thiele, K. O. (2017b). Mirror, mirror on the wall. A comparative evaluation of composite-based structural equation modeling methods. *Journal of the Academy of Marketing Science, 45*(5), 616–632.

Hair, J. F., Sarstedt, M., Ringle, C. M., & Gudergan, S. P. (2018). *Advanced issues in partial least squares structural equation modeling (PLS-SEM)*. Thousand Oaks: Sage.

Hamilton, L. C. (2013), Statistics with Stata: Version 12: Cengage Learning.

Hayton, J. C., Allen, D. G., & Scarpello, V. (2004). Factor retention decisions in exploratory factor analysis: A tutorial on parallel analysis. *Organizational Research Methods, 7*(2), 191–205.

Henson, R. K., & Roberts, J. K. (2006). Use of exploratory factor analysis in published research: Common errors and some comment on improved practice. *Educational and Psychological Measurement, 66*(3), 393–416.

Hershberger, S. L. (2005). Factor scores. In B. S. Everitt & D. C. Howell (Eds.), *Encyclopedia of statistics in behavioral science* (pp. 636–644). New York: John Wiley.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30*(2), 179–185.

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika, 36*(4), 409–426.

Kaiser, H. F. (1958). The varimax criterion for factor analytic rotation in factor analysis. *Educational and Psychological Measurement, 23*(3), 770–773.

Kaiser, H. F. (1974). An index of factorial simplicity. *Psychometrika, 39*(1), 31–36.

Kim, J. O., & Mueller, C. W. (1978). *Introduction to factor analysis: What it is and how to do it.* Thousand Oaks: Sage.

Longman, R. S., Cota, A. A., Holden, R. R., & Fekken, G. C. (1989). A regression equation for the parallel analysis criterion in principal components analysis: Mean and 95th percentile eigenvalues. *Multivariate Behavioral Research, 24*(1), 59–69.

MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods, 4*(1), 84–99.

Matsunga, M. (2010). How to factor-analyze your data right: Do's and don'ts and how to's. *International Journal of Psychological Research, 3*(1), 97–110.

Mulaik, S. A. (2009). *Foundations of factor analysis* (2nd ed.). London: Chapman & Hall.

Preacher, K. J., & MacCallum, R. C. (2003). Repairing Tom Swift's electric factor analysis machine. *Understanding Statistics, 2*(1), 13–43.

Russell, D. W. (2002). In search of underlying dimensions: The use (and abuse) of factor analysis in Personality and Social Psychology Bulletin. *Personality and Social Psychology Bulletin, 28*(12), 1629–1646.

Sarstedt, M., Schwaiger, M., & Ringle, C. M. (2009). Do we fully understand the critical success factors of customer satisfaction with industrial goods? Extending Festge and Schwaiger's

model to account for unobserved heterogeneity. *Journal of Business Market Management,* *3*(3), 185–206.

Sarstedt, M., Ringle, C. M., Raithel, S., & Gudergan, S. (2014). In pursuit of understanding what drives fan satisfaction. *Journal of Leisure Research, 46*(4), 419–447.

Sarstedt, M., Hair, J. F., Ringle, C. M., Thiele, K. O., & Gudergan, S. P. (2016). Estimation issues with PLS and CBSEM: Where the bias lies! *Journal of Business Research, 69*(10), 3998–4010.

Steiger, J. H. (1979). Factor indeterminacy in the 1930's and the 1970's some interesting parallels. *Psychometrika, 44*(2), 157–167.

Stevens, J. P. (2009). *Applied multivariate statistics for the social sciences* (5th ed.). Hillsdale: Erlbaum.

Velicer, W. F., & Jackson, D. N. (1990). Component analysis versus common factor analysis: Some issues in selecting an appropriate procedure. *Multivariate Behavioral Research, 25*(1), 1–28.

Vigneau, E., & Qannari, E. M. (2002). Segmentation of consumers taking account of external data. A clustering of variables approach. *Food Quality and Preference, 13*(7–8), 515–521.

Widaman, K. F. (1993). Common factor analysis versus principal component analysis: Differential bias in representing model parameters? *Multivariate Behavioral Research, 28*(3), 263–311.

Wold, H. O. A. (1982). Soft modeling: The basic design and some extensions. In K. G. Jöreskog & H. O. A. Wold (Eds.), *Systems under indirect observations: Part II* (pp. 1–54). Amsterdam: North-Holland.

Zwick, W. R., & Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological Bulletin, 99*(3), 432–442.