Factor Analysis

Jorge Luis García[*,†]

First Draft: May 29, 2014
This Draft: October 24, 2014

Abstract

The objective of this document is to describe a step-by-step methodology to extract factors, or underlying latent variables, from a set of observed measures. Factor analysis is arbitrary by construction. Thus, I intend to provide the exact steps for factor analyzing a set of measures and elect a consistent way of being arbitrary. The theoretical fundamentals of this document come from Gorsuch (1983). Although this author (or any other author) does not suggest any method over the other, I justify why the methods I elect are simple and transparent. Finally, I provide implementations in Python and Stata in separate files.

1 Diagonal Factor Analysis

1.1 Setting

Let $v = 1, \ldots, V$ index the number of measures and $f = 1, \ldots, F$ the number of factors in the sets $\mathcal{V}, \mathcal{F}$, with cardinaliies V and F, respectively. Let $n = 1, \ldots, I$ index the individuals observations in the set $\mathcal{I}$ with cardianlity I. X_{iv} is *observed* measure for individual i. It loads on the factors $F_{i1}, \ldots, F_{iF}$ and the factor loadings are the real numbers $w_{v1}, \ldots, w_{vF}$. Thus, the measurement system for individual i is

$$\begin{aligned} X_{i1} &= w_{11}F_{i1} + \cdots + w_{1F}F_{iF} + \eta_{i1} \\ \vdots &= \ddots \\ X_{iV} &= w_{V1}F_{i1} + \cdots + w_{VF}F_{iF} + \eta_{iV} \end{aligned} \tag{1}$$

$$\tag{2}$$

where η_{iv} is measurement error in the measure v of individual i and satisfies the mean independence property, i.e. $\mathbb{E}(\eta_{iv}|F_{if}) = 0$ for $f = 1, \ldots F$.

We observe the LHS of (1) as a set of measures and we do not observe the RHS. The objective of factor analysis is to recover $F_{i1}, \ldots, F_{iF}$, i.e. to recover a set of underlying scores defining the measures $X_{i1}, \ldots, X_{iV}$. This with the objective of, for example, obtaining a reduction in the dimension of the problem at hand.[1]

[*]Department of Economics, the University of Chicago (jorgelgarcia@uchicago.edu).

[†]I thank Sneha Elango, Tim Kautz, and Bradley Setzler for helpful comments.

[1]An example of this is Bernal and Keane (2011). They have a set of instruments shifting women's labor supply in the US. Given that welfare rules are the base of their instruments and the welfare system in the US is complex, the set of instruments is very wide and they face a "many instruments problem", i.e. 2SLS estimates are biased towards OLS estimates when the number of overidentifying instruments is large (see Stock and Yogo, 2002; Andrews

1.2 Notation

I define the basic notation throughout this document in the following lines. I do not assume that, in general, all the measures or all the factors or all the individuals are considered. That is a particular case when $n = N, f = F, v = V$ in some of the equations below, which are straightforward to recognize.[2]

1. Measures matrix (in deviation):
   The data matrix containing v measures for n individuals *in deviations from the mean* is X_{nv} .

2. Measures matrix (standardized):
   The data matrix containing v measures for n individuals *in standardized form* is Z_{nv}.

3. Factor score matrix (standardized):
   The factor score matrix containing f common factor scores for n individuals *in standardized form* is F_{nf}.

4. Factor loadings matrix:
   The factor loadings matrix containing the v weights for f factors "to recover measures from factor scores" (in a full components model in the absence of measurement error) is P_{vf}. The factor loadings vector containing the v loadings for f factors is P_{vf}.

5. Measurement error matrix:
   The measurement error matrix containing v error terms for n individuals is U_{nv}.

6. Measurement error weights:
   The measurement error weights containing the v weights for v equations is D_{vv}. In this document I assume that this matrix is equal to the identity matrix of size v, I_{vv}.

7. Matrix system:
   The standardized measurement system for n individuals, v measures, and f factors is

$$Z_{nv} = F_{nf} P'_{fv} + U_{nv} D'_{vv}. \tag{3}$$

8. Covariance matrix of the measurement system (in deviation):

$$C_{vv} := \frac{1}{N} X'_{vn} X_{nv}. \tag{4}$$

9. Correlation matrix of the measurement system (in deviation):

$$R_{vv} := S_{vv}^{-1} C_{vv} S_{vv}^{-1}. \tag{5}$$

   where S_{vv}^{-1} is a diagonal matrix and contains the standard deviation of measurement v in entry vv. Importantly, $R_{vv} = \frac{1}{N} Z'_{vn} Z_{nv}$. Thus, S_{vv}^{-1} allows to go from X_{nv} to Z_{nv}.

and Stock, 2007; Hansen et al., 2008; Anderson et al., 2010). Thus, they factor analyze their set of instruments and argue that their 2SLS estimates are similar to LIML estimates, which correct the bias in 2SLS in the case of many instruments (see Hansen et al., 2008).

[2] To ease matrix calculation and interpretation I use a subindex to indicate the dimensions of each matrix.

U_{nv} could either be a factor that is dedicated to one measure or measurement error. These are indistinguishable from the perspective of the statistician who extracts factors. There is a simplifying procedure that allows us to ignore U_{nv} in the context of Economics, and it is the following: (i) ignore the existence of U_{nv} and assume the complete measurement system is correlated; (ii) extract the factors; (iii) consider the measurement error in measurement system as part of the system for which the factors are inputs. Concretely, a method extracts factor F_{iv} for individual i while the "real" factor is $\hat{F_{iv}} + \eta_{iv}$. A standard treatment of F_{iv} as a variable with measurement error enables us to consider it in the context of regression analysis. This is why henceforth I consider the system

$$Z_{nv} = F_{nf} P'_{fv}. \tag{6}$$

1.3 Residual Factor Analysis

Gorsuch (1983) calls my preferred method *diagonal analysis*. Other literature names it triangular decomposition, sweep-out method, pivotal condensation, solid-staircase analysis, analytic factor analysis, maximal decomposition, or regression component analysis (see Gorsuch, 1983, Chapter 2). In fact, maybe the last name is the one that makes the most sense because the method has as its basic ingredient a fundamental of regression analysis, residual matrices. Residual factor analysis sounds even fancier.

These are the steps to extract the factor loadings of F factors from the measurement system with V measures from N individuals. The method to obtaining the factor scores once the factor loadings are calculated is in Section 1.4, and the method to determining the number of factors to be extracted is discussed in Section 1.5.

1. Pick the first factor: elect one of the measures in the measurement system as the first factor. There are two possibilities for doing this:

   (a) Arbitrary: elect one measure with transparent meaning. In this case the objective is to have a well-known, meaningful measure as first factor.

   (b) Maximum correlation across the measurement system: (i) compute the covariance matrix of the measurement system and square each of its entries; (ii) compute the sum of all its columns; (iii) pick the measurement with the largest sum. In this case the objective is to have the measure that correlates the most with the rest of the measurement system as first factor.

2. Compute the factor loadings for the first factor: the factor loadings for the first factor are the correlation coefficients of the first factor with the variables in the measurement system. Naturally, the factor loading of the first factor with the measure that defines it is 1. Denoting with lower case letters the entries of (5) the factor loadings for the first factor are defined as:

   $$w_{11} := r_{11}, \ldots, w_{V1} := r_{V1}. \tag{7}$$

   These loadings define P_{V1}, i.e. the vector stacking the V loadings of factor 1.

3. Residualize the correlation matrix: obtain the correlation matrix of the measurement system after making it orthogonal to the first factor. Let R^{o1}_{vv} correlation matrix of the measurement system after making it orthogonal to the first factor. Thus

   $$R^{o1}_{VV} = R_{VV} - P_{V1} P'_{V1}. \tag{8}$$

4. Obtain a second factor: repeat steps 1 and 2. Usually, the second factor is going to be chosen based on criterion (b) in step 1 because after making the system orthogonal to the first factor it is difficult to interpret what the measures mean.

5. Obtain factors $3, \ldots, F$: repeat the process making the measurement system orthogonal to factors 1 and 2 in order to obtain factor 3. Likewise, repeat the process making the system orthogonal to factors 1, 2, ..., F-1 to obtain factor F.

1.4 Obtaining the Factor Scores

In the case when we have N individuals, V measures, and F factors the measurement system is

$$Z_{NV} = F_{NF} P'_{FV}. \tag{9}$$

A simple manipulation of (9) leads to

$$F_{NF} = Z_{NV} P_{VF} \left(P'_{FV} P_{VF} \right)^{-1}, \tag{10}$$

which solves for the factor scores.

1.5 How many Factors?

I follow Guttman (1954) and elect the lower bound of the number of factors based on a simple rule of thumb. The steps are the following.

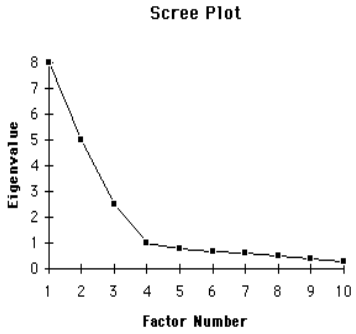
1. Calculate the absolute value of the eigenvalues of the correlation matrix of the measurement system.

2. Count the number of eigenvalues greater or equal than one.

3. Define the lower bound of the number of factors as the number of eigenvalues greater than one.

Once the lower bound is defined, as many eigenvalues as independent measures in the measurement system may be calculated. These are the inputs of the *subjective scree test*. The scree test is an eye ball test of a scatter in which the abscissas are the factor numbers and the ordinates are their corresponding eigenvalues. The lower bound of the number of factors is the actual number of factors if the plot shows a very clear pattern differentiating the factors with eigenvalues greater or equal than one from the rest. In Figure 1 the lower bound of the number of factors and the actual number of factors are the same according to my criterion because the factors with eigenvalues less than one follow a different pattern from the rest.

Figure 1: A Scree Test where the Lower Bound and the Current Number of Factors Agree



1.6 Other Annotations

1.6.1 Correcting for Attrition

It is very common for economists to correct for attrition. This is, to correct for the fact that some variables are not observed for certain individuals. A usual way to do it is to estimate a model predicting the probability of attrition based on observed characteristics. For example, if income is an outcome of interest and the researcher does not observe income for a subset of the sample but has observed characteristics for the complete sample, she can predict the probability of attrition. Then, she can use a method such as inverse probability weighting (IPW) to give greater relative weight to the observations that are more likely to have attrition (see Wooldridge, 2007).

Provided the estimated model predicts attrition, it is easy to consider an IPW scheme in which factors are extracted using the method in Section 1.3 –it is sufficient to consider the IPW scheme when calculating the correlation matrices.

1.6.2 Allowing for Correlated Factors

By construction, the method in Section 1.3 does not allow factors to be correlated. Sometimes, however, economic theory or intuition suggest that two or more sets of measurements should be considered. It is possible to apply the process in Section 1.3 to two different sets of measurements independently. If the first factor of the two sets of measurements are correlated, this procedure preserves the correlation between the two first factors. When extracting the rest of the factors for each set of measurement systems, one can make the system orthogonal to the first factor of both systems as in step 3. Thus, the first factors of the two systems will be correlated while the rest of the factors will not be correlated within or across measurement systems.[3]

2 Why Rotation Makes sense and How to Go about it?

In Section 1.6.2 I discuss how to allow for correlated factors. The procedure implies being certain about having two sets of measurement systems clearly devoted to factors of interest. For example,

[3]I thank Tim Kautz for pointing this out. His example is the following. Assume the researcher has two measurements systems: one for height and one for weight. It makes sense to allow correlation for the "primary" measures of height and weight, which would be the first factors in this case, because it is natural for height and weight to be correlated. Then, it is possible to make the rest of the systems orthogonal because the researcher is only willing to capture extra variation or information from the measures.

a measurement system could be devoted to "weight" and another to "height". The researcher arbitrarily chooses the measures with which she constructs the "weight" factor and the measures with which she constructs the "height" factor.

It could be the case that the researcher does not want to take any stand on what the measurement systems are. This could happen for two reasons: (i) the researcher does not want to take arbitrary stands on what the dedicated measurement systems are; (ii) the researcher has no idea on what the dedicated measurement systems are.

If this is the case, the alternative method is the following: (i) decide on a set of items composing the measurement system; (ii) set the number of factors (e.g., through a procedure like the one in 1.5); (iii) factor analyze the measures (e.g., through a procedure like the one in 1.3); (iv) rotate the factor axes to gain interpretability of the factors.

The first three steps are clear: the researcher decides what the measurement system is, decides the number of factor she wants to use, and factor analyses the measures according to her preferred method. Rotation is convenient in this context. In general, after factor analyzing a set of measures, the output does not have an intuitive structure. Most items load on the first few factors that explain the greatest proportion of variance. Rotation is a linear transformation intending to give a "simple structure" to the factor system. In rough terms, it seeks a structure in which items load strongly on one factor and weakly in the rest of the factor.[4] This is the way in which the researcher may be able to associate each factor to an object of economic interest, "height", "weight", etc.

References

Abdi, H. (2003). Factor Rotations in Factor Analyses. *Encyclopedia for Research Methods for the Social Sciences. Sage: Thousand Oaks, CA*, 792–795.

Anderson, T. W., N. Kunitomo, and Y. Matsushita (2010). On the Asymptotic Optimality of the LIML Estimator with Possibly many Instruments. *Journal of Econometrics 157*(2), 191–204.

Andrews, D. W. and J. H. Stock (2007). Testing with Many Weak Instruments. *Journal of Econometrics 138*(1), 24–46.

Bernal, R. and M. P. Keane (2011). Child Care Choices and Childrens Cognitive Achievement: the Case of Single Mothers. *Journal of Labor Economics 29*(3), 459–512.

Gorsuch, R. L. (1983). *Factor Analysis*. Lawrence Erlbaum Associates Publishers.

Guttman, L. (1954). Some Necessary Conditions for Common Factor Analysis. *Psychometrika 19*(2), 149–161.

Hansen, C., J. Hausman, and W. Newey (2008). Estimation with many Instrumental Variables. *Journal of Business & Economic Statistics 26*(4).

Stock, J. H. and M. Yogo (2002). Testing for Weak Instruments in Linear IV Regression.

Wooldridge, J. M. (2007). Inverse Probability Weighted Estimation for General Missing Data Problems. *Journal of Econometrics 141*(2), 1281–1301.

[4] Abdi (2003) explains the mathematical requirements for a structure to be simple and the intuition behind them.