10 Factor Analysis

Factor Analysis has two primary motivations. The **first** motivation stems from the fact that interpretation of principal components is often not transparent. A particular variable may on occasion, contribute significantly to more than one of the components (recall the wine example). Ideally we prefer each variable to contribute significantly to only one component. So one of the purposes of factor analysis is to try and clean up the interpretation of the data using something called factor rotation.

The next (and probably **primary**) motivation for factor analysis has to do with the notion that the data that we observe are somehow a function of some smaller number of unobserved variables called **factors**, that cannot be measured directly and that usually cannot be measured by a single variable (e.g. happiness). I.e., as for principal components analysis, factor analysis is a multivariate method used for data reduction purposes.

The Orthogonal Factor Model can be written algebraically as follows. If we have p variables $\mathbf{X} = (X_1, X_2, \ldots, X_p)'$ measured on a sample of n subjects, then variable X_i can be written as a linear combination of m **common factors** (**latent** variables) $F_1, F_2, \ldots, F_m$ where $m \leq p$. Hence,

$$\begin{aligned} X_1 &= l_{11}F_1 + l_{12}F_2 + \ldots + l_{1m}F_m + v_1 \\ X_2 &= l_{21}F_1 + l_{22}F_2 + \ldots + l_{2m}F_m + v_2 \\ &\ldots \\ X_p &= l_{p1}F_1 + l_{p2}F_2 + \ldots + l_{pm}F_m + v_p, \end{aligned}$$

where the l_i is the factor loading for variable i, v_i is the part of variable X_i that cannot be 'explained' by the factors and v_i affects only X_i.

In a matrix form, 10.1 can be represented as

$$X_{p\times 1} - \boldsymbol{\mu}_{p\times 1} = L_{p\times m}F_{m\times 1} + \mathbf{v}_{p\times 1}.$$

Further, we assume without much loss of generality that

$$E(F) = 0_{m\times 1}, \quad Cov(F) = I_{m\times m}, \quad E(\mathbf{v}) = 0_{p\times 1}, \quad Cov(\mathbf{v}, F) = 0_{p\times m},$$

and with serious loss of generality

$$Cov(\mathbf{v}) = \Psi = diag(\psi_1, \psi_2, \ldots, \psi_p).$$

In terms of the observable variables $\mathbf{X}$, these assumptions mean that

$$\begin{aligned} E(\mathbf{X}) &= \boldsymbol{\mu} \\ Cov(\mathbf{X}) &= \Sigma = L_{p\times m}L_{m\times p}^T + \Psi_{p\times p}, \end{aligned}$$

and usually $\mathbf{X}$ is standardized so $\Sigma = R$.

The observable $\mathbf{X}$ and the unobservable F are related by $Cov(\mathbf{X}, F) = L$.

Note that (i,i)-entry of this matrix equation $\Sigma = L_{p\times m}L_{m\times p}^T + \Psi_{p\times p}$ takes the form

$$\sigma_{i,i} = h_i^2 + \psi_i = \sum_{j=1}^{m} l_{ij}^2 + \psi_i = \text{common part} + \text{unique part},$$

where h_i^2 is the i-th **communality** and ψ_i is the i-th **specific variance.**

Note that if $T_{m\times m}$, then $(LT)(LT)^T = LL^T$, so loadings LT generate the same Σ as L, and hence loadings are not unique.

For any p, every $\Sigma_{p\times p}$ can be factorized as $\Sigma = LL^T$ for some $L_{p\times p}$, which is a factor representation with $m = p$ and $\Psi = 0$. However, $m = p$ is not of much use for us as we usually want $m \ll p$. Note we can only *approximate* Σ by $LL^T + \Psi$.

Now we consider the main steps in a factor analysis:

1. **Fit the Factor Analysis Model**

   (a) **Principal Components Factor Analysis** The spectral decomposition of Σ is

   $$\Sigma = E\Lambda E^T = \left(E\Lambda^{1/2}\right)\left(E\Lambda^{1/2}\right)^T,$$

   with $m = p$. If $\lambda_1 + \lambda_2 + \ldots \lambda_m \gg \lambda_{m+1} + \lambda_{m+2} + \ldots \lambda_p$ and $L^{(m)}$ is the first m columns of L then

   $$\Sigma \approx L^{(m)}L^{(m)T},$$

   with $\Psi = 0$.

   The remainder term $\Sigma - L^{(m)}L^{(m)T}$ is non-negative definite, so its diagonal entries are non-negative. Hence, we can get a better approximation if we consider

   $$\Sigma \approx L^{(m)}L^{(m)T} + \Psi^{(m)},$$

   where $\Psi^{(m)} = diag\left(\Sigma - L^{(m)}L^{(m)T}\right)$.

Note the **Principal Components Factor Analysis** method attempts to explain all of the variance in the variables and not just the common variance! It therefore will often have highly correlated errors. However, Hair et.al. (1998) reports that it often gives similar results to other methods if there are more than 30 variables or if most variables have communalities > 0.6.

(b) **Principal Factor Solution** The Orthogonal Factor Model is given by

$$\mathbf{X} = LF + \mathbf{v},$$

which implies $\Sigma = LL^T + \Psi$. Hence, we can use m-factor Principal Component solution to approximate Σ (or, if we standardize the variables, R) by a rank-m matrix using the spectral decomposition

$$\Sigma = \lambda_1 \mathbf{l}_1 \mathbf{l}_1^T + \ldots + \lambda_m \mathbf{l}_m \mathbf{l}_m^T + \lambda_{m+1} \mathbf{l}_{m+1} \mathbf{l}_{m+1}^T + \ldots \lambda_p \mathbf{l}_p \mathbf{l}_p^T.$$

The first m terms give the best rank-m approximation to Σ.

We can sometimes achieve higher communalities ($= diag(LL^T)$) by either:

- specifying an initial estimate of the communalities
- iterating the solution

or both.

Suppose we are working with R. Given initial communalities $h_i^{*,2}$, form the reduced correlation matrix

$$R_r = \begin{bmatrix} h_1^{*,2} & r_{1,2} & \ldots & r_{1,p} \\ r_{2,1} & h_2^{*,2} & \ldots & r_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p,1} & r_{p,2} & \ldots & h_p^{*,2} \end{bmatrix}$$

Now use the spectral decomposition of R_r to find its best rank-m approximation $R_r \approx L_r^{*(m)} L_r^{*(m),T}$. The the new communalities are

$$\tilde{h}_i^{*,2} = \sum_{j=1}^{m} l_{i,j}^{*2}$$

and Ψ is found from equating the diagonal terms:

$$\tilde{\psi}_i^* = 1 - \tilde{h}_i^{*,2}$$

or

$$\tilde{\Psi}^* = I - diag\big(L_r^{*(m)} L_r^{*(m),T}\big).$$

This is the **Principal Factor** solution. Note that the **Principal Component** solution is the special case where the initial communalities are all 1.

(c) **Iterated Principal Factor Factor Analysis** One issue with both Principal Components and Principal Factors is that if S or R is in the form $LL^T + \Psi$, neither method produces L and Ψ (unless we specify the true communalities).

The solution is **to iterate**:

i. Use the new communalities as initial communalities to get another set of Principal Factors.

ii. Repeat until nothing much changes.

(d) **Other methods** are Maximum Likelihood Factor Analysis (it requires the assumption of multivariate normality and is difficult to program but it does allow for various tests of hypotheses), Alpha Factoring, Image Factoring, Harris Factoring, Rao's Canonical Factoring, Unweighted Least Squares.

2. **Rotation**

Hence, once the (initial) factor loadings in step 1 have been calculated, the factors are rotated in order to find factors that are easier to interpret. If there are 'clusters' (groups) of variables, i.e. subgroups of variables that are strongly inter-related, then the rotation is done to try to make variables within a subgroup score as highly (positively or negatively) as possible on one particular factor while, at the same time, ensuring that the loadings for these variables on the remaining factors are as low as possible.

Ideally, we hope to see a pattern of loadings such that each variable loads highly on a single factor and delivers relatively small (or moderate) loadings on the remaining factors which implies that each row of L should have a single large entry (Johnson and Wichern, 2002).

There are two types of rotation method, **orthogonal** and **oblique** rotation. In orthogonal rotation the rotated factors remain uncorrelated whereas in oblique rotation the resulting factors will be correlated. There are a number of different methods of rotation of each type. The most common orthogonal method is called **varimax** rotation.

The **varimax Criterion** Kaiser proposed a criterion that measures interpretability. Let $\hat{L}$ is some set of loadings with communalities $\hat{h}_i^2$, $i = 1, 2, \ldots, p$ and $\hat{L}^*$ be a set of rotated loadings, $\hat{L}^* = \hat{L}T$. Let $\tilde{l}_{i,j}^* = \hat{l}_{i,j}^*/\hat{h}_i$ be scaled loadings.

Then **varimax** procedure selects the rotation to find the maximum of the sum of the squared factor loadings across the columns:

$$V = \frac{1}{p}\sum_{j=1}^{m}\left[\sum_{i=1}^{p}\tilde{l}_{i,j}^{*4} - \frac{1}{p}\left(\sum_{i=1}^{p}\tilde{l}_{i,j}^{*2}\right)^2\right].$$

This is the sample variances of the standardized loadings for each factor, summed over the m factors. The objective is to find a factor rotation that maximizes this variance.

The **Quartimax Criterion** Whereas **varimax** focuses on the columns, **quartimax** focuses on the rows.

Other popular rotation methods include **Promax**, **Oblimin**, and **Orthoblique**.

3. **Calculation of factor scores** When calculating the final factor scores (the values of the m factors, $F_1, F_2, \ldots, F_m$, for each observation), a decision needs to be made as to how many factors to include. This is usually done using one of the following methods:

   (a) Choose m such that the factors account for a particular percentage (e.g. 75%) of the total variability in the original variables.

   (b) Choose m to be equal to the number of eigenvalues over 1 (if using the correlation matrix).

   (c) Use the scree plot of the eigenvalues. This indicates whether there is an obvious cut-off between large and small eigenvalues.

4. **Interpretation** Once we have the factor loadings matrix, it is necessary to try and interpret the factors. It is common to indicate which of the loadings are actually significant by underlining or circling them (and possibly erasing the non-significant ones). Significant is measured in two ways.

   (a) **Practical Significance** Are the factor loadings large enough so that the factors actually have a meaningful effect on the variables. Hair et.al. (1998) recommends the following guidelines for practical significance:

- ± 0.3 Minimal
- ± 0.4 More Important
- ± 0.5 Practically Significant

(b) **Statistical Significance** We also would like the loading to be statistically significantly different from zero.

(c) **Further Guidance** Hatcher (1994) reports that we would like at least 3 variables loading on each factor and preferably more. Finally, it seems reasonable if you are using varimax to remove variables which load heavily on more than one factor

5. **Factor Scores**

Given a set of n observed p-dimensional vectors following the factor model:

$$\mathbf{Y}_i = \boldsymbol{\mu} + LF(i) + \mathbf{v}(i), \quad i = 1, 2, \ldots, n$$

we may wish to estimate the vectors of factor scores $\hat{F}(1), \hat{F}(2), \ldots, \hat{F}(n)$ for each observation.

There are a number of different methods that can be used for estimating factor scores from the data. These include:

(a) **Ordinary Least Squares** Recall that

$$X_{p\times 1} - \boldsymbol{\mu}_{p\times 1} = L_{p\times m} F_{m\times 1} + \mathbf{v}_{p\times 1},$$

and suppose that L is known. Then we can view this equation as a regression of $X_{p\times 1}$ on L, with coefficients F.

Hence, we can find the vector of common factors $\hat{F}$ by minimizing the sum of the squared residuals:

$$\sum_{j=1}^{p} (y_{ij} - \mu_j - l_{j1}F_1 - l_{j2}F_2 - \ldots l_{jm}F_m)^2 = (\mathbf{Y}_i - \boldsymbol{\mu} - LF)^T (\mathbf{Y}_i - \boldsymbol{\mu} - LF),$$

which translates to a case of finite samples to

$$\hat{F}(i) = \left(\hat{L}^T \hat{L}\right)^{-1} \hat{L}^T (\mathbf{Y}_i - \bar{\mathbf{Y}})$$

(b) **Bartlett's Weighted Least Squares**

Alternatively, we can account for heterogeneity of $\mathbf{v}$ with variance matrix Ψ and run a weighted least squares:

$$\hat{F}(i) = \left(\hat{L}^T \hat{\Psi}^{-1} \hat{L}\right)^{-1} \hat{L}^T \hat{\Psi}^{-1} (\mathbf{Y}_i - \bar{\mathbf{Y}})$$

(c) **Regression method**

This method is used when we employ maximum likelihood estimates to estimate factor loadings. If $\mathbf{X}$ and F have a joint multivariate normal distribution, then the conditional distribution of F given $\mathbf{X}$ is also multivariate normal. Hence, the Best Linear Unbiased Predictor is the conditional mean

$$\hat{F}(i) = \left(I + \hat{L}^T \hat{\Psi}^{-1} \hat{L}\right)^{-1} \hat{L}^T \hat{\Psi}^{-1} (\mathbf{Y}_i - \bar{\mathbf{Y}}).$$

Notice that the Bartlett Weighted Least Squares and the MLE based regression methods are related via:

$$\hat{F}(i)^{WLS} = \left[I + \left(\hat{L}^T \hat{\Psi}^{-1} \hat{L}\right)^{-1}\right] \hat{F}(i)^{Reg}$$

6. **Some Final Steps** Johnson and Wichern (2002) suggest the following to see if your solution seems reasonable: plot the factor scores against each other to look for suspicious observations and for large data sets, split them in half and perform factor analysis on each half to see if the solution is stable (i.e. crossvalidation).

Example Consider 300 hypothetical responses on 6 items from a survey of college students' favorite subject matter. The items range in value from 1 to 5, which represent a scale from Strongly Dislike to Strongly Like. Students to rate their liking of different college subject matter areas, including biology (BIO), geology (GEO), chemistry (CHEM), algebra (ALG), calculus (CALC), and statistics (STAT) (see J.M. Quick, 2011). The question we would like to answer is whether there are latent variables underlying the students responses.

Let us start from a simple assessment of a number of factors (this can be revised later). As we already studied, one way to determine the number of factors or components in a data matrix or a correlation matrix is to examine the 'scree' plot of the successive eigenvalues. '**Parallel**' analyis is an alternative technique that compares the scree of factors of the observed data with that of a random data matrix of the same size as the original.

```
> library(psych)
> fa.parallel(data)
Loading required package: parallel
Parallel analysis suggests that the number of factors =  2
and the number of components =  2
```

Another possible routine is the Very Simple Structure (VSS) criterion that allows us to compare solutions of varying complexity and for different number of factors. Graphic output indicates the "**optimal**" number of factors for different levels of complexity. Criteria include BIC, mean residual, χ^2-test, Wayne Velicer's Minimum Average Partial (MAP) criterion etc.

```
> my.vss=vss(data, n.obs=dim(data)[1])
> my.vss

Very Simple Structure
Call: vss(x = data, n.obs = dim(data)[1])
VSS complexity 1 achieves a maximimum of 0.87  with  3  factors
VSS complexity 2 achieves a maximimum of 0.93  with  3  factors

The Velicer MAP achieves a minimum of NA  with  2  factors
BIC achieves a minimum of  NA  with  2  factors
Sample Size adjusted BIC achieves a minimum of  NA  with  2  factors

Statistics by number of factors
  vss1 vss2  map dof   chisq    prob sqresid  fit RMSEA BIC SABIC complex  eChisq    eRMS
1 0.60 0.00 0.23   9 3.5e+02 2.8e-69    4.62 0.60  0.36 296 324.1     1.0 5.1e+02 2.4e-01
2 0.87 0.92 0.11   4 2.9e+00 5.7e-01    0.92 0.92  0.00 -20  -7.2     1.1 9.6e-01 1.0e-02
3 0.87 0.93 0.24   0 1.5e+00      NA    0.80 0.93    NA  NA    NA     1.1 5.9e-01 8.1e-03
4 0.80 0.91 0.46  -3 8.1e-10      NA    0.61 0.95    NA  NA    NA     1.3 1.4e-10 1.2e-07
5 0.71 0.90 1.00  -5 0.0e+00      NA    0.59 0.95    NA  NA    NA     1.6 9.8e-18 3.3e-11
6 0.71 0.90   NA  -6 0.0e+00      NA    0.59 0.95    NA  NA    NA     1.6 1.5e-26 1.3e-15
  eBIC
1  458
```

```
2  -22
3   NA
4   NA
5   NA
6   NA
```

We can also plot the results:

```
> plot(my.vss)
```

The idea of the Very Simple Structure (VSS) is to compare the original correlation matrix to that reproduced by a simplified version (S) of the original factor matrix (F), i.e. $R = SS' + U$, where S is composed of just the C greatest (in absolute value) loadings for each variable. C (or complexity) is a parameter of the model and may vary from 1 to the number of factors.

The VSS criterion compares the fit of the simplified model to the original correlations

$$VSS = 1 - sumsquares(r*)/sumsquares(r)$$

where R^* is the residual matrix $R^* = R - SS'$ and r^* and r are the elements of R^* and R respectively.

Our conclusion is that it is probably enough to look at only 2 factors. We now turn to estimation of these factors (including rotation methods).

```
> FA1=factanal(data,2,rotation="varimax")
> FA1=factanal(data,2,rotation="varimax")
> FA1

Call:
factanal(x = data, factors = 2, rotation = "varimax")

##### this is specific variances for each variable,
##### i.e. Psi
Uniquenesses:
  BIO   GEO  CHEM   ALG  CALC  STAT
```

```
0.252 0.375 0.249 0.374 0.048 0.715

Loadings:
     Factor1 Factor2
BIO  0.855   0.133
GEO  0.779   0.135
CHEM 0.865
ALG          0.791
CALC         0.971
STAT 0.170   0.506

                Factor1 Factor2
SS loadings       2.124   1.863
Proportion Var    0.354   0.311
Cumulative Var    0.354   0.665

Test of the hypothesis that 2 factors are sufficient.
The chi square statistic is 2.94 on 4 degrees of freedom.
The p-value is 0.568
```

Based on this output, we see a clear difference between Science vs Math!

Now let us see how different the results will be with a different number of factors:

```
> FA3=factanal(data,3,rotation="varimax")
> FA3

Call:
factanal(x = data, factors = 3, rotation = "varimax")

Uniquenesses:
  BIO   GEO  CHEM   ALG  CALC  STAT
```

```
0.253 0.375 0.242 0.005 0.155 0.677


Loadings:
     Factor1 Factor2 Factor3
BIO   0.850   0.153
GEO   0.775   0.151
CHEM  0.868
ALG           0.696   0.714
CALC          0.891   0.208
STAT  0.157   0.545

                Factor1 Factor2 Factor3
SS loadings       2.112   1.624   0.557
Proportion Var    0.352   0.271   0.093
Cumulative Var    0.352   0.623   0.715

The degrees of freedom for the model is 0 and the fit was 0.0045
```

Conclusions?

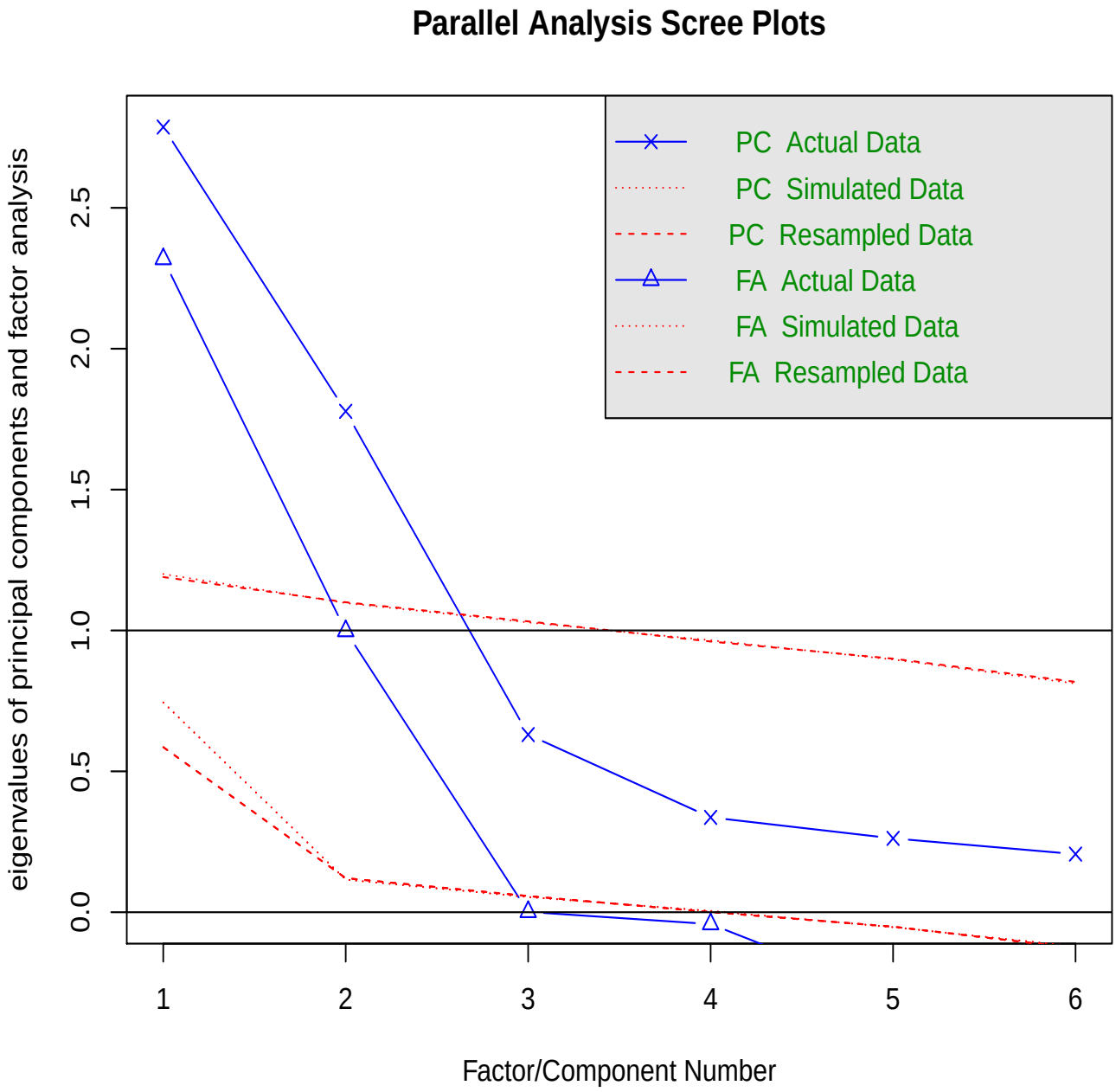

Figure 10.1: Parallel analysis plot for the student survey data.

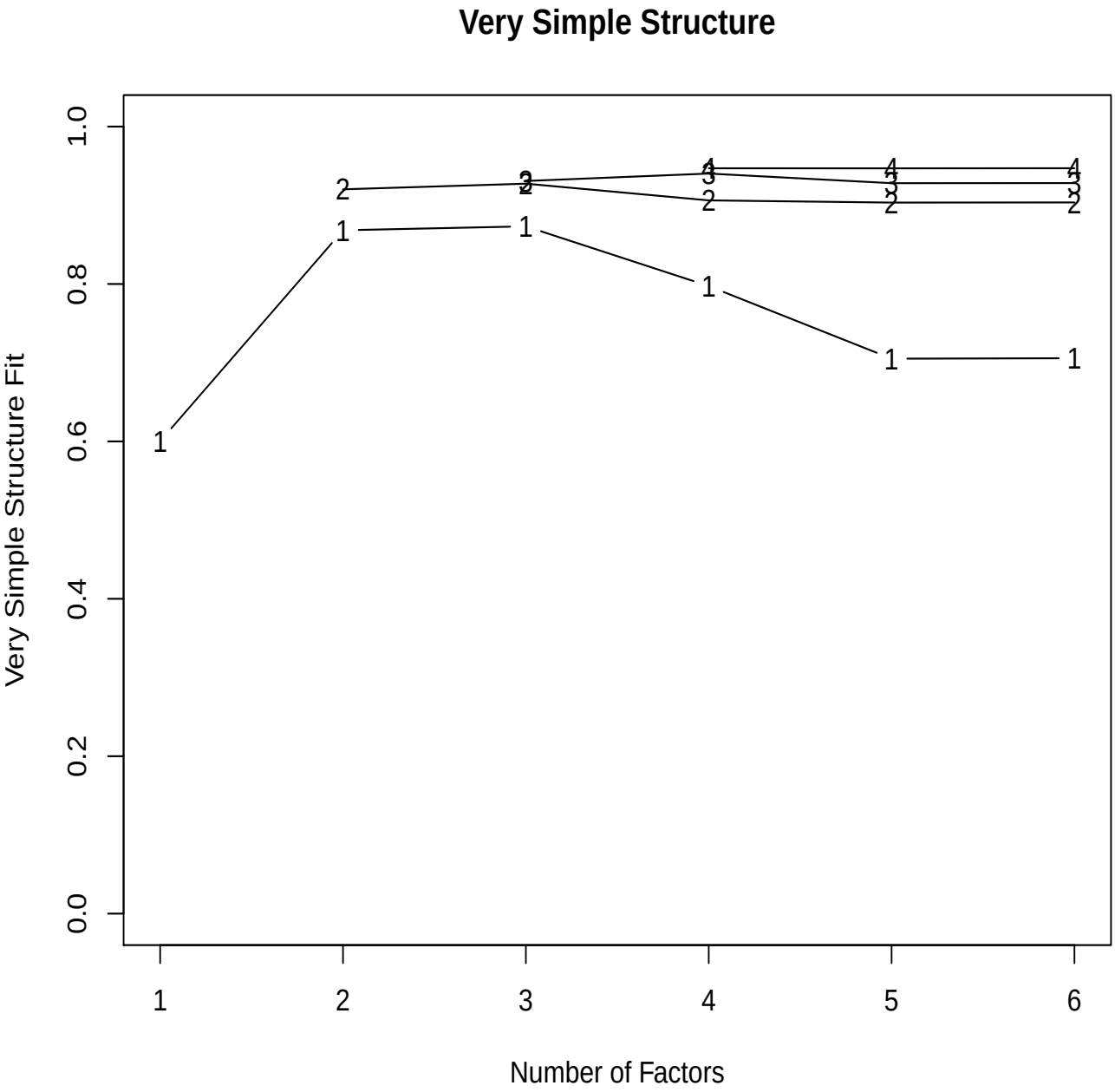

Figure 10.2: Very Simple Structure (VSS) plot for the student survey data.