FACTOR ANALYSIS

Introduction:

- Factor Analysis (EFA) is a technique which allows us to reduce a large number of correlated variables to a smaller number of 'super variables'. It does this by attempting to account for the pattern of correlations between the variables in terms of a much smaller number of **latent variables** or **factors.** A latent variable is one that cannot be measured directly, but is assumed to be related to a number of measurable, observable variables.

- These factors can be either **orthogonal** (independent and uncorrelated) or **Oblique** (they are correlated and share some variance between them). FA is used when we want to understand the relationships between a set of variables and to summarise them, rather than whether one variable has a significant effect on another.

Statistics Associated with Factor Analysis:

- **Factor model:** $X_i = A_{i1}F_1 + A_{i2}F_2 + A_{i3}F_3 + \ldots + A_{im}F_m + U_i$
- **Correlation matrix.** A correlation matrix is a lower triangle matrix showing the simple correlations, r, between all possible pairs of variables included in the analysis. The diagonal elements are all 1.
- **Communality.** Amount of variance a variable shares with all the other variables. This is the proportion of variance explained by the common factors.
- **Eigenvalue.** Represents the total variance explained by each factor.
- **Factor loadings.** Correlations between the variables and the factors.
- **Factor matrix.** A factor matrix contains the factor loadings of all the variables on all the factors
- **Factor scores.** Factor scores are composite scores estimated for each respondent on the derived factors, $F_i = W_{i1}X_1 + W_{i2}X_2 + W_{i3}X_3 + \ldots + W_{ik}X_k$
- **Percentage of variance.** The percentage of the total variance attributed to each factor.
- **Scree plot.** A scree plot is a plot of the Eigenvalues against the number of factors.

Steps in Conducting FA:

- Problem formulation
- Construct the correlation matrix
- Method of estimating the factors (ie: PCA or MLE)
- Determine the number of factors
- Rotation of factors
- Interpret the factors
- Calculate the factor scores

Example of FA with one factor:

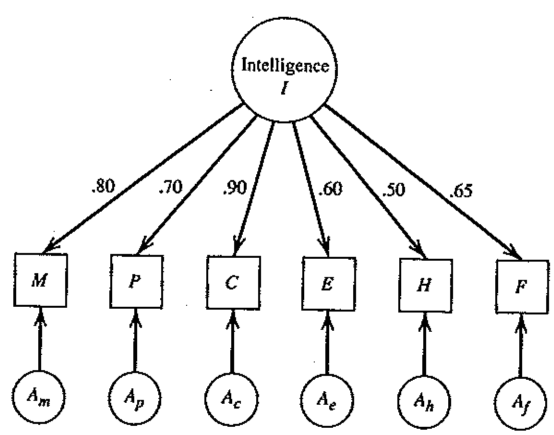
Suppose we have students' test scores (grades) for the following courses: Mathematics (M), Physics (P), Chemistry (C), English (E), History (H), and French (F). Further assume that students' performances in these courses are a function of their general intelligence level, I. In addition, it can be hypothesized that students' aptitudes for the subject areas could be different. That is, a given student may have a greater aptitude for, say, math than French. Therefore, it can be assumed that a student's grade for any given course is a function of:

1. The student's general intelligence level; and
2. The student's aptitude for a given course (i.e., the specific nature of the subject area).

For example, consider the following equations:

$$\begin{aligned} M &= .80I + A_m; & P &= .70I + A_p \\ C &= .90I + A_c; & E &= .60I + A_e \\ H &= .50I + A_h; & F &= .65I + A_f. \end{aligned} \tag{5.1}$$

It can be seen from these equations that a student's performance on any given course, say math, is a linear function or combination of the general intelligence level, I, of the student, and his/her aptitude, A_m, for the specific subject, math. The coefficients (i.e., .8, .7, .9, .6, .5, and .65) of the above equations are called *pattern loadings*.



Relationship between grades and intelligence.

Figure 1

Things to notice:

- (Variance of any variable) = (Communality) + (error variance)

  where Communality = square of pattern loading
- Correlation between *variable* and *factor* is called the loading.
- Correlation between two variables is given by their respective pattern loadings. Thus, the correlation between two variables (indicators) is due to the common factor (I)

Table 1 **Communalities, Pattern and Structure Loadings, and Correlation Matrix for One-Factor Model**

Communalities

Variable	Communality	Error or Unique Variance	Pattern Loading	Structural Loading	Shared Variance
M	.640	.360	.800	.800	.640
P	.490	.510	.700	.700	.490
C	.810	.190	.900	.900	.810
E	.360	.640	.600	.600	.360
H	.250	.750	.500	.500	.250
F	.423	.577	.650	.650	.423
Total	2.973	3.027			2.973

Correlation Matrix for One-Factor Model

	M	P	C	E	H	F
M	1.000					
P	.56	1.000				
C	.72	.63	1.000			
E	.48	.42	.54	1.000		
H	.40	.35	.45	.30	1.000	
F	.52	.46	.59	.39	.33	1.000

Objective of FA:

1. Identify the common factor that is responsible for the correlations among the indicators; and
2. Estimate the pattern and structure loadings, communalities, shared variances, and the unique variances.

In other words, the objective of factor analysis is to obtain the structure presented in Figure 5.1 and Table 5.1 using the correlation matrix. That is, the correlation matrix is the input for the factor analysis procedure and the outputs are the entries in Table 5.1.

Another Example of FA with Two Factors:

Suppose from the previous example that students grades area function of two (non correlated) factors, call them Q and V. The 2-factor model may be represented as follows:

$$\begin{aligned} M &= .800Q + .200V + A_m; & P &= .700Q + .300V + A_p \\ C &= .600Q + .300V + A_c; & E &= .200Q + .800V + A_e \\ H &= .150Q + .820V + A_h; & F &= .250Q + .850V + A_f. \end{aligned} \qquad (5.2)$$

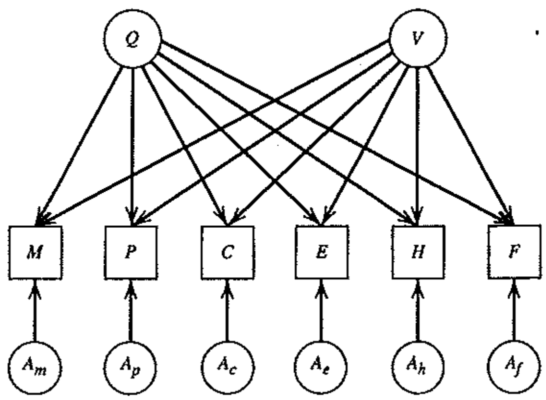


Figure 5.2 Two-factor model.

Things to notice:

- Variance of M = (M's communality with Q) + (M's communality with V) + (error variance)
  = (Total communality of M) + (error variance)

- The coefficient in equation (2) are the pattern loadings which are the correlation between each variable and the factor.

- The correlation between any two indicators is equal to the sum of the products of the respective pattern loadings for each factor. For example,

$$Corr\ (M, H) = (.800*.150) + (.200*.820) = 0.284$$

Table 5.2 Communalities, Pattern and Structure Loadings, and Correlation Matrix for Two-Factor Model

Communalities

Variable	Communalities Q	Communalities V	Communalities Total	Unique Variance
M	.640	.040	.680	.320
P	.490	.090	.580	.420
C	.360	.090	.450	.550
E	.040	.640	.680	.320
H	.023	.672	.695	.305
F	.063	.723	.786	.214
Total	1.616	2.255	3.871	2.129

Pattern and Structure Loadings and Shared Variance

Variable	Pattern Loading Q	Pattern Loading V	Structure Loading Q	Structure Loading V	Shared Variance Q	Shared Variance V
M	.800	.200	.800	.200	.640	.040
P	.700	.300	.700	.300	.490	.090
C	.600	.300	.600	.300	.360	.090
E	.200	.800	.200	.800	.040	.640
H	.150	.820	.150	.820	.023	.672
F	.250	.850	.250	.850	.063	.723
Total					1.616	2.255

Correlation Matrix

	M	P	C	E	H	F
M	1.000					
P	.620	1.000				
C	.540	.510	1.000			
E	.320	.380	.360	1.000		
H	.284	.351	.336	.686	1.000	
F	.370	.430	.405	.730	.735	1.000

<u>Interpretation of the Common Factors</u> From Table 5.2 it can be seen that the communalities or the shared variances of the variables E, H, and F with factor V are much greater than those with factor Q. Indeed, 90.24% $((.640+.672+.723)/2.255)$ of the total communality of V is due to variables E, H, and F. Therefore, one could argue that the common factor, V, measures subjects' verbal abilities. Similarly, one could argue that the common factor, Q, measures subjects' quantitative abilities because 92.20% $((.64 + .49 + .36)/1.616)$ of its communality is due to variables M, P, and C.

Factor Indeterminacy

The FA solution is not unique due to factor rotation problem.

Indeterminacy Due to the Factor Rotation Problem

Consider another two-factor model given by the following equations:

$$\begin{aligned} M &= .667Q - .484V + A_m; \quad & P &= .680Q - .343V + A_p \\ C &= .615Q - .267V + A_c; \quad & E &= .741Q + .361V + A_e \\ H &= .725Q + .412V + A_h; \quad & F &= .812Q + .355V + A_f \end{aligned} \tag{5.4}$$

Table 5.3 **Communalities, Pattern and Structure Loadings, Shared Variances, and Correlation Matrix for Alternative Two-Factor Model**

Communalities

	Communalities			
Variable	Q	V	Total	Unique Variance
M	.445	.234	.679	.321
P	.462	.118	.580	.420
C	.378	.071	.449	.551
E	.549	.130	.679	.321
H	.526	.170	.696	.304
F	.659	.126	.785	.215
Total	3.019	.849	3.868	2.132

Pattern and Structure Loadings and Shared Variance

	Pattern Loading		Structure Loading		Shared Variance	
Variable	Q	V	Q	V	Q	V
M	.667	−.484	.667	−.484	.445	.234
P	.680	−.343	.680	−.343	.462	.118
C	.615	−.267	.615	−.267	.378	.071
E	.741	.361	.741	.361	.549	.130
H	.725	.412	.725	.412	.526	.170
F	.812	.355	.812	.355	.659	.126
Total					3.019	.849

Correlation Matrix

	M	P	C	E	H	F
M	1.000					
P	.620	1.000				
C	.540	.510	1.000			
E	.320	.380	.360	1.000		
H	.284	.351	.336	.686	1.000	
F	.370	.430	.405	.730	.735	1.000

Notice that (within rounding errors), models **5.4** and **5.2** show

- The same total communalities of each variable
- The same unique variance on each variable, and
- Identical correlation matrices

Rotation of factors:

- Through rotation the factor matrix is transformed into a simpler one that is easier to interpret.
- After rotation each factor should have nonzero, or significant, loadings for only some of the variables. Each variable should have nonzero or significant loadings with only a few factors, if possible with only one.
- The rotation is called **orthogonal rotation** if the axes are maintained at right angles as in Figure 5.7 below.
- **Varimax procedure.** Axes maintained at right angles
  -Most common method for rotation.
  -An orthogonal method of rotation that minimizes the number of variables with high loadings on a factor.
  -Orthogonal rotation results in uncorrelated factors.
- **Oblique rotation.**
  -Factors are correlated.
  -Oblique rotation should be used when factors in the population are likely to be strongly correlated.

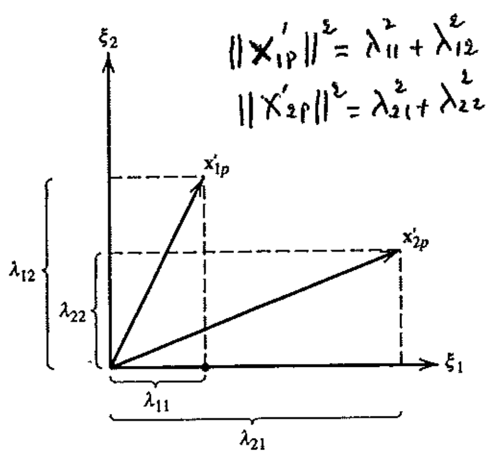


Figure 5.5 Projection of vectors onto a two-dimensional factor space.

lengths of projection vectors give communalities of the variable

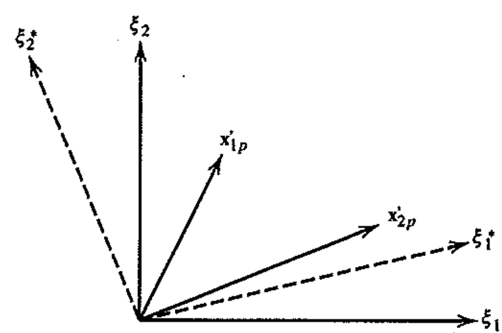


Figure 5.6 Rotation of factor solution.

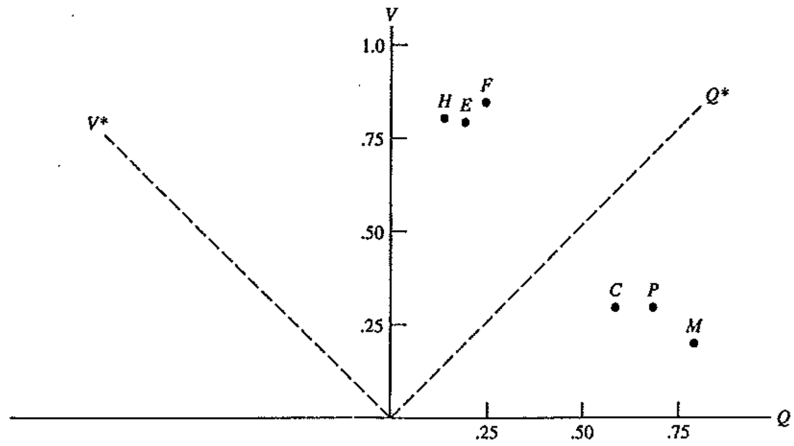


Figure 5.7 Factor solution.

Interpret Factors:

- A factor can be interpreted in terms of the variables that load high on it.

- Another useful aid in interpretation is to plot the variables, using the factor loadings as coordinates. Variables at the end of an axis are those that have high loadings on only that factor, and hence describe the factor.

Estimation of loadings and communalities:

We apply the principle component solution to the previous example.

The initial estimates of the communalities for all the variable are equal to one. Next, the correlation matrix with estimated communalities in the diagonal is subject to PCA. The 6 principle components are:

$$\begin{aligned}
\xi_1 &= .368M + .391P + .372C + .432E + .422H + .456F \\
\xi_2 &= .510M + .409P + .383C - .375E - .421H - .329F \\
\xi_3 &= -.267M - .486P + .832C - .022E - .003H - .023F \\
\xi_4 &= .728M - .665P - .152C + .065E + .012H + .035F \\
\xi_5 &= .048M - .005P - .003C - .742E + .667H + .054F \\
\xi_6 &= .042M + .039P + .024C + .343E + .447H - .824F.
\end{aligned} \tag{5.9}$$

The variances (given by the eigenvalues) of the six principal components, $\xi_1, \xi_2, \xi_3, \xi_4, \xi_5$, and ξ_6 are, respectively, 3.367, 1.194, .507, .372, .313, and .247 [1]. The above equations can be rewritten such that the principal components scores are standardized to have a variance of one. This can be done by dividing each ξ by its respective standard deviation. For example, for the first principal component

$$\frac{\xi_1}{\sqrt{3.367}} = .368M + .391P + .372C + .432E + .422H + .456F,$$

or

$$\xi_1 = .675M + .717P + .683C + .793E + .774H + .837F.$$

Standardizing each principal component results in the following equations

$$\begin{aligned}
\xi_1 &= .675M + .717P + .683C + .793E + .774H + .837F \\
\xi_2 &= .557M + .447P + .418C - .410E - .461H - .359F \\
\xi_3 &= -.190M - .346P + .592C - .015E - .002H - .016F \\
\xi_4 &= .444M - .405P - .093C + .040E + .007H + .021F \\
\xi_5 &= .027M - .003P - .002C - .415E + .373H + .030F \\
\xi_6 &= .021M + .019P + .012C + .171E + .222H - .409F.
\end{aligned} \tag{5.10}$$

Alternatively, we can write the variables as function of the 6 principal components (PC):

$$\begin{aligned}
M &= .675\xi_1 + .557\xi_2 - .190\xi_3 + .444\xi_4 + .027\xi_5 + .021\xi_6 \\
P &= .717\xi_1 + .447\xi_2 - .346\xi_3 - .405\xi_4 - .003\xi_5 + .019\xi_6 \\
C &= .683\xi_1 + .418\xi_2 + .592\xi_3 - .093\xi_4 - .002\xi_5 + .012\xi_6 \\
E &= .793\xi_1 - .410\xi_2 - .015\xi_3 + .040\xi_4 - .415\xi_5 + .171\xi_6 \\
H &= .774\xi_1 - .461\xi_2 - .002\xi_3 + .007\xi_4 + .373\xi_5 + .222\xi_6 \\
F &= .837\xi_1 - .359\xi_2 - .016\xi_3 + .021\xi_4 + .030\xi_5 - .409\xi_6.
\end{aligned} \tag{5.11}$$

Next, we need to determine which PC we should keep, based on the eigenvalue-greater-than-one rule. From previous PCA computation, it was found $\lambda_1 = 1.237, \lambda_2 = 1.105, \lambda_3 = 1.002$, and $\lambda_4 = 0.919$.

Thus we can represent the indicators as follows:

$$\begin{aligned}
M &= .675\xi_1 + .557\xi_2 + \epsilon_m \\
P &= .717\xi_1 + .447\xi_2 + \epsilon_p \\
C &= .683\xi_1 + .418\xi_2 + \epsilon_c \\
E &= .793\xi_1 - .410\xi_2 + \epsilon_e \\
H &= .774\xi_1 - .461\xi_2 + \epsilon_h \\
F &= .837\xi_1 - .359\xi_2 + \epsilon_f
\end{aligned} \tag{5.12}$$

Table 5.4 **Summary of Principal Components Factor Analysis for the Correlation Matrix of Table 5.2**

Variable	Factor Loadings ξ_1	ξ_2	Communalities	Specific Variance ϵ
M	.675	.557	.766	.234
P	.717	.447	.714	.286
C	.683	.418	.641	.359
E	.793	−.410	.797	.203
H	.774	−.461	.812	.188
F	.837	−.359	.829	.171

Notes:

1. Variance accounted for by factor ξ_1 is: 3.365 (i.e., $.675^2 + .717^2 + .683^2 + .774^2 + .837^2$).
2. Variance accounted for by factor ξ_2 is: 1.194 (i.e., $.557^2 + .447^2 + .418^2 + (-.410)^2 + (-.461)^2 + (-.359)^2$).
3. Total variance accounted for by factors ξ_1 and ξ_2 is: 4.559 (i.e., 3.365 + 1.194).
4. Total variance not accounted for by the common factors (i.e., specific variance) is: 1.441 (i.e., .234 + .286 + .359 + .203 + .188 + .171).
5. Total variance in the data is 6 (i.e., 4.559 + 1.441).

Table 5.5 gives the amount of correlation among the indicators that is due to the two factors and is referred to as the reproduced correlation matrix. The diagonal of the reproduced correlation matrix gives the communalities of each indicator.

Table 5.5 Reproduced and Residual Correlation Matrices for PCF

Reproduced Correlation Matrix

	M	P	C	E	H	F
M	.766	.733	.694	.307	.266	.365
P	.733	.714	.677	.385	.349	.440
C	.694	.677	.641	.370	.336	.422
E	.307	.385	.370	.797	.803	.811
H	.266	.349	.336	.803	.812	.813
F	.365	.440	.422	.811	.813	.829

Note: Communalities are on the diagonal.

Residual Correlation Matrix

	M	P	C	E	H	F
M	.234	−.113	−.154	.013	.018	.005
P	−.113	.285	−.167	−.005	.002	−.010
C	−.154	−.167	.359	−.010	.000	−.017
E	.013	−.005	−.010	.203	−.117	−.081
H	.018	.002	.000	−.117	.188	.079
F	.005	−.010	−.017	−.081	.078	.171

Note: Unique variances are on the diagonal.
Root mean square residual (RMSR) = .078.

The RMSR of the residual matrix $$RMSR = \sqrt{\frac{\sum_{i=1}^{p}\sum_{j=i}^{p} res_{ij}^2}{p(p-1)/2}}$$