Exploratory Factor Analysis

Introduction to Exploratory Factor Analysis

What is Factor Analysis?

- Factor analysis is a statistical modeling technique used to model the covariance structure in multivariate data.
- Factor analysis is a statistical modeling technique for estimating unobserved (or latent) relationships using observed (or measured) variables.
- As a statistical modeling technique, factor analysis has statistical assumptions.

Why do we use Factor Analysis?

- We use FA to model the correlation structure in a set of measured variables.

- We use FA to facilitate a dimension reduction from the observed measurement variables to the smaller set of unobserved latent factors.

- We use FA to improve the interpretability of our multivariate data.

When should we use Factor Analysis?

Factor analysis is most useful on problems that are 'natural factor analysis problems'. What are the characteristics of a natural factor analysis problem?

- All variable names are known, i.e. we know and understand all of the measurement variables.

- All measurement variables have been purposely selected under the guidance that they represent a measurement of a quality or trait that is recognized as important but that cannot be directly measured.

- We are able to obtain multiple measurements for each of the unmeasureable qualities (the latent factors).

Examples of Natural Factor Analysis Problems

- Measurement of physical attributes: speed, strength, agility
- Measurement of educational attainment: math, reading, problem solving
- Measurement of personality: rational, social, empathetic

Exploratory Versus Confirmatory Factor Analysis

- EFA is performed when we have no preconceived notions about the factor structure (i.e. the factor loadings) that may exist in a set of multivariate data. Since we have no preconceived notions of the factor structure, we are performing an exploratory data analysis, hence the name exploratory factor analysis.

- CFA is performed when we want to statistically test a specific factor structure. CFA will require the formal statistical assumptions of maximum likelihood estimation so that formal statistical inference can be applied to the data to draw statistical conclusions. CFA is related to other topics such as path analysis and simultaneous equations.

How Does Factor Analysis Differ from Principal Components Analysis?

- FA is a statistical model while PCA is not.
- FA is a statistical model for the correlation structure in multivariate data. PCA is not a model for the correlation strucuture. PCA is a rotation of the coordinate axes.
- FA is focused on producing a representation of the correlation structure that will provide an enhanced data interpretation.
- PCA can be used with anonymized variables, i.e. data with no known names, while FA requires well defined and known variables. FA is more subjective than PCA, and also more focused on interpretability than PCA, and hence FA is difficult to use with a large number of variables or anonymized variables.

A Modeling Process for Factor Analysis

Our Approach to Factor Analysis

- Since we are focused on statistical modeling, and not statistical theory, we will approach factor analysis from the modeling perspective.

- How do we define a *modeling process* for factor analysis?

- How different or similar will our modeling process for factor analysis be to our standard linear models approach?

A Strategy for Performing a Factor Analysis

- Perform a Principal Factor Analysis with a Varimax rotation.
- Perform an Iterative Principal Factor Analysis with a Varimax rotation.
- Perform a Maximum Likelihood Factor Analysis with a Varimax rotation.
- Compare the solutions from these three factor analyses.
  - Did each factor analysis yield roughly the same factor loadings?
  - Is one set of factors more interpretable than the other set?

A Strategy for Performing a Factor Analysis - Continued

- Evaluate the factor loadings over a range of common factors, i.e. instead of just looking at the results for $k = 4$ also consider values for k in $\{2, 3, 4, 5, 6\}$.

- If you have enough data, then evaluate your prospective factor loadings through bootstrapping or cross validation. As with any statistical relationship, for that relationship to represent a 'universal truth', then it must exist in many samples. We can effectively construct this ideal situation by employing either of these methods.

This strategy has been adapted from the advice provided in Johnson and Wichern *Applied Multivariate Analysis*.

The Common Factor Model: Model Definition and Estimation

General Assumptions of Factor Analysis

- Let $X_1, X_2, \ldots, X_p$ be observable random variables. In the context of FA we will call these variables the *response variables*.

- Let f_j denote an unobservable concept called a *common factor*.

- Let λ_{ij} denote the *factor loading* for the jth common factor f_j on the ith response X_i. Note that the factor loading λ_{ij} represents the correlation between the common factor f_j and the response variable X_i.

- Let u_j be the *specific error* or *unique factor*.

- In factor analysis we assume that each response X_i can be deconstructed into a set of common factors $f_1, f_2, \ldots, f_k$ and a unique factor u_i.

Thurstone's Common Factor Model

The Common Factor Model is defined by the following relationships.

$$X_1 = \lambda_{11} f_1 + \lambda_{12} f_2 + \cdots + \lambda_{1k} f_k + u_1 \tag{1}$$

$$X_2 = \lambda_{21} f_1 + \lambda_{22} f_2 + \cdots + \lambda_{2k} f_k + u_2 \tag{2}$$

$$X_p = \lambda_{p1} f_1 + \lambda_{p2} f_2 + \cdots + \lambda_{pk} f_k + u_p \tag{3}$$

- Each response variable X_i is assumed to be related to the unobserved factors f_j through a linear equation with the coefficients (factor loadings) given by λ_{ij}. Since the relationship will not be exact, each linear equation has a response specific error called the *unique factor* or the *specific error*.

- The factors $f_1, f_2, \ldots, f_k$ are called *common factors* because they are common to all of the response variables $X_1, X_2, \ldots, X_p$.

- The relationship between the response variables X_i and the common factors f_j looks very similar to an OLS regression model, except the common factors are not directly observable (measurable).

Not so fast my friend!

- Let's make sure that we remember the objective of factor analysis. Specifically, we do not want to be able to reproduce the response variables X_i. When using factor analysis, we want to model the correlation structure of the response variables.

- How do we define the correlation structure? We can describe the correlation structure of multivariate data using either the correlation matrix or the covariance matrix.

- We can the Common Factor Model defined by Equations (1)-(3) in matrix format.

$$\boldsymbol{X} = \boldsymbol{\Lambda f} + \boldsymbol{u} \tag{4}$$

- From this matrix representation of the Common Factor Model we can then define the covariance matrix with some additional statistical assumptions.

Not so fast my friend! - Continued

- Note: In theoretical discussions we discuss the correlation structure using the covariance matrix. However, in practice we use the correlation matrix. When the data are standardized these matrices are the same. When we perform data analysis, we (or our software) will standardize the data in order to make these matrices the same.

- It might help your understanding of the technical details if you think of the matrix $\mathbf{\Sigma}$ as the covariance matrix for standardized data so that it can be interpreted as a correlation matrix.

Needed Additional Assumptions

To define the correlation structure we need some additional asssumptions.

- The unique (specific) factor u_i has the *unique (specific) variance* ψ_i, i.e. Var(u_i) $= \psi_i$.

- The common factors f_j are mean zero, variance 1, and independent of each other, i.e. f_n and f_m are independent for all n not equal m.

- The unique (specific) factors and the common factors are independent of each other, i.e. u_i and f_j are independent for all i and j.

Defining the Correlation Structure

Using the additional statistical assumptions we can define the correlation structure for the matrix $\boldsymbol{X}$.

$$\begin{aligned} \text{Cov}(\boldsymbol{X}) &= \text{Cov}(\boldsymbol{\Lambda} f + u) \\ &= \text{Cov}(\boldsymbol{\Lambda} f) + \text{Cov}(\boldsymbol{u}) \\ &= \boldsymbol{\Lambda}\text{Cov}(\boldsymbol{f})\boldsymbol{\Lambda}^T + \text{Cov}(\boldsymbol{u}) \end{aligned} \tag{5}$$

This relationship yields the commonly displayed correlation structure

$$\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}^T + \boldsymbol{\Psi} \tag{6}$$

where $\boldsymbol{\Lambda} = \begin{bmatrix} \lambda_{11} & \cdots & \lambda_{1k} \\ \vdots & \cdots & \vdots \\ \lambda_{p1} & \cdots & \lambda_{pk} \end{bmatrix}$ and $\boldsymbol{\Psi} = \begin{bmatrix} \psi_1 & 0 & \cdots & 0 & 0 \\ 0 & \psi_2 & 0 & \cdots & 0 \\ \vdots & \cdots & \cdots & \cdots & \vdots \\ 0 & 0 & 0 & 0 & \psi_p \end{bmatrix}$.

Correlation Components

The covariance matrix $\mathbf{\Sigma}$ has several components and related quantities of interest.

- The diagonal entries of the covariance matrix are the individual variance terms.
$$\sigma_i^2 = \lambda_{i1}^2 + \cdots + \lambda_{ip}^2 + \psi_i \tag{7}$$
- The sum of the squared factor loadings is called the *common variance* or the *communality* and typically denoted by h_i^2.
$$h_i^2 = \lambda_{i1}^2 + \cdots + \lambda_{ik}^2 \tag{8}$$
- The term ψ_i is called the *unique (specific) variance*.
- The covariance between X_i and X_j is given by
$$\sigma_{ij} = \sum_{n=1}^{k} \lambda_{in}\lambda_{jn} \tag{9}$$

How do we estimated the factor loadings?

The correlation structure

$$\mathbf{\Sigma} = \mathbf{\Lambda}\mathbf{\Lambda}^T + \mathbf{\Psi} \qquad (10)$$

has one estimatable quantity and two unknowns. How can we estimate the factor loadings $\mathbf{\Lambda}$?

We can estimate $\mathbf{\Sigma}$ from the data. If we had an estimate for $\mathbf{\Psi}$ then we could estimate the **reduced covariance matrix**

$$\widehat{\mathbf{\Sigma}} - \widehat{\mathbf{\Psi}} = \widehat{\mathbf{\Lambda}}\widehat{\mathbf{\Lambda}}^T. \qquad (11)$$

How do we estimate $\mathbf{\Psi}$? Instead of estimating $\mathbf{\Psi}$ directly, most algorithms initiate the estimation with an estimate of the reduced covariance matrix $\widehat{\mathbf{\Sigma}} - \widehat{\mathbf{\Psi}}$, and then back out estimates for $\mathbf{\Psi}$.

We estimate the reduced covariance matrix $\widehat{\mathbf{\Sigma}} - \widehat{\mathbf{\Psi}}$ by computing the estimated covariance matrix $\widehat{\mathbf{\Sigma}}$ and replacing the 1's on the diagonal with $\widehat{h}_1$ - $\widehat{h}_k$ for a k factor model. These estimates $\widehat{h}_i$ are called the **prior communality estimates**.

Estimating the Reduced Covariance Matrix

There are two common methods for estimating the reduced covariance matrix $\hat{\mathbf{\Sigma}} - \hat{\mathbf{\Psi}}$.

- The most common estimation procedure is to replace the diagonal of the estimated covariance matrix with the Squared Multiple Correlation coefficients (the SMC priors option in SAS). Note that the Squared Multiple Correlation coefficient is the R-Squared value from regressing one X_i on all of the other X_j.

- The other option is to replace the diagonal of the estimated covariance matrix with the absolute value of the maximum correlation coefficient for that row (the MAX priors option in SAS).

In either case the factor loadings are then computed by performing Principal Components Analysis (as a matrix factorization technique) on the reduced covariance matrix. This estimation technique is referred to as *Principal Factor Analysis*.

Iterative Principal Factor Analysis

- Iterative Principal Factor Analysis uses an updated estimate for $\mathbf{\Psi}$ at each iteraction.

- After any iteration we can compute

  $$\hat{\mathbf{\Psi}} = \hat{\mathbf{\Sigma}} - \hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}^T. \tag{12}$$

  After updating the estimate $\hat{\mathbf{\Psi}}$, the algorithm will recompute the factor loadings using the new reduced covariance matrix until the factor loadings 'stabilize'.

- When Principal Factor Analysis is computed using an iterative estimate for $\mathbf{\Psi}$, the procedure is called *Iterative Principal Factor Analysis*.

Maximum Likelihood Factor Analysis

- Maximum Likelihood Factor Analysis is the statistical approach to factor analysis. The response variables X_i are assumed to have a multivariate normal distribution, and hence the covariance matrix $\mathbf{\Sigma}$ is assumed to have a Wishart distribution.

- Principal Factor Analysis and Iterative Principal Factor Analysis are not formal statistical models. They are not estimated from a likelihood function, and hence they do not have any means of formal inference.

- Maximum Likelihood Factor Analysis is the only formal estimation procedure for factor analysis, and hence the only estimation procedure with formal inference for factor loadings (confidence intervals) and a statistical tests for goodness-of-fit.

- Inference includes a formal test of model adequacy for the number of factors and the use of AIC as a means of model selection for the number of factors.

Model Goodness-Of-Fit

Goodness-Of-Fit for Factor Analysis

What does it mean for a factor model to fit well?

- Unfortunately, in practice the goodness-of-fit of a factor model is typically completely determined by its interpretabilty.

- All communality estimates should be less than 1, i.e. no Heywood cases.

- Factor loadings should exhibit a *simple factor structure*. This is ideal, but seldom hold in practice.

- Objective Critiques of Fit
  - Small residual matrix: The estimated matrix $\widehat{\boldsymbol{\Psi}}$ can be interpreted as a residual matrix. Componentwise metrics (matrix norms) such as Mean Absolute Error (MAE) and Mean Square Error (MSE) can be used to compare different factor models for relative fit.
  - Statistical Inference: Use MLE and statistical inference to justify your factor model.

Caveats of Goodness-Of-Fit for Factor Analysis

- Without MLE all GOF is subjective, or at least relative.
- Using MLE requires the assumption of multivariate normality. This assumption is difficult to validate from the model. One would have to validate the assumption prior to using ML FA by looking at the marginal distributions.
- Software will frequently output estimates for the Heywood case, and some people will use this output. The validity of the output in the Heywood case will be contentious.
- One item that tends to be lost in the discussion of GOF is the overall legitimacy of your factor relationships. Are they real, i.e. would they be apparent across many samples? Or are they sample specific, i.e. do they just happen to show in your sample? More importantly did you deviate your sample from a random sample and generate a spurious relationship?

Factor Rotations

Factor rotations are a common practice in factor analysis. There are two types of factor rotations: (1) orthogonal rotations and (2) oblique rotations.

- **Orthogonal** rotations will yield orthogonal factors after the rotation. The most common orthogonal rotation is the **Varimax** rotation.

- **Oblique** rotations will yield correlated factors after the rotation. The most common oblique rotation is the **Promax** rotation.

Items to note:

- Factor rotations do not improve the 'fit' of the factor model, only the factor interpretation. Factor rotations are used as a means to obtain a 'simple structure', and hence be more interpretable.

- After applying an oblique rotation the loadings matrix no longer represents the correlation structure between the observed variables and the unobserved factors.

Rotations to a Simple Structure

The primary reason to employ a factor rotation is to improve the interpretation of the factor structure by rotating to a **simple structure**.

- Each row of $\mathbf{\Lambda}$ should contain at least one zero.
- Each column of $\mathbf{\Lambda}$ should contain at least k zeros for a k factor model. Ideally, the factor grouping of the response variables is clearly identifiable.
- When comparing pairs of columns from $\mathbf{\Lambda}$, the pairs should have some elements that are zero in one column but nonzero in the other column. Again, ideally the factor groupings of the response variables are identifiable and unique to a group.

In practice simple factor structures are difficult to obtain. The user can try to fit simple factor structures through factor rotations and/or the inclusion or exclusion of response variables. In practice all of these decisions are subjective and considered to be the user's prerogative.

This advice has been adapted from *Multivariate Statistical Methods* by Morrison.

Caveats and Potential Shortcomings of Factor Analysis

- The common factor model is an *underdetermined* model (more unknowns than equations), and hence it has multiple solutions instead of a single unique solution. The existence of multiple solutions in any problem always causes confusion.

- Factor analysis will only work on a problem that is a ’factor analysis problem’, i.e. one where the user have performed a set of measurements that are intended to be related, i.e. measure an underlying concept. In this sense FA is not a general method for dimension reduction.

- In practice your factor analysis results will never be as ’nice’ as you think they should be.

Relationship Between the Covariance Matrix and the Correlation Matrix

Throughout this discussion we have been referencing the covariance matrix $\boldsymbol{\Sigma}$ and its estimate $\widehat{\boldsymbol{\Sigma}}$ (denoted by $\boldsymbol{S}$ in some books). When your data are centered and scaled, then the covariance matrix and the correlation matrix are the same matrix. Hence, when your data are not standardized, the difference between the correlation matrix is a scaling matrix.

- For two random variables X_1 and X_2 with standard deviations σ_1 and σ_2 we have

$$\text{Corr}(X_1, X_2) = \text{Cov}(X_1, X_2)/\sigma_1\sigma_2 \tag{13}$$

- Let $\boldsymbol{R}$ denote the correlation matrix. Consider the matrix representation of the problem for two random variabels X_1 and X_2.

$$\boldsymbol{R} = \begin{bmatrix} \text{Cov}(X_1, X_1)/\sigma_1^2 & \text{Cov}(X_1, X_2)/\sigma_1\sigma_2 \\ \text{Cov}(X_2, X_1)/\sigma_2\sigma_1 & \text{Cov}(X_2, X_2)/\sigma_2^2 \end{bmatrix} \tag{14}$$

Relationship Between the Covariance Matrix ... - Continued

- In terms of $\boldsymbol{\Sigma}$ we can write $\boldsymbol{R}$ as

$$\boldsymbol{R} = \text{diag}(\boldsymbol{\Sigma})^{-1/2} \cdot \boldsymbol{\Sigma} \cdot \text{diag}(\boldsymbol{\Sigma})^{-1/2}. \tag{15}$$

- Substituting the appropriate matrices yields

$$\boldsymbol{R} = \begin{bmatrix} 1/\sigma_1 & 0 \\ 0 & 1/\sigma_2 \end{bmatrix} \begin{bmatrix} \text{Cov}(X_1, X_1) & \text{Cov}(X_1, X_2) \\ \text{Cov}(X_2, X_1) & \text{Cov}(X_2, X_2) \end{bmatrix} \begin{bmatrix} 1/\sigma_1 & 0 \\ 0 & 1/\sigma_2 \end{bmatrix}. \tag{16}$$

Relationship Between the Covariance Matrix ... - Continued

- Since diagonal matrices are always invertible, any relationship specified with a covariance matrix could be respecified with a correlation matrix using an algebraic substitution. However, in practice we would simply perform factor analysis on the correlation matrix (or the software would automatically do this for us).

- Remember, although we presented all of the technical details in terms of the covariance matrix, in practice we are modeling the correlation matrix, and we are trying to find a factor structure to describe and reproduce the correlation matrix.

Review: A Strategy for Performing a Factor Analysis

A Strategy for Performing a Factor Analysis

- Perform a Principal Factor Analysis with a Varimax rotation.
- Perform an Iterative Principal Factor Analysis with a Varimax rotation.
- Perform a Maximum Likelihood Factor Analysis with a Varimax rotation.
- Compare the solutions from these three factor analyses.
  - Did each factor analysis yield roughly the same factor loadings?
  - Is one set of factors more interpretable than the other set?

A Strategy for Performing a Factor Analysis - Continued

- Evaluate the factor loadings over a range of common factors, i.e. instead of just looking at the results for $k = 4$ also consider values for k in $\{2, 3, 4, 5, 6\}$.

- If you have enough data, then evaluate your prospective factor loadings through bootstrapping or cross validation. As with any statistical relationship, for that relationship to represent a 'universal truth', then it must exist in many samples. We can effectively construct this ideal situation by employing either of these methods.

This strategy has been adapted from the advice provided in Johnson and Wichern *Applied Multivariate Analysis*.