Measurement, Design, and Analytic Techniques in Mental Health and Behavioral Sciences

Lecture 6 (April 16, 2009): Principal Components and Factor Analysis

XH Andrew Zhou

azhou@u.washington.edu

Professor, Department of Biostatistics, University of Washington

Principal component analysis and Factor analysis

- Regression models require to identify some important variables as outcome variables.
- Without identifying specific variables as outcome variables, principal component analysis and factor analysis can be used to examine the relationships among a set of variables.

Principal component analysis

- Principal component analysis technique tries to explain as much variability among a set of variables as possible in terms of a few linear combinations of the variables.

Factor analysis technique

- Factor analysis seeks to explain the relationships among a set of variables, expressed by their correlations or covariances, by a few unobserved variables, which is called factors.
- It is hoped that few factors than the original number of variables will be needed to explain the relationships among the variables.

Factor models

- Unlike measurement models for measurement error, a hypothetical construct, such as intelligence, cannot be measured directly. Instead, different aspects of intelligent are measured by different indicators or items, such as verbal, quantitative, and visual reasons.
- Answer to one particular item is a reflection of both general intelligence and an item-specific aspect, referred to as the common and specific factors.

Uni-dimensional common factor model

- A unidimensional common factor model:

$$y_{ij} = \beta_i + \lambda_i \eta_j + \epsilon_{ij},$$

  where y_{ij} is the observed response of item i on subject j, η_j is the common factor or latent trait for subject j, λ_i is a factor loading for the ith item, and ϵ_{ij} is the unique or specific factor.

- We assume that η_j and ϵ_{ij} are independent.
- Let us denote $\psi = Var(\eta_j)$ and $\theta_{ij} = Var(\epsilon_{ij})$.

Identification and equivalence

- A parametric model $g(y \mid \theta)$ is called globally identified if there are no two points θ_1 and θ_2 such that $g(y \mid \theta_1) = g(y \mid \theta_2)$.
- A parametric model $g(y \mid \theta)$ is called locally identified at θ_0 if there exists an open neighborhood of θ_0 such that there no two points θ_1 and θ_2 in the neighborhood such that $g(y \mid \theta_1) = g(y \mid \theta_2)$.

Identifibility of the unidimensional factor model

- Recall

$$y_{ij} = \beta_i + \lambda_i \eta_j + \epsilon_{ij},$$

  where $\eta_j \sim N(\gamma, \psi)$, and $\epsilon_{ij} \sim N(0, \theta_{ij})$.

- By making a linear transformation of the factor, $f_j = a\eta_j + c$, we can write the model as

$$y_{ij} = (\beta_i - \lambda_i c/a) + (\lambda_i a) f_j + \epsilon_{ij} = \beta_i^* + \lambda_i^* f_j^* + \epsilon_{ij},$$

  where $f_j^* \sim N(\gamma^*, \psi^*), \epsilon_{ij} \sim N(0, \theta_{ij})$.

- Here

$$\beta_i^* = \beta_i - \lambda_i c/a, \lambda_i^* = \lambda_i/a, \gamma^* = a\gamma + c, \psi^* = a^2\psi.$$

- Hence, different parameter points generate the same reduced form distribution and the model is not identified.

Identification restrictions

- We can either fix the scale of the common factor η_j by anchoring (typically fixing the first factor loading, $\lambda_1 = 1$) or "factor standardization' (fixing the factor variance to a positive constant, $\psi = 1$).
- Although the models from either identification restriction are equivalent, anchoring has some advantage over factor standardization from the point of view of "factorial invariance.
- For example, assume that the unidimensional factor model above holds for a population. But we consider the subpopulation of units with negative factor values. In this case the original factor loadings are recovered in the subpopulation under anchoring (with a reduced variance estimate $\hat{\psi}$) but not under factor standardization.

Unidimensional factor models in matrix notation

- Let $y_j = (y_{1j}, \ldots, y_{Ij})'$, $\Gamma = (\lambda_1, \ldots, \lambda_I)'$, $\epsilon_j = (\epsilon_{1j}, \ldots, \epsilon_{Ij})'$, $\beta = (\beta_1, \ldots, \beta_I)'$, where I is the number of items.
- The unidimensional factor model:

$$y_j = \beta + \Gamma\eta_j + \epsilon_j.$$

- The covariance structure of y_j is called a factor structure:

$$\Omega = Cov(y_j) = \Gamma\psi\Gamma' + \Theta,$$

where Θ is a diagonal matrix with the θ_{ii} placed on the diagonal.

Multidimensional factor models

- The unidimensional factor model imposes a rather restrictive structure on the covariance. In structuring $I(I+1)/2$ variances and covariances only $2I$ parameters are used.
- A less restrictive multidimensional factor models are often used:

$$y_{1j} = \beta_1 + \lambda_{11}\eta_{1j} + \ldots + \lambda_{1M}\eta_{Mj} + \epsilon_{1j}$$

$\ldots$

$$y_{Ij} = \beta_I + \lambda_{I1}\eta_{1j} + \ldots + \lambda_{IM}\eta_{Mj} + \epsilon_{Ij}$$

Multidimensional factor models in matrix form

- Let $y_j = (y_{1j}, \ldots, y_{Ij})'$,

$$\Gamma = \begin{pmatrix} \lambda_{11} & \ldots & \lambda_{1M} \\ \ldots & \ldots & \ldots \\ \lambda_{I1} & \ldots & \lambda_{IM}) \end{pmatrix},$$

  $\epsilon_j = (\epsilon_{1j}, \ldots, \epsilon_{Ij})'$, $\beta = (\beta_1, \ldots, \beta_I)'$, where I is the number of items.

- The multidimensional factor model can be written as follows:

$$y_j = \beta + \Gamma_y \eta_j + \epsilon_j,$$

  where β is a vector of constant (usually omitted if y_i is mean-centered), Γ_y is a factor loading matrix, η_j is a vector of M common factors with covariance matrix Ψ and ϵ_j a vector of unique factors with diagonal covariance matrix Θ.

- The covariance matrix of the responses become

$$\Omega = \Gamma_y \Psi \Gamma_y' + \Theta.$$

Cronbach's α - reliability

- Cronbach's α is a coefficient of reliability (or consistency).
- Cronbach's α measures how well a set of items (or variables) measures a single unidimensional latent construct.
- When data have a multidimensional structure, Cronbach's α will usually be low.
- They are referring to how well their items measure a single unidimensional latent construct.
- If you have multi-dimensional data, Cronbach's α will generally be low for all items. In this case, run a factor analysis to see which items load highest on which dimensions, and then take the alpha of each subset of items separately.

Calculation of Cronbach's α

- Cronbach's alpha can be written as a function of the number of test items AND the average inter-correlation among the items:

$$\alpha = \frac{N - \bar{r}}{1 + (N-1)\bar{r}},$$

where N is equal to the number of items and $\bar{r}$ is the average inter-item correlation among the items.

Cronbach's α, cont

- One can see from this formula that if you increase the number of items, you increase Cronbach's α. Additionally, if the average inter-item correlation is low, α will be low. As the average inter-item correlation increases, Cronbach's α increases as well.
- This makes sense intuitively - if the inter-item correlations are high, then there is evidence that the items are measuring the same underlying construct. This is really what is meant when someone says they have "high" or "good" reliability.
- Note that a reliability coefficient of .70 or higher is considered "acceptable" in most Social Science research situations)

Warnings on the use of Cronbach's α

- The first problem: α is dependent not only on the magnitude of the correlations among items, but also on the number of items in the scale.
- A scale can be made to look more 'homogenous' simply by doubling the number of items, even though the average correlation remains the same. For example, if we have two scales which each measure a distinct construct, and combine them to form one long scale, α would probably be high, although the merged scale is not a uni-dimensional.
- If α is too high, then it may suggest a high level of item redundancy; that is, a number of items asking the same question in slightly different ways.

A potential project topic

- Confidence intervals for estimated Cronbach's α.
- Koning, A.J. Franses, Ph.H.B.F. (2003). Confidence Intervals for Cronbach's Coefficient Alpha Values, https://ep.eur.nl/handle/1765/431
- Dawn Iacobucci and Adam Duhachek (2003). Advancing Alpha: Measuring Reliability With Confidence. Journal of Consumer Psychology, Vol. 13, No. 4, Pages 478-487

Confirmatory factor analysis

- If prior information is available, in terms of substantive theory, previous results, confirmative factor analysis (CFA) should be used where particular parameters are set to prescribed values, typically zero. For Γ_y is often specified as independent clusters structure where each item load one one and only one common factor.
- Confirmatory factor analysis is thus a hypotheticist procedure designed to test hypotheses about the relationship between items and factors, whose number and interpretation are determined.

An example of confirmatory factor analysis

- Mulaik (1988, Handbook of Multivariate Experimental Psychology) considered 9 subjective rating-scale variables designed to measure two "dimensions" or factors in connection with a soldier’s conception of firing a rifle in combat.
- The first factor, supposed to be "fear", had as indicators, the four scales: "frightening", " never-shaking", "terrifying", and "upsetting".
- The second factor, "optimism about outcome", had as indicators the five scales: "useful", " hopeful", "controllable", "successful", and "bearable".

An example of confirmatory factor analysis, cont

- The loadings of variables on irrelevant factors were hypothesized to be zero, whereas the factors were expected to (negatively) correlated.
- An independent clusters two-factor model where each factor is measured by three non-overlapping items:

$$\begin{pmatrix} y_{1j} \\ y_{2j} \\ y_{3j} \\ y_{4j} \\ y_{5j} \\ y_{6j} \end{pmatrix} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \\ \beta_6 \end{pmatrix} + \begin{pmatrix} 1 & 0 \\ \lambda_{21} & 0 \\ \lambda_{31} & 0 \\ \lambda_{31} & 0 \\ 0 & 1 \\ 0 & \lambda_{52} \\ 0 & \lambda_{62} \end{pmatrix} \begin{pmatrix} \eta_{1j} \\ \eta_{2j} \end{pmatrix} + \begin{pmatrix} \epsilon_{1j} \\ \epsilon_{2j} \\ \epsilon_{3j} \\ \epsilon_{4j} \\ \epsilon_{5j} \\ \epsilon_{6j} \end{pmatrix},$$

where we have fixed the scale of each factor by setting one factor loading to 1.

Exploratory factor analysis

- Exploratory factor analysis (EFA) is an inductivist method designed to discover an optimal set of factors, their number to be determined in the analysis. Each factor is then interpreted and 'named' according to the subset of items having high loadings on the factor.
- The multidimensional factor model above is not identified without some restrictions on the parameters since we can multiply the factors by an arbitrary nonsingular transformation matrix R, $\eta_j^* = R\eta_j$, and obtain the same covariance matrix as the original model by multiplying the factor loading matrix Γ_y by R^{-1},

$$\Omega = (\Gamma_y R^{-1}) R \Psi R (R^{-1} \Gamma_y') + \Theta.$$

- If R is orthogonal, the transformation is a rotation or reflection

Identifiability

- In confirmatory factor analysis, restrictions on the factor loadings serve to fix the factor rotation, and combined with constraints for the factor scales (either fixing factor variances or by fixing one factor loading for each factor) will often suffice to identify the model.
- In exploratory factor analysis, a standard but arbitrary way of identifying the model is to set the factor covariance matrix equal to the identify matrix $\Psi = I$, and fix the rotation for example by requiring that $\Gamma_y' \Theta \Gamma_y$ is diagonal.

Indeterminancy of the factor space

- Something magic about factor analysis; we are estimting coefficients of variables that are not even observed. It is hard to imagine that one can estimate it at all.
- In fact, it is not possible to uniquely estimate the F_i, but one can estimate the F_i up to a certain indeterminancy.
- Mathematically, the factors are unique except for possible linear combinations.
- Geometrically, suppose we think of the factors in a model with $k = 2$ as corresponding to values in a plane. Let us assume that this plane is a part of a three-dimensional space.

Indeterminancy of the factor space

- Within this three-dimensional space, factor analysis would determine which plane contains the two factors. However, any two perpendicular directions in the factor plane would correspond to factors that equally well fit the data in terms f explaining the covariance or correlation between the variables.
- Thus, we have factors identified up to a certain extent, but we are allowed to rotate them within a "subspace".

Methods of rotations

- This indeterminancy of the factor space allows one to play with different combinations of factors, that is, rotations, so that the factors are considered "easy to interpret".
- One of goals in the factor analysis is to find factors that represent some abstract concepts. This task is easier to accomplish when the factors are associated with some subset of he variables. That is, the factors have high loadings (in terms of absolute value) on some subset of the variables and very low (near zero in absoulte value) loeadings on the rest of the variables.

Methods of rotations

- In this case, the factor is closely associated with the subset of the variables that have high loadings. If these variables have something in common conceptually, for example they are all measures of blood pressure, or in a psychological study they all seem to be related to aggressive behavior, one might then identify the specific factor as a blood pressure factor or a aggression factor.

Visual rotation

- When it is hard to interpret the data points in the original factors, one can try to rotate the original factors by an angel θ to see whether rotated factors have an easy interpretation, that is each rotated factor associated with a subset of the variables.
- This can be done visually, called visual rotation.
- When $k > 2$, it may not be easy to use the visual rotation method.

Varimax method

- Suppose we have the loadings λ_{ij} for one selection of factors.
- Let θ_{ij} be the loadings for different set of factors, (the linear combinations of the old factors).
- Define the weighted quantities

$$\gamma_{ij} = \theta_{ij} / \sqrt{\sum_{j=1}^{p} \lambda_{ij}^2},$$

which are chosen to minimize the effect of large communalities.

Varimax method

- The Varimax method chooses the θ_{ij} to maximize the following:

$$\sum_{j=1}^{k}(\frac{1}{p}\sum_{i=1}^{p}\gamma_{ij}^{4} - \frac{1}{p^2}(\sum_{i=1}^{p}\gamma_{ij}^{2})^2).$$

- A factor analysis is said to have a general factor if there is a factor which is associated with all or almost all of the variables.
- The varimax method can be useful but does not allow general factors and should not be used when such factors may occur.

Quartimax method

- The second popular method is the quartimax method. This method, in contrast to the varimax method, tends to have one factor with large loadings on all the variables, and not many large loadings among the rest of the factors.

Steps of performing exploratory factor analysis

- The number of factors is determined based on a principal component analysis of the correlation matrix. The number of factor is typically chosen to be equal to the number of eigenvalues that are larger than one, so called Kaiser-Guttmann criteria.
- A factor analysis is performed with the chosen number of factor. It is typically difficult to ascribe meaning to the factors at this stage since most of items will have nonnegative loadings on most factor.
- An orthogonal transformation matrix R is therefore used to produce more interpretable loadings according to some criteria such as loadings either "small" or "large".
- The final step is to retain only the "salient" loadings, interpreting as zero any loadings falling below an arbitrary threshold, typically, 0.3 or 0.4.

Item response models

- The unidimensional factor model can be extended to dichotomous and ordinal responses using two different approaches.
- Factor analysis using a latent response formulation
- Item response theory

Factor analysis with latent response formulation

- A binary or ordinal-scale response y_i can be often viewed as a partial observation of a continuous latent response y_i^*. We model y_i^* by a unidimensional common factor model by

$$y_{ij}^* = \beta_i + \lambda_i \eta_j + \epsilon_{ij}'$$

- The observed response y_i takes one of S ordered response categories a_s, where $s = 1, \ldots, S$, and the relationship between observed and latent response can be written as

$$y_{ij} = \begin{cases} a_1 \text{ if } k_0 < y_{ij}^* \leq a_2 \\ a_2 \text{ if } k_1 < y_{ij}^* \leq a_3 \\ \ldots \\ a_S \text{ if } k_{S-1} < y_{ij} \leq a_S, \end{cases}$$

where $k_0 = -\infty$ and $k_S = \infty$.

Item response theory (IRT)

- Item Response Theory (aka IRT) is also sometimes called latent trait theory.
- The classical application of these IRT models is in ability testing, where item i represents questions or problems in a test and the answers are scored as right (1) or wrong (0).
- Item response theory relates characteristics of items (item parameters) and characteristics of individuals (latent traits) to the probability of a positive response.
- A variety of IRT models have been developed for dichotomous and polytomous data. In each case, the probability of answering correctly or endorsing a particular response category can be represented graphically by an item (option) response function (IRF/ORF). These functions represent the nonlinear regression of a response probability on a latent trait, such as conscientiousness or verbal ability.

One-parameter Logistic Model

- $P(y_{ij} = 1 \mid \eta)$ is the conditional probability of a positive response ($y_{ij} = 1$) to item i by subject j
- The mathematical form of the 3PL model is shown below.

$$P(y_{ij} = 1 \mid \eta_j) = \frac{1}{1 + \exp(-a(\eta_j - b_i))},$$

where:

  - η_j represents the value of the latent trait (e.g., conscientiousness or cognitive ability) for subject j,
  - The a parameter affects the steepness of the curve; as "a" increases the slope of the IRF increases. Larger "a" parameters provide better discrimination among examinees.
  - The b_i parameter represents the location of the IRF along the horizontal axis, η_j. It is commonly called the item difficulty, or threshold. Large values of b indicate "difficult" items.
  - Usually, we assume η_j has the standard normal distribution.
  - An one-parameter logistic model is just a random intercept model for dichotomous items without covariates.

The 2-parameter logistic model (2PL)

- The two-parameter logistic model allows the slope or discrimination parameter (a) to vary across items instead of being assumed to be equal as in the 1PL model.
- The mathematical form of the 2PL model is shown below.

$$P(y_{ij} = 1 \mid \eta_j) = \frac{1}{1 + \exp(-Da_i(\eta_j - b_i))},$$

  where D is a scaling constant equal to 1.702.
- Here, b_i can be interpreted as the item difficulty, giving a 50% chance of a correct answer when ability equals difficulty, whereas a_i is an item discrimination parameter determining how well the item discriminates between subjects with different abilities.

The 3-parameter logistic model (3PL)

- One assumption in the two-parameter model is that the probability of answering correctly tends to zero as ability tends to minus infinity. However, this assumption is unrealistic if multiple choice formats are used, since guessing would produce a nonzero probability of answering correctly. An extra parameter can be introduced into the two-parameter model leading to the three-parameter model.

The 3-parameter logistic model (3PL), cont

- The mathematical form of the 3PL model is shown below.

$$P_i(\theta) = c_i + (1 - c_i)\frac{1}{1 + \exp(-Da_i(\theta - b_i))},$$

  where: the "c" parameter is commonly called the "pseudo-guessing" parameter, because it indicates the probability of responding positively for examinees having very low θ.

- Note that the 2PL model can be obtained from 3PL be setting c=0; the 1PL model may be obtained by setting c=0 and a=1.
- Unfortunately, huge sample may be needed to obtain reliable estimates of this model.

Structural equation models with latent variables

- Measurement and factor models relate the latent variables to the observed variables. The structural equation models the relationships among latent variables.
- Structural equation modeling with latent variables, often refereed to as covariance structure analysis, focuses on the covariance structure whereas the mean structure is typically eliminated by subtracting the mean from each variable.
- Having defined common factor models, a structural models specifying relationships among latent variables can be constructed.
- In this structural model, there could be both latent dependent and latent explanatory variables.

A structural equation example

- Consider a structural equation model for two latent dependent variables η_{1j} and η_{2j} and two latent latent independent variables ξ_{1j} and ξ_{2j}.
- The measurement model for the dependent variable is specified as an independent clustered overlapping items, written as

$$y_j = \Gamma_Y \eta_j + \epsilon_j.$$

  Similarly, the measurement model for the explanatory variables can be written as

$$x_j = \Gamma_x \xi_j + \delta_j.$$

- The structural equation model is given as follows:

$$\eta_j = B\eta_j + \Gamma\xi_j + \zeta_j.$$

- Here we assume that ξ_{1j} and ξ_{2j} are correlated whereas the ζ_{1j} and ζ_{2j} are uncorrelated.