13/04/15.

MLA

FACTOR ANALYSIS

- Factor Analysis is a mathematical approach for attempting to explain the correlation between a large set of variables in terms of a small number of underlying factors
- Primary assumption of FA is that we cannot observe these factors directly ⇒ they are latent. It is a dimension reduction technique, like PCA except more elaborate

- Example: correlation matrix for performance of students in Irish (X_1) English (X_2) maths (X_3)

$$\begin{pmatrix} 1 & 0.8 & 0.6 \\ 0.8 & 1 & 0.85 \\ 0.6 & 0.85 & 1 \end{pmatrix}$$

The dimensionality of matrix can be reduced from $m=3$ to $m=1$ by expressing the three variables as $X_1 = \lambda_1 f + \varepsilon_1$, $X_2 = \lambda_2 f + \varepsilon_2$ $X_3 = \lambda_3 f + \varepsilon_3$

- The f's in these equations are an underlying common factor which can often be given an interpretation (in example it could be general ability)
- The λ_i's are called factor loadings, and the ε_i's are errors specific factors
- ε_i's will be of small variance if X_i is closely related to f.

- More generally we have observable random vector $X = (X_1 \dots X_m)$ with mean μ and covariance matrix Σ.
- The factor model states that X is linearly dependent upon a few unobservable random variables $f_1, \dots f_p$ called common factors, and m conditional sources of variation $\varepsilon_1, \dots \varepsilon_m$ called specific factors

We have $X_1 - \mu_1 = \lambda_{11} f_1 + \lambda_{12} f_2 + \dots \lambda_{1p} f_p + \varepsilon_1$

$$\vdots$$

$$X_m - \mu_m = \lambda_{m1} f_1 + \lambda_{m2} f_2 + \dots + \lambda_{mp} f_p + \varepsilon_m$$

$$\Rightarrow X - \mu = \Lambda f + \varepsilon$$

- The ~~term~~ λ_{ij} value is called the factor loading of the i^{th} variable on the j^{th} factor
- Λ is the matrix of factor loadings
- Note that the i^{th} specific factor ε_i is associated only with the response X_i
- Note that $f_1, f_2, \dots f_p$ $\varepsilon_1, \varepsilon_2 \dots \varepsilon_m$ are all UNOBSERVABLE random variables

Three Assumptions of factor Model are:

- $E[F]=0$, $Cov[F] = I = \begin{pmatrix} 1 & \cdots & 0 \\ 0 & \ddots & 1 \end{pmatrix}$
- $E[\varepsilon]=0$, $Cov[\varepsilon] = \Psi = \begin{pmatrix} \psi_1 & 0 & 0 \\ 0 & \psi_2 & \\ 0 & 0 & \psi_3 \end{pmatrix}$
- f and ε are independent, $Cov[F,\varepsilon]=0$

$\Rightarrow$ The above is the orthogonal factor model. If we assume $Cov[F]$ is not diagonal we have "OBLIQUE" factor model

- We have $\Sigma = Cov[X] = E[(X-\mu)(X-\mu)^T]$

$$= E[(\Lambda f+\varepsilon)(\Lambda f+\varepsilon)^T]$$
$$= E[(\Lambda f)(\Lambda f)^T + \varepsilon(\Lambda f)^T + (\Lambda f)\varepsilon^T + \varepsilon\varepsilon^T]$$
$$= \Lambda \mathbb{E}[ff^T]\Lambda^T + \mathbb{E}[\varepsilon f^T]\Lambda^T + \Lambda\mathbb{E}[f\varepsilon^T] + \mathbb{E}[\varepsilon\varepsilon^T]$$
$$= \Lambda\Lambda^T + \Psi$$

We can show $Cov[X,F] = \mathbb{E}[(X-\mu)(f-0)] = \Lambda$, hence the covariance of the observed variable X_i and the unobserved factor F_j is the factor loading λ_{ij}.

VARIENCE

- Variance of X can be split into two parts
- First portion of the variance for the i^{th} component arises from the m common factors and is referred to as the i^{th} communality
- Remainder of variance for the i^{th} component is due to the specific factor referred to as the uniqueness

$$\sigma_i^2 = \lambda_{i1}^2 \; \lambda_{i2}^2 + \dots \lambda_{ip}^2 + \psi_i$$
$$= h_i^2 + \psi_i \qquad \text{denoting } i^{th} \text{ communality as } h_i^2$$

$Var[X_i]$ = communality + unique

- The i^{th} communality is the sum of squares of the loadings of the i^{th} variable on the p common factors

MLA

FACTOR ANALYSIS

- Factor Model assumes that the ~~mxm~~ $m + m(m-1)/2$ variances and covariances of X can be reproduced from the mp factor loadings λ_{ij} and the m specific variances ψ_i.
- In the case where $p << m$ the factor model provides a simplified version of the covariance in X with fewer parameters than the $m(m+1)/2$ parameters in Σ

- Example if $X = (X_1 \ldots X_{12})$ and a factor model with $p=2$ is appropriate, then the $12 \times 13/2 = 78$ elements of Σ are described in terms of the $pm + m = 2 \times 12 + 12 = 36$ parameters λ_{ij} and ψ_i of the factor model
- Unfortunately most covariance matrices cannot be uniquely factored as $\Lambda\Lambda' + \Psi$ where $p << m$

- FA is not affected by re-scaling of the variables.
Rescaling X is the equivalent to letting $y = CX$ where $C = \mathrm{diag}(C_i)$

If the factor model holds with $\Lambda = \Lambda_X$ and $\Psi = \Psi_X$

$$Y - C\mu = C\Lambda_X f \quad C\varepsilon$$

$$\text{So,} \quad \mathrm{Var}[Y] = C\Sigma C \quad C\Lambda_X \Lambda_X^T C + C\Psi_X C$$

Adding a C in everywhere

- To demonstrate that most covariance matrices Σ cannot be uniquely factored as $\Lambda\Lambda' + \Psi$ where $p << m$, let G be any $m \times m$ orthogonal matrix $(GG^T = I)$ then

$$\begin{aligned} X - \mu &= \Lambda f + \varepsilon \\ &= \Lambda G G^T f + \varepsilon \\ &= \Lambda^* f^* + \varepsilon \end{aligned}$$

- Here $\Lambda^* = \Lambda G$ and $f^* = G^T f$
- It follows that $E[f^*] = 0$ and $\mathrm{Cov}[f^*] = I$
- Thus it is impossible, given the data X, to distinguish between Λ and Λ^* since they both generate same cov matrix Σ: $\Sigma = \Lambda\Lambda' + \Psi$
$= \Lambda G G^T \Lambda' + \Psi = \Lambda^* \Lambda^{*T} + \Psi$

4

- This leads to idea of factor rotations, since orthogonal matrices correspond to rotations of the coordinate system of x
- Usually we estimate one possibility for Λ and Ψ and rotate the resulting loading matrix (multiply by an orthogonal matrix G) so as to ease interpretation
- Once the loadings and specific variances are estimated, estimated values for the factors themselves (called factor scores) are constructed

- When the data X is assumed to be normally distributed, then estimates to Λ and Ψ can be obtained by MLE
- Normal location parameter is replaced by its MLE $\bar{x}$, whilst the log-likelihood of the data depends on Λ and Ψ through Σ
- To ensure Λ is well defined (invariance is caused by orthogonal transformations) the computationally convenient uniqueness condition is used
  $\Lambda^T \Psi^{-1} \Lambda = \Delta$ where Δ is a diagonal matrix
- Numerical optimisation of log likelihood performed to obtain MLE $\hat{\Lambda}$ and $\hat{\Psi}$

- Problems occur if $\psi_{ii} = 0$ (when uniqueness is 0) the variance in the variable X_i is completely accounted for by the factors f_i
- problem if $\psi_{ii} < 0$ → Heywood Case
  ⇒ occurs if there are too many common factors, too few common factors, not enough data or inappropriate application of model for data
  ⇒ computer resolves this by resetting ψ_{ii} to be a small positive number

Factor Rotation

- If the initial loadings are subject to an orthogonal transformation (ie. multiplied by orthogonal matrix G) the covariance can still be reproduced
- An orthogonal transformation corresponds to a rigid rotation or reflection of the axis
- Hence orthogonal transformation of factor loadings known as factor rotation
- Doesn't matter if we use Λ or $\Lambda^* = \Lambda G$ since
  $$\Sigma = \Lambda\Lambda^T + \Psi = \Lambda G G^T \Lambda^T + \Psi = \Lambda^* \Lambda^{*T} + \Psi$$
- One rotation may be more useful than another

MVLA 5

FACTOR ANALYSIS

Controversial?

- Some say we are manipulating results, others say sharpening focus. Can be used as an intermediate step to reduce dimensionality prior to other technique

Simple Structure

- One idea - rotate the factors so that each variable has a large loading on a single factor and small loadings on the other. Variables can then be split into disjoint sets, each associated with one factor
- A factor j can be interpreted as an average quality over those variables i where λ_{ij} is large

Kaiser's Varimax rotation.

- Squared loading λ_{ij}^2 is the proportion of the variance in variable i that is attributed to common factor j.

$$\mathrm{Var}[X_i] = \lambda_{i1}^2 + \lambda_{i2}^2 + \dots \lambda_{ip}^2 + \psi_i$$
$$= h_i^2 + \psi_i$$

- Aim for a rotation that makes the squared loadings λ_{ij}^2 either large or small, i.e not medium sized value
- Let $\tilde{\lambda}_{ij}^* = \lambda_{ij}^*/h_i$ be the final rotated loadings scaled by $\sqrt{\ }$ of communalities

Varimax procedure selects the orthogonal transformation G maximising the sum of the column variances across all factors $j=1 \dots p$

$$V = \frac{1}{m} \sum_{j=1}^{p} \sum_{i=1}^{m} \tilde{\lambda}_{ij}^{*4} - \frac{1}{m^2} \sum_{j=1}^{p} \left(\sum_{i=1}^{m} \tilde{\lambda}_{ij}^{*} \right)^2$$

- Maxing V corresponds to "spreading out" the squares of the loadings on each factor as much as possible hence both groups of large and negligible coefficients are found in any column of Λ^*
- Scaling the rotated loadings has the effect of giving variables with small communalities relatively more weight. After G has been determined the loadings $\tilde{\lambda}_{ij}^*$ are multiplied by h_i to ensure original communalities are preserved

OBLIQUE ROTATIONS

- An oblique rotation respond to a non rigid rotation of the coordinate system, i.e resulting axes need no longer be perpendicular
- oblique relaxes the orthogonality constraint in order to gain simplicity of interpretation
- Example: promax rotation from procrustean rotation
- less popular than orthogonal

Factor Scores

- Interest usually lies in the parameters of a factor model (ie. λ_{ij} and ε_i) but the estimated values of the common factors (ie factor scores $\hat{f}$) may be required
- location of each original obs in reduced factor space is often necessary as input for a subsequent analysis
- One method to use to estimate factor scores is regression method
For MVN theory can be shown $\hat{f} = \Lambda^T \Sigma^{-1} X$

Interpretation

- communalities are the sum of square loadings in each row of loading matrix
- Communality for each variable added to the uniqueness for each variable is the total variance for that variable (=1 as R standardised data)
- The ss loading values are the sum of squared loadings for that factor
- Proportion of total standardised sample variance due to jth factor is:

$$\frac{\hat{\lambda}_{1j}^2 + \hat{\lambda}_{2j}^2 + \dots + \hat{\lambda}_{mj}^2}{\sigma_{11} + \sigma_{22} + \dots + \sigma_{mm}} = \frac{\hat{\lambda}_{1j}^2 + \dots + \hat{\lambda}_{mj}^2}{m} \quad m = 10 \text{ etc.}$$

PCA v Factor Analysis

- PCA looks for linear combinations of data matrix X that are uncorrelated and high variance, FA seeks unobserved linear combinations of variables representing underlying fundamental quantities
- PCA makes no assumption about form of covariance matrix, FA assumes data comes from a well defined model in which specific assumptions hold eg. $E[F] = 0$
- PCA: data $\Rightarrow$ PC's FA: factors $\Rightarrow$ data
- When specific variances are large, they are absorbed into PC's. FA makes special provision for them. When specific variances are small, PCA and FA give similar results
- Two analysis often performed together. Example could use PCA to determine number of factors to extract in FA study

IMLA

08/05/15

FACTOR ANALYSIS

- Mathematical approach for attempting to explain the correlation between a large set of variables in terms of a small number of underlying factors
- Main assumption of FA is that it is not possible to observe the underlying factors directly, i.e. they are 'latent'.

- Dim reduction tech, share same obj as PCA
- Attempt to describe the relationships in a large set of variables through a smaller number of dimensions
- FA much more elaborate

Example: performance of group of children in latin, eng and maths recorded. correlation matrix for scores was:

$$\begin{pmatrix} 1 & 0.83 & 0.78 \\ & 1 & 0.67 \\ & & 1 \end{pmatrix}$$

- Dimensionality of this matrix can be reduced from $m=3$ to $m=1$ by expressing variables as

$$x_1 = \lambda_1 F + \varepsilon_1$$
$$x_2 = \lambda_2 F + \varepsilon_2$$
$$x_3 = \lambda_3 F + \varepsilon_3$$

F is the underlying common factor
λ_i's are known as factor loadings
ε_i's are errors or specific factors
- common factor will be given an interpretation e.g. general ability
- Specific factors ε_i will have small variance if x_i is closely related to general ability

The Factor Model

- The observable random vector $X = (X_1 \dots X_m)$ mean μ cov matrix Σ
- FM states that X is linearly dependent upon a few unobservable random variables $f_1, f_2 \dots f_p$ called common factors and m additional sources of variation $\varepsilon_1, \varepsilon_2 \dots \varepsilon_m$ called specific factors

$$X_1 - \mu_1 = \lambda_{11} f_1 + \lambda_{12} f_2 + \dots + \lambda_{1p} f_p + \varepsilon_1$$

$$X - \mu = \Lambda F + \varepsilon$$

λ_{ij} is factor loading of ith variable on the jth factor.
Λ = matrix of factor loadings
ith specific factor is ε_i is associated only with response X_i
Also the $f_1, f_2 \dots f_m$, $\varepsilon_1 \dots \varepsilon_m$ are all unobservable random variables

Var FM assumes:

1. $\mathbb{E}[f] = 0$. $\quad Cov[f] = I = \begin{pmatrix} 1 & 0 & 0 \\ & 1 & \\ 0 & 0 & 1 \end{pmatrix}$

2. $\mathbb{E}[\varepsilon]=0 \quad Cov[\varepsilon] = \Psi = \begin{pmatrix} \psi_1 & 0 & 0 \\ 0 & \psi_2 & 0 \\ 0 & 0 & \psi_p \end{pmatrix}$

3. f and ε are independent, $Cov[f, \varepsilon] = 0$.

Above is referred to as orthogonal factor model. However if it is assumed that $Cov[f]$ is not diagonal, then the model is referred to as the oblique factor model

- Orthogonal FM implies a specific cov structure for x: $= \Lambda\Lambda^T + \Psi$
- can also be shown that $Cov[x, f] = \mathbb{E}[(x-\mu)(f-0)] = \Lambda$ hence the cov of the observed variable x_i and unobserved factor f_j is the factor loading λ_{ij}.

Variance

- Variance can be split in 2.
- First part for i^{th} component arises from the m common factors referred to as the i^{th} communality
- The remainder of variance is due to the specific factors referred to as uniqueness.
- Denoting i^{th} communality by h_i^2 the:

$$\sigma_i^2 = \lambda_{i1}^2 + \lambda_{i2}^2 + \dots \lambda_{ip}^2 + \psi_i$$
$$= h_i^2 + \psi_i$$

- i^{th} communality is the sum of squares of loadings of the i^{th} variable on the p common factors

- FM assumes that the $m + m(m-1)/2$ variances and covariances of x can be reproduced from the mp factor loadings λ_{ij} and the m specific variances ψ_i
- In case where $p << m$ the FM provides a simplified view of the covariance in x with fewer parameters than the $m(m+1)/2$ parameters in Σ.

- Example if $x = (x_1 \dots x_{12})$ FM with $p=2$ is valid. then the $12 \times 13/2 = 78$ elements of Σ are described in terms of the $pm + m = 2 \times 12 + 12 = 36$ parameters λ_{ij} and ψ_i of the factor model

- Most cov matrices can't be uniquely factored as $\Lambda\Lambda' + \Psi$ when $p << m$

Scale Invariance

- Rescaling variance of x is equivalent to letting $y = Cx$ where $C = diag(c_i)$
- FA not affected by re-scaling of variables
- leads to idea of factor rotation

ML for FA

- When data x is assumed to be normally distributed, then estimates for Λ and Ψ can be obtained by max in likelihood
- Uniqueness condition entailed $\Lambda^T \Psi^{-1} \Lambda = \Delta$

MLA 3

09/05/15

FACTOR ANALYSIS

- Problems may occur if $\psi_{ii} = 0$
- This happens when uniqueness is zero i.e. variance in the variable X_i is completely accounted for by the factors f_i
- May find $\psi_{ii} < 0$ improper solution called Heywood case
- Occurs if there are too many common factors, too few common factors, not enough data or the inappropriate application of the model for the data
- Resolved by re-setting ψ_{ii} to a small positive number before progressing

Factor Rotation

- If initial loadings are subject to an orthogonal transformation (i.e multiplied by an orthogonal matrix G), the cov matrix Σ can still be reproduced
- An orthogonal transformation corresponds to a rigid rotation or reflection of coordinate axes
- Called factor rotation
- It is immaterial whether Λ or $\Lambda^* = \Lambda G$ is reported since

$$\Sigma = \Lambda\Lambda^T + \psi = \Lambda G G^T \Lambda^T + \psi = \Lambda^* \Lambda^{*T} + \psi$$

- Argued that this is manipulating results
- Others say it is sharpening the focus
- FA generally used as an intermediate step to reduce the dimensionality of a data set prior to other statistical analysis

A Simpler structure

- One ideal would be to rotate the factors so that each variable has a large loading on a single factor and small loadings on the other
- Variables can then be split into disjoint sets, each of which is associated with one factor
- A factor j can then be interpreted as an abstract quality of the variables i for which λ_{ij} is large

Kaiser's Varimax Rotation

- Squared loading λ_{ij}^2 is the proportion of variance in variable i that is attributable to common factor j: $\quad var[X_i] = \lambda_{i1}^2 + \lambda_{i2}^2 + \dots \lambda_{ip}^2 + \psi_i$

$$= h_i^2 + \psi_i$$

- We aim for a rotation that makes squared loadings λ_{ij}^2 either large or small i.e. few medium sized values
- Let $\lambda^*_{ij} = \lambda^*_{ij}/h_i$ be the final rotated loadings scaled by the square root of the communalities
- Varimax procedure selects the orthogonal transformation G maximising the sum of the column variances across all factors $j = 1 \dots p$
- Maximising V corresponds to spreading out the squares of the loadings on each factor as much as possible
- Scaling rotated loadings has the effect of giving variables with small communalities relatively more weight. After G has been determined the loadings $\tilde{\lambda}_{ij}$ are multiplied by h_i to ensure original communalities are preserved

Oblique Rotation

- Degree of correlation allowed between factors is generally small (2 highly correlated factors ≡ one factor)
- resulting axes need no longer be perpendicular.
- OR therefore relax the orthogonality constraint in order to gain simplicity in interpretation
- Promax rotation ⇒ derived from procrustean rotation

- Communalities are the sum of squared loadings in each row of loading matrix
- Communality for each variable added to the uniqueness for each variable is the total variance for that variable = 1 because of standardisation
- SS loadings values are the sum of squared loadings for that factor
- Proportion Var = SS loadings / m   m=10 in eg.
- Interpreted like PCA.

PCA vs Factor Analysis

- Both PCA and FA have similar aims
- PCA looks for linear combos of the data matrix X that are uncorrelated and of high variance, while FA seeks unobserved linear combos of the variables representing underlying fundamental quantities
- PCA makes no assumption about the form of cov mat. FA assumes that the data comes from a well defined model in which specific assumptions hold. E [F] = 0 etc.
- PCA data ⇒ PCs   FA: factors ⇒ data
- When specific variances are large they are absorbed into the PCs whereas FA makes special provision for them. When specific variances are small, PCA and FA give similar results
- 2 analyses often performed together. Eg. PCA first to determine number of factors to extract in a FA study.

- the linear Regression	Metric/Non metric
- Assumptions	Classical / Least Squares) Kruskal
- Variance	Stress Sammon stress
- Orthogonal Varimax Rotation	Procrustes
- Oblique	PCA vs MDS.
- PCA vs FA.	