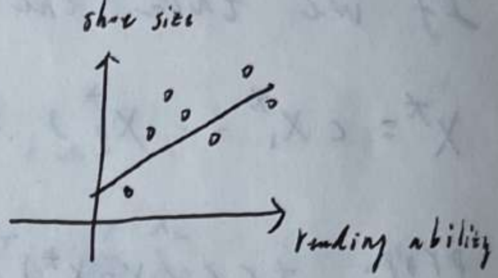
Factor Analysis

(1) <u>Background:</u>

i) A motivating example:

Shoe size and reading ability exist strong correlation



$\Rightarrow$ Latent variable: age.

ii) Purpose of Factor analysis:

- Reduce high dimensional data to a few representive variables
- Describe the relation between variables (correlation or covariance) by underlying variables.

<u>e.g.</u> Variables can be grouped by its correlation matrix.

$$\begin{pmatrix} \boxed{\text{large}} & & \text{small} \\ & \boxed{\text{large}} & \\ \text{small} & & \boxed{1} \end{pmatrix}$$

corr Matrix

<u>Rmk:</u> PCA is different from FA. Since PCA just transform the data matrix to reduce its dimension. It doesn't need modeling. But Factor analysis needs to find the underlying communal factors.

(2) Modeling:

① Orthogonal Factor Model:

$X = (X_1 \cdots X_p)^T$ observed random vector with $\begin{cases} E(X) = \mu \\ Var(X) = \Sigma. \end{cases}$

Assumptions: X is linear dependent upon a few r.v's $F_1 \cdots F_m$ called common factors and p specific factors (errors) $\varepsilon_1 \cdots \varepsilon_p$. All are unobservable.

satisfies: i) $E(\varepsilon) = 0_{p\times 1}$. $Cov(\varepsilon) = \Psi = diag\{\Psi_1 \cdots \Psi_p\}$.

ii) $E(F) = 0_{m\times 1}$ $Cov(F) = I_m$.

iii) F and ε are indept. $Cov(F, \varepsilon) = 0$.

Rmk: ii) isn't strong since we can do PCA to decompose it orthogonally.

Model: $X_i - \mu_i = \sum_{k=1}^{m} l_{ik} F_k + \varepsilon_i$. $1 \le i \le p$.

written in matrix: $X - \mu = LF + \varepsilon$.

Rmk: i) We want $m < p$ as possible.

ii) Likewise regression model: $Y = X\beta + \varepsilon$, since L is invariant, it's kind of β. And note that X consists different kinds of variables which differs a lot from regression model.

Def: l_{ij} is factor loading of i^{th} variable on j^{th} factor.

Some results: i) $Cov(X) = LL^T + \Psi$. i.e.

$$Cov(X_j, X_k) = \sum_{i=1}^{m} l_{ji} l_{ki}. \quad j \neq k.$$

$$\sigma_{ii} = \sum_{k=1}^{m} l_{ik}^2 + \psi_i =: \text{"Communality"} + \text{"Specific Var"}.$$

ii) $\sum_{1}^{p} Var(X_i) = \sum h_i^2 + \sum \psi_i$. where $h_i^2 = \sum_{k=1}^{m} l_{ik}^2$.

iii) $Cov(X, F) = L$. i.e. $Cov(X_i, F_j) = l_{ij}$.

Intepretions:

i) h_i^2 can be recognized as kind of geometric distance. It can measure the influence of $F_1 \dots F_m$ on X_i.

ii) Set $q_j^2 = \sum_{i=1}^{p} l_{ij}^2$. It represent the influence of F_j on the whole model.

Rmk: The model holds under scale transform:

i.e. For $C = diag\{c_1, \cdots c_p\}$. $c_i \neq 0$. $X^* = CX$.

$\mu^* = C\mu$. $\Sigma^* = C \Sigma C^T$. $\varepsilon^* = C\varepsilon$. Then:

$X^* = \mu^* + A^* F + \varepsilon^*$ still holds. $(A^* = CA)$

② Data Reduction:

The factor model assumes $p + \frac{p(p-1)}{2} = \frac{p(p+1)}{2}$ (w.r.t "Σ") variances. It can be reduced to pm factor loadings and p specific variances ψ_i. if $(m+1)p < \frac{p(p+1)}{2}$.

i.e. $m+1 < \frac{p+1}{2}$

Rmk: Unfortunately. Not every covariance matrix can be written in $LL^T + \psi$ form. where $m < p$.

③ Rotation Indeterminacy:

Set T is orthogonal matrix. $L^* = LT$. $F^* = T^T F$. Then:

$X - \mu = LF + \varepsilon = L^* F^* + \varepsilon$. $E(F^*) = 0$. $Cov(F^*) = I_m$.

$\Sigma = LL^T + \Psi = L^* L^{*T} + \Psi$.

i.e. L and F aren't unique. They have same properties under rotation.

Rmk: i) Var "Σ" is unaffected: because it's some kind of distance

ii) To make the L, F unique. We always impose some condition on them.

(3) Estimation:

· We will estimate L, Ψ and number m. $\Sigma = LL^T + \Psi$ will be estimated by S. (i.e. $S \simeq \hat{L}\hat{L}^T + \hat{\Psi}$)

① Principle Component Approach:

By spectral decomposition on Σ:

$$\Sigma = \sum_1^p \lambda_i e_i e_i^T = (\sqrt{\lambda_1} e_1 \cdots \sqrt{\lambda_p} e_p) \begin{pmatrix} \sqrt{\lambda_1} e_1^T \\ \vdots \\ \sqrt{\lambda_p} e_p^T \end{pmatrix} = LL^T$$

where we assume: $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$.

If the last $p-m$ eigenvalues are small. Then neglect them. $\Sigma \simeq (\sqrt{\lambda_1} e_1 \cdots \sqrt{\lambda_m} e_m) \begin{pmatrix} \sqrt{\lambda_1} e_1^T \\ \vdots \\ \sqrt{\lambda_m} e_m^T \end{pmatrix} = \tilde{L}\tilde{L}^T$

Specific variance may be taken $diag\{\Sigma - \tilde{L}\tilde{L}^T\}$.

Rmk: $\Sigma - \tilde{L}\tilde{L}^T$ may not be diagonal matrix.

Next. decompose $S = \sum_1^p \hat{\lambda}_i \hat{e}_i \hat{e}_i^T$. And set:

$\hat{L} = (\sqrt{\hat{\lambda}_1}\, \hat{e}_1 \hat{e}_1^T \cdots \sqrt{\hat{\lambda}_m}\, \hat{e}_m \hat{e}_m^T)$

i) Ψ is estimate by:

$$\hat{\Psi} = \begin{pmatrix} \hat{\psi}_1 & & 0 \\ & \ddots & \\ 0 & & \hat{\psi}_p \end{pmatrix}. \quad \hat{\psi}_i = S_{ii} - \sum_j^m \hat{l}_{ij}^2$$

ii) Communalities h_i is estimate by:

$$\hat{h}_i^2 = \sum_{k=1}^m \hat{l}_{ik}^2 .$$

Rmk: In this approach. It doesn't change the estimated loadings of given factors as the number of factors increase m to $m+1$.

iii) Select number m:

Consider $S - (\hat{L}\hat{L}^T + \hat{\Psi}) = \begin{pmatrix} 0 & & * \\ & \ddots & \\ * & & 0 \end{pmatrix} =: E_s$

$tr(E_s^T E_s)$ = sum of square entries of E_s

$\leq tr((S - \hat{L}\hat{L}^T)^T (S - \hat{L}\hat{L}^T))$ = SS of $S - \hat{L}\hat{L}^T$.

$= \sum_{m+1}^p \hat{\lambda}_k^2$ (Apply the spectral decompose)

$\Rightarrow$ Select m. s.t. $\sum_{m+1}^p \hat{\lambda}_k^2$ is small enough.

(which reduce the SSE of estimate)

Rmk: Alternatively. Set P_0. s.t. $\frac{\sum_1^m \lambda_i}{\sum_1^p \lambda_i} \geq P_0$.

Besides, we can consider the proportion of j^{th} factor $\hat{\lambda}_j / \sum_1^p S_{ii}$ to select.

② Modified PCA: Principle Factor Solution:

Ideal: The common factors should account for off diagonal as well as community portion of diagonal elements of Σ.

Algorithm: i) Guess $\hat{\Psi}$.

ii) Set L = the largest m eigenvectors of decomposition of $S - \hat{\Psi}$.

iii) Set $\hat{\Psi} = \mathrm{diag}(S - LL^T)$.

Repeat ii) and iii). untill convergence.

Rmk: i) The first step of estimate on $\hat{\Psi}$ and then set $S - \hat{\Psi}$ to decompose can ease the estimate of L on the diagonal elements. So L can better estimate the off-diagonal part of Σ.

ii) Choice of $\hat{\Psi}$ can be S^{-1}.

iii) $S - \hat{\Psi}$ may have negative eigenvalue. and estimate of $\hat{\Psi}$ in step iii) may also have negative diagonal. It's refered to Heywood case.

③ Maximum Likelihood Method:

Assumption: i) F and ε are jointly normal distribution.

ii) To make L well-def and uniqueness:

$L^T \Psi^{-1} L = \Delta$ is diagonal matrix.

$(\Leftrightarrow (\Psi^{-\frac{1}{2}} L)^T (\Psi^{-\frac{1}{2}} L) = \Delta)$

<u>Rmk</u>: ii) put $\frac{m(m-1)}{2}$ constraints to reduce the dimension of para. space to 1.

$\begin{pmatrix} \ddots & \Delta \\ 0 & \ddots \end{pmatrix}$ $\uparrow \frac{m(m-1)}{2}$

$$\Rightarrow L(\mu, \Sigma \mid X) = (2\pi)^{-\frac{np}{2}} |\Sigma|^{-\frac{n}{2}} \exp(-\frac{1}{2} tr(\Sigma^{-1} \sum_i^n (X_i - \mu)(X_i - \mu)^T))$$

To maximize it $\Leftrightarrow$ maximize $\ln|\Sigma| - \ln|S_n| + tr(\Sigma^{-1} S_n) - p$

where $\Sigma = LL^T + \Psi$. $S_n = \frac{(n-1)S}{n}$. subjects to: $L^T \Psi^{-1} L = \Delta$

MLE of L, Ψ can be obtained by numerical computation.

<u>Rmk</u>. By invariant property of MLE. The MLE of communalities are $\hat{h}_i^2 = \sum_{k=1}^m \hat{l}_{ik}^2$ of $\hat{L}$.

<u>For Standardilization:</u>

Set $Z = V^{-\frac{1}{2}}(X - \mu)$. So $\rho = V^{-\frac{1}{2}} \Sigma V^{-\frac{1}{2}}$. Set $L_Z = V^{-\frac{1}{2}} L$.

$\Psi_Z = V^{-\frac{1}{2}} \Psi V^{-\frac{1}{2}}$. Then $\rho = L_Z L_Z^T + \Psi_Z$. $(V = \begin{pmatrix} \sigma_{11} & & 0 \\ & \ddots & \\ 0 & & \sigma_{pp} \end{pmatrix})$

By invariance property: $\hat{\rho} = (\hat{V}^{-\frac{1}{2}} \hat{L})(\hat{V}^{-\frac{1}{2}} \hat{L})^T + \hat{V}^{-\frac{1}{2}} \hat{\Psi} \hat{V}^{-\frac{1}{2}}$.

where $\hat{V}^{-\frac{1}{2}}$. $\hat{L}$ are MLE of $V^{-\frac{1}{2}}$. L.

<u>Rmk</u>: i) For PCA. There're usually no relation between PC of "Σ" and "R". But in this method. They're essentially equi.

ii) The result will be very different when increase the number from m to $m+1$.

That's because we impose different assumption comparing to PCA method.

iii) Heywood case may also happen ($\hat{\psi}_i < 0$).

iv) MLE method produce F_i won't be orthogonal:

$$(\underline{\Psi}^{-\frac{1}{2}} L)^T (\underline{\Psi}^{-\frac{1}{2}} L) = \Delta$$

④ Large Sample Test for m:

Assumption: i) F and ε are jointly normal distribution.

ii) $L^T \Psi^{-1} L = \Delta$ diagonal matrix.

Test: $H_0: \Sigma_{p\times p} = LL^T + \Psi_{p\times p}$. $L \in M^{p\times m}$ v.s. $H_1: \Sigma$ any other.

i) $\sup L(\Sigma, \mu) = L(S_n, \bar{X})$

ii) $\sup\limits_{H_0} L(\Sigma, \mu) = L(\hat{L}\hat{L}^T + \hat{\Psi}, \bar{X})$, $\hat{L}, \hat{\Psi}$ are MLE of L and Ψ. $L(\hat{L}\hat{L}^T + \hat{\Psi}, \bar{x}) \propto |\hat{L}\hat{L}^T + \hat{\Psi}|^{-\frac{n}{2}} \exp(-\frac{n}{2} tr((\hat{L}\hat{L}^T + \hat{\Psi})^{-1} S_n))$

Rmk: $tr((\hat{L}\hat{L}^T + \hat{\Psi})^{-1} S_n) = p$.

$$\Rightarrow -2\ln\Lambda = n\ln\left(\frac{|\hat{L}\hat{L}^T + \hat{\Psi}|}{|S_n|}\right) \sim \chi^2(df).$$

$$df = \nu - \nu_0 = \frac{1}{2}p(p+1) - [p(m+1) - \frac{1}{2}m(m-1)]$$

$= \#$ para. of Σ $- [\#$ para. of L, Ψ $- \#$ Constraint$]$

$$= \frac{1}{2}[(p-m)^2 - p - m]$$

Rmk: For $df > 0 \Rightarrow m < \frac{1}{2}(2p+1-\sqrt{8p+1})$

Bartlett show the converge of approxi. can be improved if replace n by $n-1-(2p+4m+5)/6$

$$\Rightarrow R = \{ (n-1-\frac{2p+4m+5}{6}) \ln \frac{|\hat{L}\hat{L}^T + \hat{\Psi}|}{|S_n|} > \chi^2_{df}(\alpha) \}$$

(4) Factor Rotation:

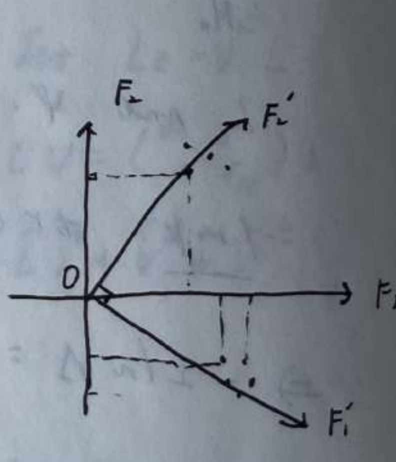
Since the original loading may not be readily interpretable. e.g. some l_{ij} are positive and some are negative which are large.

To achieve a simple structure. We need rotate L and F.

Rmk: Dimension of rotation for factor loading is $\frac{m(m-1)}{2}$. For $T = (T_1 \cdots T_m)$. Constraint: $T_1 \perp T_2$. $T_1 \perp T_3$. $T_2 \perp T_3$. $\cdots$ $1+2+\cdots+(m-1)$.

① Simple Structure:

Ideally. We like to see a pattern that each variable loads highly on a single factor and loads small on the remaining factors.



e.g. Rotate $F_1 0 F_2$ to $F_1' 0 F_2'$. A partition into mutually exclusive groups would be desirable.

⑦ Varimax Criteria:

· suppose $\hat{L}^* = (\hat{l}^*_{ij})$ is $\hat{L}$ after rotation

Def: $\tilde{l}^*_{ij} = \hat{l}^*_{ij} / \hat{h}_i$. (i.e. scaling. consider the weight)

$$\lambda_{ij}^2 = \tilde{l}^{*2}_{ij} . \quad \bar{\lambda}_j = \sum_{i=1}^{p} \lambda_{ij}^2 / p .$$

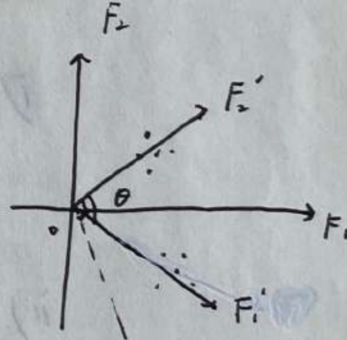
$$\Rightarrow V_j = \sum_{i=1}^{p} (\lambda_{ij}^2 - \bar{\lambda}_j)^2 / p = \frac{1}{p} \left[\sum_{i=1}^{p} \frac{\hat{l}^{*4}_{ij}}{\hat{h}_i^4} - \left(\sum_{j=1}^{p} \frac{\hat{l}^{*2}_{ij}}{\hat{h}_i^2} \right)^2 / p \right]$$

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

Select T st. make V as large as possible. i.e. make the data loading of F_j disperse enough.

Intepretion: $V \propto \sum_{j=1}^{m} \left(\begin{array}{c} \text{Var of square of scaled} \\ \text{loading for } j^{th} \text{ factor} \end{array} \right)$

Rmk: In some case. even orthogonal rotate don't provide an easy intapretation. It's possible to use the oblique rotations at expense of orthogonality of factors. i.e. $Cov(F^*) \neq I_m$

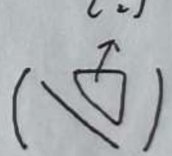


e.g. $m=2$. $A = \begin{pmatrix} a_{11} & a_{12} \\ & \vdots & \\ a_{p1} & a_{p2} \end{pmatrix}$. set $T = \begin{pmatrix} \cos\varphi & -\sin\varphi \\ \sin\varphi & \cos\varphi \end{pmatrix}$

$B = AT = \begin{pmatrix} b_{11} & b_{12} \\ & \vdots & \\ b_{p1} & b_{p2} \end{pmatrix}$. $V_t = \frac{1}{p^2} \left[p \sum_{i=1}^{p} b_{it}^4 / h_i^4 - \left(\sum_{i}^{p} \frac{b_{it}^2}{h_i^2} \right)^2 \right]$

$$\frac{\partial V}{\partial \varphi} = \frac{\partial (V_1 + V_2)}{\partial \varphi} = 0 \Rightarrow \text{obtain } \varphi$$

$\binom{m}{2}$



For $m > 2$. consider: (F_i, F_j). $i \neq j$. $\binom{m}{2}$ times

(4) Factor Scores:

We will predict / estimate the unobservable random factors $F_1, F_2 \cdots F_m$.

① Weighted Least Square:

$X = \mu + LF + \varepsilon$. $\varepsilon \sim N(0, \Psi)$ Assume. L, Ψ, μ are known.

minimize: $\varepsilon^T \Psi^{-1} \varepsilon = (X-\mu-LF)^T \Psi^{-1} (X-\mu-LF)$

$\Rightarrow$ Weighted LSE is $\hat{F} = (L^T \Psi^{-1} L)^{-1} L^T \Psi^{-1} (X-\mu)$

Rmk: Practically, L, Ψ, μ are unknown. We take estimate $\hat{L}, \hat{\Psi}, \hat{\mu} = \bar{X}$ to replace it in WLSE. $\Rightarrow \hat{F}_j = (\hat{L}^T \hat{\Psi}^{-1} \hat{L})^{-1} \hat{L}^T \hat{\Psi}^{-1} (x_j - \bar{x})$

i) MLE method:

$\hat{L}^T \hat{\Psi}^{-1} \hat{L} = \hat{\Delta} \Rightarrow \hat{F}_j = \hat{\Delta}^{-1} \hat{L}^T \hat{\Psi}^{-1} (x_j - \bar{x})$

ii) Correlation matrix factors:

$\hat{F}_j = (\hat{L}_z^T \hat{\Psi}_z^{-1} \hat{L}_z)^{-1} \hat{L}_z^T \hat{\Psi}_z^{-1} z_j, \quad z_j = V^{-\frac{1}{2}} (x_j - \bar{x})$

② Regression Method:

i) Thompson Factors:

Express F in X: $F_i = \sum_{k=1}^{p} \beta_{ik} (X_k - \mu) =: \sum_{1}^{p} \beta_{ik} \tilde{X}_k$.

Develop model: $F_i = \sum_{k=1}^{p} b_{ik} \tilde{X}_k$. $1 \le i \le m$.

$\Rightarrow l_{ij} = Cov(F_j, X_i) = \sum_{k=1}^{p} b_{jk} \sigma_{ik}$.

$\therefore B = L^T \Sigma^{-1}$. We obtain: $\hat{F} = L^T \Sigma^{-1} (X - \mu)$.

$\Rightarrow \hat{F} = \hat{L}^T (\hat{L}\hat{L}^T + \hat{\Psi})^{-1} (X - \bar{X})$.

ii) <u>Bayesian Method:</u>

Assume F, ε are jointly normal distribution.

$$\begin{pmatrix} X-\mu \\ F \end{pmatrix} \sim N_{m+p}(0, \Sigma^*),\quad \Sigma^* = \begin{pmatrix} \Sigma & L \\ L^T & I \end{pmatrix},\quad \Sigma = LL^T + \Psi.$$

$$\Rightarrow \begin{cases} E(F|X) = E(F) + L^T\Sigma^{-1}(X-\mu) = L^T\Sigma^{-1}(X-\mu) \\ Cov(F|X) = I - L^T\Sigma^{-1}L. \end{cases}$$

$\therefore \hat{F} = \hat{L}^T(\hat{L}^T\hat{L} + \hat{\Psi})^{-1}(X - \bar{X})$. identical with i).

③ <u>Comparison:</u>

Note that: $\hat{L}^T(\hat{L}\hat{L}^T + \hat{\Psi})^{-1} = (I + \hat{L}^T\hat{\Psi}^{-1}\hat{L})^{-1}\hat{L}^T\hat{\Psi}^{-1}$.

(By calculate $\begin{pmatrix} \hat{\Psi} & \hat{L} \\ \hat{L}^T & I \end{pmatrix}^{-1}$)

$$\Rightarrow \hat{F}^{LS} = (\hat{L}^T\hat{\Psi}^{-1}\hat{L})^{-1}(I + \hat{L}^T\hat{\Psi}^{-1}\hat{L})\hat{F}^R$$
$$= (I + (\hat{L}^T\hat{\Psi}^{-1}\hat{L})^{-1})\hat{F}^R$$

<u>Rmk:</u> In MLE method, if $\hat{\Delta} = (\hat{L}^T\hat{\Psi}\hat{L}) \approx 0$.

then $\hat{F}^{LS} = \hat{F}^R$.

i) $E(\hat{F}^{LS}|F) = F$. unbiased.

$E(\hat{F}^R|F) = (I + L^T\Psi^{-1}L)^{-1}L^T\Psi^{-1}LF$. biased.

ii) $E((\hat{F}^{LS} - F)(\hat{F}^{LS} - F)^T) = (L^T\Psi L)^{-1}$.

$E((\hat{F}^R - F)(\hat{F}^R - F)^T) = (I + L^T\Psi^{-1}L)^{-1}$.

(6) Strategies:

i) Perform PCA

ii) Perform MLE

iii) Compare Solution of i). ii).

iv) Repeat i). ii). iii) for other number of common factors m.

v) For large data sets. split them into half and perform FA.

Rmk: i) To reduce the effect of incorrect determination of number m. S is often used to estimate $\hat{\Sigma}$ rather than $\hat{L}\hat{L}^T + \hat{\Psi}$.

ii) After rotation $\hat{L}^* = T\hat{L}$. Then $\hat{F}^* = T^T\hat{F}$. should be used.

iii) After rotation, difference between estimate of MLE and PCA may be little. Since it breaks the constraints.