## Factor Analysis

$\underset{\sim}{X} = \underset{\sim}{\mu} + L\underset{\sim}{F} + \underset{\sim}{\varepsilon}$ $\quad \text{Cov}(\underset{\sim}{\varepsilon}) = \Psi$

$\downarrow$

$\text{Cov}(\underset{\sim}{F}) = I$

$\text{Cov}(\underset{\sim}{X}) = LL^T + \Psi$

$= \underset{\sim}{\mu} + (LQ)(Q^T\underset{\sim}{F}) + \underset{\sim}{\varepsilon}$

$= \underset{\sim}{\mu} + L^*\underset{\sim}{F}^* + \underset{\sim}{\varepsilon}$

Last time: Principal factors estimation.
- iterative algorithm computing $\hat{L}$ based on $\hat{\Psi}$ and updating $\hat{\Psi}$ based on $\hat{L}$.

## Maximum Likelihood estimation.

R function: factanal

Assume $\underset{\sim}{X} \sim N_p(\underset{\sim}{\mu}, LL^T + \Psi)$

MLE of $\underset{\sim}{\mu} = \hat{\underset{\sim}{\mu}} = \frac{1}{n}\sum_{i=1}^{n} \underset{\sim}{x}_i$

$\ln\mathcal{L}(L, \Psi) = -\frac{n}{2}[\ln(|LL^T + \Psi|) + \text{trace}((LL^T + \Psi)^{-1}\hat{C}) + \cdots]$

$\downarrow$ log likelihood

Where $\hat{C} = \frac{1}{n}\sum_{i=1}^{n}(\underset{\sim}{x}_i - \bar{\underset{\sim}{x}})(\underset{\sim}{x}_i - \bar{\underset{\sim}{x}})^T = \frac{n-1}{n}S$

- maximizing $\ln\mathcal{L}$ is not particularly easy.
  - need to assume $\hat{\Psi}$ has positive diagonal elements etc.
- likelihood function can be used to examine / test adequacy of $r$ factor model.

Test $H_0$: $C = LL^T + \Psi$ for some $p \times r$ $L$ and diagonal $\Psi$.
$H_1$: $C$ is arbitrary pos. def.

$V$ = 2 × difference in max'ed $\ln\mathcal{L}$

$\sim \chi^2(\nu)$ where $\nu = \frac{1}{2}(p-r)^2 - \frac{1}{2}(p+r)$

e.g. $p=5, r=2$, then $\nu = \frac{9}{2} - \frac{7}{2} = 1$

~~$p=5, r=3, \nu = \frac{4}{2} - \frac{3}{2} < 0$~~

Observe $V = v_{obs}$ then p-value $\doteq P(\chi^2(\nu) \geq v_{obs})$
- small p-value indicates $r$ factor model not adequate.
- larger p-value indicates $r$ factor model is adquate.

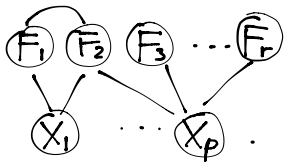
## Choosing rotations

Graphical model representation

$\underset{\sim}{Y} = \begin{pmatrix} \underset{\sim}{X} \\ \underset{\sim}{F} \end{pmatrix}$ $\quad \text{Cov}(\underset{\sim}{Y}) = \begin{pmatrix} LL^T + \Psi & L \\ L^T & I \end{pmatrix} = C_+$

Concentration matrix ← diagonal

$$K_+ = C_+^{-} = \begin{pmatrix} \Psi^{-1} & -\Psi^{-1}L \\ -L^T\Psi^{-1} & I + L^T\Psi^{-1}L \end{pmatrix}$$




Graphical structure is simplest if $L$ has many 0 elements

$$L = \begin{pmatrix} l_{11} & l_{12} & \cdots & l_{1r} \\ \vdots & & & \vdots \\ l_{p1} & & \cdots & l_{pr} \end{pmatrix} = (\underset{\sim}{l}_1 \cdots \underset{\sim}{l}_r)$$

no link/edge between $X_i$ & $F_j$ if $l_{ij} = 0$.

$\underset{\sim}{l}_i^T \Psi^{-1} \underset{\sim}{l}_j = 0$ then have no link between $F_i$ & $F_j$

$\underset{\sim}{l}_i = \begin{pmatrix} 0 \\ \vdots \end{pmatrix}$ $\underset{\sim}{l}_j = \begin{pmatrix} \vdots \\ 0 \\ 0 \end{pmatrix}$ This occurs if $l_{ki} l_{kj} = 0$ for all $k = 1, \cdots, p$

## Type of rotations

① orthogonal $L \to LQ$ where $QQ^T = I$
② Oblique $L \to LQ$ where $Q$ is not orthogonal
Note: Assume variables standardized → mean 0, variance 1
⇒ this allows loadings to be comparable.

## Varimax rotation

Idea: Maximize variance of squared elements of each $\underset{\sim}{l}_1, \cdots, \underset{\sim}{l}_r$.

maximize

$$\sum_{j=1}^{i} \sum_{i=1}^{p} \left[ \frac{l_{ij}^2}{\hat{\psi}_i} - \frac{1}{p} \sum_{k=1}^{p} \frac{l_{kj}^2}{\hat{\psi}_k} \right]^2$$

- forces squared elements of each $\underset{\sim}{l}_j$ to be as dispersed as possible.
- "encourages" 0 loadings
- varimax is the default in factanal.

## Oblique rotation

Take $Q$ non-orthogonal with inverse $Q^{-1}$.

Model: $\underset{\sim}{X} = \underset{\sim}{\mu} + L\underset{\sim}{F} + \underset{\sim}{\varepsilon} = \underset{\sim}{\mu} + (LQ)(Q^{-1}\underset{\sim}{F}) + \underset{\sim}{\varepsilon}$

$$= \underset{\sim}{\mu} + L^* \underset{\sim}{F}^* + \underset{\sim}{\varepsilon}$$

$$Cov(\underset{\sim}{F}^*) = Q^{-1} \underbrace{Cov(\underset{\sim}{F})}_{I} (Q^{-1})^T = Q^{-1}(Q^{-1})^T \neq I$$

$F_1^*, \cdots, F_r^*$ are no longer uncorrelated.

Problem? No practical reason to exclude correlated factors.

Example: Promax rotation.
- effectively raises varimax loadings to some power greater than 1 (e.g. 2, 4, 6)

$\Rightarrow$ emphasises large loadings, reduces small loadings

## Application to athletics records data

$p=8$ events, $n=55$ countries.
Look at 1+2 factor models with various rotations (see Blackboard)

(r=1) uniqueness $(\psi_1, \dots, \psi_8)$

| | 100m | 200m | 400m | 800m | 1500m | 5000m | 10000m | marathon |
|---|---|---|---|---|---|---|---|---|
| $\psi_1, \dots, \psi_8 \rightarrow$ | 0.447 | 0.389 | 0.242 | 0.114 | 0.059 | 0.104 | 0.369 | 0.205 |
| loadings $\rightarrow$ | 0.748 | 0.782 | 0.871 | 0941 | 0.970 | 0.946 | 0.794 | 0.892 |

But p-value $= 1.63 \times 10^{-16} \Rightarrow$ 1 factor model is not adequate

(r=2)

| | 100m | 200m | 400m | 800m | 1500m | 5000m | 10000m | marathon |
|---|---|---|---|---|---|---|---|---|
| $\psi_1, \dots, \psi_8 \rightarrow$ | 0.076 | 0.123 | 0.147 | 0.128 | 0.081 | 0.037 | 0.333 | 0.092 |
| loadings $\rightarrow$ | 0.295 | 0.378 | 0.552 | 0.711 | 0.807 | 0.904 | 0.738 | 0.913 |
| (varimax) | 0.915 | 0.857 | 0.740 | 0.605 | 0.517 | 0.382 | 0.350 | 0.217 |

p-value $= 0.306 \Rightarrow$ 2 factor model is ok.

Blackboard: 2 other rotations
① rotation = "none" $\rightarrow$ close to 1st 2 PC loadings
② promax.

## Cluster Analysis

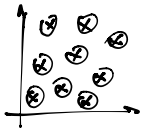
Data $\underline{x}_1, \dots, \underline{x}_n$ (often labelled by country or individual)
Goal: Try to group together (cluster) "similar" observations.
How?
Define distance function between pairs of points: Euclidean distance. $\left(\sum_{i=1}^{n} (x_i - y_i)^2\right)^{\frac{1}{2}} = d(\underline{x}, \underline{y})$
Manhatten distance $d(\underline{x}, \underline{y}) = \sum_{i=1}^{n} |x_i - y_i|$

Describe inter-observation distances using a distance matrix.

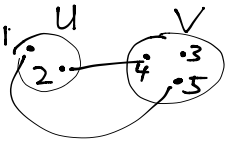
$$D = \begin{pmatrix} 0 & d_{12} & d_{13} & \cdots & d_{1n} \\ d_{21} & \cdot & & & \vdots \\ \vdots & & \cdot & & \vdots \\ & & & \cdot & \\ d_{n1} & \cdots & & & 0 \end{pmatrix} \quad \text{where } d_{ij} = d(\underline{x}_i, \underline{x}_j)$$

## Hierarchical clustering methods

Idea: Start with n clusters each with 1 observation.
- Group together similar observations i.e. observations with small distances.



Need some way to measure distances between clusters
Three approaches:
① Single linkage
② complete linkage
③ average linkage



① Single linkage: $d(U,V) = d_{24}$

$$d(U,V) = \min[d(x_i, x_j) : x_i \in U,\ x_j \in V]$$

② complete linkage: $d(U,V) = d_{15}$

$$d(U,V) = \max[d(x_i, x_j) : x_i \in U,\ x_j \in V]$$

③ average linkage: $d(U,V) = \frac{1}{6}(d_{14} + d_{15} + d_{13} + d_{23} + d_{24} + d_{25})$

$$d(U,V) = \text{average}[d(x_i, x_j) : x_i \in U,\ x_j \in V]$$

Note: $d(U,V) = d(V,U)$

Triangle inequality? $d(U,V) \le d(U,W) + d(W,V)$

General algorithm:

1. Start with n clusters each with single observation ⟶ distance matrix D
2. Search D to find nearest pair of clusters (depends on approach), call these U+V
3. Merge U+V into a single cluster and redefine D.
4. Repeat steps 2+3 until we have a single cluster

Example: Single linkage

$$D = \begin{array}{c} \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 9 & 0 & & & \\ 3 & 7 & 0 & & \\ 6 & 5 & 9 & 0 & \\ 11 & 10 & 2 & 8 & 0 \end{pmatrix}$$

Link observations 3 + 5 into a new cluster (call it "35")

$$D' = \begin{array}{c} 35 \\ 1 \\ 2 \\ 4 \end{array} \begin{pmatrix} 35 & 1 & 2 & 4 \\ 3 & 0 & & \\ 7 & 9 & 0 & \\ 8 & 6 & 5 & 0 \end{pmatrix}$$

Link (35) & 1 ⟹ (1,3,5)

$$D'' = \begin{array}{c} 135 \\ 2 \\ 4 \end{array} \begin{pmatrix} 135 & 2 & 4 \\ 7 & 0 & \\ 6 & 5 & 0 \end{pmatrix}$$

Link 2+4 ⟶ 24

$$D''' = \begin{array}{c} 135 \\ 24 \end{array} \begin{pmatrix} 135 & 24 \\ 6 & 0 \end{pmatrix}$$

This is called a Dendogram: