Factor Analysis

- Factor analysis is used to make inferences about unobservable quantities such as intelligence, musical ability, patriotism, consumer attitudes, that cannot be measured directly.

- The goal of factor analysis is to describe correlations between p variables in terms of variation in a few underlying and unobservable *factors*. This unobserved factors are also known as *latent variables*.

- Changes across subjects in the values of one or more of the unobserved factors could affect the values of an entire subset of variables and cause them to be highly correlated.

Objectives of Factor Analysis

- Identification of underlying factors:
  - cluster variables into homogeneous groups
  - create new variables (e.g., factors)
  - allows one to gain insight into categories

- screening of variables
  - select few variables to represent a larger set
  - handling collinearity in regression

Motivating Example with a Factor Analysis Model

Spearman considered a sample of children's examination marks in three subjects: Classics (x_1), French (x_2), and English (x_3) with the following correlation matrix

$$\begin{pmatrix} 1.00 & & \\ 0.83 & 1.00 & \\ 0.78 & 0.67 & 1.00 \end{pmatrix}.$$

A *single-factor model* can be assumed as follows:

$$x_1 = \ell_1 f + \epsilon_1,$$
$$x_2 = \ell_2 f + \epsilon_2,$$
$$x_3 = \ell_3 f + \epsilon_3.$$

Orthogonal Factor Model (OFM)

Given a $p-$dimensional random vector $\mathbf{X} = (X_1, X_2, \ldots, X_p)'$ with population mean vector $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_p)$ and population covariance matrix Σ, the *Orthogonal Factor Model* is

$$\begin{array}{rcl} X_1 - \mu_1 & = & \ell_{11}F_1 + \ell_{12}F_2 + \cdots + \ell_{1m}F_m + \epsilon_1 \\ X_2 - \mu_2 & = & \ell_{21}F_1 + \ell_{22}F_2 + \cdots + \ell_{2m}F_m + \epsilon_2 \\ \vdots & & \vdots \\ X_p - \mu_p & = & \ell_{p1}F_1 + \ell_{p2}F_2 + \cdots + \ell_{pm}F_m + \epsilon_p \end{array},$$

- μ_j is the *mean* of the jth variable.
- F_i's are the *common factors* (latent variables): $F_i \sim (0, 1)$ and $cov(F_j, F_k) = 0$ for $j \neq k$.
- ℓ_{ij}'s are the *factor loadings*.
- ϵ_i's are measurement errors (also called specific factor): $cov(\epsilon_j, \epsilon_k) = 0$ for $j \neq k$.

Matrix Notation

In matrix notation,

$$(\mathbf{X} - \boldsymbol{\mu})_{p\times 1} = L_{p\times m}\mathbf{F}_{m\times 1} + \boldsymbol{\epsilon}_{p\times 1}$$

with

$$L = \begin{bmatrix} \ell_{11} & \cdots & \ell_{1m} \\ \ell_{21} & \cdots & \ell_{2m} \\ \vdots & & \vdots \\ \ell_{p1} & \cdots & \ell_{pm} \end{bmatrix}, \mathbf{F} := (F_1, F_2, \ldots, F_m)'$$

- Assumptions:

$$E(\mathbf{F}) = \mathbf{0}, \qquad Var(\mathbf{F}) = E(\mathbf{F}\mathbf{F}') = I_{m\times m},$$

$$E(\boldsymbol{\epsilon}) = \mathbf{0}, \qquad Var(\boldsymbol{\epsilon}) = \Psi = \text{diag}\{\psi_i\}$$

  and $\mathbf{F}, \boldsymbol{\epsilon}$ are independent, so that $Cov(\mathbf{F}, \boldsymbol{\epsilon}) = 0_{m\times p}$.
- How to get $Var(\mathbf{X})$ and $Cov(\mathbf{X}, \mathbf{F})$?

OFM for Sample Data

Suppose that we have n observations $\mathbf{x}_i, i = 1, \ldots, n$. Then the OFM for $\mathbf{x}_i$'s is

$$\begin{array}{rcl} x_{i1} - \mu_1 & = & \ell_{11}F_{i1} + \ell_{12}F_{i2} + \cdots + \ell_{1m}F_{im} + \epsilon_{i1} \\ x_{i2} - \mu_2 & = & \ell_{21}F_{i1} + \ell_{22}F_{i2} + \cdots + \ell_{2m}F_{im} + \epsilon_{i2} \\ \vdots & & \vdots \\ x_{ip} - \mu_p & = & \ell_{p1}F_{i1} + \ell_{p2}F_{i2} + \cdots + \ell_{pm}F_{im} + \epsilon_{ip} \end{array},$$

- In matrix notation:

$$\underset{p\times 1}{\mathbf{x}_i} - \underset{p\times 1}{\boldsymbol{\mu}} = \underset{p\times m}{L}\underset{m\times 1}{\mathbf{F}_i} + \underset{p\times 1}{\boldsymbol{\epsilon}_i}.$$

- We often consider OFM for standardized data $\mathbf{z}_i := (z_{i1}, \ldots, z_{ip})'$ where $z_{ij} := (x_{ij} - \bar{x}_j)/(\sqrt{s_{jj}})$.

Terminologies

- Under the orthogonal factor model:

$$\begin{aligned}\text{Var}(X_i) &= \sigma_{ii} = \ell_{i1}^2 + \ell_{i2}^2 + \cdots + \ell_{im}^2 + \psi_i \\ \text{Cov}(X_i, X_k) &= \sigma_{ik} = \ell_{i1}\ell_{k1} + \ell_{i2}\ell_{k2} + \cdots + \ell_{im}\ell_{km}.\end{aligned}$$

- The *communality* of X_i (denoted by h_i is portion of the variance of X_i that is explained by the m common factors:

$$h_i^2 := \ell_{i1}^2 + \ell_{i2}^2 + \cdots + \ell_{im}^2.$$

- $Var(X_i) = h_i^2 + \psi_i$, where ψ_i is part of variance that is not explained by the common factors and is called the *uniqueness* (to X_i).

- If X_i has been standardized, then $\psi_i = 1 - h_i^2$.

Rotation of Factor Loadings

- When $m > 1$, there is no unique set of loadings and thus there is ambiguity associated with the factor model.
- Consider any $m \times m$ orthogonal matrix T such that $TT' = T'T = I$. We can rewrite our model as

$$\mathbf{X} - \boldsymbol{\mu} = L\mathbf{F} + \boldsymbol{\epsilon} = LTT'\mathbf{F} + \boldsymbol{\epsilon} = L^*\mathbf{F}^* + \boldsymbol{\epsilon},$$

with $L^* = LT$ and $\mathbf{F}^* = T'\mathbf{F}$.

- In addition,

$$\begin{aligned} E(\mathbf{F}^*) &= T'E(\mathbf{F}) = \mathbf{0}, \\ \mathsf{Var}(\mathbf{F}^*) &= T'\mathsf{Var}(\mathbf{F})T = T'T = I, \\ LL' + \Psi &= LTT'L' + \Psi = L^*L^{*'} + \Psi. \end{aligned}$$

- Conclusion: the loadings are *not uniquely* determined.

Rotation of Factor Loadings

- How to resolve this ambiguity?

- Typically, we first obtain the matrix of loadings (recognizing that it is not unique) and then rotate it by mutliplying by an orthogonal matrix.

- We choose the orthogonal matrix to optimize some desired criterion. For example, a *varimax rotation* of the factor loadings results in a set of loadings with maximum variability, i.e., some coefficients are large (positive or negative) and some are close to zero).

- The varimax rotation is the default rotation for some R functions, but other types of rotations may sometimes be available.

Estimation in Orthogonal Factor Models

- Consider a random sample of n observations $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n$ from some population.
- We often assume the OFM for the standardized data $\mathbf{z}_i$:
- For the m factors, we want to estimate the factor loading matrix L and the factors $F_i, i = 1, \ldots, n$.
- We use the sample correlation matrix R as an estimate of the population correlation matrix.
- Estimation:
  - First step: estimate L (factor loadings), where a rotation of L may be used to improve interpretation.
  - Second step: estimate F_i (factor scores) given L.

Estimation in Orthogonal Factor Models

- Common methods for estimation of factor loadings are:
  - The principal component method
  - The iterative principal factor method
  - Maximum likelihood estimation (assumes normality)

- The last two methods focus on using variation in common factors to describe correlations among measured traits. Principal component analysis gives somewhat more attention to variances.

- Estimated factor loadings from any of those methods can be rotated, as explained later, to facilitate interpretation of results.

The Principal Component Method

- The factor model is

$$\underset{p\times 1}{\mathbf{z}_i} = \underset{p\times 1}{L}_{p\times m}\underset{m\times 1}{\mathbf{F}_i} + \underset{p\times 1}{\boldsymbol{\epsilon}_i}$$

- Let $(\widehat{\lambda}_i, \widehat{\mathbf{e}}_i)$ denote the eigenvalues and eigenvectors of R and recall the spectral decomposition

$$R \;=\; \widehat{\lambda}_1\widehat{\mathbf{e}}_1\widehat{\mathbf{e}}_1' + \widehat{\lambda}_2\widehat{\mathbf{e}}_2\widehat{\mathbf{e}}_2' + \cdots + \widehat{\lambda}_p\widehat{\mathbf{e}}_p\widehat{\mathbf{e}}_p'.$$

- Choose m such that $\widehat{\lambda}_1 + \ldots + \widehat{\lambda}_m$ is much larger than $\widehat{\lambda}_{m+1} + \ldots + \widehat{\lambda}_p$.
- The first m terms gives the best rank-m approximation to R. The m rank approximation matrix is

$$R^{(m)} := \widehat{\lambda}_1\widehat{\mathbf{e}}_1\widehat{\mathbf{e}}_1' + \widehat{\lambda}_2\widehat{\mathbf{e}}_2\widehat{\mathbf{e}}_2' + \cdots + \widehat{\lambda}_p\widehat{\mathbf{e}}_m\widehat{\mathbf{e}}_m'$$

The Principal Component Method

- Estimate L by $\widehat{L} = L^{(m)}$, where $L^{(m)}$ is the $p \times m$ matrix with the columns $\sqrt{\widehat{\lambda}_1}\widehat{\mathbf{e}}_1, \ldots, \sqrt{\widehat{\lambda}_m}\widehat{\mathbf{e}}_m$.
- Estimate Ψ by $\widehat{\Psi} = diag(R - \widehat{L}\widehat{L}')$.

Principal Component Estimation

- The estimated specific variances are given by the diagonal elements of $S - \tilde{L}\tilde{L}'$, so

$$\tilde{\Psi} = \begin{bmatrix} \tilde{\psi}_1 & 0 & \cdots & 0 \\ 0 & \tilde{\psi}_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \tilde{\psi}_p \end{bmatrix}, \quad \tilde{\psi}_i = s_{ii} - \sum_{j=1}^{m} \tilde{\ell}_{ij}^2.$$

- Communalities are estimated as

$$\tilde{h}_i^2 = \tilde{\ell}_{i1}^2 + \tilde{\ell}_{i2}^2 + \cdots + \tilde{\ell}_{im}^2.$$

- If variables are standardized, then we substitute R for S and substitute 1 for each s_{ii}.

Principal Component Estimation

- In many applications of factor analysis, m, the number of factors, is decided prior to the analysis.

- If we do not know m, we can try to determine the 'best' m by looking at the results from fitting the model with different values for m.

- Examine how well the off-diagonal elements of S (or R) are reproduced by the fitted model $\tilde{L}\tilde{L}' + \tilde{\Psi}$ because by definition of $\tilde{\psi}_i$, diagonal elements of S are reproduced exactly but the off-diagonal elements are not. The chosen m is appropriate if the *residual matrix*

$$S - (\tilde{L}\tilde{L}' + \tilde{\Psi})$$

has small off-diagonal elements.

Principal Component Estimation

- Another approach to deciding on m is to examine the contribution of each potential factor to the total variance.

- The contribution of the k-th factor to the sample variance for the i-th trait, s_{ii}, is estimated as $\tilde{\ell}^2_{ik}$.

- The contribution of the k-th factor to the total sample variance $s_{11} + s_{22} + ... + s_{pp}$ is estimated as

$$\tilde{\ell}^2_{1k} + \tilde{\ell}^2_{2k} + ... + \tilde{\ell}^2_{pk} = (\sqrt{\hat{\lambda}_k}\hat{\mathbf{e}}_k)'(\sqrt{\hat{\lambda}_k}\hat{\mathbf{e}}_k) = \hat{\lambda}_k.$$

Principal Component Estimation

- As in the case of PCs, in general

$$\left(\begin{array}{c}\text{Proportion of total sample}\\ \text{variance due to jth factor}\end{array}\right) = \frac{\hat{\lambda}_j}{s_{11} + s_{22} + \cdots + s_{pp}},$$

  or equals $\hat{\lambda}_j/p$ if factors are extracted from R.

- Thus, a 'reasonable' number of factors is indicated by the minimum number of PCs that explain a suitably large proportion of the total variance.
- Also examine a scree plot
- Also consider if a meaningful interpretation can be given to each of the m rotated factors.

Example: Life Expectation

Data are life expectancies in years at four points of the lives of men and women in various countries (Keyfitz and Flieger, 1971)

- m0: life expectation for newborn males
- m25: life expectation for 25 year old men
- m50: life expectation for 50 year old men
- m75: life expectation for 75 year old men
- f0: life expectation for newborn females
- f25: life expectation for 25 year old women
- f50: life expectation for 50 year old women
- f75: life expectation for 75 year old women

R code demonstration

Principal Factor Method

- The principal factor method for estimating factor loadings can be viewed as an iterative modification of the principal components methods that allows for greater focus on explaining correlations among observed traits.

- The principal factor approach begins with a guess about the communalities and then iteratively updates those estimates until some convergence criterion is satisfied. (Try several different sets of starting values.)

- The principal factor method provides factor loadings that more closely reproduce correlations. Principal components are more heavily influenced by accounting for variances and will explain a higher proportion of the total variance.

Principal Factor Method

- The estimated loading matrix should provide a good approximation for all of the correlations and part of the variances as follows

$$LL' \approx R - \Psi \approx \begin{bmatrix} (h_1)^2 & r_{12} & r_{13} & \cdots & r_{1p} \\ r_{21} & (h_2)^2 & r_{23} & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ r_{p1} & r_{p2} & r_{p3} & \cdots & (h_p)^2 \end{bmatrix}$$

  where $(h_i)^2 = 1 - \Psi_{ii}$ is the estimated communality for the i-th factor

- Find a factor loading matrix L so that LL' is a good approximation for $R - \Psi$, rather than trying to make LL' a good approximation for R.

- This is an iterative process

Principal Factor Method

- Note that $R - \Psi^*$ is generally not positive definite and some eigenvalues can be negative.

- The results are sensitive to the choice of the number of factors m

- If m is too large, some communalities can be larger than one, which would imply that variation in factor values accounts for more than 100 percent of the variation in the measured traits. There are options to deal with this:
  - **HEYWOOD:** Set any estimated communality larger than one equal to one and continue iterations with the remaining variables.
  - **ULTRAHEYWOOD:** Continue iterations with all of the variables and hope that iterations eventually bring you back into the allowable parameter space.

Maximum Likelihood Estimation

- To implement the ML method, we need to include some assumptions about the distribution of the p-dimensional vector of observations, X_j, for the j-th subject and the m-dimensional vector of unobservable factor values F_j:

$$\mathbf{X}_j \sim \mathrm{N}_p(\boldsymbol{\mu}, \Sigma), \;\; \mathbf{F}_j \sim \mathrm{N}_m(\mathbf{0}, I_m), \;\; \boldsymbol{\epsilon}_j \sim \mathrm{N}_p(\mathbf{0}, \Psi_{p\times p}),$$

  where $\mathbf{X}_j = L\mathbf{F}_j + \boldsymbol{\epsilon}_j$, $\Sigma = LL' + \Psi$, and $\mathbf{F}_j$ is independent of $\boldsymbol{\epsilon}_j$. Also, Ψ is a diagonal matrix.

- The log-likelihood function for the data is

$$\begin{aligned}\ell(\boldsymbol{\mu}, \Sigma) &:= -\frac{n}{2}\log|2\pi\Sigma| - \frac{1}{2}\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu})'\Sigma^{-1}(\mathbf{x}_i - \boldsymbol{\mu}) \\ &= -\frac{n}{2}\log|2\pi\Sigma| - \frac{n}{2}\mathrm{tr}(\Sigma^{-1}S) - \frac{n}{2}(\bar{\mathbf{x}} - \boldsymbol{\mu})'\Sigma^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}).\end{aligned}$$

Maximum Likelihood Estimation

- Because the L that maximizes the likelihood for this model is not unique, we need another set of restrictions that lead to a unique maximum of the likelihood function:

$$L'\Psi^{-1}L = \Delta,$$

  where Δ is a diagonal matrix.

- Maximizing the likelihood function with respect to $(\boldsymbol{\mu}, L, \Psi)$ is not easy because the likelihood depends in a very non-linear fashion on the parameters.

- Efficient numerical algorithms exist to maximize the likelihood iteratively and obtain ML estimates

$$\widehat{L}_{p\times m} = \{\widehat{\ell}_{ij}\}, \quad \hat{\Psi} = \text{diag}\{\widehat{\psi}_i\}, \quad \hat{\boldsymbol{\mu}} = \bar{\mathbf{x}}.$$

Maximum Likelihood Estimation

- The maximum likelihood estimates (MLEs) of the communalities for the p variables (with m factors) are

$$\widehat{h}_i^2 = \sum_{j=1}^{m} \widehat{\ell}_{ij}^2, \quad i = 1, ..., p,$$

- The proportion of the total variance accounted for by the jth factor is given by

$$\frac{\sum_{i=1}^{p} \widehat{\ell}_{ij}^2}{\text{trace}(S)}$$

if using S, or dividing by $p = \text{trace}(R)$ if using R.

How Many Factors to Choose?

- We wish to test whether the m factor model appropriately describes the covariances (correlations) among the p variables.

- We test the null hypothesis that m factors are sufficient, i.e,

$$H_0 : \Sigma_{p\times p} = L_{p\times m}L'_{m\times p} + \Psi_{p\times p}$$

versus the alternative that the covariance (correlation) matrix can be any positive definite matrix

$$H_a : \Sigma_{p\times p} \text{ is a positive definite matrix}$$

Likelihood Ratio Test for Number of Factors

- The test statistic ,

$$-2\ln\Lambda = (n-1-(2p+4m+5)/6)\log\frac{|\hat{L}\hat{L}'+\hat{\Psi}|}{|\hat{\Sigma}|},$$

  has an approximate chi-square distribution when the null hypothesis is true (when m factors are sufficient to describe the correlations).

- The degrees of freedom for the large sample chi-square approximation to this test statistic are

$$(1/2)[(p-m)^2-p-m]$$

- This test is known as a *likelihood ratio test*.

Likelihood Ratio test for Number of Factors

- We reject H_0 at level α if

$$[n-1-(2p+4m+5)/6]\log\left(\frac{|\widehat{L}\widehat{L}'+\widehat{\Psi}|}{|\widehat{\Sigma}|}\right) > \chi^2_{(df),1-\alpha}$$

  with $df = \frac{1}{2}[(p-m)^2-p-m]$ for large n and large $n-p$.

- To have $df > 0$, we must have $m < \frac{1}{2}(2p+1-\sqrt{8p+1})$.
- If the data are not a random sample from a multivariate normal distribution, this test tends to indicate the need for too many factors.
- Use this test as a guideline for selecting the number of factors, but also consider other things such as the interpretation of the factors, proportion of variance explained, and the desire for simplicity (using a few factors).
- R code demonstration

Measures of Sampling Adequacy

- In constructing a survey or other type of instrument to examine ”factors” like political attitudes, social constructs, mental abilities, consumer confidence, that cannot be measured directly, it is common to include a set of questions or items that should vary together as the values of the factor vary across subjects.

- In a job attitude assessment, for example, you might include questions of the following type in the survey:
  1. How well do you like your job?
  2. How eager are you to go to work in the morning?
  3. How professionally satisfied are you?
  4. What is your level of frustration with assignment to meaningless tasks?

Measure of Sampling Adequacy

- If the responses to these and other questions had strong positive or negative correlations, you may be able to identify a ”job satisfaction” factor.

- In marketing, education, and behavioral research, the existence of highly correlated responses to a battery of questions is used to defend the notion that an important factor exists and to justify the use of the survey form or measurement instrument.

- Measures of sampling adequacy are based on correlations and partial correlations among variables (e.g. responses to different questions on a survey)

Measure of Sampling Adequacy

- Kaiser (1970, *Psychometrica*) proposed a statistic called the Measure of Sampling Adequacy (MSA).

- MSA measures the relative sizes of the pairwise correlations to the partial correlations between all pairs of variables as follows:

  $$MSA = 1 - \frac{\sum_j \sum_{k \neq j} q_{jk}^2}{\sum_j \sum_{k \neq j} r_{jk}^2}$$

  where r_{jk} is the marginal sample correlation between variables j and k and q_{jk} is the *partial correlation* between the two variables after accounting for all other variables in the data set.

- If the r's are relatively large and the q's are relatively small, MSA tends to 1. This indicates that more than two variables are changing together as the values of the "factor" vary across subjects.

Measure of Sampling Adequacy

- The MSA can take on values between 0 and 1 and Kaiser proposed the following guidelines:

MSA range	Interpretation	MSA range	Interpretation
0.9 to 1	Marvelous data	0.6 to 0.7	Mediocre data
0.8 to 0.9	Meritorious data	0.5 to 0.6	Miserable
0.7 to 0.8	Middling	0.0 to 0.5	Unacceptable

- The KMO function in the ”psych” package in R computes an overall MSA and an individual MSA for each variable in the data set

Cronbach's Alpha

- A set of observed variables $\mathbf{X} = (X_1, X_2, \ldots, X_p)'$ that all "measure" the same latent trait should all have high positive pairwise correlations

- Define an "average" correlation

$$\bar{r} = \frac{\frac{1}{\frac{p(p-1)}{2}} \sum \sum_{i<j} \hat{Cov}(X_i, X_j)}{\frac{1}{p} \sum_{i=1}^{p} \hat{Var}(X_i)}$$

- Cronbach's Alpha is

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

Cronbach's Alpha

- In the extreme case where all pairwise correlations are 1, we have

$$\alpha = 1$$

- When $\bar{r} = 0$, then $\alpha = 0$

- Be sure that the scores for all items are orientated in the same direction so all correlations are positive

- R code demonstration

Factor Rotation

- As mentioned earlier, multiplying a matrix of factor loadings by any orthogonal matrix leads to the same approximation to the covariance (or correlation) matrix.

- This means that mathematically, it does not matter whether we estimate the loadings as $\hat{L}$ or as $\hat{L}^* = \hat{L}T$ where $T'T = TT' = I$.

- The estimated residual matrix matrix also remains unchanged:

$$S - \hat{L}\hat{L}' - \hat{\Psi} = S - \hat{L}TT'\hat{L}' - \hat{\Psi} = S - \hat{L}^*\hat{L}^{*'} - \hat{\Psi}$$

- The specific variances $\hat{\psi}_i$ and therefore the communalities also remain unchanged.

Factor Rotation

- We rotate factors in order to better interpret results.
- There are many choices for a rotation matrix T, and to choose, we first establish a mathematical criterion and then see which T can best satisfy the criterion.
- One possible objective is to have each one of the p variables load highly on only one factor and have moderate to negligible loads on all other factors.
- *Varimax* rotation tries to acheive this goal
- It is not always possible to achieve this type of result.

Varimax Rotation

- Define $\tilde{\ell}^*_{ij} = \hat{\ell}^*_{ij}/\hat{h}_i$ as the "scaled" loading of the i-th variable on the j-th rotated factor.

- Compute the variance of the squares of the scaled loadings for the j-th rotated factor

$$\frac{1}{p}\sum_{i=1}^{p}\left((\tilde{\ell}^*_{ij})^2 - \frac{1}{p}\sum_{k=1}^{p}(\tilde{\ell}^*_{kj})^2\right)^2 = \frac{1}{p}\left(\sum_{i=1}^{p}(\tilde{\ell}^*_{ij})^4 - \frac{1}{p}\left[\sum_{k=1}^{p}(\tilde{\ell}^*_{kj})^2\right]^2\right)$$

- The varimax procedure finds the orthogonal transformation of the loading matrix that maximizes the **sum** of those variances, summing across all m rotated factors.

- After rotation each of the p variables should load highly on at most one of the rotated factors

Maximum Likelihood Estimation: Life Expectancy Data

- Varimax rotation of $m = 2$ factors

Variable	MLE factor 1	MLE factor 2	Rotated factor 1	Rotated factor 2
m0	0.917	-0.367	0.972	0.179
m25	0.743	-0.079	0.670	0.329
m50	0.753	0.294	0.480	0.651
m75	0.509	0.578	0.122	0.760
w0	0.927	-0.355	0.973	0.194
w25	0.990	0.089	0.790	0.603
w50	0.914	0.386	0.567	0.815
w75	0.630	0.652	0.185	0.888
prop of var	0.661	0.159	0.446	0.374

Quartimax Rotation

- The varimax rotation will destroy an "overall" factor

- The quartimax rotation tries to

  1. Preserve an overall factor such that each of the p variables has a high loading on that factor

  2. Create other factors such that each of the p variables has a high loading on at most one factor

- There are many other rotations that have been proposed. In R many rotations can be performed by various functions in the "GPArotation" package.

Quartimax Rotation

- Define $\tilde{\ell}^*_{ij} = \hat{\ell}^*_{ij}/\hat{h}_i$ as the "scaled" loading of the i-th variable on the j-th rotated factor.

- Compute the variance of the squares of the scaled loadings for the i-th variable

$$\frac{1}{m}\sum_{j=1}^{m}\left((\tilde{\ell}^*_{ij})^2 - \frac{1}{m}\left[\sum_{k=1}^{m}(\tilde{\ell}^*_{ik})\right]^2\right)^2 = \frac{1}{m}\left(\sum_{j=1}^{m}(\tilde{\ell}^*_{ij})^4 - \frac{1}{m}\left[\sum_{k=1}^{m}(\tilde{\ell}^*_{ik})^2\right]^2\right)$$

- The quartimax procedure finds the orthogonal transformation of the loading matrix that maximizes the **sum** of those variances, summing across all p variables.

PROMAX Transformation

- The varimax and quartimax rotations produce uncorrelated factors

- PROMAX is a non-orthogonal (oblique) transformation that

  1. is not a rotation

  2. can produce correlated factors

  3. tries to force each of the p variables to load highly on only one of the factors.

PROMAX Transformation

1. First perform a varimax rotation to obtain loadings L^*
2. Construct another $p \times m$ matrix Q such that

$$\begin{aligned} q_{ij} &= |\ell^*_{ij}|^{k-1}\ell^*_{ij} && \text{for } \ell^*_{ij} \neq 0 \\ &= 0 && \text{for } \ell^*_{ij} = 0 \end{aligned}$$

   where $k > 1$ is selected by trial and error, usually $k < 4$.

3. Find a matrix U such that each column of L^*U is close to the corresponding column of Q. Choose the j-th column of U to minimize

$$(\mathbf{q}_j - L^*\mathbf{u}_j)'(\mathbf{q}_j - L^*\mathbf{u}_j)$$

   This yields

$$U = [(L^*)'(L^*)]^{-1}(L^*)'Q$$

PROMAX Transformation

1. Rescale U so that the transformed factors have unit variance. Compute $D^2 = diag[(U'U)^{-1}]$ and $M = UD$.

2. The PROMAX factors are obtained from

$$L^*\mathbf{F} = L^*MM^{-1}\mathbf{F} = L^P\mathbf{F}^P$$

- The PROMAX transformation yields factors with loadings

$$L^P = L^*M$$

- The correlation matrix for the transformed factors is $(M'M)^{-1}$

- R code demonstration

Factor Analysis of Test Scores

- The sample correlation between $p = 6$ test scores collected from $n = 220$ students is given below:

$$R = \begin{bmatrix} 1.0 & 0.439 & 0.410 & 0.288 & 0.329 & 0.248 \\ & 1.0 & 0.351 & 0.354 & 0.320 & 0.329 \\ & & 1.0 & 0.164 & 0.190 & 0.181 \\ & & & 1.0 & 0.595 & 0.470 \\ & & & & 1.0 & 0.464 \\ & & & & & 1.0 \end{bmatrix}.$$

- An $m = 2$ factor model was fitted to this correlation matrix using ML.

Example: Test Scores

- Estimated factor loadings and communalities are

Variable	Loadings on factor 1	Loadings on factor 2	Communalities $\widehat{h}_i^2$
Gaelic	0.553	0.429	0.490
English	0.568	0.288	0.406
History	0.392	0.450	0.356
Arithmetic	0.740	-0.273	0.623
Algebra	0.724	-0.211	0.569
Geometry	0.595	-0.132	0.372

Example: Test Scores

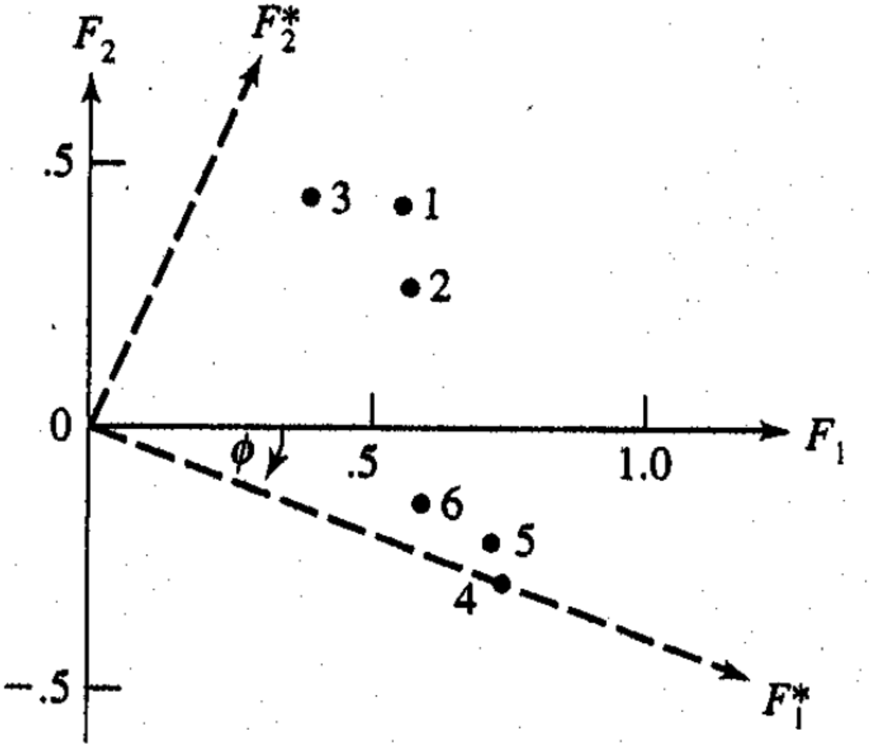
- All variables load highly on the first factor. We call that a 'general intelligence' factor.

- Half of the loadings are positive and half are negative on the second factor. The positive loadings correspond to the 'verbal' scores and the negative correspond to the 'math' scores.

- Correlations between scores on verbal tests and math tests tend to be lower than correlations among scores on math tests or correlations among scores on verbal tests, so this is a 'math versus verbal' factor.

- We plot the six loadings for each factor $(\widehat{\ell}_{i1}, \widehat{\ell}_{i2})$ on the original coordinate system and also on a rotated set of coordinates chosen so that one axis goes through the loadings $(\widehat{\ell}_{41}, \widehat{\ell}_{42})$ of the fourth variable on the two factors.

Factor Rotation for Test Scores



Varimax Rotation for Test Scores

- Loadings for rotated factors using the varimax criterion are as follows:

	Loadings on	Loadings on	Communalities
Variable	F_1^*	F_2^*	$\widehat{h}_i^2$
Gaelic	0.232	0.660	0.490
English	0.321	0.551	0.406
History	0.085	0.591	0.356
Arithmetic	0.770	0.173	0.623
Algebra	0.723	0.215	0.569
Geometry	0.572	0.213	0.372

- F_1^* is primarily a 'mathematics ability factor' and F_2^* is a 'language/history factor'.

PROMAX Rotation for Test Scores

- F_1^p is more clearly a 'mathematics ability factor' and F_2^p is more clearly a 'language/history factor'.

Variable	Loadings on F_1^p	Loadings on F_2^p	Communalities $\widehat{h}_i^2$
Gaelic	0.059	0.668	0.490
English	0.191	0.519	0.406
History	-0.084	0.635	0.356
Arithmetic	0.809	0.041	0.623
Algebra	0.743	0.021	0.569
Geometry	0.575	0.064	0.372

- These factors have correlation 0.505

Estimation of Factor Scores

- Sometimes, we require estimates of the factor scores for each respondent in the sample.

- Let $\widehat{\mathbf{f}}_j$ denote the vector of m factor scores for the j-th respondent.

- A regression estimate is obtained from the conditional mean of $\mathbf{f}_j$ given the values of the explanatory variables

$$\widehat{\mathbf{f}}_j = \widehat{E}(\mathbf{F}_j|\mathbf{x}_j) = \hat{L}'(\hat{L}\hat{L}' + \widehat{\Psi})^{-1}(\mathbf{x}_j - \bar{\mathbf{x}}).$$

- Sometimes, to reduce the errors that may be introduced if the number of factors m is not quite appropriate, S is used in place of $(\hat{L}\hat{L}' + \widehat{\Psi})$ in the expression for $\widehat{\mathbf{f}}_j$.

Heywood Cases and Other Potential Problems

- When loadings are estimated iteratively (as in the Principal Factor or ML methods) some estimates of communalities may be larger than one.
  - If any communality exceeds 1.0, then more than 100% of the variation in that variable is explained by the variation in the factor scores, which is impossible.
  - The corresponding specific variance is negative.
- The validity of the factor analysis is called into question.

Heywood Cases and Other Potential Problems

- When can these things happen?

  1. When we start with bad prior communality estimates for iterating.

  2. When there are too many common factors.

  3. When we do not have enough data to obtain stable estimates.

  4. When the common factor model is just not appropriate for the data.

- These problems can occur even in seemingly ’good’ datasets and create serious problems when we try to interpret results.

Heywood Cases and Other Potential Problems

- When a final communality estimate equals 1, we refer to this as a Heywood case.

- When a final communality estimate exceeds 1, we have an Ultra Heywood case.

- There is some concern about the validity of results in Heywood cases (and ultra Heywood cases).

- Large sample results, such as the large sample chi-square approximation to likelihood ratio test for the number of factors, will not be reliable.

Heywood Cases and Other Potential Problems

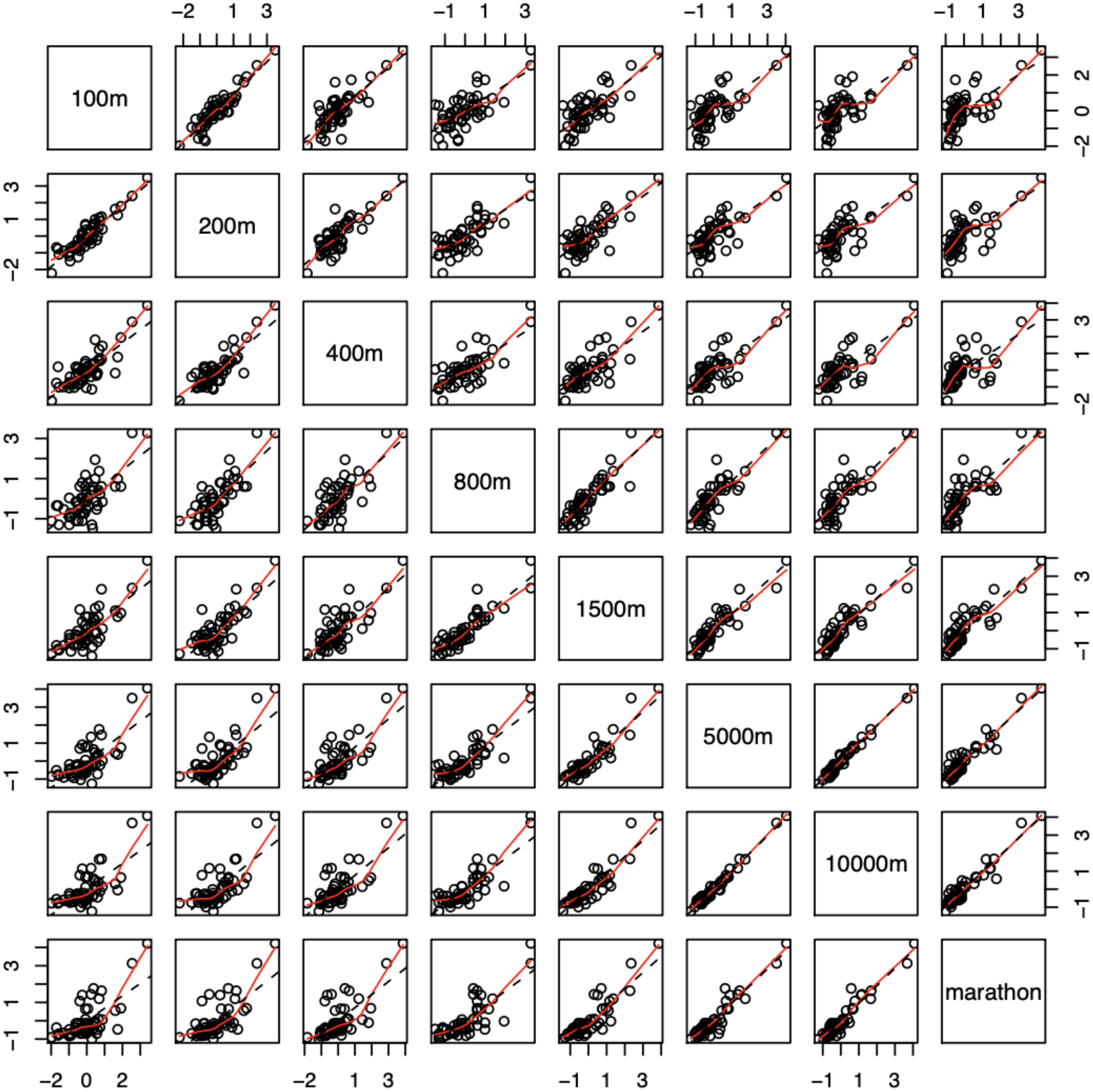
- When iterative algorithms fail to converge, specification of a Heywood option or an ultra Heywood option is allowed by some software (not factanal in R)
  - The Heywood option fixes to 1 any communality that goes above 1 and continues iterating on the rest of the variables.
  - The ultra Heywood option continues iterating on all variables, regardless of the size of the communalities, hoping that the solution eventually comes back inside the parameter space (this rarely happens).
- When factor loadings are estimated using the principal components method, none of these problems arise (assuming that R is of full column rank).

Example: track records for men and women

- Use maximum likelihood estimation to fit factor models to **standardized** observations on national track records for men and women in $n = 54$ countries
- The dataset for women includes seven variables $x_1, ..., x_7$ that correspond to national records on 100m, 200m, 400m, 800m, 1500m, 3000m and marathon races.
- The dataset for men includes eight variables $x_1, ..., x_8$ that correspond to national records on 100m, 200m, 400m, 800m, 1500m, 5000m, 10000m and marathon races.
- The first three times are in seconds and the remaining times are in minutes.

Example: track records for men and women

- First examine scatterplots and correlation matrices.
- Test for the appropriate number of factors using likelihood ratio tests.
- The factanal function in R does not seem to let specific variances go below 0.005. There is no warning of a Heywood condition.


100m
200m
400m
800m
1500m
5000m
10000m
marathon
−2
1
3
−1
1
3
−1
1
3
−1
1
3
−2
0
2
−2
1
3
−1
1
3
−1
1
3
−2
0
2
3
1
−2
3
1
−2
3
1
−1
3
1
−1
3
1
−1
3
1
−1
3
1
−1

Example: track records for men

- The scatter plot matrix reveals a couple of extreme countries with slow times but they conform to the correlation pattern. All relationships are approximately straight line relationships.

- Results indicate that the two factor model may not be sufficient (p-value=.0173).

- Maximum likelihood estimation for the three factor solution has communalities less than one.

- Likelihood ratio test for 3 factors has p-value=0.223

Track Records for Men: 2 factors

Race	Varimax Factor 1	Varimax Factor 2	Promax Factor 1	Promax Factor 2	$\widehat{h}_i^2$
100m	0.402	0.839	0.097	0.868	0.865
200m	0.409	0.892	0.068	0.933	0.963
400m	0.515	0.711	0.300	0.643	0.772
800m	0.671	0.581	0.568	0.393	0.788
1500m	0.748	0.554	0.682	0.316	0.866
5000m	0.886	0.450	0.915	0.109	0.988
10000m	0.900	0.424	0.946	0.069	0.989
Marathon	0.865	0.402	0.910	0.063	0.912

- The Promax factors have correlation 0.697

- There is a distance race factor and a sprint factor

Track Records for Men: 3 factors

Race	Promax Factor 1	Promax Factor 2	Promax Factor 3	$\hat{h}_i^2$
100m	0.024	1.038	-0.131	0.918
200m	-0.047	0.906	0.119	0.931
400m	0.202	0.633	0.096	0.771
800m	0.112	0.023	0.887	0.995
1500m	0.505	0.179	0.327	0.890
5000m	0.918	0.061	0.033	0.985
10000m	0.999	0.021	-0.023	0.994
Marathon	0.965	0.010	-0.021	0.914

- There is a distance race factor, a sprint factor, and a middle distance factor

Example: track records for women

- The two factor model does not appear to fit for women (p-value$<$.0001).

- Using SAS to fit the three-factor model needed the Heywood option to converge to a boundary solution. (The reported p-value of 0.0295 for the likelihood ratio test is unreliable.)

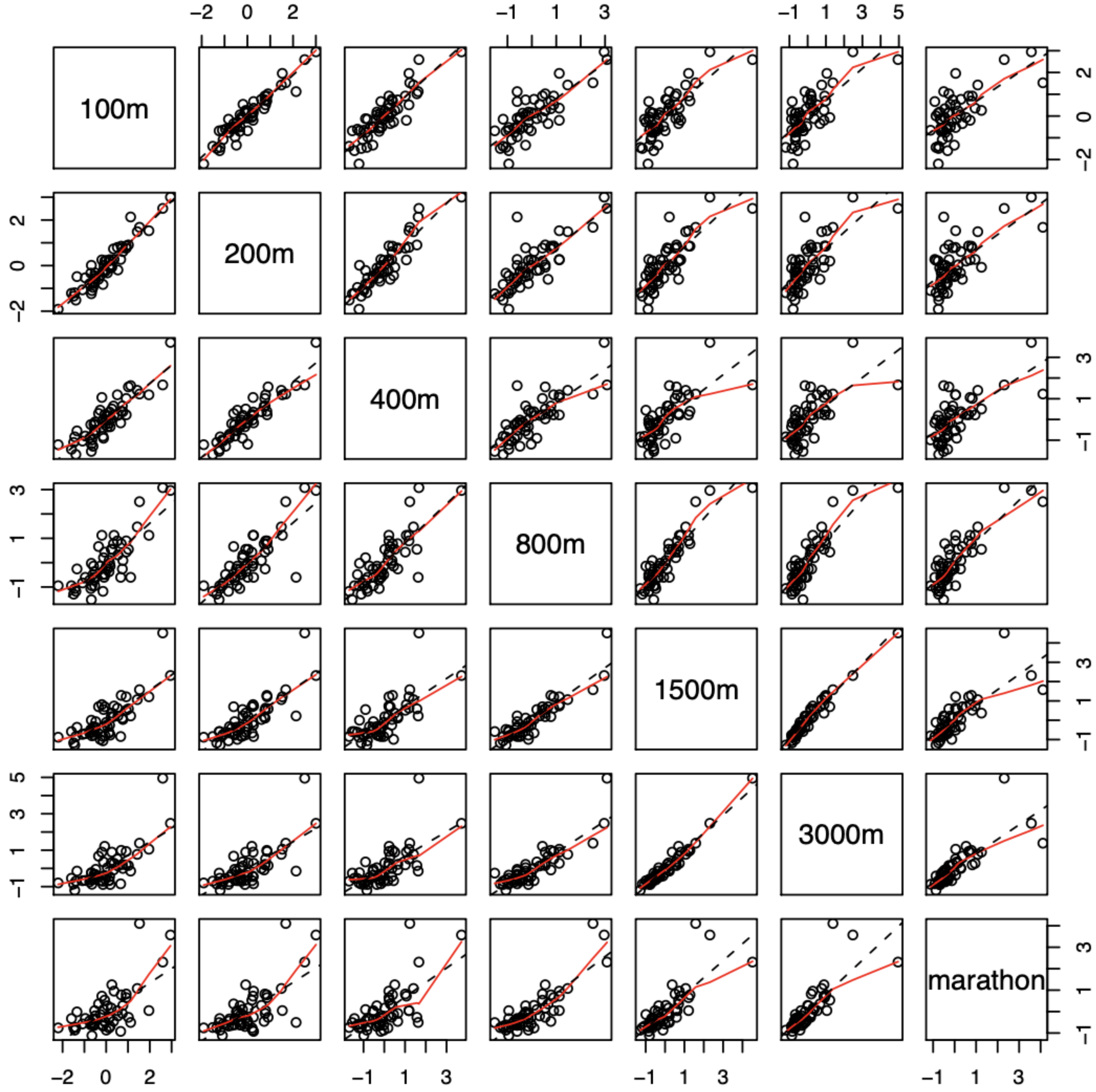
- Two varimax rotated factors have a similar interpretation to the rotated factors for men, but the marathon does not load so highly on the distance factor

Example: track records for women

- Three varimax rotated factors (using a Heywood option in SAS) allow for a sprint factor, a 1500m and 3000m factor, and a third factor for which the 800m and the marathon have relatively high loadings.

- A four factor solution (using a Heywood option) does not split up the loadings on the third factor for the 800m and marathon races (factanal in R will not fit 4 factors).

- The test for the number of factors could have been adversely affected by some extreme cases (outliers)


100m
200m
400m
800m
1500m
3000m
marathon
−2
0
2
−1
1
3
5

Example: track records for women

Race	Varimax Factor 1	Varimax Factor 2	Promax Factor 1	Promax Factor 2	$\widehat{h}_i^2$
100m	0.455	0.836	0.151	0.841	0.906
200m	0.449	0.880	0.120	0.901	0.976
400m	0.395	0.832	0.075	0.867	0.848
800m	0.728	0.571	0.639	0.359	0.856
1500m	0.879	0.460	0.891	0.139	0.984
3000m	0.915	0.367	0.985	0.000	0.972
Marathon	0.670	0.432	0.662	0.200	0.662

- The Promax factors have correlation 0.693
- There is a distance race factor and a sprint factor
- When a third factor is included it reflects the three countries with very slow 800m and marathon times

Example: track records for women

Race	Varimax Factor 1	Varimax Factor 2	Promax Factor 1	Promax Factor 2	$\widehat{h}_i^2$
100m	0.455	0.836	0.151	0.841	0.906
200m	0.449	0.880	0.120	0.901	0.976
400m	0.395	0.832	0.075	0.867	0.848
800m	0.728	0.571	0.639	0.359	0.856
1500m	0.879	0.460	0.891	0.139	0.984
3000m	0.915	0.367	0.985	0.000	0.972
Marathon	0.670	0.432	0.662	0.200	0.662

- The Promax factors have correlation 0.693
- There is a distance race factor and a sprint factor
- When a third factor is included it reflects the three countries with very slow 800m and marathon times