- factor rotation
  - Recall
    - a solution of factor loading is not unique
    - all factor loadings obtained from the initial loadings by an orthogonal transformation have the same ability to reproduce the covariance/correlation matrix
  - If $\hat{\mathbf{L}}$ is the $p \times m$ matrix of estimated factor loadings obtained by any method (principal component, maximum likelihood, and so forth) then

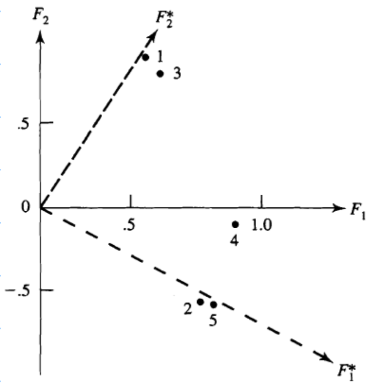
$$\hat{\mathbf{L}}^* = \hat{\mathbf{L}}\mathbf{T}, \quad \text{where } \mathbf{T}\mathbf{T}' = \mathbf{T}'\mathbf{T} = \mathbf{I}$$

is a $p \times m$ matrix of "rotated" loadings.

- **Q**: why is it called "rotation"?
- **Q**: dimension of *all* rotated factor loadings = ?
- Note: estimated covariance/correlation matrix, residual matrix, estimated specific variances and communalities remains unchanged after rotation

- **Q**: Why need rotation?
  - since the original loading may not be readily interpretable (e.g., some factors may have several large loadings and some loadings may be positive while other may be negative) it's usual practice to rotate them until a "simple structure " is achieved

- **Q**: What is a "simple structure"?
  - Ideally, we should like to see a pattern of loadings such that each variable loads highly on a *single* factor and has small to moderate loading on the remaining factors
  - When there are two factors, we can plot the loadings and visually rotation the data to find an interpretable solution



  - For more than two factors, some automation is necessary, which should make:
    - each factor have a few high positive loadings and the remainder be small
    - pairs of factors have few large loading in common. A partition into mutually exclusive groups would be desirable

Variable	Estimated factor loadings F_1	F_2	Rotated estimated factor loadings F_1^*	F_2^*	Communalities $\tilde{h}_i^2$
1. Taste	.56	.82	.02	.99	.98
2. Good buy for money	.78	−.52	.94	−.01	.88
3. Flavor	.65	.75	.13	.98	.98
4. Suitable for snack	.94	−.10	.84	.43	.89
5. Provides lots of energy	.80	−.54	.97	−.02	.93
Cumulative proportion of total (standardized) sample variance explained	.571	.932	.507	.932	

➢ varimax criterion

- Define $\tilde{\ell}^*_{ij} = \hat{\ell}^*_{ij}/\hat{h}_i$ (the scaling has the effect of giving variables with small communalities relatively more weight in rotation)
- varimax procedure selects the orthogonal transformation $\mathbf{T}$ that makes

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

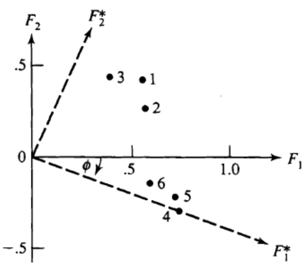
as large as possible.

- interpretation of the varimax criterion:

$$V \propto \sum_{j=1}^{m}\begin{pmatrix}\text{variance of squares of (scaled) loadings for} \\ j\text{th factor}\end{pmatrix}$$

  - maximizing V corresponds to "spreading out" the squares of the loadings on each factor as much as possible

➢ <u>Remark</u>: In some cases, even orthogonal rotations do not provide an easy interpretation of the solution. It is possible to allow non-orthogonal rotations, called *oblique* rotation. This allows for possible simplicity at the expense of losing the orthogonality of the factors



• factor scores (predict values of the unobserved random factors $F_1, \ldots, F_m$)

➢ (c.f.) the scores in principal component analysis

➢ weighted least squares method

- Suppose first that the mean vector $\boldsymbol{\mu}$, the factor loadings $\mathbf{L}$, and the specific variance $\boldsymbol{\Psi}$ are known for the factor model

$$\underset{(p\times 1)}{\mathbf{X}} - \underset{(p\times 1)}{\boldsymbol{\mu}} = \underset{(p\times m)}{\mathbf{L}}\underset{(m\times 1)}{\mathbf{F}} + \underset{(p\times 1)}{\boldsymbol{\varepsilon}}$$

- The sum of the squares of the errors, weighted by the reciprocal of their variances, is

$$\sum_{i=1}^{p}\frac{\varepsilon_i^2}{\psi_i} = \boldsymbol{\varepsilon}'\boldsymbol{\Psi}^{-1}\boldsymbol{\varepsilon} = (\mathbf{x} - \boldsymbol{\mu} - \mathbf{L}\mathbf{f})'\boldsymbol{\Psi}^{-1}(\mathbf{x} - \boldsymbol{\mu} - \mathbf{L}\mathbf{f}) \qquad (*)$$

- choosing the estimates $\hat{\mathbf{f}}$ of $\mathbf{f}$ to minimize $(*)$. The solution is

$$\hat{\mathbf{f}} = (\mathbf{L}'\boldsymbol{\Psi}^{-1}\mathbf{L})^{-1}\mathbf{L}'\boldsymbol{\Psi}^{-1}(\mathbf{x} - \boldsymbol{\mu})$$

- take the estimates $\hat{\mathbf{L}}$, $\hat{\boldsymbol{\Psi}}$, and $\hat{\boldsymbol{\mu}} = \bar{\mathbf{x}}$ as the true values and obtain the factor scores for the jth case as

$$\hat{\mathbf{f}}_j = (\hat{\mathbf{L}}'\hat{\boldsymbol{\Psi}}^{-1}\hat{\mathbf{L}})^{-1}\hat{\mathbf{L}}'\hat{\boldsymbol{\Psi}}^{-1}(\mathbf{x}_j - \bar{\mathbf{x}})$$

  - When MLE method is used, $\hat{\mathbf{L}}'\hat{\boldsymbol{\Psi}}^{-1}\hat{\mathbf{L}} = \hat{\boldsymbol{\Delta}}$ is a diagonal matrix

$$\hat{\mathbf{f}}_j = \hat{\boldsymbol{\Delta}}^{-1}\hat{\mathbf{L}}'\hat{\boldsymbol{\Psi}}^{-1}(\mathbf{x}_j - \bar{\mathbf{x}})$$

  - if the correlation matrix is factored

$$\hat{\mathbf{f}}_j = (\hat{\mathbf{L}}_{\mathbf{z}}'\hat{\boldsymbol{\Psi}}_{\mathbf{z}}^{-1}\hat{\mathbf{L}}_{\mathbf{z}})^{-1}\hat{\mathbf{L}}_{\mathbf{z}}'\hat{\boldsymbol{\Psi}}_{\mathbf{z}}^{-1}\mathbf{z}_j = \hat{\boldsymbol{\Delta}}_{\mathbf{z}}^{-1}\hat{\mathbf{L}}_{\mathbf{z}}'\hat{\boldsymbol{\Psi}}_{\mathbf{z}}^{-1}\mathbf{z}_j$$

where $\mathbf{z}_j = \mathbf{D}^{-1/2}(\mathbf{x}_j - \bar{\mathbf{x}})$

➢ regression method (conditional normal distribution)

- Suppose that the common factors and the specific factors are jointly normally distributed
- $\mathbf{X} - \boldsymbol{\mu} = \mathbf{LF} + \boldsymbol{\varepsilon}$ has an $N_p(\mathbf{0}, \mathbf{LL}' + \boldsymbol{\Psi})$ distribution
- Moreover, the joint distribution of $(\mathbf{X} - \boldsymbol{\mu})$ and $\mathbf{F}$ is $N_{m+p}(\mathbf{0}, \boldsymbol{\Sigma}^*)$, where

$$\underset{(m+p)\times(m+p)}{\boldsymbol{\Sigma}^*} = \left[\begin{array}{c|c} \underset{(p\times p)}{\boldsymbol{\Sigma} = \mathbf{LL}' + \boldsymbol{\Psi}} & \underset{(p\times m)}{\mathbf{L}} \\ \hline \underset{(m\times p)}{\mathbf{L}'} & \underset{(m\times m)}{\mathbf{I}} \end{array}\right]$$

- the conditional distribution of $\mathbf{F}|\mathbf{x}$ is multivariate normal with

$$\text{mean} = E(\mathbf{F}|\mathbf{x}) = \mathbf{L}'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) = \mathbf{L}'(\mathbf{LL}' + \boldsymbol{\Psi})^{-1}(\mathbf{x} - \boldsymbol{\mu})$$

$$\text{covariance} = \text{Cov}(\mathbf{F}|\mathbf{x}) = \mathbf{I} - \mathbf{L}'\boldsymbol{\Sigma}^{-1}\mathbf{L} = \mathbf{I} - \mathbf{L}'(\mathbf{LL}' + \boldsymbol{\Psi})^{-1}\mathbf{L}$$

- the jth factor score vector is given by

$$\hat{\mathbf{f}}_j = \hat{\mathbf{L}}'\hat{\boldsymbol{\Sigma}}^{-1}(\mathbf{x}_j - \bar{\mathbf{x}}) = \hat{\mathbf{L}}'(\hat{\mathbf{L}}\hat{\mathbf{L}}' + \hat{\boldsymbol{\Psi}})^{-1}(\mathbf{x}_j - \bar{\mathbf{x}})$$

◆ Denote the scores generated by the weighted least squares by $\hat{\mathbf{f}}_j^{LS}$ and those by the regression method by $\hat{\mathbf{f}}_j^{R}$. Because

$$\underset{(m\times p)}{\hat{\mathbf{L}}'}(\hat{\mathbf{L}}\hat{\mathbf{L}}' \underset{(p\times p)}{+} \hat{\boldsymbol{\Psi}})^{-1} = (\mathbf{I} + \underset{(m\times m)}{\hat{\mathbf{L}}'\hat{\boldsymbol{\Psi}}^{-1}\hat{\mathbf{L}}})^{-1} \underset{(m\times p)}{\hat{\mathbf{L}}'} \underset{(p\times p)}{\hat{\boldsymbol{\Psi}}^{-1}} \quad \text{(exercise 9.6, textbook)}$$

$$\hat{\mathbf{f}}_j^{LS} = (\hat{\mathbf{L}}'\hat{\boldsymbol{\Psi}}^{-1}\hat{\mathbf{L}})^{-1}(\mathbf{I} + \hat{\mathbf{L}}'\hat{\boldsymbol{\Psi}}^{-1}\hat{\mathbf{L}})\hat{\mathbf{f}}_j^{R} = (\mathbf{I} + (\hat{\mathbf{L}}'\hat{\boldsymbol{\Psi}}^{-1}\hat{\mathbf{L}})^{-1})\hat{\mathbf{f}}_j^{R}$$

For maximum likelihood estimates $(\hat{\mathbf{L}}'\hat{\boldsymbol{\Psi}}^{-1}\hat{\mathbf{L}})^{-1} = \hat{\boldsymbol{\Delta}}^{-1}$ and if the elements of this diagonal matrix are close to zero, the regression and generalized least squares methods will give nearly the same factor scores.

◆ In an attempt to reduce the effect of a (possible) incorrect determination of the number of factors, $\mathbf{S}$ is often used for $\hat{\boldsymbol{\Sigma}}$, rather than $\hat{\mathbf{L}}\hat{\mathbf{L}}' + \hat{\boldsymbol{\Psi}}$, i.e,

$$\hat{\mathbf{f}}_j = \hat{\mathbf{L}}'\mathbf{S}^{-1}(\mathbf{x}_j - \bar{\mathbf{x}})$$

or, if a correlation matrix is factored,

$$\hat{\mathbf{f}}_j = \hat{\mathbf{L}}_{\mathbf{z}}'\mathbf{R}^{-1}\mathbf{z}_j$$

where $\mathbf{z}_j = \mathbf{D}^{-1/2}(\mathbf{x}_j - \bar{\mathbf{x}})$

➢ If rotated loadings $\hat{\mathbf{L}}^* = \hat{\mathbf{L}}\mathbf{T}$ are used in place of the original loadings the subsequent factor scores, $\hat{\mathbf{f}}_j^*$, are related to $\hat{\mathbf{f}}_j$ by $\hat{\mathbf{f}}_j^* = \mathbf{T}'\hat{\mathbf{f}}_j$

- A strategy for factor analysis
  1. perform a principal component factor analysis
  2. perform a maximum likelihood factor analysis
  3. compare the solutions obtained from the 2 factor analyses
  4. repeat the 1st 3 steps for other number of common factors m
  5. for large data sets, split them in half and perform a FA on each part

❖ **Reading**: Textbook, 9.4, 9.5, 9.6