- Then, the distribution of

$$\underbrace{\mathbf{X}}_{\sim N_p(\boldsymbol{\mu},\, \boldsymbol{\Sigma}=\mathbf{LL}'+\boldsymbol{\Psi})} = \boldsymbol{\mu} + \mathbf{L}\underbrace{\mathbf{F}}_{\sim N_m(\mathbf{0},\, \mathbf{I})} + \underbrace{\boldsymbol{\varepsilon}}_{\sim N_p(\mathbf{0},\, \boldsymbol{\Psi})}$$

is multivariate normal and its likelihood is

$$L(\boldsymbol{\mu},\boldsymbol{\Sigma}) = (2\pi)^{-\frac{np}{2}}|\boldsymbol{\Sigma}|^{-\frac{n}{2}}e^{-\left(\frac{1}{2}\right)\mathrm{tr}\left[\boldsymbol{\Sigma}^{-1}\left(\sum_{j=1}^{n}(\mathbf{x}_j-\bar{\mathbf{x}})(\mathbf{x}_j-\bar{\mathbf{x}})'+n(\bar{\mathbf{x}}-\boldsymbol{\mu})(\bar{\mathbf{x}}-\boldsymbol{\mu})'\right)\right]}$$

$$= (2\pi)^{-\frac{(n-1)p}{2}}|\boldsymbol{\Sigma}|^{-\frac{(n-1)}{2}}e^{-\left(\frac{1}{2}\right)\mathrm{tr}\left[\boldsymbol{\Sigma}^{-1}\left(\sum_{j=1}^{n}(\mathbf{x}_j-\bar{\mathbf{x}})(\mathbf{x}_j-\bar{\mathbf{x}})'\right)\right]}$$

$$\times (2\pi)^{-\frac{p}{2}}|\boldsymbol{\Sigma}|^{-\frac{1}{2}}e^{-\left(\frac{n}{2}\right)(\bar{\mathbf{x}}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}}-\boldsymbol{\mu})}$$

Cf. the Σ in principal component approach.

where $\boldsymbol{\Sigma} = \mathbf{LL}' + \boldsymbol{\Psi}$ ⇒ parameters, m, L, Ψ, μ.

- To make $\mathbf{L}$ well defined, we can impose the computationally convenient uniqueness condition

how many constraints? # of 0's $= \frac{m^2-m}{2}$

$$(\Psi^{-1/2}L)'(\Psi^{-1/2}L) = \mathbf{L}'\underbrace{\boldsymbol{\Psi}^{-1}}_{\Psi^{-1/2}\Psi^{-1/2}}\mathbf{L} = \boldsymbol{\Delta} \qquad \text{a diagonal matrix} \Rightarrow \begin{bmatrix} \ddots & 0 \\ 0 & \ddots \end{bmatrix}$$

(Q: dimension reduction of the parameter space caused by the condition $= \frac{m(m-1)}{2}$)

- Maximizing the likelihood is equivalent to minimizing

$$\ln|\boldsymbol{\Sigma}| - \ln|\underbrace{\mathbf{S}_n}_{\text{MLE of }\Sigma}| + \mathrm{tr}(\boldsymbol{\Sigma}^{-1}\mathbf{S}_n) - p \qquad (❖)$$

(for details, see Supplement 9A).

subject to $\mathbf{L}'\boldsymbol{\Psi}^{-1}\mathbf{L} = \boldsymbol{\Delta}$, a diagonal matrix.

  - Note: the function take the value zero if $\boldsymbol{\Sigma} = \mathbf{LL}' + \boldsymbol{\Psi}$ equals $\mathbf{S}_n$
  - The MLE of $\mathbf{L}$ and $\boldsymbol{\Psi}$ can be obtained by numerical maximization

- **Result 9.1.** Let $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ be a random sample from $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} = \mathbf{LL}' + \boldsymbol{\Psi}$ is the covariance matrix for the m common factor model The maximum likelihood estimators $\hat{\mathbf{L}}$, $\hat{\boldsymbol{\Psi}}$, and $\hat{\boldsymbol{\mu}} = \bar{\mathbf{x}}$ ~~maximize~~ minimize (❖) subject to $\hat{\mathbf{L}}'\hat{\boldsymbol{\Psi}}^{-1}\hat{\mathbf{L}}$ being diagonal.

The maximum likelihood estimates of the communalities are

By the invariance property of MLE.

$$\hat{h}_i^2 = \hat{\ell}_{i1}^2 + \hat{\ell}_{i2}^2 + \cdots + \hat{\ell}_{im}^2 \qquad \text{for } i = 1, 2, \ldots, p$$

$\theta \qquad \tau = g(\theta)$

$\hat{\theta}_{MLE} \qquad \hat{\tau}_{MLE} = g(\hat{\theta}_{MLE})$

so

$$\left(\begin{array}{c}\text{Proportion of total sample}\\ \text{variance due to } j\text{th factor}\end{array}\right) = \frac{\hat{\ell}_{1j}^2 + \hat{\ell}_{2j}^2 + \cdots + \hat{\ell}_{pj}^2}{s_{11} + s_{22} + \cdots + s_{pp}}$$

- If the variables are standardized so that $\mathbf{Z} = \mathbf{V}^{-1/2}(\mathbf{X} - \boldsymbol{\mu})$, then the covariance matrix $\boldsymbol{\rho}$ of $\mathbf{Z}$ has the representation

$\mathbf{V} = \begin{bmatrix} s_{11} & & 0 \\ & \ddots & \\ 0 & & s_{pp} \end{bmatrix}$

correlation matrix of X

$$\boldsymbol{\rho} = \mathbf{V}^{-1/2}\underbrace{\boldsymbol{\Sigma}}_{= LL'+\Psi}\mathbf{V}^{-1/2} = \underbrace{(\mathbf{V}^{-1/2}\mathbf{L})}_{L_z}\underbrace{(\mathbf{V}^{-1/2}\mathbf{L})'}_{L_z'} + \underbrace{\mathbf{V}^{-1/2}\boldsymbol{\Psi}\mathbf{V}^{-1/2}}_{\Psi_z}$$

Thus, $\boldsymbol{\rho}$ has a factorization with loading matrix $\mathbf{L}_z = \mathbf{V}^{-1/2}\mathbf{L}$ and specific variance matrix $\boldsymbol{\Psi}_z = \mathbf{V}^{-1/2}\boldsymbol{\Psi}\mathbf{V}^{-1/2}$

$= L_z L_z' + \Psi_z$

- By the invariance property of MLE, the MLE of $\boldsymbol{\rho}$ is

$$\hat{\boldsymbol{\rho}} = (\hat{\mathbf{V}}^{-1/2}\hat{\mathbf{L}})(\hat{\mathbf{V}}^{-1/2}\hat{\mathbf{L}})' + \hat{\mathbf{V}}^{-1/2}\hat{\boldsymbol{\Psi}}\hat{\mathbf{V}}^{-1/2}$$

$$= \hat{\mathbf{L}}_z\hat{\mathbf{L}}_z' + \hat{\boldsymbol{\Psi}}_z$$

MLE of loading matrix from covariance.

where $\underbrace{\hat{\mathbf{V}}}_{= \begin{bmatrix} S_{11} & & \\ & \ddots & \\ & & S_{nn} \end{bmatrix}}^{-1/2}$ and $\hat{\mathbf{L}}$ are the maximum likelihood estimators of $\mathbf{V}^{-1/2}$ and $\mathbf{L}$.

Why? ∵ they impose different constraints

◆ Note: There is usually no relationship between the principal components of the sample correlation matrix and the sample covariance matrix. For maximum likelihood factor analysis, however, the results of analyzing either matrix are essentially equivalent (this is not true of principal factor analysis)

↳ Note: not $\hat{L}_z = \hat{L}$

- The MLE method could produce very different results when $m \to m+1$
- The MLE approach can also experience difficulties with Heywood cases $\hat{\psi}_i < 0$.
- comparison of the PC and MLE approaches (Example: stock-price data)

Variable	Maximum likelihood: Estimated factor loadings (not orthogonal) F_1	F_2	Specific variances $\hat{\psi}_i = 1 - \hat{h}_i^2$	Principal components: Estimated factor loadings (orthogonal) F_1	F_2	Specific variances $\tilde{\psi}_i = 1 - \tilde{h}_i^2$
1. J P Morgan	.115	.755	.42	.732	−.437	.27
2. Citibank	.322	.788	.27	.831	−.280	.23
3. Wells Fargo	.182	.652	.54	.726	−.374	.33
4. Royal Dutch Shell	1.000	−.000	.00	.605	.694	.15
5. Texaco	.683	−.032	.53	.563	.719	.17
Cumulative proportion of total (standardized) sample variance explained	.323	.647		.487	.769	

$\frac{\Sigma 0}{5} = 1 - 0.647$

$0.647 - 0.323 = 0.324$

$0.769 - 0.487 = 0.282$

larger

- orthogonality
- estimated value of loading (very different ← ∵ different constraints)
- total variance explained.

◆ residual matrix $(\Sigma - LL' - \Psi)$

MLE

$$\begin{bmatrix} 0 & .001 & -.002 & .000 & .052 \\ .001 & 0 & .002 & .000 & -.033 \\ -.002 & .002 & 0 & .000 & .001 \\ .000 & .000 & .000 & 0 & .000 \\ .052 & -.033 & .001 & .000 & 0 \end{bmatrix}$$

PC

$$\begin{bmatrix} 0 & -.099 & -.185 & -.025 & .056 \\ -.099 & 0 & -.134 & .014 & -.054 \\ -.185 & -.134 & 0 & .003 & .006 \\ -.025 & .014 & .003 & 0 & -.156 \\ .056 & -.054 & .006 & -.156 & 0 \end{bmatrix}$$

— n is large enough (note: not p large)

• A large sample test for the number of common factors (i.e., m)

➢ Recall: choosing m might be done by examining the total variance explained — may be subjective

➢ Assume the common factors (F) and the specific factors (ε) are jointly normally distributed

➢ null and alternative hypotheses

$$H_0: \underset{(p\times p)}{\boldsymbol{\Sigma}} = \underset{(p\times m)}{\mathbf{L}} \; \underset{(m\times p)}{\mathbf{L}'} + \underset{(p\times p)}{\boldsymbol{\Psi}}$$ subject to $L'\Psi^{-1}L = \Delta$.

H_1: $\boldsymbol{\Sigma}$ any other positive definite matrix.

- under alternative, $\hat{\boldsymbol{\mu}} = \bar{\mathbf{x}}$ and $\hat{\boldsymbol{\Sigma}} = ((n-1)/n)\mathbf{S} = \mathbf{S}_n$, the maximized likelihood is proportional to $|\mathbf{S}_n|^{-n/2}e^{-np/2}$
- under null, $\hat{\boldsymbol{\mu}} = \bar{\mathbf{x}}$ and $\hat{\boldsymbol{\Sigma}} = \hat{\mathbf{L}}\hat{\mathbf{L}}' + \hat{\boldsymbol{\Psi}}$, where $\hat{\mathbf{L}}$ and $\hat{\boldsymbol{\Psi}}$ are the MLE of $\mathbf{L}$ and and $\boldsymbol{\Psi}$, the maximized likelihood is proportional to

$$\underbrace{|\hat{\mathbf{L}}\hat{\mathbf{L}}' + \hat{\boldsymbol{\Psi}}|}_{\hat{\Sigma}}^{-n/2} \exp\left(-\tfrac{1}{2}n\, \boxed{\mathrm{tr}\left[(\hat{\mathbf{L}}\hat{\mathbf{L}}' + \hat{\boldsymbol{\Psi}})^{-1}\mathbf{S}_n\right]}\right)$$ $= P$ (supplement 9A)

- likelihood ratio statistic

$$-2\ln\Lambda = -2\ln\left[\frac{\text{maximized likelihood under } H_0}{\text{maximized likelihood}}\right] = n\ln\left(\frac{|\hat{\boldsymbol{\Sigma}}|}{|\mathbf{S}_n|}\right)$$

$\hat{L}\hat{L}'+\hat{\psi}$

$\sim \chi^2_{dim(H_1)-dim(H_0)}$

with degrees of freedom,

\# of parameters in a covariance matrix

\# of parameters in L, ψ.

$$v - v_0 = \boxed{\tfrac{1}{2}p(p+1)} - \left[\boxed{p(m+1)} - \boxed{\tfrac{1}{2}m(m-1)}\right]$$
$$= \tfrac{1}{2}[(p-m)^2 - p - m]$$

→ # of constraints in $L'\psi^{-1}L = \Delta$.

- Bartlett (1954) has shown that the chi-square approximation can be improved by replacing n with $(n - 1 - (2p + 4m + 5)/6)$
- Using Bartlett's correction, we reject H_0 at the α significant level if

$$(n - 1 - (2p + 4m + 5)/6)\ln\frac{|\hat{\mathbf{L}}\hat{\mathbf{L}}' + \hat{\boldsymbol{\Psi}}|}{|\mathbf{S}_n|} > \chi^2_{\underbrace{[(p-m)^2-p-m]/2}_{>0}}(\alpha)$$

- Note: because the number of degrees of freedom must be positive, it follows that

$$m < \tfrac{1}{2}\left(2p + 1 - \sqrt{8p+1}\right)$$

in order to apply the test

❖ **Reading**: Textbook, 9.3, Supplement 9A

• 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

add constraints

a point in $\mathbb{R}^m$

L: $F_1 \cdots F_m$; x_1: $\ell_{11} \cdots \ell_{1m}$; $\vdots$; x_p

T: $t_1 \cdots t_m \in \mathbb{R}^m$

orthogonal unit vector

- 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

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

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

t_1: arbitrary.
t_2: → $t_2 \perp t_1$ → 1
t_3: → $t_3 \perp t_2, t_3 \perp t_1$ → 2
$\vdots$
t_m: → → $m-1$

$\frac{m(m-1)}{2}$

- **Q**: why is it called "rotation"?
- **Q**: dimension of *all* rotated factor loadings = $\frac{m(m-1)}{2}$?
- Note: estimated covariance/correlation matrix, residual matrix, estimated specific variances and communalities remain unchanged after rotation

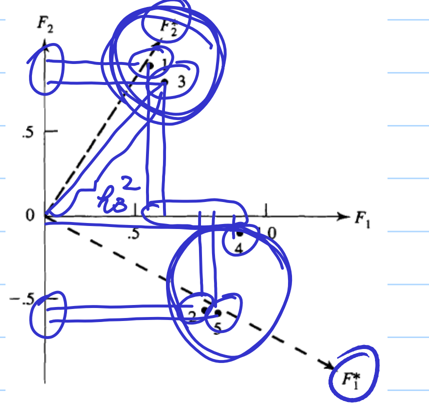
→ $\hat{\Sigma} = \hat{L}\hat{L}' + \hat{\psi} = \hat{L}^*\hat{L}^{*\prime} + \hat{\psi}$

- **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

often seen in PC approach.

➢ **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	

same

same

different

➢ varimax criterion

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

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

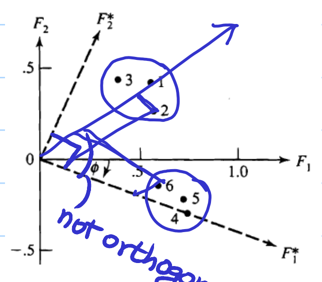
as large as possible. $= \sum_j \left\{ \underbrace{\frac{1}{p}(\Sigma(\ell^2)^2)}_{EY^2} - \underbrace{\left[\frac{1}{p}(\Sigma\ell^2)\right]^2}_{(EY)^2} \right\}.$

- interpretation of the varimax criterion: $X \to \ell$, $Y = X^2$

$$V \propto \sum_{j=1}^{m}\left(\boxed{\text{variance}}\text{ of }\boxed{\text{squares}}\text{ of (scaled) loadings for } j\text{th factor}\right)$$

  - 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 $cov(F^*) \neq I$.



• factor scores (predict values of the unobserved random factors $F_1, \ldots, F_m$) similar to $\hat\varepsilon$ (regression)

➢ (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 $Y = X\beta + \varepsilon$ minimize $(Y - X\beta)'\Psi^{-1}(Y - X\beta)$
$$\underset{(p\times 1)}{\mathbf{X}} - \underset{(p\times 1)}{\boldsymbol{\mu}} = \underset{(p\times m)}{\mathbf{L}}\underset{(m\times 1)}{\mathbf{F}} + \underset{(p\times 1)}{\boldsymbol{\varepsilon}}$$
  - The sum of the squares of the errors, weighted by the reciprocal of their variances, is
$$\sum_{i=1}^{p}\frac{\varepsilon_i^2}{\psi_i} = \boldsymbol{\varepsilon}'\boldsymbol{\Psi}^{-1}\boldsymbol{\varepsilon} = (\mathbf{x} - \boldsymbol{\mu} - \mathbf{Lf})'\underset{\psi^{-1/2}\psi^{-1/2}}{\boldsymbol{\Psi}^{-1}}(\mathbf{x} - \boldsymbol{\mu} - \mathbf{Lf}) \qquad (*)$$
$$= [\psi^{-1/2}(x - \mu - Lf)]'[\psi^{-1/2}(x - \mu - Lf)]$$
  - choosing the estimates $\hat{\mathbf{f}}$ of $\mathbf{f}$ to minimize $(*)$. The solution is
$$\hat{\mathbf{f}} = (\mathbf{L}'\boldsymbol{\Psi}^{-1}\mathbf{L})^{-1}\mathbf{L}'\boldsymbol{\Psi}^{-1}(\mathbf{x} - \boldsymbol{\mu}) = [(\psi^{-1/2}L)'(\psi^{-1/2}L)]^{-1}(\psi^{-1/2}L)'\psi^{-1/2}(x - \mu)$$
  - take the estimates $\hat{\mathbf{L}}$, $\hat{\boldsymbol{\Psi}}$, and $\hat{\boldsymbol{\mu}} = \bar{\mathbf{x}}$ as the true values and obtain the ~~factor~~ X scores Y 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 $\mathbf{z}_j = \mathbf{D}^{-1/2}(\mathbf{x}_j - \bar{\mathbf{x}})$