因子分析

张伟平

zwp@ustc.edu.cn

Office: 东区管理科研楼 1006

Phone: 63600565

课件 http://staff.ustc.edu.cn/~zwp/

论坛 http://fisher.stat.ustc.edu.cn

简介

1.1 简介

- 因子分析起源于 20 世纪初, K. 皮尔逊 (Pearson) 和 C. 斯皮尔曼 (Spearman) 等学者为定义和测定智力所作的努力，主要是由对心理测量学有兴趣的科学家们培育和发展了因子分析。
- 因子分析常用于对不能直接观测的变量, 例如智力, 音乐能力, 爱国主义, 消费者态度等等, 进行推断.
- 个体在一个或多个因子上取值的变化可以影响可观测变量中的多个变量 (某个变量子集), 从而导致该子集内变量之间高度相关.
- 因子分析常包括**探索性因子分析**(Exploratory FA) 和**验证性因子分析**(Confirmatory FA) 两类.
- 探索性因子分析的目的是为了降维, 降维的方式是试图用少数

几个潜在的, 不可观测的随机变量来描述原始变量间的协方差关系.

- 验证性因子分析则是以对事先的模型假设进行检验为重心.

一家市场研究公司希望了解消费者如何选择光顾哪家商店.

↑Example
↓Example

- 随机对光顾每家商店的消费者进行了包含 $80(p = 80)$ 道问题的问卷调查
- 市场研究人员提出假设 **消费者的选择是基于几个潜在的因子: 商店人员的友好程度, 消费者服务水平, 商店气氛, 产品种类, 产品质量和一般的价格水平.**
- 因子分析使用 80 个问题的响应值之间的相关性来决定 80 个变量是否可以分为 6 组, 每组分别反映所假设的一个因子.

1.2 正交因子模型

- $\mathbf{X} = (X_1, \ldots, X_p)'$ 为 p 维观测变量, 其均值和协方差矩阵分别为 μ 和 Σ.
- 因子分析模型假定 X 可以表示为 m 个**公共因子** (common factors)$F_1, \ldots, F_m$ 和 p 个**特殊因子** (unique factors)$\epsilon_1, \ldots, \epsilon_p$ 的线性组合

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

其中 l_{ij} 称为第 i 个变量在第 j 个因子上的**因子负荷** (factor loading).

- 表达成矩阵形式: $\mathbf{X}_{p\times 1} = \boldsymbol{\mu} + \mathbf{L}_{p\times m}\mathbf{F} + \boldsymbol{\epsilon}$, $\mathbf{L}$ 称为因子负荷阵.

  上述模型和线性模型形式非常相像, 但是注意右边每个量我们均不能直接观测到, 即公共因子和特殊因子均为随机变量且不能直接观测到. 因此需要假定一些结构才能进行推断.

正交因子模型 (Orthogonal factor model) 假设

(1). $E\mathbf{F} = 0, \quad Var(\mathbf{F}) = E(\mathbf{F}\mathbf{F}') = I_m$

(2). $E\boldsymbol{\epsilon} = 0, \quad Var(\boldsymbol{\epsilon}) = E(\boldsymbol{\epsilon}\boldsymbol{\epsilon}') = \Psi = diag\{\psi_1, \ldots, \psi_p\}$

(3). $cov(\mathbf{F}, \boldsymbol{\epsilon}) = 0$

- 在正交因子模型假设下

$$\Sigma = Var(\mathbf{X}) = Var(\mathbf{L}\mathbf{F} + \boldsymbol{\epsilon}) = \mathbf{L}\mathbf{L}' + \Psi$$

特别,

$$\sigma_{ii} = Var(X_i) = \sum_{j=1}^{m} l_{ij}^2 + \psi_i := h_i^2 + \psi_i$$

$$\sigma_{ij} = cov(X_i, X_j) = \sum_{k=1}^{m} l_{ik} l_{kj}, i \neq j$$

- 方差 σ_{ii} 由两部分构成: 由 m 个公共因子贡献的部分, h_i^2, 称为**共性**或**共性方差**(communality); 由公共因子不能解释 (由特殊因子解释) 的部分称为**特殊方差**(uniqueness, specific variance).

- 由协方差结构假设, 正交因子模型假定了 $p(p+1)/2$ 个方差参数可以通过 $pm + p$ 个参数表达.

- 出于降维的需要, 我们常常希望 m 要比 p 小得多, 这样分解式 $\Sigma = \mathbf{L}\mathbf{L}' + \Psi$ 通常只能近似成立. 一般来说, m 选取得越小, 上

述近似效果就越差, 即因子模型拟合得越不理想. 拟合得太差的因子模型是没有什么实际意义的.

- 注意不是所有的协方差矩阵 Σ 都可以分解为 $\mathbf{L}\mathbf{L}' + \Psi$(见课本例 9.2).
- 公共因子 $\mathbf{F}$ 和负荷阵 $\mathbf{L}$ 不唯一: 对任意正交矩阵 $\mathbf{T}$ 有

$$X - \boldsymbol{\mu} = \mathbf{L}\mathbf{F} + \epsilon = (\mathbf{L}\mathbf{T})(\mathbf{T}'\mathbf{F}) + \boldsymbol{\epsilon} = \mathbf{L}^*\mathbf{F}^* + \boldsymbol{\epsilon}$$

$\mathbf{L}^*, \mathbf{F}^*$ 满足正交因子模型的所有假设. 因此因子负荷阵

$$\mathbf{L}^* = \mathbf{L}\mathbf{T} \text{ 和 } \mathbf{L}$$

是同样的表达.

- 由正交矩阵的性质, 称正交变换 $\mathbf{L}\mathbf{T}, \mathbf{T}'\mathbf{F}$ 为因子旋转. 在合适的准则下对因子负荷阵 $\mathbf{L}$ 进行旋转, 以期得到更易解释的结果.

1.3 参数估计

- 对可观测变量 $\mathbf{X}$, 假设我们有一组样本: $\mathbf{x}_1, \ldots, \mathbf{x}_n$
- 由 $\Sigma = \mathbf{L}\mathbf{L}' + \Psi$ 结构, 需要估计 $\mathbf{L}$ 和 Ψ
- 当 $\mathbf{L}$ 和 Ψ 被估计出来后, 使用线性模型理论可得因子 $\mathbf{F}$ 的估计 (称为**因子得分**).
- 记 $\mathbf{S}$ 为样本协方差矩阵, 则 $\mathbf{S}$ 为 Σ 的估计. 于是首先我们**需要研究 p 个变量之间是否存在足够大的相关以进行因子分析**. 如果 $\mathbf{S}$ 的非对角元都约为零, 则特殊因子方差 ψ_i 占控制地位, 因此此时我们不能识别公共因子.
- 常见的估计方法包括
  - 主成分法
  - 迭代主因子法

– 最大似然法 (假设正态)

- 第一种方法对方差关注更多, 后两种方法关注如何使用公共因子的波动来描述观测性状之间的相关性.

1.3.1 主成分法

(The principal component method)

- 由 Σ 的非负定性, 可以得到正交分解

$$\Sigma = \lambda_1 \mathbf{e}_1 \mathbf{e}_1' + \cdots + \lambda_p \mathbf{e}_p \mathbf{e}_p := \mathbf{e} \Lambda \mathbf{e}'$$

其中 $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$ 为特征根, $\mathbf{e}_1, \ldots, \mathbf{e}_p$ 为相应的特征向量. $\mathbf{e} = [\mathbf{e}_1, \ldots, \mathbf{e}_p]$, $\Lambda = diag\{\lambda_1, \ldots, \lambda_p\}$.

- 记 $\mathbf{Y} = \mathbf{e}'(\mathbf{X} - \boldsymbol{\mu})$, 则 $\mathbf{Y}$ 为总体主成分. 于是

$$\mathbf{X} - \boldsymbol{\mu} = \mathbf{eY} = \sum_{j=1}^{m} \mathbf{e}_j Y_j + \sum_{j=m+1}^{p} \mathbf{e}_j Y_j$$

$$:= \mathbf{LF} + \boldsymbol{\epsilon},$$

其中 $\mathbf{L} = [\sqrt{\lambda_1}\mathbf{e}_1, \ldots, \sqrt{\lambda_m}\mathbf{e}_m]$, $\mathbf{F} = (Y_1/\sqrt{\lambda_1}, \ldots, Y_m/\sqrt{\lambda_m})'$, $\boldsymbol{\epsilon} = \sum_{j=m+1}^{p} \mathbf{e}_j Y_j$.

- 可以验证正交因子模型的假设条件 (1) 和 (3) 成立, 但是 (2) 未必成立.

- 相应地,

$$\Sigma = \mathbf{LL}' + Var(\boldsymbol{\epsilon})$$

若特殊因子 $\boldsymbol{\epsilon}$ 对协方差的贡献很小, 则 $\Sigma \approx \mathbf{LL}'$, 从而对协方差的一个良好近似为

$$\Sigma \approx \mathbf{LL}' + diag\{\psi_1, \ldots, \psi_p\}$$

其中 $\psi_i = \sigma_{ii} - \sum_{j=1}^{m} l_{ij}^2 = \sigma_{ii} - h_i^2, i = 1, \ldots, p.$

- 使用样本数据 $\mathbf{x}_1, \ldots, \mathbf{x}_n$ 估计上述参数. 记 $(\hat{\lambda}_i, \hat{\mathbf{e}}_j), j = 1, \ldots, p$ 为样本协方差矩阵 $\mathbf{S}$ 的特征根和特征向量对, 且 $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_p \geq 0$.

因子模型的主成分解:

$$\hat{\mathbf{L}} = [\sqrt{\hat{\lambda}_1}\hat{\mathbf{e}}_1, \ldots, \sqrt{\hat{\lambda}_m}\hat{\mathbf{e}}_m],$$

$$\hat{\Psi} = diag(\mathbf{S} - \hat{\mathbf{L}}\hat{\mathbf{L}}') = diag\{\hat{\psi}_1, \ldots, \hat{\psi}_p\}, \hat{\psi}_i = s_{ii} - \sum_{j=1}^{m} \hat{l}_{ij}$$

- 在一些问题里, 因子个数 m 是事先取定的

- 若 m 不能事先取定, 则可以基于不同 m 下拟合的模型结果来寻找 “最好” 的 m:
  - 选择 m, 使得
    $$\mathbf{S} - (\hat{\mathbf{L}}\hat{\mathbf{L}}' + \hat{\Psi})$$
    有较小的非对角元.
  - 类似主成分个数的选择方法, 选择 m 使得
    $$\frac{\sum_{i=1}^{m} \hat{\lambda}_i}{\hat{\lambda}_1 + \cdots \hat{\lambda}_p}$$
    较大.
  - 基于相关系数矩阵分析时候, 选择特征根大于 1 的个数.
- 对标准化变量, 此时协方差矩阵为相关系数阵, 相应的分析和基于协方差的分析完全类似.

1.3.2 迭代主因子法

(Iterated principal factor method)

- ◇ 主因子法是对主成分方法的修正. 我们通过样本相关系数矩阵 R 来说明. 同样的过程也可以用于样本协方差矩阵.
- ◇ 由于总体相关系数矩阵 $\boldsymbol{\rho} = \mathbf{L}\mathbf{L}' + \Psi$, 因此 $\boldsymbol{\rho} - \Psi = \mathbf{L}\mathbf{L}'$. 对角元 $\rho_{ii} = h_i^2 + \psi_i = 1$.
- ◇ 从而若 Ψ 已知, 则可对 $\boldsymbol{\rho} - \Psi$ 进行矩阵分解得到 $\mathbf{L}$; 有了 $\mathbf{L}$ 后可以由 $\Psi = \rho - \mathbf{L}\mathbf{L}'$ 得到 Ψ.
- ◇ 因此, 使用样本相关系数矩阵 R 代替 $\boldsymbol{\rho}$, 使用 Ψ 的一个初始估计代替, 则可以迭代求解 $\hat{\mathbf{L}}, \hat{\Psi}$. 此即为迭代主因子法.

计算步骤如下:

1. 若我们有 ψ_i 的估计 ψ_i^*, 则使用 $1-\psi_i^* = h_i^{*2}$ 代替 $\rho_{ii}=1$ 后, 得到一个 “ 缩减” 的样本相关系数矩阵

$$\mathbf{R}_r = \begin{pmatrix} h_1^{*2} & r_{12} & \cdots & r_{1p} \\ r_{21} & h_2^{*2} & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & h_p^{*2} \end{pmatrix}$$

对 m 个因子, 类似主成分方法将 $\mathbf{R}_r$ 因子化得到

$$\mathbf{R}_r \approx \mathbf{L}_r^* \mathbf{L}_r^{*\prime}$$

其中 $\mathbf{L}_r^*$ 由 $\mathbf{R}_r$ 的前 m 个特征根和特征向量构成.

2. 有了 $\mathbf{L}_r^*$ 后, 更新 Ψ 的估计为 $\Psi^* = diag\{\mathbf{R} - \mathbf{L}_r^* \mathbf{L}_r^{*\prime}\}$

3. 重复上面 1-2 步直至满足收敛准则.

特殊方差 ψ_i(或共性方差 h_i^2) 的常用初始估计有如下几种:

- 取 $\psi_i^* = 1/r^{ii}$, 其中 r^{ii} 为 R^{-1} 的第 i 个对角元, 此时 $h_i^{*2} = 1 - \psi_i^*$ 为 X_i 和其它 $p-1$ 个变量间的样本负相关系数的平方. 这种初始估计方法最常用.
- 取 $h_i^{*2} = \max_{j \neq i} |r_{ij}|$, 此时 $\psi^* = 1 - h_i^{*2}$
- 取 $h_i^{*2} = 1$, 此时 $\psi^* = 0$, 得到的 $\mathbf{L}_r^*$ 为主成分解.

对主因子法, 需要注意:

- 缩减的相关系数矩阵 $\mathbf{R}_r$ 未必总是正定的, 因此其一些特征根可能是负的
- 主因子方法的结果对因子个数 m 比较敏感, 不同 m 下得到的结果可能差异较大.
- 如果 m 太大, 一些共性方差估计 h_i^{*2} 可能会大于 1.

1.3.3 最大似然方法

- 假定 $\mathbf{F} \sim N_m(0, I_m), \boldsymbol{\epsilon} \sim N_p(0, \Psi)$ 且两者相互独立. 从而 $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \mathbf{L}\mathbf{L}' + \Psi)$.
- 记 $l(\boldsymbol{\mu}, \mathbf{L}, \Psi)$ 为对数似然函数, 因此

$$\begin{aligned}(\hat{\boldsymbol{\mu}}, \hat{\mathbf{L}}, \hat{\Psi}) &= \arg\max l(\boldsymbol{\mu}, \mathbf{L}, \Psi) \\ &= \arg\max \Big\{ -\frac{n}{2} log|\mathbf{L}\mathbf{L}' + \Psi| \\ &\qquad -\frac{1}{2}\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu})'(\mathbf{L}\mathbf{L}' + \Psi)^{-1}(\mathbf{x}_i - \boldsymbol{\mu})\Big\}\end{aligned}$$

- 计算得到 $\hat{\boldsymbol{\mu}} = \bar{\mathbf{x}}$, $\hat{\mathbf{L}}, \hat{\Psi}$ 满足

$$\begin{cases} \hat{\Sigma}\hat{\Psi}^{-1}\hat{\mathbf{L}} = \hat{\mathbf{L}}(I_m + \hat{\mathbf{L}}'\hat{\Psi}^{-1}\hat{\mathbf{L}}) \\ \hat{\Psi} = diag(\hat{\Sigma} - \hat{\mathbf{L}}\hat{\mathbf{L}}') \end{cases}$$

其中 $\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})$.

- 由于对 $\hat{\mathbf{L}}$ 的解不唯一 (对任意正交阵 T, $\hat{\mathbf{L}}T$ 也满足等式), 因此为了得到唯一解, 常常附加计算上的限制条件

$$\hat{\mathbf{L}}'\hat{\Psi}^{-1}\hat{\mathbf{L}} \text{ 为对角阵}$$

- 对最大似然解, 当因子个数增加时候, 原来因子的估计负荷会发生变化, 这与主成分解和主因子解不同.

因子个数的似然比检验

- 使用似然比检验方法, 可以检验因子分析模型对 p 元变量的协方差矩阵是否合适:

$$H_0 : \Sigma_{p\times p} = \mathbf{L}_{p\times m}\mathbf{L}' + \Psi \leftrightarrow H_a : \Sigma > 0$$

- 计算得到检验 H_0 的对数似然比统计量

$$-2\log\Lambda = -2\log\frac{H_0\text{下最大似然}}{\text{最大似然}} = n\log\left(\frac{|\hat{\mathbf{L}}\hat{\mathbf{L}}' + \hat{\Psi}|}{|\hat{\Sigma}|}\right) \rightsquigarrow_{H_0} \chi^2_{df}$$

其中 $df = dim\Theta_a - dim\Theta_0 = (p + p(p+1)/2) - (p + pm + p - m(m-1)/2) = [(p-m)^2 - p - m]/2$.

- Bartlett 修正 $-2log\Lambda$ 为如下式以更好的逼近 χ^2:

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

- 对给定的检验水平 α, 拒绝域为 (对较大的 n 和 $n-p$)

$$-2log\Lambda > \chi^2_{df}(\alpha)$$

- 为保证 $df > 0$, 必须满足 $m < \frac{1}{2}(2p + 1 - \sqrt{8p+1})$.

- 当 n 较大, m 相对于 p 较小时候, 此时假设 H_0 通常会被拒绝, 这导致保留更多的因子. 然而此时 $\hat{\mathbf{L}}\hat{\mathbf{L}}' + \hat{\Psi}$ 可能已经和 $\mathbf{S}$ 足够接近. 因此选择 m 时必须加以判断.

1.4 因子旋转

- 前面我们已经看到, 对因子负荷阵乘以一个正交矩阵可以得到对协方差矩阵同样的逼近.
- 这意味着, 我们使用 $\hat{\mathbf{L}}$ 或 $\hat{\mathbf{L}}T$ 来估计因子负荷阵都是可以的, 其中 T 为任意正交矩阵.
- 估计的残差矩阵 $\mathbf{S}-\hat{\mathbf{L}}\hat{\mathbf{L}}'-\hat{\Psi}=\mathbf{S}-(\hat{\mathbf{L}}T)(\hat{\mathbf{L}}T)'-\hat{\Psi}$ 保持不变, 而且估计的特殊方差和共性方差均不变.
- 我们希望通过旋转因子来更好对结果进行解释: 使因子负荷的平方两极分化, 要么接近 0, 要么接近 1.
- 因子旋转方法主要有正交旋转和斜交变换两类.

正交旋转方法使得旋转后的因子仍然保持独立性. 常用的正交旋转方法包括: varimax, quartimax, equamax 等.

正交旋转方法 (Orthogonal rotation)

- **最大方差旋转法 (varimax)**(Harman, 1976)
  - 直观上使得 p 个变量中的每个变量仅在一个因子上有很大的负荷, 在其他因子上有适中到很小的负荷. 即在每个因子上的负荷方差最大.
  - 该方法简化因子 (因子负荷阵的列, 是最常用的因子旋转方法.
  - 记 $\hat{\mathbf{L}}^* = (\hat{l}^*_{ij})$ 为旋转的因子负荷阵, 令 $l^*_{ij} = \hat{l}^*_{ij}/\hat{h}_i$, 以及

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

  - 从而选择 T, 使得 V 达到最大.

- **最大四次方法 (quartimax)**
  - varimax 旋转方法会破坏"overall" 因子: 每个变量在该因子上都有高负荷.
  - quartimax 旋转目的在于
    * 保留"overall" 因子
    * 建立其它因子, 使得每个变量在至多一个 (其它) 因子上有高负荷.
  - 最大化

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

  - 可使得解释每个变量所需的因子最少。该方法简化变量 (因子负荷阵的行)。

- **最大平衡值法 (Equamax)**
  - 它是最大方差旋转法和最大四次方法的一种权衡.
  - 目的是同时简化行和列.

斜交变换 (Oblique transformation)

正交旋转适合于公共因子假定是相互独立的因子模型. 除正交旋转外, 许多研究者也考虑斜交 (非正交) 变换. 此时允许因子之间相关, 这在社会科学里是常见的, 因为研究者很少同时研究人类行为一些完全独立的方面 (因子), 关心的因子更多是相关的. 常用的斜交旋转方法包括 Direct oblimin, Promax, quartimin 等.

- **直接最小斜交变换 (Direct oblimin)**
  - 得到的因子是相关的
- **最优斜交变换 (Promax)**(Hendrickson and White, 1964)

  - 得到的因子是相关的
  - 在正交旋转的基础上, 再进行斜交变换.
  - 该变换可比直接最小斜交变换更快地计算出来，因此适用于大型数据集。

- 斜交变换下特殊方差保持不变.

使用哪个?

- 一般来说选择哪种方法不是非常关键的
- 正交旋转容易解释
- 斜交变换下结构可能更加简单, 但是因子之间的相关性难以解释

1.5 因子得分

- 公共因子的估计值, 称为因子得分 (factor scores), 有时候也是感兴趣的, 比如把得到的因子作为自变量来进行回归.
- 因子得分是对不可观测的随机变量 (因子) 的估计, 不同于未知参数的估计
- 假设 $\mathbf{x}_1, \ldots, \mathbf{x}_n$ 为 p 元变量 $\mathbf{X}$ 的一组观测样本, 满足正交因子模型

$$\mathbf{x}_i - \mu = Lf_i + e_i, i = 1, \ldots, n$$

Bartlett **方法**(加权最小二乘方法)

- 视特殊因子为随机误差, L 为回归设计阵, f_i 为回归系数, 则由于 $Var(e_i) = \Psi$, 因此由广义最小二乘方法, 得到

$$\hat{f}_i = (L'\Psi^{-1}L)^{-1}L'\Psi^{-1}(\mathbf{x}_i - \mu), i = 1, \ldots, n$$

– (最大似然方法) 使用 L, μ, Ψ 的最大似然估计 $\hat{L}, \hat{\mu} = \bar{\mathbf{x}}, \hat{\Psi}$ 代替, 得到

$$\hat{f}_i^{LS} = (\hat{L}'\hat{\Psi}^{-1}\hat{L})^{-1}\hat{L}'\hat{\Psi}^{-1}(\mathbf{x}_i - \bar{\mathbf{x}}), i = 1, \ldots, n$$

$\hat{e}_i = \mathbf{x}_i - \bar{\mathbf{x}} - \hat{L}\hat{f}_i$ 为 e_i 的估计.

– (主成分法) 使用 L, μ, Ψ 的主成分估计 $\tilde{L}, \hat{\mu} = \bar{\mathbf{x}}, \tilde{\Psi}$ 代替, 此时常使用平凡的最小二乘方法 (不加权), 得到

$$\tilde{f}_i = (\tilde{L}'\tilde{L})^{-1}\tilde{L}'(\mathbf{x}_i - \bar{\mathbf{x}}), i = 1, \ldots, n$$

$\tilde{e}_i = \mathbf{x}_i - \bar{\mathbf{x}} - \tilde{L}\tilde{f}_i$ 为 e_i 的估计.

回归方法

– 若公共因子和特殊因子服从多元正态分布, 则 $\mathbf{X} - \boldsymbol{\mu} = \mathbf{LF} + \boldsymbol{\epsilon}$ 有多元正态分布 $N_p(0, \mathbf{LL}' + \Psi)$.

– 由于

$$\Sigma^* = Var\begin{pmatrix} \mathbf{X} \\ \mathbf{F} \end{pmatrix} = Var\begin{pmatrix} \mathbf{LF}+\boldsymbol{\epsilon} \\ \mathbf{F} \end{pmatrix} = \begin{pmatrix} \mathbf{LL}'+\Psi & \mathbf{L} \\ \mathbf{L}' & I_m \end{pmatrix}$$

即 $(\mathbf{X}-\boldsymbol{\mu}, \mathbf{F}) \sim N(0, \Sigma^*)$.

– 在多元正态下, 易知 $\mathbf{F}|\mathbf{X}=\mathbf{x}$ 有多元正态分布, 其条件均值为

$$E(\mathbf{F}|\mathbf{X}=\mathbf{x}) = 0 + \mathbf{L}'(\mathbf{LL}'+\Psi)^{-1}(\mathbf{x}-\boldsymbol{\mu})$$

以及条件协方差为

$$Var(\mathbf{F}|\mathbf{X}=\mathbf{x}) = I_m - \mathbf{L}'(\mathbf{LL}'+\Psi)^{-1}\mathbf{L} = (\mathbf{L}'\Psi^{-1}\mathbf{L}+I_m)^{-1}$$

– 给定观测 $\mathbf{x}_i$, 使用最大似然估计 $\hat{L}, \hat{\Psi}, \hat{\boldsymbol{\mu}} = \bar{\mathbf{x}}$ 代替, 得到

$$\begin{aligned} \hat{f}_i^R &= \hat{L}'(\hat{L}\hat{L}'+\hat{\Psi})^{-1}(\mathbf{x}_i-\bar{\mathbf{x}}) \\ &= (I_m+\hat{L}'\hat{\Psi}\hat{L})^{-1}\hat{L}'\hat{\Psi}^{-1}(\mathbf{x}_i-\bar{\mathbf{x}}), i=1,\ldots,n \end{aligned}$$

– 加权最小二乘估计 $\hat{f}_i^{LS}$ 和回归估计 $\hat{f}_i^R$ 之间有关系

$$\hat{f}_i^{LS} = (\hat{L}'\hat{\Psi}^{-1}\hat{L})^{-1}(I_m + \hat{L}'\hat{\Psi}^{-1}\hat{L})\hat{f}_i^R = (I_m + (\hat{L}'\hat{\Psi}^{-1}\hat{L})^{-1})\hat{f}_i^R$$

– 为降低 (可能的) 错误指定因子个数造成的影响, 实际中也常使用样本协方差矩阵 S 来代替 $\hat{\Sigma} = \hat{L}\hat{L}' + \hat{\Psi}$, 此时

$$\hat{f}_i = \hat{L}S^{-1}(\mathbf{x}_i - \bar{\mathbf{x}}), i = 1, \ldots, n$$

- 若使用因子旋转 $\hat{L}^* = \hat{L}T$, 则因子得分为

$$\hat{f}_i^* = T'\hat{f}_i, i = 1, \ldots, n$$

因子分析的步骤

- 判断数据是否可以使用因子分析方法
  - 相关系数矩阵: 变量之间有较强的相关系数, 则表明可能可以将它们归为同一因子的原因. 一般需要存在 0.3 以上的相关系数 (Tabachnick & Fidell, 2007)
  - 样本量: 越大越好. 经验上, 对 p 个变量, 样本量至少要 $5p$
  - 因子分析的充分性检验:
    * KMO(Kaiser-Meyer-Olkin) index: 用于检验因子分析是否充分. 一般需要 $KMO > 0.5$[1] .

$$KMO = \frac{\sum_{i \neq j} r_{ij}^2}{\sum_{i \neq j} r_{ij}^2 + \sum_{i \neq j} a_{ij}^2}$$

[1] Kaiser (1974) 建议的标准, 0.5 以下是不能接受的, [0.5, 0.6) 太少, [0.6, 0.7) 普通的, [0.7,0.8) 中等的, [0.8, 0.9) 值得称赞的, [0.9, 1.0) 不可思议的.

其中 $R = (r_{ij})$ 为相关相关系数矩阵, $PR = (a_{ij})$ 为偏相关系数矩阵.

* Bartlett's test of sphericity: 检验总体相关系数矩阵是否为单位阵, 即 $H_0 : \boldsymbol{\rho} = I_p$ 则 Bartlett 检验统计量 (修正的似然比) 为

$$\chi^2 = -\Big(n - 1 - \frac{2p+5}{6}\Big) log|R| \rightsquigarrow_{H_0} \chi^2_{p(p-1)/2}$$

- 进行因子模型的估计
- 决定因子的个数
- 因子旋转
- 得到最终的模型结果并解释