多元统计期末复习

因子分析

$$
\begin{gathered}
X-\mu=LF+\epsilon \\
E(F)=0, Cov(F)=I \\
E(\epsilon)=0, Cov(\epsilon)=\Psi\text{是对角阵} \\
F\text{和}\epsilon\text{独立}
\end{gathered}
$$

载荷矩阵L, 隐变量F.

$$
\begin{gathered}
\Sigma=LL'+\Psi \\
Cov(X,F)=L
\end{gathered}
$$

L可旋转，不可伸缩。

h^2共性方差 commonalities，是LL'的对角元

$1-h^2$ 特殊方差 uniquenesses，Ψ的的对角元

Heywood cases: 解没有统计意义，如Ψ的对角元小于0

PCA方法

$$
\begin{gathered}
L=[\sqrt{\lambda_1}e_1,\sqrt{\lambda_2}e_2,\cdots,\sqrt{\lambda_m}e_m]_{(p\times m)} \\
\text{样本总方差归因于因子}i\text{的比例}=\frac{\lambda_i}{\sum_{j=1}^p s_{jj}}
\end{gathered}
$$

载荷估计不随因子数量改变而改变。

$$
||S-(LL'+\Psi)||_F^2\le||S-LL'||_F^2=\sum_{i=m+1}^p\hat{\lambda}_i^2
$$

```
library(psych)
principal(data, nfactors, rotate = 'none')
```

MLE方法

假设F,ϵ都是正态的，并假设$L'\Psi^{-1}L=\Delta$.

$$
\begin{gathered}
\hat{h}_i^2=\sum_{j=1}^m\hat{l}_{ij}^2 \\
\text{样本总方差归因于因子}i\text{的比例}=\frac{\sum_{i=1}^p\hat{l}_{ij}^2}{\sum_{j=1}^p s_{jj}}
\end{gathered}
$$

载荷估计随因子数量改变。

```
factanal(x, factors, covmat = NULL,
         scores = c("none", "regression", "Bartlett"),
         rotation = c("varimax", "none"))
```

在标准化下，

$$\hat{L} = \hat{V}^{1/2}\hat{L}_z$$
$$\hat{\Psi} = \hat{V}^{1/2}\hat{\Psi}_z\hat{V}^{1/2}$$

其中$V^{1/2}$是标准差矩阵

$$V^{1/2} = diag(\sqrt{\sigma_{11}}, \sqrt{\sigma_{22}}, \cdots, \sqrt{\sigma_{pp}})$$

对比

1. 是否满足正态性，如果没有明显拒绝，则mle，pc都可用，否则只能用pc。
2. 若mle，pc都可用，看mle结果里如果方差明显有很大的，则mle更合适，更符合模型设定，说明pc强行约束了比较小的方差、不符合数据特征。
3. mle方法能更好地拟合数据特征，而pc方法运算更简便。

FA中，不是一味追求解释总方差的比例，而是更看重因子本身的实际含义是否合理。模型本身允许因子的特殊方差可以很大。**由于pc方法中因子所解释的方差的比例更大而选择该方法是不合理的。**

因子旋转

motivation：便于解释

方法：varimax

旋转改变样本总方差归因于因子的比例，但是不改变Σ、共性方差、特殊方差。

因子得分

加权最小二乘：无偏，误差大

回归：有偏，误差小

$$\text{rotation: } L* = LT, f^* = T'f$$

FA和PCA的对比

PCA是找全部主成分，使得投影到任意维数r)空间，主成分确定的r维平面都是最 优的。而FA只找一个r维平面，所以当固定维数的时候，可以在该平面内旋转，不影响投影结果，但却可以有很好的解释度。

典型相关分析

$$Cov(X^{(1)}) = \Sigma_{11}, Cov(X^{(2)}) = \Sigma_{22}, Cov(X^{(1)}, X^{(2)}) = \Sigma_{12}$$

典型相关变量对$U_k = e_k'\Sigma_{11}^{-1/2}X^{(1)} = a_k'X^{(1)}, V_k = f_k'\Sigma_{22}^{-1/2}X^{(2)} = b_k'X^{(2)}$.

$$Cov(U_k, V_k) = \rho_k^*$$

$\rho_1^{*2} \geq \rho_2^{*2} \geq \cdots \geq \rho_p^{*2}$是$\Sigma_{11}^{-1/2}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1/2}$的特征值，$e_1, e_2, \cdots, e_k$是相应特征向量。$f_1, f_2, \cdots, f_k$是$\Sigma_{22}^{-1/2}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1/2}$的特征向量。

$$Cov(U) = Cov(V) = I$$
$$Cov(U, V) = diag(\rho_1^*, \cdots, \rho_p^*)$$

更多的相关关系：

$$
\begin{aligned}
\rho_{U,X^{(1)}} &= A\Sigma_{11}V_{11}^{-1/2} \\
\rho_{U,X^{(2)}} &= A\Sigma_{12}V_{22}^{-1/2} \\
\rho_{V,X^{(1)}} &= B\Sigma_{21}V_{11}^{-1/2} \\
\rho_{V,X^{(2)}} &= B\Sigma_{22}V_{21}^{-1/2}
\end{aligned}
$$

其中$A = [a_1, \cdots, a_p]', B = [b_1, \cdots, b_q]'$.

判别与分类

最小ECM法则：

$$
\begin{aligned}
R_1 &: \frac{f_1(x)}{f_2(x)} \geq \frac{c(1|2)p_2}{c(2|1)p_1} \\
R_2 &: \frac{f_1(x)}{f_2(x)} < \frac{c(1|2)p_2}{c(2|1)p_1}
\end{aligned}
$$

假设$\pi_1 : N(\mu_1, \Sigma_1), \pi_2 : N(\mu_2, \Sigma_2)$.

LDA

设$\Sigma_1 = \Sigma_2 = \Sigma$, ECM法则化简为

$$
R_1 : (\mu_1 - \mu_2)'\Sigma^{-1}x - \frac{1}{2}(\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 + \mu_2) \geq \ln \frac{c(1|2)p_2}{c(2|1)p_1}
$$

对于样本，Σ用S_{pooled}代替。

记$a' = (\bar{x}_1 - \bar{x}_2)'S_{pooled}^{-1}$, 当不等式右边为0时，

$$
R_1 : a'x \geq \frac{1}{2}a'(\bar{x}_1 + \bar{x}_2)
$$

更一般的：

$$
\begin{aligned}
a'x - \frac{1}{2}a'(\bar{x}_1 + \bar{x}_2) \geq \ln \frac{c(1|2)p_2}{c(2|1)p_1} \\
a'x - m \geq \ln \frac{c(1|2)p_2}{c(2|1)p_1}
\end{aligned}
$$

$y = a'x, m = \dfrac{\bar{y}_1 + \bar{y}_2}{2}$

QDA

$\Sigma_1 \neq \Sigma_2$

Result 11.4

The regions R_1 and R_2 that minimize the ECM are defined by the values x for which the following inequalities hold:

$$R_1: -\frac{1}{2}x'(\Sigma_1^{-1}-\Sigma_2^{-1})x+(\mu_1'\Sigma_1^{-1}-\mu_2'\Sigma_2^{-1})x-k \ge \ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right]$$

$$R_2: -\frac{1}{2}x'(\Sigma_1^{-1}-\Sigma_2^{-1})x+(\mu_1'\Sigma_1^{-1}-\mu_2'\Sigma_2^{-1})x-k < \ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right]$$

where $k=\frac{1}{2}\ln\left(\frac{|\Sigma_1|}{|\Sigma_2|}\right)+\frac{1}{2}(\mu_1'\Sigma_1^{-1}\mu_1-\mu_2'\Sigma_2^{-1}\mu_2)$.

评估指标

Total probability of misclassification

需要已知总体。

$$TPM = p_1P(2|1) + p_2P(1|2)$$

令$p_1 = p_2$，对于LDA，记$\Delta^2 = (\mu_1-\mu_2)'\Sigma^{-1}(\mu_1-\mu_2)$

$$TPM = \Phi(-\frac{\Delta}{2})$$

The Apparent Error Rate

$$APER = \frac{n_{1M}+n_{2M}}{n_1+n_2}$$

聚类分析

层次聚类

- single: can handle non-elliptical shapes, but is sensitive to noise and outliers.
- complete: more robust to noise and outliers, but tends to break large clusters.
- average
- ward

```
res <- hclust(dist(d), 'average')  # 'single', 'complete', 'ward.D'
plot(res)
cutree(res, k=3)
```

优势：不需要假设类的数量，易于展示。

缺点：复杂度高，$O(n^3)/O(n^2)$，各种方法都有缺点。

K-means

选择k，经验规则：$\sqrt{n}/2$.

$$SSE = \sum_{i=1}^{K}\sum_{x\in C_i} dist^2(m_i, x)$$

m_i是C_i类的中心。

```
library(cluster)
res <- pam(d, k = 2, medoids = c(1, 3))
memb <- res$clustering
```

优势：易于计算

劣势：对异常点敏感，难以处理非凹的聚类。

EM

假设$Y \sim Multinomial(1, \pi), X|Y \sim N(\mu_l, \Sigma_l)$.

选择k的依据

$$BIC = -2 \log L + m \log n$$

```
library(mclust)
res <- Mclust(d, 3)
memb <- res$classification
```