第六章 主成分分析

主成分分析的目的：

在尽量少损失信息的前提下，将多个指标转化为少数几个综合指标，进而达到对数据降维的效果.

主成份分析的思路：

选取变量的线性组合，使其方差最大化，尽量接近各分量方差之和，进而代表总体的散布程度.

例子：如何制定成年男子上衣服装号码的案例.

成年男子上衣的8个人体部位尺寸的均值与标准差

(样本量: 5115, 单位: cm)

部位	均值	标准差
身高	167.48	6.09
颈椎点高	142.91	5.60
腰围高	100.58	4.44
坐姿颈椎点高	65.61	2.67
颈围	36.83	2.11
胸围	87.53	5.55
后肩横弧	43.24	2.75
臂全长	54.53	3.04

目标：找出一个或几个指标来制定成年男子上衣的号型

成年男子上衣的8个人体部位尺寸的协方差阵

	身高	颈椎点高	腰围高	坐姿颈椎点高	颈围	胸围	后肩横弧	臂全长
身高	37.115							
颈椎点高	33.069	31.314						
腰围高	24.631	22.624	19.739					
坐姿颈椎点高	12.364	11.506	7.119	7.131				
颈围	2.695	2.593	1.217	1.575	4.437			
胸围	11.155	11.177	6.163	5.334	7.013	30.784		
后肩横弧	7.367	7.075	4.030	3.229	2.084	7.472	7.554	
臂全长	12.597	11.911	9.322	3.573	0.577	4.049	2.340	9.246

方差最大法： 取方差大的一个或几个变量作为代表

变量的方差越大, 其散布就越大. 散布越大, 以此对人群分类,

不同类之间的个体差别就越显著.

方差最大法选取的变量：**身高、颈椎点高、胸围、腰围高**

成年男子上衣的8个人体部位尺寸的相关阵

	身高	颈椎点高	腰围高	坐姿颈椎点高	颈围	胸围	后肩横弧	臂全长
身高	1.00							
颈椎点高	0.97	1.00						
腰围高	0.91	0.91	1.00					
坐姿颈椎点高	0.76	0.77	0.60	1.00				
颈围	0.21	0.22	0.13	0.28	1.00			
胸围	0.33	0.36	0.25	0.36	0.60	1.00		
后肩横弧	0.44	0.46	0.33	0.44	0.36	0.49	1.00	
臂全长	0.68	0.70	0.69	0.44	0.09	0.24	0.28	1.00

身高、颈椎点高和腰围高之间具有很强的相关性

合适的变量：身高和胸围

我国的标准：按成人男子身高制定上衣号型

由方差最大法选出的代表变量身高和胸围仍有相关性，
应该可以对它们进一步压缩，以选出更具代表性的指标.

问题：是否有更具代表性一个或少数指标?

代表性标准：方差最大

方法：主成分分析

记 X 是 p 维的随机向量, $p > 1$, $Cov(X) = \Sigma$.

我们希望, 基于 X 找到一个变量 Y, 使得其方差尽可能大, 足以代表 X 的散布.

一个自然选择是：变量 X 的线性组合, 即 $a'X$, $a \in R^p$.

由于 $Cov(X) = \Sigma$, $Var(a'X) = a'Cov(X)a = a'\Sigma a$,
若不对 a 加约束, 则 $a'X$ 的最大方差可以达无穷.

一个自然的约束:

正则化约束: $a'a = 1.$

优化问题: $\sup_{a'a=1} Var(a'X) = \sup_{a'a=1} a'\Sigma a.$

令

$$\mathbf{a}_1 = \text{argmax}_{a'a=1} a'\Sigma a,$$

则称 $\mathbf{a}_1'X$ 为 X 的**第一主成份**.

$\mathbf{a}_1'X$ 是正则化系数下方差最大的 X 的线性组合.

$\mathbf{a}_1'X$ 的散布程度最接近 X, 是代表 X 的首选.

令 Σ 的特征根为 $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$, 与这些特征根对应的正则正交特征向量为 $\alpha_1, \cdots, \alpha_p$.

易知 $\mathbf{a}_1 = \alpha_1 = (\alpha_{11}, \cdots, \alpha_{1p})'$, $Var(\mathbf{a}_1' X) = \mathbf{a}_1' \Sigma \mathbf{a}_1 = \lambda_1$.

则,

第一主成份方向是: 总体协差阵的最大特征根所对应的正则特征向量,

第一主成份的方差是: 总体协差阵的最大特征根.

通过成人男子8个人体部位尺寸的协方差阵知,

$$\begin{aligned} \lambda_1 &= 100.5771, \\ \mathbf{a}_1 &= (0.5920, 0.5469, 0.4052, 0.2062, 0.0638, 0.2680, 0.1416, 0.2183)'. \end{aligned}$$

第一主成份 $\approx 0.5 \times$ (身高+颈椎点高+腰围高).

成年男子上衣第一主成份

部位(x_i)	系数(α_{1i})
身高	0.5920
颈椎点高	0.5469
腰围高	0.4052
坐姿颈椎点高	0.2062
颈围	0.0638
胸围	0.2680
后肩横弧	0.1416
臂全长	0.2183

$$Var(\mathbf{a}_1'X) = \lambda_1 = 100.5771 > Var(x_1) = Var(\textbf{身高}) = 37.115.$$

第一主成份 $\mathbf{a}_1'X$ 的方差(散布程度)更大.

将其作为成年男子上衣的第一基本特征更具有代表性, 以此对人群进行划分将更细致.

国外确定服装号型：第一主成份.

设 p 维随机向量 $X = (x_1, \cdots, x_p)'$, 其**离散程度的信息**可用

$$Var(x_1) + \cdots + Var(x_p)$$

表示, 即向量各分量方差的总和.

第一主成份的作用:

将 X 所含离散程度的信息最大化地用一个其线性组合变量 $a'X$ 所含离散程度的信息来代替.

$$\textbf{第一主成份离散程度信息的贡献率} = \frac{Var(\mathbf{a}_1' X)}{\sum_{i=1}^{p} Var(x_i)} \times 100\%.$$

$$\text{成年男子上衣第一主成份的贡献率} = \frac{100.5771}{147.32} = 68.3\%.$$

问题：第一主成份代表性是否足够？或第一主成份贡献率是否足够?

寻找第二主成份: $\mathbf{a}_2'X$.

一个自然要求： 第二主成份应该与第一主成份正交,
从而不含有第一主成份的信息.

正则化约束： $a'a = 1;$

正交化约束： $a'\mathbf{a}_1 = 0;$

优化问题： $\sup_{\substack{a'a=1 \\ a'\mathbf{a}_1=0}} Var(a'X) = \sup_{\substack{a'a=1 \\ a'\mathbf{a}_1=0}} a'\Sigma a.$

不难知道, $\mathbf{a}_2 = \alpha_2 = (\alpha_{21}, \cdots, \alpha_{2p})'$, $Var(\mathbf{a}_2'X) = \mathbf{a}_2'\Sigma\mathbf{a}_2 = \lambda_2$.

即, 第二主成份方向是: 总体协差阵的第二大特征根所对应的
正则特征向量; 第二主成份的方差是: 总体协差阵的第二大特征根.

第一主成份与第二主成份的正交性：

$$Cov(\mathbf{a}_1'X, \mathbf{a}_2'X) = \mathbf{a}_1'\Sigma\mathbf{a}_2 = \lambda_2\alpha_1'\alpha_2 = 0.$$

因此, **正态总体下**, 第一主成份与第二主成份相互独立.

第二主成份离散程度信息的贡献率 $= \dfrac{Var(\mathbf{a}_2'X)}{\sum_{i=1}^{p} Var(x_i)} \times 100\%.$

通过成人男子8个人体部位尺寸的协方差阵知,

$$\begin{aligned} \lambda_2 &= 28.4471, \\ \mathbf{a}_2 &= (0.1849, 0.1362, 0.2028, -0.0083, -0.2320, -0.9003, -0.1867, 0.0831)'. \end{aligned}$$

第二主成份 $\approx$ 胸围.

成年男子上衣第二主成份

部位(x_i)	系数(α_{1i})
身高	0.1849
颈椎点高	0.1362
腰围高	0.2028
坐姿颈椎点高	-0.0083
颈围	-0.2320
胸围	-0.9003
后肩横弧	-0.1867
臂全长	0.0831

$$Var(\mathbf{a}_2'X) = \lambda_2 = 28.4471.$$

$$\text{成年男子上衣第二主成份的贡献率} = \frac{28.4471}{147.32} = 19.3\%.$$

第一、第二主成份的累计贡献率 $= \dfrac{Var(\mathbf{a}_1'X) + Var(\mathbf{a}_2'X)}{\sum_{i=1}^{p} Var(x_i)} \times 100\%.$

成年男子上衣第一、第二主成份的累计贡献率 $= \dfrac{129.0242}{147.32} = 87.6\%.$

问题：第一、第二主成份代表性是否足够？

停止？或类似地继续寻找更多的主成份？

6.1 总体主成分分析

设 $X \overset{d}{\sim} N_p(\mu, \Sigma)$. 令 Σ 的特征根为 $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$,
与这些特征根对应的正则正交特征向量为 $\alpha_1, \cdots, \alpha_p$.

令 $T = (\alpha_1, \cdots, \alpha_p)$, 则 T 是正交阵, 且

$$T'\Sigma T = \Lambda, \quad \Lambda = diag(\lambda_1, \cdots, \lambda_p).$$

(1) 令 $Y = T'X$, $Y = (y_1, \cdots, y_p)'$. 则称 Y 为 X 的**主成份**.

令 $\alpha_i = (\alpha_{1i}, \cdots, \alpha_{pi})'$, 则称 $y_i = \alpha_i' X = \sum_{j=1}^{p} \alpha_{ji} x_j$ 为 X 的

第 i 主成份, 其方差为 $Var(\alpha_i' X) = \alpha_i' \Sigma \alpha_i = \lambda_i$, $1 \leq i \leq p$.

(2) Y 的协方差阵为 $Cov(Y) = T'\Sigma T = \Lambda$. 因此有

(2.1) X 的第 i 个主成份的方差为 $Var(y_i) = \lambda_i,\ 1 \le i \le p.$

(2.2) 记 $\Sigma = (\sigma_{ij})_{p\times p}$, 则

$$\sum_{i=1}^{p} Var(y_i) = \sum_{i=1}^{p} \lambda_i = tr(\Sigma) = \sum_{i=1}^{p} \sigma_{ii} = \sum_{i=1}^{p} Var(x_i),$$

即 Y 与 X 具有相同的散布程度.

(2.3) 任意两个主成份都相互独立.

(3) 称 $\lambda_k / \sum_{i=1}^{p} \lambda_i$ 为第 k 个主成份 y_k 的贡献率,

它表示第 k 个主成份保留总体 X 散布程度信息的比例;

称 $\sum_{i=1}^{k} \lambda_i / \sum_{i=1}^{p} \lambda_i$ 是前 k 个主成份 $(y_1, \cdots, y_k)$ 的累计贡献率,

它表示前 k 个主成份保留总体散布程度信息的比例.

主成分与总体的相关性

记 Σ 的第 j 个行向量为 $(\sigma_{j1}, \cdots, \sigma_{jp})$, $1 \le j \le p$.

由于 $\Sigma\alpha_k = \lambda_k\alpha_k$, 因而有

$$\sum_{i=1}^{p} \sigma_{ji}\alpha_{ik} = \lambda_k\alpha_{jk},$$

$$\sum_{i=1}^{p} \sigma_{ij}\alpha_{ik} = \lambda_k\alpha_{jk}.$$

因此,

$$Cov(y_k, x_j) = Cov(\sum_{i=1}^{p} \alpha_{ik}x_i, x_j) = \sum_{i=1}^{p} \sigma_{ij}\alpha_{ik} = \lambda_k\alpha_{jk}.$$

X 的第 k 个主成份与 X 的第 j 个分量 x_j 的**相关系数**为

$$\rho_{y_k,\,x_j}=\frac{Cov(y_k,x_j)}{\sqrt{Var(y_k)}\sqrt{Var(x_j)}}=\frac{\lambda_k\alpha_{jk}}{\sqrt{\lambda_k}\sqrt{\sigma_{jj}}}=\frac{\alpha_{jk}\sqrt{\lambda_k}}{\sqrt{\sigma_{jj}}}.$$

称 $\rho_{y_k,\,x_j}$ 为第 k 个主成份 y_k 中变量 x_j 的因子负荷量.

主成份与 X 分量的复相关系数

令 $\rho_{Y,\,x_j}$ 为 Y 与 x_j 的**复相关系数**, $1\le j\le p$, 则

$$\rho^2_{Y,\,x_j}=\sum_{k=1}^{p}\rho^2_{y_k,\,x_j}=\frac{1}{\sigma_{jj}}\sum_{k=1}^{p}\lambda_k\alpha^2_{jk},\ 1\le j\le p.$$

由 $T'\Sigma T = \Lambda$, 知 $\Sigma = T\Lambda T'$, 即有 $\sigma_{jj} = \sum_{k=1}^{p} \lambda_k \alpha_{jk}^2$, $1 \le j \le p$.

则主成份 Y 与 X 的分量 x_j 的复相关系数 $\rho_{Y,x_j} = 1$, $1 \le j \le p$.

这说明主成份中含有分量 x_j 的离散程度的全部信息.

事实上, 有 $X = TY$, 即知

$$x_j = \sum_{k=1}^{p} \alpha_{jk} y_k, \quad 1 \le j \le p.$$

(4) 称 $\rho_{y_k,x_j}=\alpha_{jk}\sqrt{\lambda_k}/\sqrt{\sigma_{jj}}$ 为**第 k 个主成份 y_k 中变量 x_j 的因子负荷量**, 且 $\sum_{k=1}^{p}\rho^2_{y_k,x_j}=\sum_{k=1}^{p}\lambda_k\alpha^2_{jk}/\sigma_{jj}=1$.

(5) 称 $\rho^2_{y_k,x_j}=\lambda_k\alpha^2_{jk}/\sigma_{jj}$ 是**第 k 个主成份 y_k 的对于 X 的第 j 个分量 x_j 的贡献率**. 它表示第 k 个主成份保留 x_j 离散程度的信息的比例.

称 $\sum_{i=1}^{k}\rho^2_{y_i,x_j}=\sum_{i=1}^{k}\lambda_i\alpha^2_{ji}/\sigma_{jj}$ 是**前 k 个主成份 $(y_1,\cdots,y_k)$ 的对于 X 的第 j 个分量 x_j 的累计贡献率**. 它表示前 k 个主成份保留 x_j 离散程度的信息的比例.

部位(x_i)	第一主成份对各部位的贡献率($\lambda_1\alpha_{i1}^2/\sigma_{ii}$)
身高	95.04%
颈椎点高	95.93%
腰围高	83.77%
坐姿颈椎点高	59.99%
颈围	9.20%
胸围	23.45%
后肩横弧	26.67%
臂全长	51.86%

第一主成份对身高有关的部位贡献率很高, 但对与胸围有关的部位(颈围、胸围和后肩横狐)贡献率不大.

有必要提取第二主成份.

部位(x_i)	第二主成份对各部位的贡献率($\lambda_2\alpha_{i2}^2/\sigma_{ii}$)
身高	2.62%
颈椎点高	1.69%
腰围高	5.93%
坐姿颈椎点高	0.03%
颈围	34.51%
胸围	74.90%
后肩横弧	13.13%
臂全长	2.12%

第二主成份对胸围的贡献率很高.

第一和第二主成份对身高和胸围的累计贡献率分别达约98%,

故可以认为它们已具有足够代表性.

6.2 R 主成分分析

主成份分析主要是对随机向量的协方差矩阵进行分析, 将向量投影到方差大的方向以获得重要的主成份.

问题: 变量的量纲影响变量的方差. 有必要消除量纲对方差的影响.

方法: 对变量进行标准化处理. 即令

$$X^* = diag(\sigma_{11}^{-1/2}, \cdots, \sigma_{pp}^{-1/2})X = (\sigma_{11}^{-1/2}x_1, \cdots, \sigma_{pp}^{-1/2}x_p)',$$

则

$$Cov(X^*) = diag(\sigma_{11}^{-1/2}, \cdots, \sigma_{pp}^{-1/2})\Sigma\, diag(\sigma_{11}^{-1/2}, \cdots, \sigma_{pp}^{-1/2}) = R,$$

其中 R 是 X 的相关阵.

X^* 的主成份与量纲无关.

R 主成分分析: 处理量纲的主成份分析.

(1) 设 R 的特征根为 $\lambda_1^* \geq \cdots \geq \lambda_p^* \geq 0$, 与这些特征根对应的正则正交特征向量为 $\alpha_1^*, \cdots, \alpha_p^*$. 令 $T^* = (\alpha_1^*, \cdots, \alpha_p^*)$,

$$Y^* = (T^*)' X^* = (T^*)' (\sigma_{11}^{-1/2} x_1, \cdots, \sigma_{pp}^{-1/2} x_p)',$$

则称 Y^* 为 X 的**R主成份**.

令 $Y^* = (y_1^*, \cdots, y_p^*)'$, $\alpha_i^* = (\alpha_{1i}^*, \cdots, \alpha_{pi}^*)'$, 则称

$$y_i^* = \alpha_i' X^* = \sum_{j=1}^{p} \alpha_{ji}^* \sigma_{jj}^{-1/2} x_j$$

为 X 的第 i **个R主成份**, $1 \leq i \leq p$.

(2) Y^* 的协方差阵为 $Cov(Y^*) = \Lambda^* = diag(\lambda_1^*, \cdots, \lambda_p^*)$, $\sum_{i=1}^{p} \lambda_i^* = p$.

(3) 称 λ_k^*/p 为第 k 个**R**主成份 y_k^* 的贡献率,

$\sum_{i=1}^{k} \lambda_i^*/p$ 为前 k 个**R**主成份 $(y_1^*, \cdots, y_k^*)$ 的累计贡献率.

(4) 称 $\alpha_{jk}^* \sqrt{\lambda_k}$ 为第 k 个**R**主成份 y_k^* 中变量 x_j 的因子负荷量,

$$\sum_{k=1}^{p} \lambda_k^* (\alpha_{jk}^*)^2 = 1, \quad 1 \le j \le p.$$

(5) 称 $\sum_{i=1}^{k} \lambda_k^* (\alpha_{jk}^*)^2$ 为前 k 个**R**主成份 $(y_1^*, \cdots, y_k^*)$ 的对于 X 的第 j 个分量 x_j 的累计贡献率.

6.3 样本主成分分析

样本主成份分析: 基于观测数据的主成份分析.

假设总体 $X \overset{d}{\sim} N_p(\mu, \Sigma)$, 其观测样本为 $x_1, \cdots, x_n$.
则 (μ, Σ) 的极大似然估计为

$$\hat{\mu} = \bar{x} = n^{-1} \sum_{i=1}^{n} x_i,$$
$$\hat{\Sigma} = S = n^{-1} \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})'.$$

样本主成份分析也就是基于样本协方差阵 S 的主成份分析.
它也等价于某个分布下的总体主成份分析.

记 S 的特征根分别为 $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_p \geq 0$, 与这些特征根对应的正则正交特征向量为 $\hat{\alpha}_1, \cdots, \hat{\alpha}_p$.

令 $\hat{Y} = \hat{T}'X$, $\hat{Y} = (\hat{y}_1, \cdots, \hat{y}_p)'$, 其中 $\hat{T} = (\hat{\alpha}_1, \cdots, \hat{\alpha}_p)$, 则称 $\hat{Y}$ 为 X 的**样本主成份**.

记 $\hat{\alpha}_k = (\hat{\alpha}_{1k}, \cdots, \hat{\alpha}_{pk})'$, 则称 $\hat{y}_k = \hat{\alpha}_k' X = \sum_{i=1}^{p} \hat{\alpha}_{ik} x_i$ 为 X 的**第 k 样本主成份**, $1 \leq k \leq p$.

$\hat{y}_k = \hat{\alpha}_k' X$, $\hat{\alpha}_k$ 和 $\hat{\lambda}_k$ 分别是 X 的第 k 主成份 $y_k = \alpha_k' X$, 第 k 主成分系数 α_k 和第 k 主成份的方差 λ_k 的极大似然估计, $1 \leq k \leq p$.

相应地, 可以得到主成份对**总体的贡献率**、对总体分量的**因子负荷量**以及**总体分量的贡献率**的极大似然估计.

经验总体下的总体主成分分析

定义随机向量 X^*, 它服从离散分布, 分布函数为

$$P(X^* = x_i) = \frac{1}{n}, \quad 1 \le i \le n.$$

则 X^* 的分布就是样本 $x_1, \cdots, x_n$ 的经验分布.

显然有,

$$\begin{aligned} E(X^*) &= n^{-1} \sum_{i=1}^{n} x_i = \bar{x}, \\ Cov(X^*) &= E[(X^* - E(X^*))(X^* - E(X^*)'] \\ &= n^{-1} \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})' = S. \end{aligned}$$

(1) 求 X^* 的第一主成份:

有 $\hat{\alpha}_1 = argmax_{\alpha'\alpha=1} Var(a'X^*) = argmax_{\alpha'\alpha=1} a'Sa,$

则 $\hat{\alpha}_1 X^*$ 是 X^* 的第一主成份.

(2) 求 X^* 的第二主成份:

有 $\hat{\alpha}_2 = argmax_{\substack{\alpha'\alpha=1 \\ \alpha'\hat{\alpha}_1=0}} Var(a'X^*) = argmax_{\substack{\alpha'\alpha=1 \\ \alpha'\hat{\alpha}_1=0}} a'Sa,$

则 $\hat{\alpha}_2 X^*$ 是 X^* 的第二主成份.

(3) 依次求解 X^* 的第三到第 p 主成份.

因此, X^* 的主成份系数与 X 的样本主成份系数是一致的.

且 $Var(\hat{\alpha}_i' X^*) = \hat{\lambda}_i,\ 1 \le i \le p.$

6.4 样本 R 主成分分析

基于样本相关阵的主成份分析就是样本 R 主成份分析.

记 $\hat{R} = diag(s_{11}^{-1/2}, \cdots, s_{pp}^{-1/2}) S \, diag(s_{11}^{-1/2}, \cdots, s_{pp}^{-1/2})$,

即 $\hat{R}$ 是样本相关阵. 基于 $\hat{R}$ 进行主成份分析即可.

此外, 令

$$x_i^* = diag(s_{11}^{-1/2}, \cdots, s_{pp}^{-1/2})(x_i - \bar{x}), \quad 1 \le i \le n,$$

那么 $x_1^*, \cdots, x_n^*$ 的样本协差阵也是 $x_1, \cdots, x_n$ 的样本相关阵 $\hat{R}$.

则对 $x_1^*, \cdots, x_n^*$ 进行主成份分析即是样本 R 主成份分析.

注意：

(1) (总体)主成份分析与R主成份分析的结论可能不一致.

(2) 样本主成份分析与样本R主成份分析的结论可能不一致.

6.5 主成分的统计推断

对实际数据进行的主成份分析时, 事先会设定一个主成份贡献率的阈值 $(1-\delta)$.

得到样本的主成份后, 可以计算前 k 个样本主成份的贡献率

$$\sum_{i=1}^{k} \hat{\lambda}_i / \sum_{i=1}^{p} \hat{\lambda}_i.$$

如果 $\sum_{i=1}^{k} \hat{\lambda}_i / \sum_{i=1}^{p} \hat{\lambda}_i > (1-\delta)$, 是否就可以认为

$$\sum_{i=1}^{k} \lambda_i / \sum_{i=1}^{p} \lambda_i > (1-\delta)?$$

需要对协差阵的特征根 $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$ 进行统计推断.

首先假定 $\Sigma > 0$, 则参数 (μ, Σ) 的似然函数为

$$\frac{1}{|\Sigma|^{n/2}} \exp\left\{-\frac{1}{2} tr[\Sigma^{-1}(V + n(\bar{x}-\mu)(\bar{x}-\mu)')]\right\},$$

由于 $\Sigma = T\Lambda T'$, 即 $(\lambda_1, \cdots, \lambda_p)$ 和 $(\alpha_1, \cdots, \alpha_p)$ 仅与 Σ 有关, 其似然函数为

$$\begin{aligned} L(\lambda_1, \cdots, \lambda_p, \alpha_1, \cdots, \alpha_p) &= \frac{1}{|\Sigma|^{n/2}} \exp\left\{-\frac{1}{2} tr(\Sigma^{-1} V)\right\} \\ &= \frac{1}{|T\Lambda T'|^{n/2}} \exp\left\{-\frac{1}{2} tr(T\Lambda^{-1} T' V)\right\} \\ &= (\prod_{i=1}^{p} \lambda_i)^{-n/2} \exp\left\{-\frac{1}{2}\left(\sum_{i=1}^{p} \frac{\alpha_i' V \alpha_i}{\lambda_i}\right)\right\}. \end{aligned}$$

为简单起见, 再假定 $\lambda_1 > \cdots > \lambda_p > 0$, 即所有特征根都不等.
此时 $\lambda_1, \cdots, \lambda_p$ 与 $\alpha_1, \cdots, \alpha_p$ 无关.

因为由 Σ 的任意性, 在给定 $\lambda_1, \cdots, \lambda_p$ 下,
正交矩阵 $T = (\alpha_1, \cdots, \alpha_p)$ 也是任意的.

事实上, 考虑参数的自由度: 在 $\lambda_1 > \cdots > \lambda_p > 0$ 下

$$dim(\Lambda) = p, \quad dim(T) = p^2 - p - \frac{p(p-1)}{2},$$

$$dim(\Lambda) + dim(T) = \frac{p(p+1)}{2} = dim(\Sigma).$$

Fisher信息阵与极大似然估计的渐近正态性

假设 $x_1, \cdots, x_n$ 是服从密度函数为 $p(x, \theta)$ 的独立样本.

记 $\hat{\theta}$ 为 θ 的极大似然估计, $X^{(n)} = (x_1, \cdots, x_n)$. 对数似然函数为

$$l(\theta | X^{(n)}) = \sum_{i=1}^{n} \log p(x_i, \theta).$$

则**Fisher信息阵**为

$$\begin{aligned} I_n(\theta) &= Var_\theta \left[\frac{\partial l(\theta | X^{(n)})}{\partial \theta} \right] \quad (\text{一般性的定义}) \\ &= -E_\theta \left[\frac{\partial^2 l(\theta | X^{(n)})}{\partial \theta^2} \right] \\ &= -nE_\theta \left[\frac{\partial^2}{\partial \theta^2} \log p(x_1, \theta) \right] \quad (\text{独立同分布下}) \\ &\triangleq nI(\theta). \end{aligned}$$

$\hat{\theta}$ **的渐近正态性**(一般情形):

$$(I_n(\theta))^{1/2}(\hat{\theta}-\theta) \xrightarrow{d} N(0, I_p) \ (n \to \infty).$$

在独立同分布情形下, 有

$$\sqrt{n}(\hat{\theta}-\theta) \xrightarrow{d} N(0, I^{-1}(\theta)) \ (n \to \infty).$$

考虑 $(\hat{\lambda}_1, \cdots, \hat{\lambda}_p)$ 的渐近分布

有对数似然函数

$$\begin{aligned} l(\lambda_1, \cdots, \lambda_p, \alpha_1, \cdots, \alpha_p) &= \sum_{i=1}^{p} \left(-\frac{n}{2} \log \lambda_i - \frac{\alpha_i' V \alpha_i}{2\lambda_i} \right) \\ &\triangleq \sum_{i=1}^{p} l(\lambda_i, \alpha_i). \end{aligned}$$

因此对任意的 $i \neq j$, 有

$$\frac{\partial^2 l(\theta | X^{(n)})}{\partial \theta^2} = \frac{\partial^2 l(\lambda_1, \cdots, \lambda_p, \alpha_1, \cdots, \alpha_p)}{\partial \lambda_i \partial \lambda_j} = 0.$$

那么由Fisher信息阵的结构, 知 $\hat{\lambda}_1, \cdots, \hat{\lambda}_p$ 的极限分布是相互独立的正态分布.

下面计算 $\hat{\lambda}_1, \cdots, \hat{\lambda}_p$ 的渐近方差.

由于 $V \overset{d}{\sim} W_p(n-1, \Sigma)$, 等价地有

$$V \overset{d}{=} \sum_{k=1}^{n-1} Z_k Z_k',$$

其中, $Z_1, \cdots, Z_{n-1}$ 是 $i.i.d.$ 的正态 $N_p(0, \Sigma)$ 随机向量.

因此, 对 $1 \leq i \leq p$, 有

$$\alpha_i' V \alpha_i \overset{d}{=} \sum_{k=1}^{n-1} (\alpha_i' Z_k)^2.$$

由于 $\alpha_i' \Sigma \alpha_i = \lambda_i$, 知 $\alpha_i' Z_1, \cdots, \alpha_i' Z_{n-1}$ 是独立同分布的 $N_1(0, \lambda_i)$ 随机变量. 因此

$$\alpha_i' V \alpha_i \overset{d}{\sim} \lambda_i \chi^2(n-1).$$

计算 $\hat{\lambda}_i$ 的Fisher信息, 为:

$$
\begin{aligned}
-E_{(\lambda_i,\alpha_i)}\left[\frac{\partial^2 l(\lambda_i)}{\partial \lambda_i^2}\right] &= -E_{(\lambda_i,\alpha_i)}\left[\frac{n}{2\lambda_i^2} - \frac{\alpha_i' V \alpha_i}{\lambda_i^3}\right] \\
&= -\frac{n}{2\lambda_i^2} + \frac{(n-1)\lambda_i}{\lambda_i^3} \\
&= \frac{n-2}{2\lambda_i^2}.
\end{aligned}
$$

则 $(\hat{\lambda}_1, \cdots, \hat{\lambda}_p)$ 的Fisher信息阵为

$$I_n = diag(\frac{n-2}{2\lambda_1^2}, \cdots, \frac{n-2}{2\lambda_p^2}).$$

由极大似然估计的渐近正态性知

$$I_n^{1/2} \begin{pmatrix} \hat{\lambda}_1 - \lambda_1 \\ \vdots \\ \hat{\lambda}_n - \lambda_n \end{pmatrix} \xrightarrow{d} N(0, I_p) \ (n \to \infty),$$

即当样本量 n 趋于无穷大时, 有

$$\sqrt{n-2} \begin{pmatrix} \hat{\lambda}_1 - \lambda_1 \\ \vdots \\ \hat{\lambda}_n - \lambda_n \end{pmatrix} \xrightarrow{d} N\left(0, diag(2\lambda_1^2, \cdots, 2\lambda_p^2)\right).$$

当 Σ 的特征根有重根时, 情况比较复杂.

由极大似然估计的渐近正态性可以构造 λ_i 的渐近置信区间

$$\hat{\lambda}_i \left[1+\sqrt{2/(n-2)}Z_{1-\beta/2}\right]^{-1} \leq \lambda_i \leq \hat{\lambda}_i \left[1-\sqrt{2/(n-2)}Z_{1-\beta/2}\right]^{-1},$$

也可通过方差齐性变换, 导出

$$\sqrt{n-2}\left(\ln(\hat{\lambda}_i^{\sqrt{2}/2}) - \ln(\lambda_i^{\sqrt{2}/2})\right) \xrightarrow{d} N(0,1),$$

可得 λ_i 的另一个置信水平为 $(1-\beta)$ 的渐近置信区间

$$\hat{\lambda}_i \exp\left\{-\frac{2}{n-2}Z_{1-\beta/2}\right\} \leq \lambda_i \leq \hat{\lambda}_i \exp\left\{\frac{2}{n-2}Z_{1-\beta/2}\right\}.$$

6.5.1 与主成份分析有关的检验问题

检验问题 I：

$$\mathbf{H}_0 : \lambda_{k+1} + \cdots + \lambda_p \leq \gamma.$$

检验统计量的构造：由 $(\hat{\lambda}_1, \cdots, \hat{\lambda}_p)$ 的渐近正态性, 有

$$\sqrt{n-2}\left(\sum_{i=k+1}^{p} \hat{\lambda}_i - \sum_{i=k+1}^{p} \lambda_i\right) \xrightarrow{d} N(0, \sum_{i=k+1}^{p} 2\lambda_i^2),$$

进而可得

$$\frac{\sqrt{n-2}\left(\sum_{i=k+1}^{p} \hat{\lambda}_i - \sum_{i=k+1}^{p} \lambda_i\right)}{\sqrt{\sum_{i=k+1}^{p} 2\hat{\lambda}_i^2}} \xrightarrow{d} N(0,1).$$

则当

$$\sum_{i=k+1}^{p} \hat{\lambda}_i > \gamma + \frac{\sqrt{\sum_{i=k+1}^{p} 2\hat{\lambda}_i^2}}{\sqrt{n-2}} Z_{1-\alpha}$$

时拒绝零假设, 它犯第一类错误的概率渐近不超过 α.

检验问题 II： 前 k 个主成份的累计贡献率是否大于给定的值 δ?

$$\mathbf{H}_0: \quad \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{p} \lambda_i} \leq \delta.$$

考虑如下的统计量的渐近分布

$$\sqrt{n-2}\left(\frac{\sum_{i=1}^{k}\hat{\lambda}_i}{\sum_{i=1}^{p}\hat{\lambda}_i}-\frac{\sum_{i=1}^{k}\lambda_i}{\sum_{i=1}^{p}\lambda_i}\right).$$

定义如下的函数

$$f(\lambda_1,\cdots,\lambda_p)=\sum_{i=1}^{k}\lambda_i/\sum_{i=1}^{p}\lambda_i,$$

那么由**Cramér**定理有

$$\sqrt{n-2}\left(f(\hat{\lambda}_1,\cdots,\hat{\lambda}_p)-f(\lambda_1,\cdots,\lambda_p)\right)\xrightarrow{d}N(0,v^2),$$

其中

$$v^2 = \frac{2\left[(\sum_{i=k+1}^{p}\lambda_i)^2(\sum_{i=1}^{k}\lambda_i^2) + (\sum_{i=1}^{k}\lambda_i)^2(\sum_{i=k+1}^{p}\lambda_i^2)\right]}{(\sum_{i=1}^{p}\lambda_i)^4}.$$

事实上, 若记 $\lambda = (\lambda_1, \cdots, \lambda_p)'$, 则有

$$\frac{\partial f(\lambda)}{\partial \lambda_i} = \frac{I(1 \le i \le k)}{\sum_{j=1}^{p}\lambda_j} - \frac{\sum_{j=1}^{k}\lambda_j}{(\sum_{j=1}^{p}\lambda_j)^2}, \quad 1 \le i \le p.$$

因此

$$\begin{aligned} v^2 &= \left(\frac{\partial f(\lambda)}{\partial \lambda}\right)' diag(2\lambda_1^2, \cdots, 2\lambda_p^2)\frac{\partial f(\lambda)}{\partial \lambda} \\ &= 2\sum_{i=1}^{p}\lambda_i^2\left(\frac{I(1 \le i \le k)}{\sum_{j=1}^{p}\lambda_j} - \frac{\sum_{j=1}^{k}\lambda_j}{(\sum_{j=1}^{p}\lambda_j)^2}\right)^2. \end{aligned}$$

将极大似然估计 $\hat{\lambda}_1, \cdots, \hat{\lambda}_p$ 代入 v^2 即得估计 $\hat{v}^2$,

$$\hat{v}^2 = \frac{2\left[(\sum_{i=k+1}^{p}\hat{\lambda}_i)^2(\sum_{i=1}^{k}\hat{\lambda}_i^2) + (\sum_{i=1}^{k}\hat{\lambda}_i)^2(\sum_{i=k+1}^{p}\hat{\lambda}_i^2)\right]}{(\sum_{i=1}^{p}\hat{\lambda}_i)^4}.$$

因此有

$$\frac{\sqrt{n-2}}{\hat{v}}\left(\frac{\sum_{i=1}^{k}\hat{\lambda}_i}{\sum_{i=1}^{p}\hat{\lambda}_i} - \frac{\sum_{i=1}^{k}\lambda_i}{\sum_{i=1}^{p}\lambda_i}\right) \xrightarrow{d} N(0,1).$$

故检验问题II的解是：

$$\text{当} \quad \frac{\sum_{i=1}^{k} \hat{\lambda}_i}{\sum_{i=1}^{p} \hat{\lambda}_i} \geq \delta + \frac{\hat{v}}{\sqrt{n-2}} Z_{1-\alpha} \text{ 时},$$

拒绝零假设, 即认为 $\dfrac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{p} \lambda_i} > \delta,$

它犯第一类错误的概率渐近不超过 α.

例子：成年男子上衣服装号码的案例续(统计检验)

(1) 计算得样本协差阵的特征根从大到小依次为

$$\hat{\lambda}_1 = 100.5771, \quad \hat{\lambda}_2 = 28.4471, \quad \hat{\lambda}_3 = 5.7489, \quad \hat{\lambda}_4 = 4.4522,$$
$$\hat{\lambda}_5 = 3.1978, \quad \hat{\lambda}_6 = 2.5854, \quad \hat{\lambda}_7 = 1.3834, \quad \hat{\lambda}_8 = 0.9280.$$

(2) 设定累计贡献率的阈值 $\delta = 0.85$,
显著性水平设定为 $\alpha = 0.05$.

(3) 由于

$$\frac{\sum_{i=1}^{2}\hat{\lambda}_i}{\sum_{i=1}^{8}\hat{\lambda}_i} = 87.6\%,$$

我们把零假设设定为:

$$\mathbf{H}_0: \quad \frac{\sum_{i=1}^{2}\lambda_i}{\sum_{i=1}^{8}\lambda_i} \leq \delta = 0.85.$$

(4) 计算

$$\begin{aligned}\hat{v}^2 &= \frac{2\left[(\sum_{i=3}^{8}\hat{\lambda}_i)^2(\sum_{i=1}^{2}\hat{\lambda}_i^2)+(\sum_{i=1}^{2}\hat{\lambda}_i)^2(\sum_{i=3}^{8}\hat{\lambda}_i^2)\right]}{(\sum_{i=1}^{8}\hat{\lambda}_i)^4}\\ &= 0.0207.\end{aligned}$$

(5) 计算检验临界值, 其中 $n = 5115$, $Z_{1-\alpha} = Z_{0.95} = 1.6449$,

$$C_r = \delta + \frac{\hat{v}}{\sqrt{n-2}} Z_{1-\alpha} = 0.8533,$$

(6) 由于

$$\frac{\sum_{i=1}^{2}\hat{\lambda}_i}{\sum_{i=1}^{8}\hat{\lambda}_i} = 0.876 > C_r = 0.8533,$$

故拒绝零假设, 即认为

$$\frac{\sum_{i=1}^{2}\lambda_i}{\sum_{i=1}^{8}\lambda_i} > 0.85,$$

亦两个主成分已满足代表原总体散度的要求.

6.5.2 R主成份分析的检验

由于在R主成份分析中, 样本相关阵的特征根 $\hat{\lambda}_1^*, \cdots, \hat{\lambda}_p^*$ 要满足约束条件 $\sum_{i=1}^{p} \hat{\lambda}_i^* = 1$. 因此, $\hat{\lambda}_1^*, \cdots, \hat{\lambda}_p^*$ 不再是渐近独立的.

此外, $(\lambda_1^*, \cdots, \lambda_p^*)$ 与 $(\alpha_1^*, \cdots, \alpha_p^*)$ 不再是无关的, 因此

有关主成份分析的渐近理论对R主成份分析不再成立.