一. 数据的组织

1.1 阵列

对于 n 个样本，选择 p 个特征进行记录，以

$$x_{jk} = \text{第}k\text{个变量的第}j\text{项的观测值}$$

那么这 n 个观测值可表示为：

$$\begin{array}{ccccc}
x_{11} & x_{12} & \cdots & x_{1k} & \cdots & x_{1p} \\
x_{21} & x_{22} & \cdots & x_{2k} & \cdots & x_{2p} \\
\vdots & \vdots & & \vdots & & \vdots \\
x_{j1} & x_{j2} & \cdots & x_{jk} & \cdots & x_{jp} \\
\vdots & \vdots & & \vdots & & \vdots \\
x_{n1} & x_{n2} & \cdots & x_{nk} & \cdots & x_{np}
\end{array}$$

使用矩阵存储：

$$X = \begin{bmatrix}
x_{11} & x_{12} & \cdots & x_{1k} & \cdots & x_{1p} \\
x_{21} & x_{22} & \cdots & x_{2k} & \cdots & x_{2p} \\
\vdots & \vdots & & \vdots & & \vdots \\
x_{j1} & x_{j2} & \cdots & x_{jk} & \cdots & x_{jp} \\
\vdots & \vdots & & \vdots & & \vdots \\
x_{n1} & x_{n2} & \cdots & x_{nk} & \cdots & x_{np}
\end{bmatrix}$$

1.2 描述统计量

(1) 样本均值：在特征方向上对样本进行平均

$$\bar{x}_k = \frac{1}{n}\sum_{j=1}^{n} x_{jk} \qquad (k=1,2,\cdots,p)$$

(2) 样本方差：类似地每个特征的方差为

$$s_k^2 = \frac{1}{n}\sum_{j=1}^{n}(x_{jk}-\bar{x}_k)^2 \qquad (k=1,2,\cdots,p)$$

一般我们也记 $s_k^2 := s_{kk}$

(3) 样本协方差：任意2个特征之间的协方差

$$s_{ik} = \frac{1}{n}\sum_{j=1}^{n}(x_{ji}-\bar{x}_i)(x_{jk}-\bar{x}_k) \qquad i,k \in \{1,2,\cdots,p\}$$

(4) 样本相关系数：对样本协方差进行归一化

$$r_{ik} = \frac{s_{ik}}{\sqrt{s_{ii}}\sqrt{s_{kk}}} \qquad i,k \in \{1,2,\cdots,p\}$$

下面采用矩阵方式：

样本均值向量 $$\bar{x} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{bmatrix}$$

样本协方差矩阵 $$S = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{bmatrix}$$

样本相关系数矩阵 $$R = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{bmatrix}$$

二、统计量计算

对于 $\bar{x}$, S, R 可以通过样本观测值 X 求得：

(1) $\bar{x} = \frac{1}{n} X^T \cdot \vec{1}$

(2) $S = \frac{1}{n-1} X^T \left(I - \frac{1}{n} \cdot \vec{1} \cdot \vec{1}^T\right) X$ （S_n 为 $\frac{1}{n}$，S 为 $\frac{1}{n-1}$）

(3) $R = D^{-\frac{1}{2}} \cdot S \cdot D^{-\frac{1}{2}}$

其中：I 为对角单位阵，$D^{\frac{1}{2}} = \mathrm{diag}(\sqrt{\sigma_{11}}, \cdots \sqrt{\sigma_{pp}})$

Proof (1) $X^T \vec{1} = \begin{bmatrix} x_{11} & \cdots & x_{n1} \\ \vdots & & \vdots \\ x_{1p} & \cdots & x_{np} \end{bmatrix}_{p\times n} \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}_{n\times 1} = \begin{bmatrix} \sum_{i=1}^{n} x_{i1} \\ \vdots \\ \sum_{i=1}^{n} x_{ip} \end{bmatrix}_{p\times 1}$

故 $\bar{x} = \frac{1}{n} X^T \vec{1} \in \mathbb{R}^p$

(2) $\vec{1} \cdot \bar{x}^T = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}_{n\times 1} [\bar{x}_1, \cdots, \bar{x}_p]_{1\times p} = \begin{bmatrix} \bar{x}_1 & \bar{x}_2 & \cdots & \bar{x}_p \\ \vdots & \vdots & & \vdots \\ \bar{x}_1 & \bar{x}_2 & \cdots & \bar{x}_p \end{bmatrix}_{n\times p}$

故 $X - \vec{1} \cdot \bar{x}^T = \{x_{jk} - \bar{x}_k\}_{n\times p} \quad j \in \{1, 2, \cdots, n\} \quad k \in \{1, \cdots, p\}$

$\Rightarrow S_n = \frac{1}{n} (X - \vec{1} \cdot \bar{x}^T)^T (X - \vec{1} \cdot \bar{x}^T)$

$= \frac{1}{n} (X - \frac{1}{n} \vec{1} \vec{1}^T X)^T (X - \frac{1}{n} \vec{1} \vec{1}^T X) \quad \because$ 代入(1)

$= \frac{1}{n} (X^T - \frac{1}{n} X^T \vec{1} \vec{1}^T)(X - \frac{1}{n} \vec{1} \vec{1}^T X)$

$= \frac{1}{n} X^T (I - \frac{1}{n} \vec{1} \vec{1}^T)(I - \frac{1}{n} \vec{1} \vec{1}^T) X$

$= \frac{1}{n} X^T (I - \frac{1}{n} \vec{1} \vec{1}^T) X \quad \square$

最后一步因为：

$(I - \frac{1}{n} \vec{1} \vec{1}^T)(I - \frac{1}{n} \vec{1} \vec{1}^T) = I - \frac{2}{n} \vec{1} \vec{1}^T + \frac{1}{n^2} \vec{1} \vec{1}^T \vec{1} \vec{1}^T$

注意到 $\vec{1}^T \vec{1} = n$ 故上式 $= I - \frac{1}{n} \vec{1} \vec{1}^T \quad \square$

(3) $r_{ik} = \frac{S_{ik}}{\sqrt{S_{ii}} \sqrt{S_{kk}}}$ 类似二次型，易证