

《多元统计分析》试卷 (2023)

1. 法国 Decastar 巡回赛是世界上最大的国际田联十项全能 (decathlon) 赛事，比赛次序如下：

第一天：100 米，跳远，铅球，跳高，400 米；

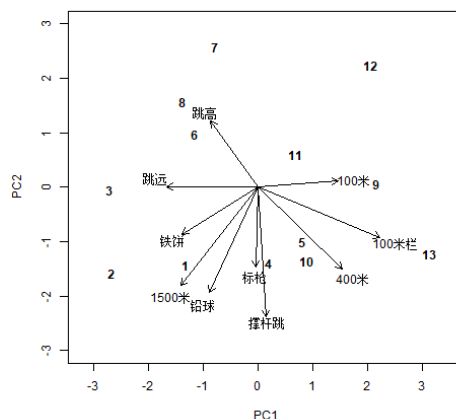
第二天：100 米栏，铁饼，撑杆跳，标枪，1500 米。

其中径赛跑步类以时间度量成绩（单位：秒，时间值越小越好），田赛包括跳跃和投掷项目，以高度或距离度量成绩（单位：米，值越大越好）。对 2004 年该赛事前 13 名运动员 10 项成绩的**相关系数矩阵**做主成分分析，前两个主成分 PC1 和 PC2 的载荷（即主成分方向）如下表所示

	100 米	跳远	铅球	跳高	400 米	100 米栏	铁饼	撑杆跳	标枪	1500 米
PC1	0.48	-0.37	-0.20	-0.19	0.34	0.50	-0.31	0.03	-0.01	-0.31
PC2	0.03	0.00	-0.43	0.27	-0.33	-0.20	-0.19	-0.53	-0.33	-0.40

请回答如下问题：

- (a) 根据上述载荷，解释第一主成分 PC1 的含义。
- (b) 已知前两个主成分 PC1 和 PC2 的标准差分别为 1.76, 1.42, 计算它们的累计方差贡献率。
- (c) 分析运动员的 PC 散点图（右图，数字代表运动员，大小代表名次），第一名（数字 1）的投掷类成绩如何（好、中、差）？他的 4 个径赛项目表现各如何（好、中、差）？最后一名（数字 13）有什么特点？
- (d) 跳高和撑杆跳是正相关还是负相关？图中跳远和 100 米方向相反说明了什么？1500 米作为径赛项目与其它径赛项目类似吗？简单解释原因。



2. 国际天文学联合会于 2006 年将冥王星 (Pluto) 移除出了太阳系九大行星之列，这主要是因为此前冥王星的质量被错误地高估了。冥王星所属的柯伊伯带 (Kuiper) 在太阳系的外围，聚集了大量小行星和彗星，有观点认为即使在该区域内冥王星在若干指标上也不是很突出，冥王星的降级在某种意义上说明了星体类型的定义不是一件容易的事情。一项研究收集了柯伊伯带 10 个较大行星 (包含冥王星) 的 5 个指标：

Brightness: 校正距离后的亮度； Albedo: 光反射率； Distance: 轨道与太阳之间最远的距离； Diameter: 星体直径； Year: 发现年份（年份越早代表越容易被发现。冥王星被发现于 1930 年，其它行星在 2003 年左右被发现）。

假设两因子模型，基于**相关系数矩阵**得到因子载荷的极大似然估计如下：

变量	Brightness	Albedo	Diameter	Distance	Year
因子 F_1	0.99	0.97	0.81	0.12	-0.39
因子 F_2	0.00	0.12	-0.58	0.42	0.87

- (a) 请解释两个因子的含义。
- (b) 试计算各个变量的特殊方差。两因子模型对哪个或哪些变量拟合较差？
- (c) 总方差等于多少？试计算第因子 F_1, F_2 分别解释总方差的比例及两者的累计方差解释比例。

3. (聚合层次聚类方法在初始步骤将每个物件看作一个类, 然后不断合并距离最小的类, 其中两个类的单连结距离 (single-linkage) 定义为两个类元素之间距离的最小值。现假设 5 个物件 $a-e$ 的距离矩阵如下

$$D = \begin{matrix} & \begin{matrix} a & b & c & d & e \end{matrix} \\ \begin{matrix} a \\ b \\ c \\ d \\ e \end{matrix} & \begin{pmatrix} 0 & 1 & 3 & 4 & 4 \\ 1 & 0 & 2 & 3 & 5 \\ 3 & 2 & 0 & 3 & 5 \\ 4 & 3 & 3 & 0 & 2 \\ 4 & 5 & 5 & 2 & 0 \end{pmatrix} \end{matrix}$$

试应用单连结聚合层次聚类方法进行聚类, 画出树图。

4. 假设 X 是 $n \times p$ 矩阵 (行为样本, 列为变量), 假设 $\mathbf{u} \in R^n$ 为 n 个样本的某种重要性计分, $\mathbf{v} \in R^p$ 为 p 个变量的重要性计分, $\|\mathbf{u}\| = \|\mathbf{v}\| = 1$ 。假设存在非 0 常数 c_1, c_2 使得

$$X\mathbf{v} = c_1\mathbf{u}, X^T\mathbf{u} = c_2\mathbf{v},$$

证明 $c_1 = c_2$, 其最大值为 X 的最大奇异值, 并说明何时达到最大。

5. 假设 $W = (w_{ij})$ 是元素非负的 $n \times p$ 矩阵, $m = n + p$, $r = \text{rank}(W)$ 。构造对称矩阵 A 如下

$$A_{m \times m} = \begin{pmatrix} 0 & W \\ W^T & 0 \end{pmatrix}.$$

- (a) 假设 W 有奇异值分解 $W = U\Gamma V^T$, 其中 $\Gamma = \text{diag}(\gamma_1, \dots, \gamma_r)$ 是 $r \times r$ 对角矩阵, $\gamma_1 \geq \dots \geq \gamma_r > 0$ 为 W 的奇异值, U, V 分别是 $n \times r$ 和 $p \times r$ 矩阵, 满足 $U^T U = V^T V = I_r$ 。证明 A 的所有非零特征根为 $\gamma_1 \geq \dots \geq \gamma_r > 0 > -\gamma_r \geq \dots \geq -\gamma_1$, 且 $\begin{pmatrix} U \\ V \end{pmatrix}$ 和 $\begin{pmatrix} U \\ -V \end{pmatrix}$ 的各列分别为正特征根和负特征根对应的特征向量。
- (b) 记 $\mathbf{1}_m = (1, \dots, 1)^T \in R^m$, 度数向量 $\mathbf{d} = A\mathbf{1}_m$, 度数对角矩阵 $D = \text{diag}(\mathbf{d}) > 0$ 。令拉普拉斯矩阵 $L = D - A$ 。对任何 $\mathbf{x} = (\mathbf{u}^T, \mathbf{v}^T)^T \in R^m$, 其中 $\mathbf{u} = (u_1, \dots, u_n)^T \in R^n, \mathbf{v} = (v_1, \dots, v_p)^T \in R^p$, 证明

$$\mathbf{x}^T L \mathbf{x} = \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^p w_{ij} (u_i - v_j)^2,$$

由此说明 L 是非负定矩阵, 且最小特征根为 0, $\mathbf{1}_m$ 是对应的特征向量。

- (c) 记号同 (b), 令无符号拉普拉斯矩阵 $Q = D + A$, 证明 Q 是非负定矩阵, 最小特征根为 0, 特征向量为 $(\mathbf{1}_n^T, -\mathbf{1}_p^T)^T$ 。

6. 假设 $W_{p \times p} \sim W_p(n)$, $n \geq p$, $\mathbf{x}_{p \times 1} \sim N_p(\mathbf{0}, I_p)$, 且 W 与 $\mathbf{x}$ 独立。试证明

$$\frac{|W|}{|W + \mathbf{x}\mathbf{x}^T|} \sim \text{Beta}\left(\frac{n-p+1}{2}, \frac{p}{2}\right),$$

其中 $|A|$ 表示方阵 A 的行列式, $\text{Beta}(\alpha, \beta)$ 代表参数为 $\alpha > 0, \beta > 0$ 的 beta 分布。