HW5

EX 1. 假设 $x_{11},\ldots,x_{1n_1} \overset{iid}{\sim} N_p(\mu_1,\Sigma)$，$x_{21},\ldots,x_{2n_2} \overset{iid}{\sim} N_p(\mu_2,\Sigma)$ 两组样本独立，两组的样本均值分别是 $\bar{x}_1,\bar{x}_2$，样本方差矩阵分别是 S_1,S_2。定义 $S_{pooled}=[(n_1-1)S_1+(n_2-1)S_2]/(n-2)$，$n=n_1+n_2$。考虑两样本检验问题 $H_0:\mu_1-\mu_2=0_{p\times 1}$，两样本 Hotelling T^2 检验统计量定义为

$$T^2=\frac{n_1n_2}{n_1+n_2}(\bar{x}_1-\bar{x}_2)^\top S_{pooled}^{-1}(\bar{x}_1-\bar{x}_2).$$

另一方面，两样本问题作为多组 MANOVA 的最简单的情况，其检验也可使用 MANOVA 的 Wilks 检验 $\Lambda^*=|W|/|W+B|$ (W,B 的定义，第 9 讲 P3)，试验证

$$-n\log\Lambda^*=n\log(1+T^2/(n-2)).$$

【Solution】

1. 推导：

在两样本 MANOVA 中，组间平方和矩阵为秩 1 矩阵 $B=\frac{n_1n_2}{n}(\bar{x}_1-\bar{x}_2)(\bar{x}_1-\bar{x}_2)^\top$。利用 $|A+uv^\top|=|A|(1+v^\top A^{-1}u)$，可得：

$$|W+B|=|W|\left(1+\frac{n_1n_2}{n}(\bar{x}_1-\bar{x}_2)^\top W^{-1}(\bar{x}_1-\bar{x}_2)\right)$$

由于组内平方和矩阵 $W=(n-2)S_{pooled}$，代入 Hotelling T^2 的定义：

$$\frac{|W+B|}{|W|}=1+\frac{1}{n-2}\left[\frac{n_1n_2}{n}(\bar{x}_1-\bar{x}_2)^\top S_{pooled}^{-1}(\bar{x}_1-\bar{x}_2)\right]=1+\frac{T^2}{n-2}$$

即 $\Lambda^*=(1+T^2/(n-2))^{-1}$。两边取对数并乘以 $-n$ 即得证。

2. 解释：

Wilks Λ^* 从方差分解的角度出发，衡量的是组内散布度占据总散布度的广义比例。而 Hotelling T^2 的核心部分正是两组样本均值向量在多元空间中的**马氏距离（Mahalanobis Distance）**。这证明了在比较组间差异时，方差的广义几何比率（Λ^*）与标准化几何距离（T^2）是单调等价的。

EX 2. 假设 $x_1, \ldots, x_n \overset{iid}{\sim} N_p(\mu, \Sigma)$ 我们考虑如下球对称方差假设

$$H_0 : \Sigma = \sigma^2 I_p, \quad \sigma^2 > 0 \text{ 未知,}$$

记 S 为样本协方差矩阵，似然比统计量 $\Lambda = \frac{\max L(\mu, \gamma I_p)}{\max L(\mu, \Sigma)}$。
(a) 证明 Wilks 统计量

$$\Lambda^* = \Lambda^{2/n} = \frac{\det(S)}{(\mathrm{tr}(S)/p)^p}$$

(b) 已知 $-2\log(\Lambda) = -n\log(\Lambda^*) \to \chi^2_f$，求出自由度 f。
(c) 证明原假设成立时，$\Lambda^* \perp\!\!\!\perp \mathrm{tr}(S)$。

【Solution】
(a) 推导： 无约束条件下的极大似然估计为 $\hat{\Sigma} = \frac{n-1}{n} S$。在 H_0 下 σ^2 的 MLE 为 $\hat{\sigma}^2 = \frac{(n-1)\mathrm{tr}(S)}{np}$。似然比 $\Lambda = (|\hat{\Sigma}|/|\hat{\sigma}^2 I_p|)^{n/2}$，取 $2/n$ 次方后，$(n-1)/n$ 等常数项相互抵消，得到 $\Lambda^* = \frac{\det(S)}{(\mathrm{tr}(S)/p)^p}$。
(b) 自由度： 在无约束条件下，均值 μ 有 p 个参数，协方差矩阵 Σ 包含 $\frac{p(p+1)}{2}$ 个独立参数。$H_0 : \Sigma = \sigma^2 I_p$ 的球对称约束下，均值 μ 仍有 p 个参数，但协方差矩阵仅剩 1 个独立参数（即未知的 σ^2）。均值参数的个数在两步中相互抵消，因此多出的限制条件个数为：

$$f = \frac{p(p+1)}{2} - 1$$

(c) 独立: 在 H_0 下，$S \sim W_p(n-1, \sigma^2 I_p)/(n-1)$。根据指数族性质，$\mathrm{tr}(S)$ 是未知扰动方差 σ^2 的**完备充分统计量**。观察 Λ^* 的结构可知，它本质上是协方差矩阵行列式的几何平均与算术平均之比，是一个尺度不变的**辅助统计量（Ancillary Statistic）**，其分布与 σ^2 无关。由数理统计中的 **Basu 定理**可知，完备充分统计量与辅助统计量必然独立，即得 $\Lambda^* \perp\!\!\!\perp \mathrm{tr}(S)$。
(c) 证法 2(利用球对称性): $\Lambda^* = \frac{\det(S)}{(\mathrm{tr}(S)/p)^p} = \det\left(\frac{S}{\mathrm{tr}(S)}\right) p^p$，实际上，我们可以证明更强的结论：

$$\frac{S}{\mathrm{tr}(S)} \perp\!\!\!\perp \mathrm{tr}(S).$$

不妨假设 $\sigma^2 = 1$, 原假设下 $S \sim W_p(m), m = n-1$，故不妨设 $S = Z^\top Z$, 其中

$$Z = (\mathbf{z}_1, ..., \mathbf{z}_m)^\top = \begin{pmatrix} \mathbf{z}_1^\top \\ \vdots \\ \mathbf{z}_m^\top \end{pmatrix},$$

其中 $\mathbf{z}_1, ..., \mathbf{z}_m$ iid $\sim N_p(0, I_p)$。注意到 $\text{tr}(S) = \text{tr}(Z^\top Z) = \|Z\|^2 = \sum_{i=1}^m \|\mathbf{z}_i\|^2$, 根据 Z 的球对称性 (参见后面注解), $\frac{Z}{\|Z\|} \perp\!\!\!\perp \|Z\| = \sqrt{tr(S)}$, 所以

$$\frac{S}{\text{tr}(S)} = \frac{Z^\top Z}{\|Z\|^2} = \left(\frac{Z}{\|Z\|}\right)^\top \left(\frac{Z}{\|Z\|}\right) \perp\!\!\!\perp \|Z\| = \sqrt{tr(S)}$$

注: Z 的球对称性可以这样理解: 因为 Z 的所有元素独立且服从 $N(0,1)$, $vec(Z^\top) = \begin{pmatrix} \mathbf{z}_1 \\ \vdots \\ \mathbf{z}_m \end{pmatrix} \sim N_{mp}(0, I_{mp})$, 这是球对称分布, 从而 $vec(Z^\top)/\|vec(Z)\| = vec(Z^\top)/\|Z\| \perp\!\!\!\perp \|Z\|$, 从而 $Z/\|Z\|$ 作为 $vec(Z^\top)/\|Z\|$ 的重排也与 $\|Z\|$ 独立。

EX 3. 假设数据矩阵为 $X_{n\times p} = (x_1, \ldots, x_n)^\top$ 且 X 已经中心化, 假设样本协方差阵 $S = X^\top X/(n-1)$ 的谱分解为 $S = V\Lambda V^\top$ 其中 $V^\top V = VV^\top = I_p$, $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_p)$, $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$。主成分矩阵 $Y_{n\times p} = XV$ 的第 j 列为所有 n 个样本点的第 j 个主成分, $j = 1, \ldots, p$。

(a) 证明 Y 是中心化的矩阵, 其样本协方差矩阵为 $Y^\top Y/(n-1) = \Lambda$, Y 与 X 的样本协方差矩阵为 $Y^\top X/(n-1) = \Lambda V^\top$ (即第 k 个变量与第 j 个主成分之间的样本协方差为 $v_{kj}\lambda_j$, v_{kj} 是 V 的 (k,j) 元)。

(b) 证明第 j 主成分和第 k 个变量的样本相关系数 $r_{jk} = v_{kj}\sqrt{\frac{\lambda_j}{s_{kk}}}$, 其中 s_{kk} 为第 k 个变量的样本方差, 即 S 的 (k,k) 元。

【Solution】

1. 证明:

(a) Y 的均值向量 $\mathbf{1}^\top Y = (\mathbf{1}^\top X)V = \mathbf{0}^\top V = \mathbf{0}^\top$, 故 Y 已中心化。

Y 的协方差阵 $\frac{1}{n-1}Y^\top Y = V^\top \left(\frac{X^\top X}{n-1}\right) V = V^\top S V = V^\top V\Lambda V^\top V = \Lambda$。

Y 与 X 的交叉协方差阵 $\frac{1}{n-1}Y^\top X = V^\top S = V^\top V\Lambda V^\top = \Lambda V^\top$。

(b) y_j 与 x_k 的协方差为 $\Lambda V^\top$ 的第 (j,k) 元, 即 $\lambda_j v_{kj}$。代入相关系数公式:

$$r_{jk} = \frac{\text{Cov}(y_j, x_k)}{\sqrt{\text{Var}(y_j)\text{Var}(x_k)}} = \frac{\lambda_j v_{kj}}{\sqrt{\lambda_j s_{kk}}} = v_{kj}\sqrt{\frac{\lambda_j}{s_{kk}}}$$

2. 解释:

(a) 问中 $Y^\top Y = \Lambda$ 在数学上严格证明了 PCA 实现了**特征的完全解耦 (去相关)**。(b) 问推导出的 r_{jk} 在因子分析中被称为 **"主成分载荷 (Loading)"**, 其平方 r_{jk}^2 表示第 j 个主成分能解释第 k 个原变量多少比例的方差。这是下一题进行载荷分析的理论基础。

EX 4. 假设 $x_1, \ldots, x_n \in R^p$ 的样本方差-协方差矩阵具有如下形式

$$S = \sigma^2 I_p + \mathbf{1}_p \mathbf{1}_p^\top + ee^\top,$$

其中 $\mathbf{1}_p = (1, 1, \ldots, 1)^\top$，$e = (1, -1, 0, \ldots, 0)^\top$。求 $x_i, i = 1, \ldots, n$ 的第一主成分和第二主成分，以及它们的方差在总方差中所占的累积比例。

【Solution】

1. 求解：

要使第一主成分、第二主成分顺序成立, 应有 $p \geq 3$, 易证 $\mathbf{1}_p^\top e = 0$，即基向量正交。考察特征方程 $Sv = \lambda v$：令 $v_1 = \mathbf{1}_p$，则 $S\mathbf{1}_p = \sigma^2 \mathbf{1}_p + p\mathbf{1}_p + 0 = (\sigma^2 + p)\mathbf{1}_p$，特征值为 $\lambda_1 = \sigma^2 + p$。令 $v_2 = e$，则 $Se = \sigma^2 e + 0 + 2e = (\sigma^2 + 2)e$，特征值为 $\lambda_2 = \sigma^2 + 2$。总方差为 $\mathrm{tr}(S) = p\sigma^2 + p + 2$。当 $p \geq 3$ 时 $\lambda_1 > \lambda_2 > \sigma^2$。

因此, 第一主成分 $y_1 = \frac{1}{\sqrt{p}} \sum_{i=1}^{p} x_{ij}$, 第二主成分 $y_2 = \frac{1}{\sqrt{2}}(x_{i1} - x_{i2})$。累积比例为 $\frac{2\sigma^2+p+2}{p\sigma^2+p+2}$。

2. 解释：

这个矩阵模拟了心理学数据中常见的结构。$\mathbf{1}_p\mathbf{1}_p^\top$ 导致了各变量的全面正相关（正流形），产生了**全正权重的第一主成分**（一般能力因子）。$\mathbf{ee}^\top$ 引入局部对比，受正交约束产生**正负交替的第二主成分**（双极/形状因子）。

EX 5. Thurstone (1933) 关于 PMA(primary mental abilities) 的研究中, 对 $n = 213$ 进行了 9 项与语言表达能力有关的测试。9 项测试分别是 Sentences, Vocabulary, Sentence Completion, First Letters, Four Letter Words, Suffixes, Letter Series, Pedigrees, Letter group。

上述 9 项测试任务的目的分别是希望测试语言理解能力 (Verbal comprehension, 1-3)，用词流利程度 (Word fluency, 4-6) 和推理逻辑能力 (Reasoning, 7-9)。因为没有原始数据，只有相关系数矩阵，我们进行总体主成分分析 (主因子分析)，得到如下输出结果：

```
Call:
princomp(covmat = Thurstone)
Standard deviations:
Comp.1  Comp.2  Comp.3  Comp.4  Comp.5  Comp.6  Comp.7  Comp.8  Comp.9
2.202   1.044   1.019   0.690   0.669   0.612   0.567   0.484   0.409
```

前 3 个主成分的载荷:

	Sent.	Vocab	S.C.	F.L.	4.L.	Suf.	L.S.	Ped.	L.G.
Comp.1	0.37	0.38	0.36	0.33	0.32	0.30	0.30	0.32	0.29
Comp.2	0.38	0.33	0.36	-0.44	-0.45	-0.33	-0.04	0.23	-0.24
Comp.3	0.15	0.21	0.19	0.21	0.13	0.36	-0.57	-0.33	-0.52

(a) 试计算前 3 个主成分的方差贡献率。
(b) 试根据成分载荷解释前三个主成分的含义，与 Thurstone 的目标是否一致?

【Solution】

(a) 方差贡献率计算：

由于是对相关系数矩阵做 PCA，总方差为变量数 $p = 9$。由标准差可求特征值：$\lambda_1 = 2.202^2 = 4.849$，$\lambda_2 = 1.044^2 = 1.090$，$\lambda_3 = 1.019^2 = 1.038$。以总方差 9 为基数，各 PC 贡献率分别为：53.9%, 12.1%, 11.5%。前三个主成分的累积方差贡献率为 77.5%。**(b) 含义解释与目标一致性分析：**

结合前四题推导的理论基础，解读如下：

1. **PC1（一般智力因子）：**测试项目间存在全面正相关（正流形），根据 Perron-Frobenius 定理，第一主成分的载荷必然全为正。它代表了测试对象的**综合语言基础认知能力**（即 Spearman 的 g 因子）。

2. **PC2 与 PC3（双极对比因子）：**受限于 PCA 提取成分必须相互正交的数学约束（如习题 4 所揭示），后续成分的载荷必然呈正负交替状态：

- **PC2**：语言理解（1-3）为正，用词流利（4-6）为负。这反映了被试在**“深度理解”与“快速提取流利度”之间的差异**。
- **PC3**：推理逻辑（7-9）具有绝对负载荷。这反映了**“逻辑推理”与常规语言表现之间的相对独立性**。

与 Thurstone 目标的一致性反思：表面上看，PCA 成功分出了 Thurstone 预设的三种能力维度特征。然而，**这并不完全契合 Thurstone 的最终目标**。Thurstone 强烈反对这种正负交替的“双极对比”，他提出了“简单结构（Simple Structure）”理论。他认为，真正的基本心理能力应该是相互独立的。因此，直接的主成分分析并非最终结论，而是需要进一步采用**因子旋转（如 Varimax 正交旋转）**，消除这些此消彼长的负荷，使得每项测试只在一个纯粹的能力维度上具有高载荷。
