多元统计分析2025年5月23日主成分分析

第六章 主成分分析

第一节 引言

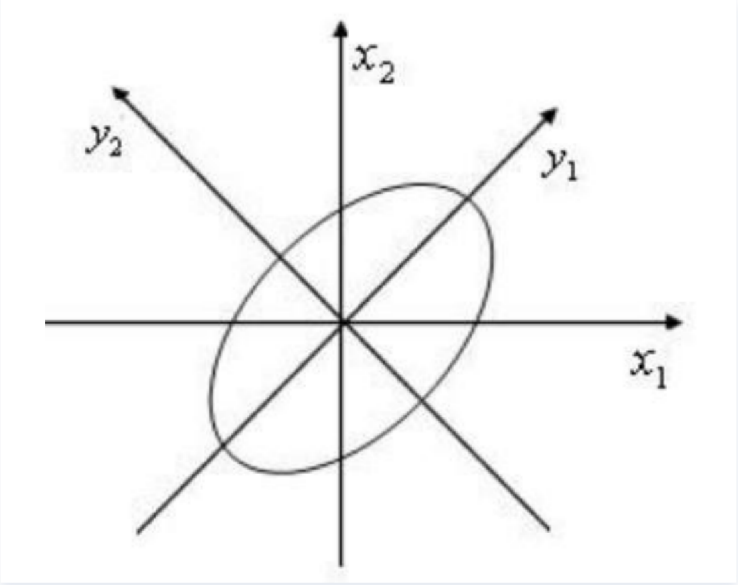
1. **多元统计分析的核心问题**：
   - 处理的是多变量（多指标）问题。
   - 由于变量较多，增加了分析问题的复杂性。
   - 实际问题中，变量之间可能存在一定的相关性，因此多变量中可能存在信息的重叠。
2. **PCA 的基本思想**：
   - **降维思想**：用较少的变量来代替原来较多的变量，并可以反映原来多个变量的大部分信息。
   - 主成分分析（Principal Component Analysis, PCA）是由 Hotelling 于1933年首先提出的。
   - **核心思路**：
     - 多个变量之间往往存在一定程度的相关性。
     - 通过线性组合的方式，从这些指标中尽可能多地提取信息。
       - 当第一个线性组合（第一主成分）不能提取更多的信息时（已最大化方差），
       - 考虑用第二个线性组合（第二主成分，与第一主成分正交）继续提取过程，
       - 依此类推，直到所提取的信息与原指标相差不多时为止。
     - 目标是用较少的主成分得到较多的信息量，得到一个更低维的随机向量。

- 主成分分析既可以降低数据"维数"又保留了原数据的大部分信息。

3. **信息量的度量**：
   - 变量（属性、指标）的信息量：
     - 当一个变量只取一个常数值时，提供的信息量非常有限。
     - 取一系列不同数据时，可以从中读出最大值、最小值、平均数等信息。
     - 变量的变异性越大，说明它对各种场景的"遍历性"越强，提供的信息就更加充分，信息量就越大。
   - 主成分分析中的信息，通常指**指标的变异性**，用**标准差或方差**表示。
4. **主成分分析的数学模型初步**：
   - 设 p 个变量构成的 p 维随机向量为 $X=(X_1,\ldots,X_p)'$。
   - 对 X 作正交变换，令 $Y=T'X$，其中 T 为正交阵。
   - 要求 Y 的各分量是不相关的。
   - 并且 Y 的第一个分量的方差是最大的，第二个分量的方差次之，以此类推。
   - 为了保持信息不丢失（指总变异信息），Y 的各分量方差和与 X 的各分量方差和相等。

第二节 主成分的几何意义及数学推导

一、主成分的几何意义

1. **正交变换与坐标旋转**：正交变换在几何上可以理解为坐标系的旋转。
2. **考虑二维空间**：
   - 假设共有 n 个样品，每个样品都测量了两个指标 (X_1,X_2)，它们大致分布在一个椭圆内（如下图所示）。
   - 

   - 散点的分布总有可能沿着某一个方向略显扩张，这个方向就把它看作椭圆的长轴方向。
   - 在坐标系 x_1Ox_2 中，单独看这 n 个点的分量 X_1 和 X_2，它们沿着 x_1 方向和 x_2 方向都具有较大的离散性，其离散的程度可以分别用 X_1 的方差和 X_2 的方差测定。

- 如果仅考虑 X_1 或 X_2 中的任何一个分量，那么包含在另一分量中的信息将会损失，因此，直接舍弃某个分量不是“降维”的有效办法。

3. **坐标旋转**：

- 将该坐标系按逆时针方向旋转某个角度 θ 变成新坐标系 y_1Oy_2，这里 y_1 是椭圆的长轴方向，y_2 是椭圆的短轴方向。
- 旋转公式为：

$$\begin{cases} Y_1 = X_1 \cos\theta + X_2 \sin\theta \\ Y_2 = -X_1 \sin\theta + X_2 \cos\theta \end{cases}$$

- 新变量 Y_1 和 Y_2 是原变量 X_1 和 X_2 的线性组合，它的矩阵表示形式为：

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = T'X$$

- T' 为旋转变换矩阵，它是正交矩阵，即有 $T' = T^{-1}$ 或 $T'T = I$。

> 这个旋转矩阵在高等代数下也是重要的。

1. **新坐标系下的特性**：

- n 个点在新坐标系下的坐标 Y_1 和 Y_2 几乎不相关。
- Y_1 和 Y_2 为原始变量 X_1 和 X_2 的综合变量。
- n 个点在 y_1 轴上的方差达到最大，在此方向上包含了有关 n 个样品的最大量信息。
- 欲将二维空间的点投影到某个一维方向上，则选择 y_1 轴方向能使信息的损失最小。
- 称 Y_1 为第一主成分，称 Y_2 为第二主成分。
- 第一主成分的效果与椭圆的形状有很大的关系：
  - 椭圆越是扁平，n 个点在 y_1 轴上的方差就相对越大，在 y_2 轴上的方差就相对越小。
  - 用第一主成分代替所有样品所造成的信息损失也就越小。

2. **两种极端情形**：

- **椭圆的长轴与短轴的长度相等，即椭圆变成圆**：
  - 第一主成分只含有二维空间点约一半信息。
  - 若仅用这一个综合变量，则将损失约50%的信息。
  - 原因是：原始变量 X_1 和 X_2 的相关程度几乎为零，它们所包含的信息几乎不重叠。
- **椭圆扁平到了极限，变成 y_1 轴上的一条线**：
  - 第一主成分包含二维空间点的全部信息。
  - 仅用这一个综合变量代替原始数据不会有任何的信息损失。
  - 主成分分析效果最理想。第二主成分不包含任何信息，舍弃它没有信息损失。

二、主成分的数学推导

1. 基本设定：

- 设 $X=(X_1,\ldots,X_p)'$ 为一个 p 维随机向量，并假定存在二阶矩。
- 其均值向量与协方差阵分别记为： $\mu=E(X),\Sigma=D(X)$。
- 考虑如下线性变换：

$$\begin{aligned}Y_1&=t_{11}X_1+t_{12}X_2+\cdots+t_{1p}X_p=T_1'X\\Y_2&=t_{21}X_1+t_{22}X_2+\cdots+t_{2p}X_p=T_2'X\\&\vdots\\Y_p&=t_{p1}X_1+t_{p2}X_2+\cdots+t_{pp}X_p=T_p'X\end{aligned}$$

- 用矩阵表示为： $Y=T'X$，其中 $Y=(Y_1,Y_2,\ldots,Y_p)',T=(T_1,T_2,\ldots,T_p)$。

2. 目标与约束：

- 希望这组新的变量 $Y_1,\ldots,Y_m$ $(m\le p)$ 可以充分地反映原变量 $X_1,\ldots,X_p$ 的信息，而且相互独立（或不相关）。
- 注意到，对于 $Y_1,\ldots,Y_m$ 有：
  - 方差： $D(Y_i)=D(T_i'X)=T_i'D(X)T_i=T_i'\Sigma T_i,\ i=1,2,\ldots,m$
  - 协方差： $\mathrm{Cov}(Y_i,Y_k)=\mathrm{Cov}(T_i'X,T_k'X)=T_i'\mathrm{Cov}(X,X)T_k=T_i'\Sigma T_k,\ i,k=1,2,\ldots,m$
- 问题就转化为，在新的变量 $Y_1,\ldots,Y_m$ 相互独立的条件下（即 $\mathrm{Cov}(Y_i,Y_k)=T_i'\Sigma T_k=0$ for $i\ne k$），求 T_i 使得 $D(Y_i)=T_i'\Sigma T_i$ $(i=1,2,\ldots,m)$ 达到最大（依次最大）。

3. 借助投影寻踪 (Projection Pursuit) 思想：

- 注意到，使得 $D(Y_i)$ 达到最大的线性组合，用常数乘以 T_i 后，$D(Y_i)$ 也随之增大。为了消除这种不确定性，不妨假设 T_i 满足 $T_i'T_i=1$（即 T_i 为单位向量）。
- **第一主成分 $Y_1=T_1'X$**：满足 $T_1'T_1=1$，使得 $D(Y_1)=T_1'\Sigma T_1$ 达到最大。
- **第二主成分 $Y_2=T_2'X$**：满足 $T_2'T_2=1$，且 $\mathrm{Cov}(Y_2,Y_1)=\mathrm{Cov}(T_2'X,T_1'X)=T_2'\Sigma T_1=0$，使得 $D(Y_2)=T_2'\Sigma T_2$ 达到最大。
- **第 k 主成分 $Y_k=T_k'X$**：满足 $T_k'T_k=1$，且 $\mathrm{Cov}(Y_k,Y_i)=\mathrm{Cov}(T_k'X,T_i'X)=T_k'\Sigma T_i=0$（对于 $i<k$），使得 $D(Y_k)=T_k'\Sigma T_k$ 达到最大。

4. 求解第一主成分：

- 构造目标函数（拉格朗日函数）：
  $\varphi_1(T_1,\lambda)=T_1'\Sigma T_1-\lambda(T_1'T_1-1)$
- 对 T_1 求导数并令其为零：
  $\frac{\partial\varphi_1}{\partial T_1}=2\Sigma T_1-2\lambda T_1=0$
- 即 $(\Sigma-\lambda I)T_1=0$。这是一个特征值问题。
- 两边左乘 T_1' 得到： $T_1'\Sigma T_1-\lambda T_1'T_1=0\Rightarrow T_1'\Sigma T_1=\lambda$ (因为 $T_1'T_1=1$)。
- $D(Y_1)=\lambda$。为使方差最大，λ 应取 Σ 的最大特征值 λ_1。
- 由于 X 的协差阵 Σ 为非负定的，其特征值均大于等于零。不妨设 $\lambda_1\ge\lambda_2\ge\cdots\ge\lambda_p\ge0$。

- 那么，Y_1 的最大方差值为 λ_1，其相应的单位化特征向量为 T_1。

5. **求解第二主成分**：
   - $\mathrm{Cov}(Y_2, Y_1) = T_2'\Sigma T_1 = T_2'(\lambda_1 T_1) = \lambda_1 T_2' T_1 = 0$。若 $\lambda_1 \neq 0$，则 $T_2' T_1 = 0$。
   - 构造目标函数：
     $\varphi_2(T_2, \lambda, \rho) = T_2'\Sigma T_2 - \lambda(T_2' T_2 - 1) - 2\rho(T_2' T_1)$
   - 对 T_2 求导数：
     $\frac{\partial \varphi_2}{\partial T_2} = 2\Sigma T_2 - 2\lambda T_2 - 2\rho T_1 = 0$
   - 用 T_1' 左乘上式有：$T_1'\Sigma T_2 - \lambda T_1' T_2 - \rho T_1' T_1 = 0$。
   - 由于 $T_1'\Sigma T_2 = (T_2'\Sigma T_1)' = 0$（因为 $T_2'\Sigma T_1 = 0$），且 $T_1' T_2 = (T_2' T_1)' = 0$，那么 $\rho T_1' T_1 = 0$。因为 $T_1' T_1 = 1$，所以 $\rho = 0$。
   - 从而 $(\Sigma - \lambda I)T_2 = 0$，且 $T_2'\Sigma T_2 = \lambda$。为使 $D(Y_2)$ 最大且与 Y_1 正交，λ 应取 Σ 的第二大特征值 λ_2，T_2 是对应的单位特征向量。

6. **求解第 k 主成分**：
   - 这说明：如果 X 的协差阵 Σ 的特征根为 $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$。
   - 则 Y_2 的最大方差值为第二大特征根 λ_2，其相应的单位化的特征向量为 T_2。
   - 针对一般情形，第 k 主成分 $Y_k = T_k' X$ 应该是在 $T_k' T_k = 1$ 且 $T_k' T_i = 0$（或 $T_k'\Sigma T_i = 0$ for $i < k$，更准确地是要求 T_k 与 $T_1, \ldots, T_{k-1}$ 正交）的条件下，使得 $D(Y_k) = T_k'\Sigma T_k$ 达到最大的。
   - 构造目标函数：
     $\varphi_k(T_k, \lambda, \rho_i) = T_k'\Sigma T_k - \lambda(T_k' T_k - 1) - 2\sum_{i=1}^{k-1} \rho_i (T_k' T_i)$
   - 对 T_k 求导数：
     $\frac{\partial \varphi_k}{\partial T_k} = 2\Sigma T_k - 2\lambda T_k - 2\sum_{i=1}^{k-1} \rho_i T_i = 0$
   - 用 T_j' $(j < k)$ 左乘上式，并利用 $T_j'\Sigma T_k = T_j'\lambda_k T_k = \lambda_k T_j' T_k = 0$（对于不同的特征向量）和 $T_j' T_i = \delta_{ji}$（Kronecker delta），可推得 $\rho_j = 0$ for all $j < k$。
   - 从而 $(\Sigma - \lambda I)T_k = 0$ 且 $T_k'\Sigma T_k = \lambda$。
   - 对于 X 的协差阵 Σ 的特征根 $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$，则有 Y_k 的最大方差值为第 k 大特征根 λ_k，其相应的单位化的特征向量为 T_k。

7. **总结**：
   - 设 $X = (X_1, \ldots, X_p)'$ 的协差阵为 Σ，其特征根为 $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$，相应的单位化的特征向量为 $T_1, T_2, \ldots, T_p$。
   - 那么，由此所确定的主成分为 $Y_1 = T_1' X, Y_2 = T_2' X, \ldots, Y_m = T_m' X$（通常取 $m \leq p$ 个主成分），其方差分别为 Σ 的特征根 $\lambda_1, \lambda_2, \ldots, \lambda_m$。 ==

> 可以将整个寻找T的过程看作一个优化问题。

第三节 主成分的性质

一、主成分的一般性质

设 $Y=(Y_1,Y_2,\dots,Y_p)'$ 是 X 的主成分，由 Σ 的所有特征根构成的对角阵为

$$\Lambda=\begin{pmatrix}\lambda_1 & & 0\\ & \ddots & \\ 0 & & \lambda_p\end{pmatrix}$$

主成分可表示为 $Y=T'X$，其中 $T=(T_1,\dots,T_p)$ 是由特征向量构成的正交矩阵。

1. 性质1：主成分的协方差矩阵是对角阵。

- 证明：
  - $E(Y)=E(T'X)=T'E(X)=T'\mu$
  - $D(Y)=T'D(X)T=T'\Sigma T=\Lambda$
    (因为 T 的列是 Σ 的正交特征向量，所以 $\Sigma T=T\Lambda$，则 $T'\Sigma T=T'T\Lambda=I\Lambda=\Lambda$)。
- 这意味着主成分 $Y_1,\dots,Y_p$ 互不相关，且 $D(Y_i)=\lambda_i$。

2. 性质2：主成分的总方差等于原始变量的总方差。

- 证明：由矩阵"迹"的性质知
  $\mathrm{tr}(\Lambda)=\mathrm{tr}(T'\Sigma T)=\mathrm{tr}(\Sigma TT')=\mathrm{tr}(\Sigma I)=\mathrm{tr}(\Sigma)$
- 所以
  $\sum_{i=1}^p\lambda_i=\sum_{i=1}^p\sigma_{ii}$
  其中 $\sigma_{ii}=D(X_i)$。
- 或
  $\sum_{i=1}^p D(Y_i)=\sum_{i=1}^p D(X_i)$
- 这表明主成分变换保持了原始数据的总变异性。

3. 性质3：主成分 Y_k 与原始变量 X_i 的相关系数为

$\rho(Y_k,X_i)=\frac{\sqrt{\lambda_k}}{\sqrt{\sigma_{ii}}}t_{ki}$
其中 t_{ki} 是特征向量 T_k 的第 i 个分量。这个相关系数称之为**因子载荷量 (factor loading)**。

- 证明：
  $\rho(Y_k,X_i)=\frac{\mathrm{Cov}(Y_k,X_i)}{\sqrt{D(Y_k)D(X_i)}}=\frac{\mathrm{Cov}(T_k'X,e_i'X)}{\sqrt{\lambda_k\sigma_{ii}}}$
  其中 $e_i=(0,\dots,0,1,0,\dots,0)'$ (第 i 个位置为1)。
  $\mathrm{Cov}(T_k'X,e_i'X)=T_k'\Sigma e_i=e_i'\Sigma T_k=e_i'(\lambda_kT_k)=\lambda_k(e_i'T_k)=\lambda_kt_{ki}$。
  所以
  $\rho(Y_k,X_i)=\frac{\lambda_kt_{ki}}{\sqrt{\lambda_k\sigma_{ii}}}=\frac{\sqrt{\lambda_k}t_{ki}}{\sqrt{\sigma_{ii}}}$

4. 性质4：

$\sum_{i=1}^p\rho^2(Y_k,X_i)\cdot\sigma_{ii}=\lambda_k,\quad(k=1,2,\dots,p)$

- 证明：将性质3的 $\rho(Y_k,X_i)$ 代入：
  $\sum_{i=1}^p\left(\frac{\sqrt{\lambda_k}t_{ki}}{\sqrt{\sigma_{ii}}}\right)^2\sigma_{ii}=\sum_{i=1}^p\frac{\lambda_kt_{ki}^2}{\sigma_{ii}}\sigma_{ii}=\sum_{i=1}^p\lambda_kt_{ki}^2=\lambda_k\sum_{i=1}^pt_{ki}^2$
  因为 T_k 是单位特征向量，所以 $\sum_{i=1}^pt_{ki}^2=T_k'T_k=1$。
  因此 $\sum_{i=1}^p\rho^2(Y_k,X_i)\cdot\sigma_{ii}=\lambda_k\cdot1=\lambda_k$。

- 如果原始变量已标准化（即 $\sigma_{ii}=1$），则 $\sum_{i=1}^{p}\rho^2(Y_k,X_i)=\lambda_k$。如果进一步 PCA 是基于相关矩阵做的，则 λ_k 是相关矩阵的特征值。

二、主成分的方差贡献率

1. 概念：

- 由主成分的性质2可以看出，主成分分析把 p 个原始变量 $X_1,X_2,\dots,X_p$ 的总方差 $\mathrm{tr}(\Sigma)$ 分解成了 p 个相互独立的变量 $Y_1,Y_2,\dots,Y_p$ 的方差之和 $\sum_{k=1}^{p}\lambda_k$。
- 主成分分析的目的是减少变量的个数，所以一般不会使用所有主成分的，忽略一些带有较小方差的主成分将不会给总方差带来太大的影响。

2. 贡献率计算：

- 称 $\varphi_k=\lambda_k/\sum_{j=1}^{p}\lambda_j$ 为第 k 个主成分 Y_k 的**贡献率**。
- 第一主成分的贡献率最大，这表明 $Y_1=T_1'X$ 综合原始变量 $X_1,X_2,\dots,X_p$ 的能力最强。
- $Y_2,Y_3,\dots,Y_p$ 的综合能力依次递减。
- 若只取 $m(<p)$ 个主成分，则称 $\psi_m=\sum_{k=1}^{m}\lambda_k/\sum_{j=1}^{p}\lambda_j$ 为主成分 $Y_1,\dots,Y_m$ 的**累计贡献率**。
- 累计贡献率表明 $Y_1,\dots,Y_m$ 综合 $X_1,X_2,\dots,X_p$ 的能力。
- 通常取 m，使得累计贡献率达到一个较高的百分数（如 85% 以上）。

3. 例子：

- 设 $X=(X_1,X_2,X_3)'$ 的协方差矩阵为
  $$\Sigma=\begin{pmatrix}1&-2&0\\-2&5&0\\0&0&2\end{pmatrix}$$
- 其特征值为 $\lambda_1=5.83,\lambda_2=2.00,\lambda_3=0.17$。总方差 $=5.83+2.00+0.17=8.00$。
- 相应的特征向量为 (根据 Σ 对称性及特征值，可以验证或重新计算得到)
  $$t_1\approx\begin{pmatrix}0.383\\-0.924\\0.000\end{pmatrix},t_2=\begin{pmatrix}0\\0\\1\end{pmatrix},t_3\approx\begin{pmatrix}0.924\\0.383\\0.000\end{pmatrix}$$
  (注意：t_1,t_3 主要由左上角 2×2 子矩阵决定，因为 X_3 与 X_1,X_2 不相关。)
- 若只取一个主成分，则贡献率为 $\varphi_1=5.83/8.00=0.72875=72.875\%$。
- **因子载荷量表** (示例数据，根据 Σ 和特征向量计算 $\rho(Y_k,X_i)$)

i	$\rho(Y_1,X_i)$	$\rho(Y_2,X_i)$
1	0.925	0.000
2	-0.998	0.000
3	0.000	1.000

- Y_1 对第三个变量 X_3 的因子载荷量为零。这是因为 X_3 与 X_1 和 X_2 都不相关，在 Y_1 中未包含有关 X_3 的信息 (Y_1 主要由 X_1,X_2 构成)。
- 仅取一个主成分就显得不够了，故应再取 Y_2。

- 累计贡献率为 $(5.83+2.00)/8.00=7.83/8.00=0.97875=97.875\%$。

三、主成分的解释

1. **意义的重要性**：主成分分析成功与否取决于主成分是否有意义。
2. **载荷**：
   - 由 $Y_k=t_{k1}X_1+t_{k2}X_2+\cdots+t_{kp}X_p$ 称 t_{ki} (特征向量的分量) 为第 k 主成分 Y_k 在第 i 个原始变量 X_i 上的载荷，它度量了 X_i 对 Y_k 的重要程度。(注意：有时因子载荷量 $\rho(Y_k,X_i)$ 也被称为载荷，特别是在基于相关矩阵的PCA中，$\rho(Y_k,X_i^*)=\sqrt{\lambda_k^*}t_{ki}^*$)
   - 在解释主成分时，需要考察载荷大小。绝对值较大的载荷表明对应的原始变量对该主成分的贡献较大。
3. **方差与主成分的联系**：
   - 方差大的那些变量与具有大特征值的主成分（如 Y_1）有较密切的联系。
   - 而方差小的另一些变量与具有小特征值的主成分有较强的联系。
   - 通常取前几个主成分，因此所取主成分会过于照顾方差大的变量，而对方差小的变量却照顾得不够。**这也是为什么通常建议对原始数据进行标准化的原因之一**。
4. **例子**：
   - 设 $X=(X_1,X_2,X_3)'$ 的协方差矩阵为
     $$\Sigma=\begin{pmatrix}16&2&30\\2&1&4\\30&4&100\end{pmatrix}$$
   - 经计算，Σ 的特征值及特征向量为：
     $$\lambda_1=109.793,\lambda_2=6.469,\lambda_3=0.738$$
     $$t_1=\begin{pmatrix}0.305\\0.041\\0.951\end{pmatrix},t_2=\begin{pmatrix}0.944\\0.120\\-0.308\end{pmatrix},t_3=\begin{pmatrix}-0.127\\0.992\\-0.002\end{pmatrix}$$
   - 相应的主成分分别为：
     $$Y_1=0.305X_1+0.041X_2+0.951X_3$$
     $$Y_2=0.944X_1+0.120X_2-0.308X_3$$
     $$Y_3=-0.127X_1+0.992X_2-0.002X_3$$
   - 解释：
     - 方差大的原始变量 X_3 (其方差 $\sigma_{33}=100$) 在很大程度上控制了第一主成分 Y_1 (载荷0.951)。
     - 方差小的原始变量 X_2 (其方差 $\sigma_{22}=1$) 几乎完全控制了第三主成分 Y_3 (载荷0.992)。
     - 方差介于中间的 X_1 (其方差 $\sigma_{11}=16$) 则基本控制了第二主成分 Y_2 (载荷0.944)。
   - Y_1 的贡献率为：
     $$\frac{\lambda_1}{\lambda_1+\lambda_2+\lambda_3}=\frac{109.793}{109.793+6.469+0.738}=\frac{109.793}{117}\approx0.938$$
   - 高贡献率归因于 X_3 的方差比 X_1 和 X_2 的方差大得多。
   - 另外，Y_1 与 X_1,X_3 的相关系数远大于与 X_2 的相关系数。

第四节 主成分分析应用中应注意的问题

一、实际应用中主成分分析的出发点

1. 变量单位的影响：

- 主成分计算从协方差阵 Σ 出发，变量单位的改变会产生不同的主成分。例如，将长度单位从米换为厘米，该变量的方差会增大 100^2 倍，从而可能主导主成分的构成。
- **“大数吃小数”**：主成分倾向于多归纳方差大的变量的信息。

2. 标准化处理：

- 为了消除量纲和数量级的影响，通常对原始数据进行标准化处理：
  $X_i^* = \frac{X_i - E(X_i)}{\sqrt{D(X_i)}} \quad i = 1, \ldots, p$
- $X^* = (X_1^*, \ldots, X_p^*)'$ 的协方差矩阵就是 X 的**相关系数矩阵 R**。
- 实际应用中，X 的相关系数矩阵 $\mathbf{R}$ 可以利用样本数据来估计。

3. 基于协方差阵 vs. 相关阵：

- 从相关阵求得的主成分与协差阵求得的主成分一般情况是不相同的。这种差异有时很大。
- **实际应用中**：
  - 如果各指标之间的数量级相差悬殊，特别是各指标有不同的物理量纲的话，较为合理的做法是使用相关矩阵 $\mathbf{R}$ 代替协方差矩阵 Σ进行主成分分析。
  - 采用 $\mathbf{R}$ 代替 Σ 后，可以看作是用标准化的数据做分析，这样使得主成分有现实意义，便于剖析实际问题，又可以避免突出数值大的变量。

4. 基于相关矩阵的PCA步骤：

1. 将原始数据标准化 $X_i \to X_i^*$。
2. 建立变量的相关系数阵 $\mathbf{R}$ (即 X^* 的协方差阵)。
3. 求 $\mathbf{R}$ 的特征根为 $\lambda_1^* \geq \cdots \geq \lambda_p^* \geq 0$，相应的特征向量为 $T_1^*, T_2^*, \ldots, T_p^*$。(注意 $\sum \lambda_i^* = p$)
4. 由累积方差贡献率确定主成分的个数 (m)，并写出主成分为 $Y_i^* = (T_i^*)' X^*$，$i = 1, 2, \ldots, m$。

5. 上例（P28-29）化为相关阵出发计算：

- X 的相关矩阵 (假设由原始数据计算得到)
  $$\mathbf{R} = \begin{pmatrix} 1 & 0.5 & 0.75 \\ 0.5 & 1 & 0.4 \\ 0.75 & 0.4 & 1 \end{pmatrix}$$
- $\mathbf{R}$ 的特征值及特征向量为：
  $\lambda_1^* = 2.114, \lambda_2^* = 0.646, \lambda_3^* = 0.240$
  $$t_1^* = \begin{pmatrix} 0.627 \\ 0.497 \\ 0.600 \end{pmatrix}, t_2^* = \begin{pmatrix} -0.241 \\ 0.856 \\ -0.457 \end{pmatrix}, t_3^* = \begin{pmatrix} -0.741 \\ 0.142 \\ 0.656 \end{pmatrix}$$

- 相应的主成分（基于标准化变量 X_i^*）分别为：
  $Y_1^* = 0.627X_1^* + 0.497X_2^* + 0.600X_3^*$
  $Y_2^* = -0.241X_1^* + 0.856X_2^* - 0.457X_3^*$
  $Y_3^* = -0.741X_1^* + 0.142X_2^* + 0.656X_3^*$
- Y_1^* 的贡献率为 $\lambda_1^*/p = 2.114/3 \approx 0.705$。
- Y_1^* 和 Y_2^* 累计贡献率为 $(\lambda_1^* + \lambda_2^*)/p = (2.114 + 0.646)/3 \approx 0.920$。
- 从 $\mathbf{R}$ 出发的 Y_1^* 的贡献率 0.705 明显小于从 Σ 出发的 Y_1 的贡献率 0.938。
- 原始变量方差之间的差异越大，这一点倾向越明显。
- Y_1^* 用标准化前的原变量表达如下（使用 P28 的 $\sigma_1 = \sqrt{16} = 4, \sigma_2 = \sqrt{1} = 1, \sigma_3 = \sqrt{100} = 10$）：
  $Y_1^* = 0.627\left(\frac{X_1-\mu_1}{4}\right) + 0.497\left(\frac{X_2-\mu_2}{1}\right) + 0.600\left(\frac{X_3-\mu_3}{10}\right)$
  $= 0.157(X_1-\mu_1) + 0.497(X_2-\mu_2) + 0.060(X_3-\mu_3)$
  类似地可以写出 Y_2^* 和 Y_3^*。
- Y_i^* 在原变量 X_1, X_2, X_3 上的载荷（指这里的系数 $0.157, 0.497, 0.060$）相对大小与上例中 Y_i 在 X_1, X_2, X_3 上的载荷（指 t_{i1}, t_{i2}, t_{i3}）相对大小之间有着非常大的差距。
- **标准化后的结论完全可能会发生很大的变化**。

二、主成分的合理选择与解释

1. 基本原则：

- 在主成分分析中，首先应保证所提取的前几个主成分的累计贡献率达到一个较高的水平。
- 其次对这些被提取的主成分必须都能够给出具有意义的解释。
- 主成分的含义一般多少带点模糊性，不像原始变量的含义那么清楚、确切，这是变量降维过程中不得不付出的代价。
- 提取的主成分个数 m 通常应明显小于原始变量个数 p (除非 p 本身较小)，否则维数降低的“利”可能抵不过主成分含义不如原始变量清楚的“弊”。

2. 解释的困难与成功要素：

- 如果原始变量之间具有较高的相关性，则前面少数几个主成分的累计贡献率通常就能达到一个较高水平，此时的累计贡献率通常较易得到满足。
- 主成分分析的困难之处在于要如何给出主成分的解释，所提取的主成分中如有一个主成分解释不了，整个主成分分析也就失败了。
- 主成分分析是变量降维的一种重要、常用的方法，但该方法要应用得成功，一是靠原始变量的合理选取，二是靠“运气”（指能否得到易于解释的主成分）。

3. 例子：男子身材指标分析

- 对128名成年男子的6项身材指标进行PCA（基于相关矩阵 $\hat{R}$）。
  - X_1: 身高, X_2: 坐高, X_3: 胸围, X_4: 手臂长, X_5: 肋围, X_6: 腰围
- 相关阵 $\hat{R}$ 的前三个特征值、特征向量（载荷 t_j^*）及贡献率:

特征向量 (X_i^*)	$\hat{t}_1^*$	$\hat{t}_2^*$	$\hat{t}_3^*$
X_1^*: 身高	0.469	-0.365	0.092
X_2^*: 坐高	0.404	-0.397	0.613
X_3^*: 胸围	0.394	0.397	-0.279
X_4^*: 手臂长	0.408	-0.365	-0.705
X_5^*: 肋围	0.337	0.569	0.164
X_6^*: 腰围	0.427	0.308	0.119
特征值	**3.287**	**1.406**	**0.459**
贡献率	**0.548**	**0.234**	**0.077**
累计贡献率	**0.548**	**0.782**	**0.859**

- 前三个主成分（基于标准化变量 X_i^*）：
$$Y_1^* = 0.469X_1^* + 0.404X_2^* + 0.394X_3^* + 0.408X_4^* + 0.337X_5^* + 0.427X_6^*$$
$$Y_2^* = -0.365X_1^* - 0.397X_2^* + 0.397X_3^* - 0.365X_4^* + 0.569X_5^* + 0.308X_6^*$$
$$Y_3^* = 0.092X_1^* + 0.613X_2^* - 0.279X_3^* - 0.705X_4^* + 0.164X_5^* + 0.119X_6^*$$
- **解释**：
  - 前两个主成分的累计贡献率为78.2%，前三个主成分的累计贡献率达85.9%，因此可以考虑只取前面两个或三个主成分。
  - **第一主成分 Y_1^***：对所有（标准化）原始变量都有近似相等的正载荷，故称第一主成分为（身材）**大小成分**。
  - **第二主成分 Y_2^***：在 X_3^*, X_5^*, X_6^* (胸围、肋围、腰围，代表“围度”)上有中等程度的正载荷，而在 X_1^*, X_2^*, X_4^* (身高、坐高、手臂长，代表“长度”)上有中等程度的负载荷，称第二主成分为**形状成分**（或胖瘦成分）。
  - **第三主成分 Y_3^***：在 X_2^* (坐高)上有大的正载荷，在 X_4^* (手臂长)上有大的负载荷，而在其余变量上的载荷都较小，可称第三主成分为**臂长成分**（可能反映上肢相对于躯干的比例）。
  - 由于第三主成分的贡献率不高（7.65%）且实际意义也不太重要，因此也可考虑取前两个主成分。

三、利用主成分分析进行综合评价

1. 基本思路：

- 评价指标体系的选择与综合。
- 加权：权重如何选取?
- 主成分分析能从选定的指标体系中归纳出大部分信息。
- 根据主成分提供的信息进行综合评价。
- 利用主成分进行综合评价是将原有的信息进行综合。
- 权重根据它们的**方差贡献率**来确定（即主成分的信息含量）。

2. 综合评价函数：

- 设 $Y_1, Y_2, \ldots, Y_p$ 是所求出的 p 个主成分，它们的特征根（方差）分别是 $\lambda_1, \lambda_2, \ldots, \lambda_p$。

- 将特征根"归一化"作为权重（如果选取前 m 个主成分）：
  $w_i = \frac{\lambda_i}{\sum_{j=1}^{m} \lambda_j} \quad i = 1, 2, \ldots, m$
  (注意：幻灯片中分母为 $\sum_{i=1}^{m} \lambda_i$，这意味着权重是相对于选取的 m 个主成分的总方差而言的。如果使用所有 p 个主成分，分母是 $\sum_{j=1}^{p} \lambda_j$)
- 记为 $W = (w_1, w_2, \ldots, w_m)'$，$Y_{sel} = (Y_1, \ldots, Y_m)'$。
- 构造综合评价函数为：
  $Z = w_1 Y_1 + w_2 Y_2 + \cdots + w_m Y_m = W' Y_{sel}$
  如果 $Y_{sel} = T'_{sel} X$ (其中 T_{sel} 是前 m 个特征向量构成的矩阵)，则
  $Z = W' T'_{sel} X = (T_{sel} W)' X$
  令 $T_{sel} W = w^*$，有 $Z = (w^*)' X$。
- 综合评价函数是对原始指标计算主成分，再对之加权，经过两次线性运算得到的线性综合。

第五节 实例分析与计算机实现

一、主成分分析实例 (某市工业部门经济指标)

- **数据**：表6.1是某市工业部门13个行业的8项重要经济指标的数据。
  - X1: 年末固定资产净值 (万元)
  - X2: 职工人数据 (人)
  - X3: 工业总产值 (万元)
  - X4: 全员劳动生产率 (元/人年)
  - X5: 百元固定资产原值实现产值 (元)
  - X6: 资金利税率 (%)
  - X7: 标准燃料消费量 (吨)
  - X8: 能源利用效果 (万元/吨)
- **问题**：如何从这些经济指标出发，对各工业部门进行综合评价与排序?
- **方法**：先计算这些指标的主成分（通常基于标准化数据，即相关矩阵），然后通过主成分的大小（或加权得分）进行排序。
  - 表6.2 特征根和累计贡献率 (基于相关矩阵)

序号	特征根	方差贡献率%	累计贡献率%
1	3.1049	38.8114	38.8114
2	2.8974	36.2180	75.0294
3	0.9302	11.6277	86.6571
...	...	...	...
(前3个主成分累计贡献率已达86.6571%)			

- 表6.3 特征向量 (部分)

  $Y_1^* = 0.476X_1^* + 0.473X_2^* + 0.424X_3^* - 0.213X_4^* - 0.388X_5^* - 0.352X_6^* + 0.215X_7^* + 0.055X_8^*$

  $Y_2^* = 0.296X_1^* + 0.278X_2^* + 0.378X_3^* + 0.451X_4^* + 0.331X_5^* + 0.403X_6^* - 0.377X_7^* + 0.273X_8^*$

  ...

- **综合得分计算**：以特征根（即各主成分的方差，代表信息量）为权，对8个主成分进行加权综合。综合得分的计算公式是 (这里 $m = p = 8$):

  $Y_{\text{综合}} = \frac{\lambda_1}{\sum_{i=1}^{8}\lambda_i}Y_1^* + \frac{\lambda_2}{\sum_{i=1}^{8}\lambda_i}Y_2^* + \cdots + \frac{\lambda_8}{\sum_{i=1}^{8}\lambda_i}Y_8^*$

- 表6.4 各行业主成分得分及排序 (部分)

行业	Y_1^*	Y_2^*	...	Y_8^*	综合得分	排序
冶金	1.475	0.759	...	0.004	0.911	2
机器	4.528	2.262	...	0.023	2.589	1
...	...	...	...	...	...	...

- **结论**：机器行业在该地区的综合评价排在第一。从前两个主成分得分上看，该行业也排在第一位。排在最后三位的分别是皮革行业、电力行业和煤炭行业。

二、利用 R 进行主成分分析（非重点）

- **数据**：表6.5 分地区城镇居民家庭收支基本情况 (示例数据)
- **R code 示例 (基于协方差矩阵)**:

```r
#read data
# data.frame=read.table(file="datasets/pca.dat",header=T) # 假设数据已读入
# #calculate the covariance matrix of the data set
# # 假设 data 是一个包含数值变量的矩阵，例如前30行，第2到6列
# data=data.matrix(data.frame[c(1:30),c(2:6)])
# data.cov=cov(data)
# head(data.cov)

# #By using the function eigen the eigenvalues and eigenvectors of the
# #covariance matrix are computed
# Eigenvalues ← eigen(data.cov)$values
# Eigenvectors ← eigen(data.cov)$vectors

# #Principal Components can be estimated via a matrix multiplication
# PC ← as.matrix(data) %*% Eigenvectors

# #As a check of the result, we compute the covariance matrix of PC.
# #The variances of cov(PC) should be equal to the Eigenvalues and the
# #covariances should be 0 (aside from rounding errors) since the
# #Principal Components have to be uncorrelated.
# cov(PC) # 对角线元素应接近 Eigenvalues，非对角线元素应接近0

# #We do this for the first three Eigenvalues
# Eigenvalues[1:3]
# cov(PC)[1:3, 1:3] # 检查前3个主成分的协方差

```

```
# #We calculate the proportions of the variation explained by the various
# #components:
# print(round(Eigenvalues/sum(Eigenvalues) * 100, digits = 2)) # 各主成分贡献率
# round(cumsum(Eigenvalues)/sum(Eigenvalues) * 100, digits = 2) # 累计贡献率
```

- **R code 示例 (使用 `prcomp` 函数，推荐)**:

```
# #PCA using prcomp
# #The best way to do PCA with R is to use the function prcomp from the
# #package stats.
# #prcomp with the argument scale = TRUE (default: scale =FALSE) the
# #variables #can be scaled to a unit variance before the analysis takes place.

# #read data (同上)
# # data.frame=read.table(file="datasets/pca.dat",header=T)
# # data=data.matrix(data.frame[c(1:30),c(2:6)])

# data.pca <- prcomp(data, scale. = FALSE) # scale. = FALSE 基于协方差矩阵
                                           # scale. = TRUE 基于相关矩阵 (推荐)
# #Note: To reproduce our previous calculation we use the default case (scale =
# #FALSE). The PrintOutput of data.pca gives us the estimated standard
# #deviations as well as the rotations (loadings).
# print(data.pca) # 查看结果，包含标准差(sdev)和旋转(rotations，即特征向量)
# summary(data.pca) # 查看各主成分的重要性，包括贡献率和累计贡献率

# data.pca.var <- data.pca$sdev^2 # 方差即标准差的平方
# data.pca.var[1:3] # 前三个主成分的方差
# #which are identical to the Eigenvalues Eigenvalues[1:3] (如果与手动计算一致)

# plot(data.pca, type="l") # 碎石图 (Scree plot)，帮助判断主成分个数
# biplot(data.pca)      # 双标图，展示样本点和原始变量在主成分空间的投影
```

- **输出示例 (Total Variance Explained)**:
  类似于SPSS的输出，R中的 `summary(data.pca)` 也会提供类似信息。

Component	Total (Eigenvalue)	% of Variance	Cumulative %
1	2.576	51.520	51.520
2	1.389	27.790	79.310
3	.961	19.222	98.532

Component	Total (Eigenvalue)	% of Variance	Cumulative %
4	.047	.932	99.465
5	.027	.535	100.000
Extraction Method: Principal Component Analysis. (这是基于相关矩阵的结果，因为总方差为5，即变量个数)			

详细讲讲这个因子载荷量

在主成分分析的语境下，

> **因子载荷量**主要指的是**主成分与原始变量之间的相关系数**。

它衡量了一个特定的原始变量与一个特定的主成分之间的线性关联程度。

让我们回顾一下：

> **性质3：** 主成分 Y_k 与原始变量 X_i 的相关系数为
> $\rho(Y_k, X_i) = \frac{\sqrt{\lambda_k}}{\sqrt{\sigma_{ii}}} t_{ki}$
> 并称之为**因子载荷量**。
>
> 这里：
>
> - Y_k 是第 k 个主成分。
> - X_i 是第 i 个原始变量。
> - λ_k 是协方差矩阵 Σ 的第 k 个特征值（也是 Y_k 的方差）。
> - σ_{ii} 是原始变量 X_i 的方差。
> - t_{ki} 是构成第 k 个主成分 $Y_k = t_{k1}X_1 + t_{k2}X_2 + \cdots + t_{kp}X_p$ 时，X_i 前面的系数（即对应于 λ_k 的特征向量 T_k 的第 i 个分量）。

因子载荷量的含义和解释：

1. **相关性的度量**：
   - 因子载荷量的取值范围是 $[-1, 1]$，与普通的相关系数一样。
   - **绝对值大小**：其绝对值越接近 1，表示该原始变量 X_i 与该主成分 Y_k 之间的线性相关性越强。这意味着该主成分能够很好地代表该原始变量所包含的变异信息的一部分。
   - **绝对值大小**：其绝对值越接近 0，表示该原始变量 X_i 与该主成分 Y_k 之间的线性相关性越弱。这意味着该主成分与该原始变量的关联不大。

- **正负号**：
  - 正的因子载荷量表示该原始变量与该主成分呈正相关。当该原始变量的值增加时，该主成分的值也倾向于增加。
  - 负的因子载荷量表示该原始变量与该主成分呈负相关。当该原始变量的值增加时，该主成分的值倾向于减少。

2. **解释主成分的构成**：
   - 因子载荷量是**解释和命名主成分的关键**。通过查看哪些原始变量在一个主成分上具有较高的（绝对值）因子载荷量，我们可以理解该主成分主要概括了哪些原始变量的信息，或者说该主成分代表了原始变量集合中的哪种潜在的“共同特征”或“对比关系”。
   - 例如，在幻灯片P41的男子身材例子中：
     - 第一主成分 Y_1^* 对所有原始变量（身高、坐高、胸围等）都有近似相等的正载荷（这里 $\hat{t}_1^*$ 的值直接反映了因子载荷量的相对大小，因为是基于相关矩阵）。这表明 Y_1^* 是一个衡量“整体大小”的指标，所以被称为“（身材）大小成分”。
     - 第二主成分 Y_2^* 在胸围、肋围、腰围上有中等程度的正载荷，而在身高、坐高、手臂长上有中等程度的负载荷。这表明 Y_2^* 反映了“围度”指标与“长度”指标之间的一种对比关系，因此被称为“形状成分”（或胖瘦成分）。

3. **与载荷系数 (t_{ki}) 的关系**：
   - 系数 t_{ki} 是原始变量 X_i 在线性组合成主成分 Y_k 时的**权重**。它直接决定了 X_i 在数值上对 Y_k 的贡献大小。
   - 因子载荷量 $\rho(Y_k, X_i)$ 是 Y_k 与 X_i 之间的**相关性**。
   - 从公式 $\rho(Y_k, X_i) = \frac{\sqrt{\lambda_k}}{\sqrt{\sigma_{ii}}} t_{ki}$ 可以看出，因子载荷量是权重 t_{ki} 经过 $\frac{\sqrt{\lambda_k}}{\sqrt{\sigma_{ii}}}$ 调整后的结果。这个调整因子考虑了主成分的方差（信息量）和原始变量自身的方差。

4. **在基于相关矩阵的PCA中**：
   - 当主成分分析是基于**相关矩阵 $\mathbf{R}$** 进行时（即原始数据已经标准化，每个 X_i^* 的均值为0，方差 $\sigma_{ii}^* = 1$），公式变为：
     $\rho(Y_k^*, X_i^*) = \frac{\sqrt{\lambda_k^*}}{\sqrt{1}} t_{ki}^* = \sqrt{\lambda_k^*} t_{ki}^*$
     其中 λ_k^* 是相关矩阵 $\mathbf{R}$ 的特征值，t_{ki}^* 是相应的特征向量分量。
   - 在这种情况下，因子载荷量 $\rho(Y_k^*, X_i^*)$ 与权重 t_{ki}^* 成正比，比例因子是 $\sqrt{\lambda_k^*}$。这意味着 t_{ki}^* 的绝对值大小顺序与 $\rho(Y_k^*, X_i^*)$ 的绝对值大小顺序是一致的。因此，在基于相关矩阵的PCA中，人们常常直接用特征向量的分量（即载荷系数 t_{ki}^*）来解释主成分，因为它们直接反映了原始标准化变量对主成分的贡献，并且它们的相对大小也反映了与主成分的相关性强弱。但严格来说，相关系数 $\rho(Y_k^*, X_i^*)$ 才是标准的因子载荷量。

5. **因子载荷量的平方的意义**：
   - 因子载荷量的平方 $\rho^2(Y_k, X_i)$ 表示第 k 个主成分 Y_k 的方差中，可以由原始变量 X_i 解释的比例（或者反过来说，原始变量 X_i 的方差中，可以由主成分 Y_k 解释的比例，这取决于具体的因子分析模型语境，但在PCA中，前者更常见）。
   - 对于基于相关矩阵的PCA，我们有 $\sum_{i=1}^{p} \rho^2(Y_k^*, X_i^*) = \lambda_k^*$。这意味着第 k 个主成分 Y_k^* 的方差（即 λ_k^*）等于它与所有原始标准化变量 X_i^* 的因子载荷量的平方和。这表明了主成分是如何从所有原始变量中“提取”方差的。

总结一下，因子载荷量在主成分分析中的核心作用是：

- **量化主成分与原始变量的线性关系强度和方向。**
- **作为解释和命名主成分的主要依据，帮助我们理解每个主成分的实际意义。**

在阅读PCA的结果报告时，因子载荷量（或在相关矩阵PCA中，特征向量的各分量）是非常重要的信息，它们告诉我们降维后的新变量（主成分）究竟代表了原始数据的哪些方面。

再强调一下Y_k与X_i的相关系数这回事

在主成分分析中，有：

$Y_k = \sum_{i=1}^{p} t_{ki} X_i$

> 这里的t_{ki}并不是$\rho(Y_k, X_k)$，原因有：
>
> 1. $\{X_i\}$之间可能有相关性；
> 2. $\rho \in [-1, 1]$可能不满足。

因此，$\rho(Y_k, X_i) = \frac{\sqrt{\lambda_k}}{\sqrt{\sigma_{ii}}} t_{ki}$才是它们的相关系数，虽然和$t_{ki}$有关，但需要乘以一个比例系数。

> 肖宇翔
>
> 2025年5月23日