第13章 数据降维

主成分分析

- 主成分的基本思想与理论
- 主成分的几何意义
- 总体主成分及其性质
- 样本主成分的导出

- 主成分分析是把各变量之间互相关联的复杂关系进行简化分析的方法。
- 在社会经济的研究中，为了全面系统的分析和研究问题，必须考虑许多经济指标，这些指标能从不同的侧面反映我们所研究的对象特征，但许多指标在某种程度上都存在信息重叠，具有一定的相关性。

- 主成分分析试图在力保数据信息丢失最少的原则下，对这种多变量的截面数据表进行最佳综合简化，也就是说，对高维变量空间进行降维处理。
- 很显然，识辨系统在一个低维空间要比在一个高维空间容易得多。

一、主成分分析的基本思想

多元统计分析方法运用降维的思想，在损失很少信息的情况下，把原来具有一定相关性的指标（如p个指标）进行线性变换，形成少数几个（如m个）互不相关的综合指标，用主成分来代替原来指标进行分析。

每一个主成分都是原始变量的线性组合，

主成分分析又称为主分量分析或主轴分析。

主要目的是数据的压缩（降维）和数据的解释

二、主成分分析的基本理论

设随机变量**X**= $(X_1, X_2, \cdots, X_p)'$，均值为**μ**，协方差为∑。

对**X**进行线性变换，可以形成新的综合变量，用**Y**表示

$$\begin{cases} Y_1 = u_{11}X_1 + u_{12}X_2 + \cdots + u_{1p}X_p = u_1'X \\ Y_2 = u_{21}X_1 + u_{22}X_2 + \cdots + u_{2p}X_p = u_2'X \\ \cdots\cdots\cdots \\ Y_p = u_{p1}X_1 + u_{p2}X_2 + \cdots + u_{pp}X_p = u_p'X \end{cases}$$

上述线性变换的约束条件：

- 每个主成分的系数平方和为1。
- 主成分之间不相关，即无重叠的信息。
- 主成分的方差依次递减，重要性依次递减.

1、$u_i' u_i$ **=1**，即 $u_{i1}^2 + u_{i2}^2 + \cdots + u_{ip}^2 = 1$，$(i = 1, 2, 3, \cdots p)$；

2、Y_i 与 Y_j 相互无关 $(i \neq j, i, j = 1, 2, 3, \cdots p)$；

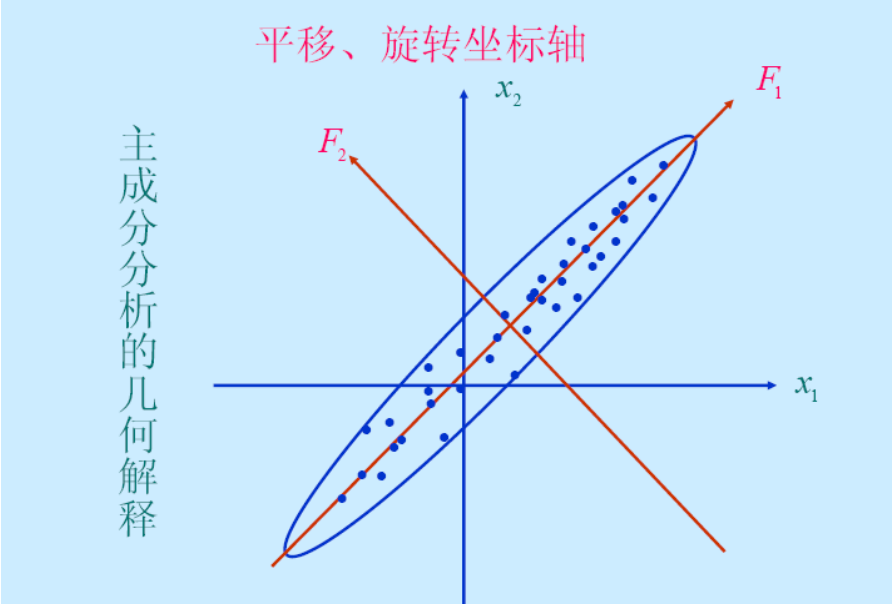
3、Y_1 是 X_1，X_2，…，X_p 的一切满足原则 **1** 的线性组合中方差最大者；Y_2 是与 Y_1 不相关的 X_1，X_2，…，X_p 所有的线性组合中方差最大者；…，Y_p 是与 $Y_1, Y_2, \cdots, Y_{p-1}$ 都不相关的 X_1，X_2，…，X_p 的所有线性组合中方差最大者。

主成分与原始变量之间有如下基本关系：

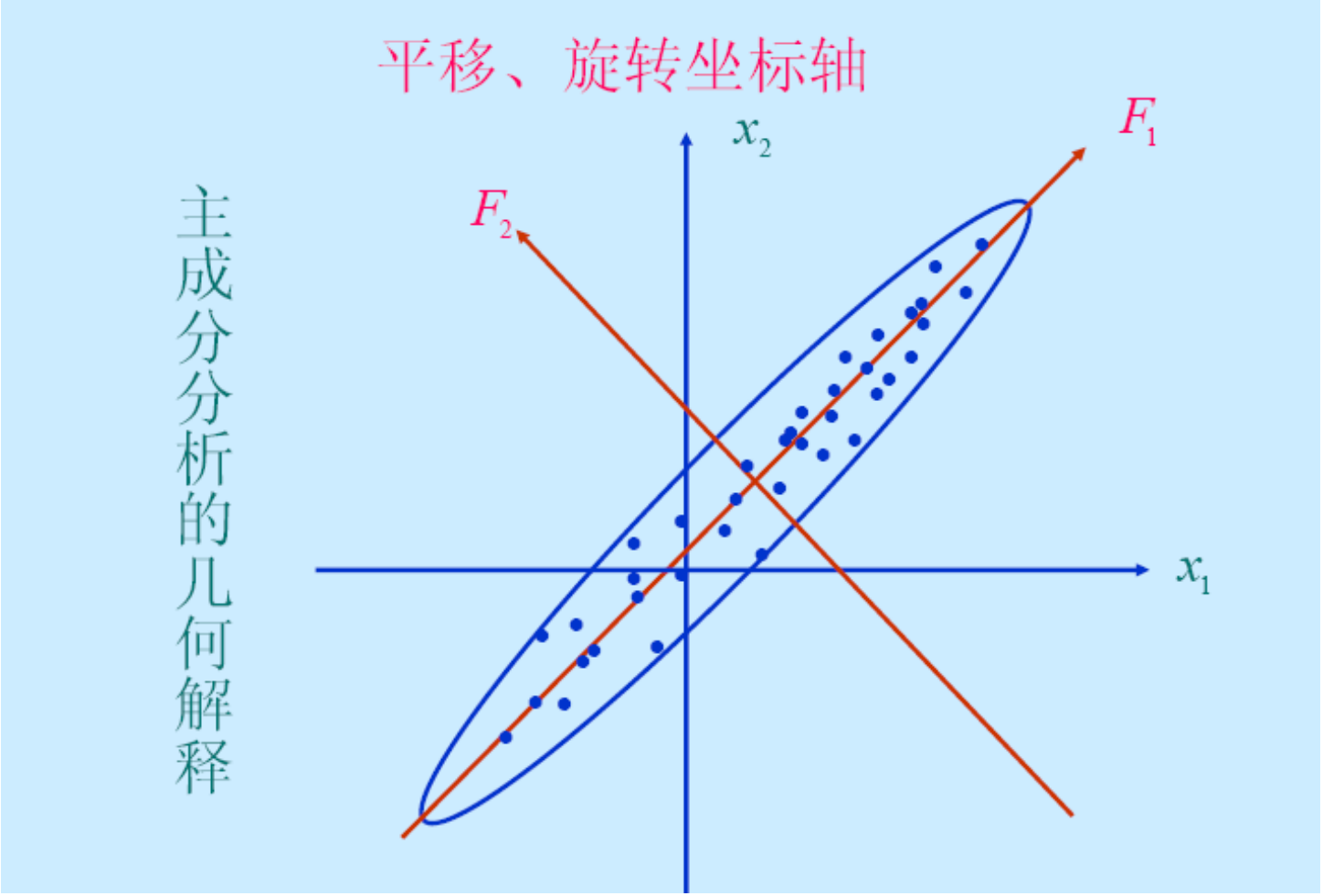
- 每一主成分都是各原始变量的线性组合
- 主成分的数目大大少于原始变量的数目
- 主成分保留了原始变量绝大多数的信息
- 各主成分之间互不相关

第2节 主成分的几何意义

为了方便，我们在二维空间中讨论主成分的几何意义。设有N个样本，每个样本有两个观测变量X1、X2，在由二维变量组成的坐标空间中，*N*个样本散布在一个椭圆内（阴影部分），当两个变量的相关性越大，则这个椭圆就越扁 .由图可以看出这n个样本点无论是沿着x1 轴方向或x2轴方向都具有较大的离散性，其离散的程度可以分别用观测变量x1 的方差和x2 的方差定量地表示。显然，如果只考虑xl和x2 中的任何一个，那么包含在原始数据中的经济信息将会有较大的损失。

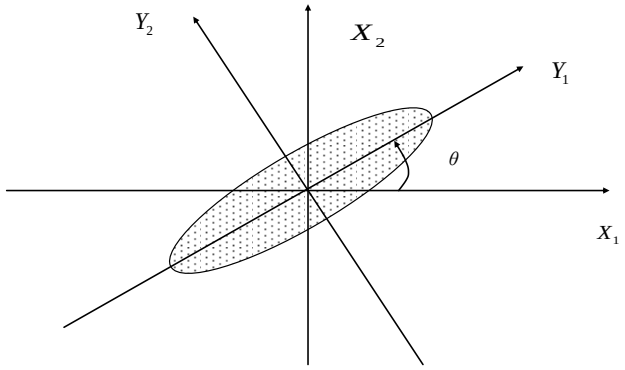


如果我们将xl 轴和x2轴先平移，再同时按逆时针方向旋转θ角度，得到新坐标轴Fl和F2。Fl和F2是两个新变量。



针方向旋转 θ 角度，得到新坐标轴 Y_1 和 Y_2 (这相当于是新坐标系下点的坐标)，坐标轴旋转公式如下：

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



其矩阵形式为：

$$\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = U \bullet X$$

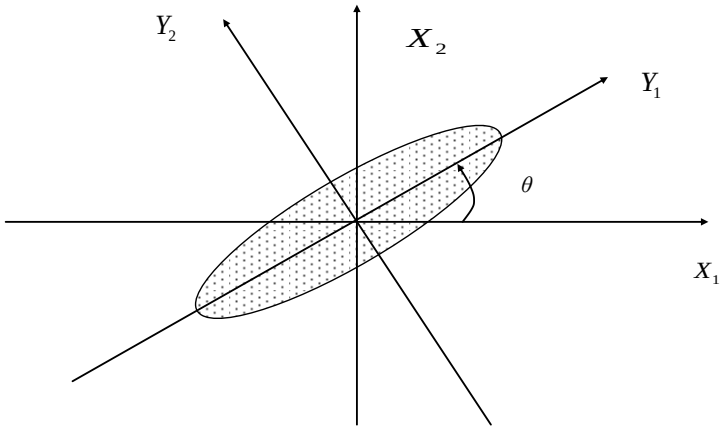
式中，U 为旋转变换矩阵，有上式可知它是正交矩阵，即满足：

$$U' = U^{-1}, U' \bullet U = I$$

经过这样的旋转后， N 个样品点在 Y_1 轴上的离散程度最大，变量 Y_1 代表了绝大多数信息

从代数学的角度来看，主成分就是p个原始变量的线性组合；而从几何的角度来看，这些线性组合正是把由 $X_1, X_2, \cdots, X_p$ 构成的坐标系经旋转而产生新的坐标系，新坐标使之通过样本方差最大的方向（或者说具有最大的样本方差)。

主成分分析在几何上就直观地表现为寻找P维空间中椭球的主轴问题。



第3节 主成分的推导及性质

线性代数中有关定理的回顾

1、若A是P×P阶实对称阵，则一定可以找到正交阵P使

$$P^{-1}AP=\Lambda=\begin{pmatrix}\lambda_1 & & 0\\ & \ddots & \\ 0 & & \lambda_p\end{pmatrix}$$

其中：$\lambda_1,\lambda_2,\cdots\lambda_p$ 是A的特征根

2. 若上述实对称矩阵A的特征根所对应的单位特征向量为 $\gamma_1,\gamma_2,\cdots\gamma_p$

$$P=(\gamma_1,\gamma_2,\cdots\gamma_p)=\begin{pmatrix}\gamma_{11} & \gamma_{12} & \cdots & \gamma_{1p}\\ \gamma_{21} & \gamma_{22} & \cdots & \gamma_{2p}\\ \vdots & \vdots & \vdots & \vdots\\ \gamma_{p1} & \gamma_{p2} & \cdots & \gamma_{pp}\end{pmatrix}$$

则不同特征根所对应的特征向量是正交的，即

$$\gamma_i'\gamma_j=0,(i\neq j)$$

当然，$\gamma_i'\gamma_i=1,(i=1,2,3,\cdots,p)$

即P为正交方阵，其性质有：$PP'=P'P=I$

$$P^{-1}=P'$$

推导 $A=P\Lambda P^{-1}=P\Lambda P'$

3、谱分解式

（1）设A是一个P×P对称阵，则A的谱分解式为：

$$A=\sum_{i=1}^{p}\lambda_i\gamma_i\gamma_i'=\lambda_1\gamma_1\gamma_1'+\lambda_2\gamma_2\gamma_2'+\cdots+\lambda_p\gamma_p\gamma_p'$$

其中：$\lambda_1,\lambda_2,\cdots\lambda_p$ 是A的特征值，$\gamma_1,\gamma_2,\cdots\gamma_p$ 是相对应的标准正交特征向量

（2）平方根矩阵的谱分解式为：

$$A^{\frac{1}{2}}=\sum_{i=1}^{p}\sqrt{\lambda_i}\gamma_i\gamma_i'=P\Lambda^{\frac{1}{2}}P'$$

平方根矩阵的性质：

$$A^{\frac{1}{2}}A^{\frac{1}{2}}=A$$

$$(A^{\frac{1}{2}})'=A^{\frac{1}{2}}$$

主成分的推导及性质

设 $$Y = u_1X_1 + u_2X_2 + \cdots + u_pX_p = u'X$$

其中，$u = (u_1, u_2, \cdots u_p)'$ $X = (X_1, X_2, \cdots X_p)'$

求主成分就是寻找X的线性函数u'X使相应的方差最大，即使

$$Var(u'X) = E(u'X - E(u'X))(u'X - E(u'X))'$$

$$= u'E(X - E(X))(X - E(X))'u$$

$$= u'\Sigma u$$

最大,且 $u'u = 1$ 。

结论：

主成分方差的最大值 在Σ的最大特征值所对应的特征向量处达到。依次类推，使第P个主成分方差$Var(Y_P)$ 达到最大值是在Σ的第P个特征值所对应的特征向量处达到。

主成分的推导
（一）第一主成分

设 X 的自协方差阵为Σ。由于Σ为非负定的对称阵，
则利用线性代数的知识可得，必存在正交阵 P，

使得 $P'\sum P=\begin{bmatrix}\lambda_1 & & 0\\ & \ddots & \\ 0 & & \lambda_p\end{bmatrix}$

其中，$\lambda_1,\lambda_2,\cdots\lambda_p$ 为Σ的特征根，不妨假设 $\lambda_1\geq\lambda_2\geq,\cdots\geq\lambda_p$ 。

而 P 恰好是由特征根相对应的特征向量所组成的正交阵。

$$P=(\gamma_1,\cdots\gamma_p)=\begin{bmatrix}\gamma_{11} & \gamma_{12} & \cdots & \gamma_{1p}\\ \gamma_{21} & \gamma_{22} & \cdots & \gamma_{2p}\\ \vdots & & & \vdots\\ \gamma_{p1} & \gamma_{p2} & \cdots & \gamma_{pp}\end{bmatrix}$$

$\gamma_i=(\gamma_{1i},\cdots\gamma_{pi})'$ **i=1,2,3 p**

也就是说，$\gamma_1,\gamma_2,\cdots\gamma_p$ 为矩阵Σ各特征值对应的标准正交特征向量。

则第 i 个主成分就是：$Y_i=\gamma_{1i}X_1+\gamma_{2i}X_2+\cdots+\gamma_{pi}X_p$

下面我们来看，是否由P的第一列元素所构成的原始变量的线性组合有最大的方差？

设有 P 维正交向量，$\gamma_1=(\gamma_{11},\gamma_{21},\cdots\gamma_{p1})'$，

则第 1 个主成分就是：

$$Y_1=\gamma_{11}X_1+\gamma_{21}X_2+\cdots+\gamma_{p1}X_p=\gamma_1'X$$

$$Var(Y_1)=\gamma_1'\Sigma\gamma_1=\gamma_1'P\begin{bmatrix}\lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_p\end{bmatrix}P'\gamma_1$$

$$=\gamma_1'[\gamma_1,\gamma_2,\cdots\gamma_p]\begin{bmatrix}\lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_p\end{bmatrix}\begin{bmatrix}\gamma_1' \\ \gamma_2' \\ \vdots \\ \gamma_p'\end{bmatrix}\gamma_1$$

$$= \sum_{i=1}^{p} \lambda_i \gamma_1' \gamma_i \gamma_i' \gamma_1$$

$$= \sum_{i=1}^{p} \lambda_i (\gamma_1' \gamma_i)^2$$

$$\leq \lambda_1 \sum_{i=1}^{p} (\gamma_1' \gamma_i)^2$$

$$= \lambda_1 \sum_{i=1}^{p} \gamma_1' \gamma_i \gamma_i' \gamma_1$$

$$= \lambda_1 \gamma_1' P P' \gamma_1 = \lambda_1 \gamma_1' \gamma_1 = \lambda_1$$

因此，当且仅当$u_1 = \gamma_1$时，即$Y_1 = \gamma_{11} X_1 + \gamma_{21} X_2 + \cdots + \gamma_{p1} X_p$时，有最大方差$\lambda_1$

如果第一主成分的信息不够，则需要寻找第二主成分。

（二）第二主成分

在约束条件 $\mathrm{cov}(Y_1, Y_2) = 0$ 下，

寻找第二主成分：

$$Y_2 = \gamma_{12} X_1 + \gamma_{22} X_2 + \cdots + \gamma_{p2} X_p$$

因为 $\mathrm{cov}(Y_1, Y_2) = \mathrm{cov}(\gamma_1' X, \gamma_2' X) = \gamma_2' \Sigma \gamma_1 = \lambda_1 \gamma_2' \gamma_1 = 0$

所以 $\gamma_2' \gamma_1$ =0，

则，对 **P** 维向量 γ_2，有

$$\mathrm{var}(Y_2) = \gamma_2'\Sigma\gamma_2 = \sum_{i=1}^{p}\lambda_i\gamma_2'\gamma_i\gamma_i'\gamma_2$$

$$= \sum_{i=1}^{p}\lambda_i(\gamma_2'\gamma_i)^2 \le \lambda_2\sum_{i=2}^{p}(\gamma_2'\gamma_i)^2 = \lambda_2\sum_{i=1}^{p}\gamma_2'\gamma_i\gamma_i'\gamma_2$$

$$= \lambda_2\gamma_2'PP'\gamma_2 = \lambda_2\gamma_2'\gamma_2 = \lambda_2$$

因此，第二主成分 **Y2** 有次大方差 λ_2

其余主成分的推导同理。

写为矩阵形式：

$$\mathbf{Y=U'X=}\begin{bmatrix} u_{11} & u_{12} & \cdots & u_{1p} \\ u_{21} & u_{22} & \cdots & u_{2p} \\ \vdots & & & \vdots \\ u_{p1} & u_{p2} & \cdots & u_{pp} \end{bmatrix}\begin{bmatrix} X_1 \\ X_2 \\ \\ X_p \end{bmatrix}=\begin{bmatrix} \gamma_1' \\ \gamma_2' \\ \\ \gamma_p' \end{bmatrix}\begin{bmatrix} X_1 \\ X_2 \\ \\ X_p \end{bmatrix}$$

其中：

$$X=(X_1,X_2,\cdots,X_p)'$$

主成分的性质

1、期望

2、方差

- **性质1，Y的协方差阵是对角阵 Λ**

- **性质2，** 记 $\Sigma=(\sigma_{ij})_{p\times p}$，有 $\sum_{i=1}^{p}\lambda_i=\sum_{i=1}^{p}\sigma_{ii}$，

  通常称 $\sum_{i=1}^{p}\sigma_{ii}$ 为原总体 X 的总方差（或称总惯量）。

3、精度分析

称 $a_k = \dfrac{\lambda_k}{\sum_{i=1}^{p} \lambda_i} (k = 1, 2, 3, \cdots, p)$ 为第 k 个主成分 Y_k 的方差贡献率；

又称 $\dfrac{\sum_{i=1}^{m} \lambda_i}{\sum_{i=1}^{p} \lambda_i}$ 为主成分 $Y_1, Y_2, \cdots, Y_m$ $(m < p)$ 的累积贡献率。

- 我们进行主成分分析的目的之一是希望用尽可能少的主成分Y_1，Y_2，…，Y_k（$k \leq p$）代替原来的P个指标。
- 到底应该选择多少个主成分，在实际工作中，主成分个数的多少取决于能够反映原来变量80%以上的信息量为依据，即当累积贡献率≥80%时的主成分的个数就足够了。最常见的情况是主成分为2到3个。

4、原始变量与主成分之间的相关系数

第k个主成分Y_k与原始变量X_i的相关系数$\rho(Y_k, X_i)$称作因子载荷量。

因子载荷量是主成分分析中非常重要的解释依据，因子负荷量的绝对值大小刻画了该主成分的主要意义及其成因。由相关性质我们可知，因子载荷量与系数向量成正比。

- 性质3， $\rho(Y_k, X_i) = u_{ki}\sqrt{\lambda_k}\Big/\sqrt{\sigma_{ii}}$

$$k, i = 1, 2, \cdots, p$$

由此可知，因子负荷量与系数向量成正比，与X_i的标准差成反比关系。

5、原始变量被主成分的提取率

- 前面我们讨论了主成分的贡献率和累计贡献率，他度量了Y1，Y2，……，Ym分别从原始变量X1，X2，……XP中提取了多少信息。
- 那么X1，X2，……XP各有多少信息分别Y1，Y2，……，Ym被提取了。应该用什么指标来度量？我们考虑到当讨论Y1分别与X1，X2，……XP的关系时，可以讨论Y1分别与X1，X2，……XP的相关系数，但是由于相关系数有正有负，所以只有考虑相关系数的平方。

- 性质4，

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

- 性质5，

$$\sum_{i=1}^{p} \rho^2(Y_k, X_i) = \frac{1}{\sigma_{ii}} \sum_{i=1}^{p} \lambda_k u_{ki}^2 = 1$$

定义3：

将前 m 个主成分 $Y_1, Y_2, \cdots, Y_m$ 对原始变量 X_i 的贡献率 $\nu_i^{(m)}$

定义为 X_i 与 $Y_1, Y_2, \cdots, Y_m$ 的相关系数的平方，它等于：

$$\nu_i^{(m)} = \sum_{k=1}^{m} \lambda_k u_{ik}^2 \Big/ \sigma_{ii}$$

这一定义说明了前m 个主成分提取了原始变量Xi中Vi的信息。

第4节 样本主成分的导出

当总体协方差Σ与相关阵R通常是未知的，需要通过样本数据估计。

设有n个样品，每个样品有p个指标，这样共得到np个数据，

原始资料矩阵如下：

$$X \equiv \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

记：

$$S = \frac{1}{n-1}\sum_{k=1}^{n}(x_{ki}-\bar{x}_i)(x_{ki}-\bar{x}_i)' \qquad R = (r_{ij})_{p\times p}$$

$$\bar{x}_i = \frac{1}{n}\sum_{k=1}^{n} x_{ki} \qquad r_{ij} = \frac{S_{ij}}{\sqrt{S_{ii}S_{jj}}}$$

若原始资料阵 X 是经过标准化处理的，则由矩阵 X 求得的协方差阵就是相关矩阵，即 S 与 R 完全相同。

- **由相关阵R出发求解主成分：**

  根据总体主成分的定义，主成分Y的协方差是：

$$\mathrm{cov}(Y) = u' \mathrm{cov}(X) u = u \Sigma u' = \Lambda$$

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 & 0 & \cdots & 0 \\ 0 & \lambda_2 & 0 & \cdots & 0 \\ 0 & 0 & \lambda_3 & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & \lambda_p \end{bmatrix}$$

- 假定资料矩阵为已标准化的数据矩阵，则可由相关矩阵代替协方差矩阵，于是上式可表示为：

$$u'Ru = \Lambda$$

用u'左乘上式，得

$$Ru' = u'\Lambda$$

即

$$\begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1p} \\ r_{21} & r_{22} & \cdots & r_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ r_{p1} & r_{p2} & \cdots & r_{pp} \end{bmatrix} \begin{bmatrix} u_{11} & u_{12} & \cdots & u_{1p} \\ u_{21} & u_{22} & \cdots & u_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ u_{p1} & u_{p2} & \cdots & u_{pp} \end{bmatrix}$$

$$= \begin{bmatrix} u_{11} & u_{12} & \cdots & u_{1p} \\ u_{21} & u_{22} & \cdots & u_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ u_{p1} & u_{p2} & \cdots & u_{pp} \end{bmatrix} \begin{bmatrix} \lambda_1 & & & 0 \\ & \lambda_2 & & \\ & & \ddots & \\ 0 & & & \lambda_p \end{bmatrix}$$

把上式全部展开得到 p^2 个方程，这里只是考虑在矩阵

乘积中由第一列得到的 p 个方程：

$$\begin{cases} r_{11}u_{11}+r_{12}u_{12}+\cdots+r_{1p}u_{1p}=u_{11}\lambda_1 \\ r_{21}u_{11}+r_{22}u_{12}+\cdots+r_{2p}u_{1p}=u_{12}\lambda_1 \\ \cdots\cdots \\ r_{p1}u_{11}+r_{p2}u_{12}+\cdots+r_{pp}u_{1p}=u_{1p}\lambda_1 \end{cases}$$

整理得到：

$$\begin{cases} (r_{11}-\lambda_1)u_{11}+r_{12}u_{12}+\cdots+r_{1p}u_{1p}=0 \\ r_{21}u_{11}+(r_{22}-\lambda_1)u_{12}+\cdots+r_{2p}u_{1p}=0 \\ \cdots\cdots \\ r_{p1}u_{11}+r_{p2}u_{12}+\cdots+(r_{pp}-\lambda_1)u_{1p}=0 \end{cases}$$

为了得到上面齐次方程的非零解，根据线性方程组的理论知，要求线性系数矩阵行列式为0，即：

$$\begin{vmatrix} r_{11}-\lambda_1 & r_{12} & \cdots & r_{1p} \\ r_{21} & r_{22}-\lambda_1 & \cdots & r_{2p} \\ \vdots & \vdots & & \vdots \\ r_{p1} & r_{p2} & \cdots & r_{pp}-\lambda_1 \end{vmatrix} = 0$$

即 $$\left|R-\lambda_1 I\right| = 0$$

对于 $\lambda_2,\cdots,\lambda_p$ 完全可以得到类似的方程，于是，

所求的新的综合变量（主成分）的方差 λ_i（$i=1,2,\cdots,p$）是 $\left|R-\lambda I\right|=0$

的 p 个根，λ 为相关矩阵的特征值，相应的各个 u_{ij} 是其特征向量的分量。

因为 R 为正定矩阵，所以其特征根都是非负实数，

将它们按大小顺序排列 $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$，其相应

的特征向量记为 $\gamma_1, \gamma_1, \cdots, \gamma_p$，则相对于 Y_1 的方差为：

$$\mathrm{var}(Y_1) = \mathrm{var}(\gamma_1' X) = \lambda_1$$

同理有

$$\mathrm{var}(Y_i) = \mathrm{var}(\gamma_i' X) = \lambda_i$$

即对于 Y_1 有最大方差，Y_2 有次大方差，---，并且协方差为：

$$\mathrm{cov}(Y_i, Y_j) = \mathrm{cov}(\gamma_i' X, \gamma_j' X) = \gamma_i' R \gamma_j'$$

$$= \gamma_i' \left(\sum_{\alpha=1}^{p} \lambda_\alpha \gamma_\alpha \gamma_\alpha' \right) \gamma_j' = \sum_{\alpha=1}^{p} \lambda_\alpha (\gamma_i' \gamma_\alpha)(\gamma_\alpha' \gamma_j') = 0 \quad (i \neq j)$$

由此可有新的综合变量（主成分）$Y_1, Y_2, \cdots, Y_p$ 彼此不相关，并且 Y_i 的方差为 λ_i，则 $Y_1 = \gamma_1' X, Y_2 = \gamma_2' X, \cdots, Y_p = \gamma_p' X$ 分别被称为第一、第二、…、第 p 个主成分。

主成分分析的相关问题

1.数据是否标准化

- 对度量单位不同的指标或取值范围彼此差异非常大的指标，应考虑先将数据标准化，再由其协方差阵出发进行主成分分析。
- 对同度量或是取值范围在同量级的数据，可直接从未标准化数据的协方阵求主成分。

2.主成分分析不要求数据来自正态总体

主成分分析是对矩阵结构（协方差阵或相关阵）的分析，主要用到矩阵运算、矩阵对角化和矩阵谱分解技术，未涉及总体分布问题。

3.主成分分析对重叠信息的剔除是无能为力的

主成分分析对原始变量的重叠信息无法剔除，这就要求在选取初始变量时要避免选取有过多存在重叠信息的变量，对高度多重共线性的变量要注意其主成分分析结果。

当然，主成分分析适用于大部分变量相关系数大于0.3，小于0.3时，主成分分析效果差。

主成分分析并不完美！

因子分析

引言

因子分析(factor analysis)是一种数据简化的技术。它是利用降维的思想，通过研究原始变量相关矩阵内部的依赖关系，把一些具有错综复杂关系的变量归结为少数几个综合因子的一种多变量统计分析方法。

这几个综合因子能够反映原来众多变量的主要信息。原始的变量是可观测的显性变量，而综合的因子是**不可观测的潜在变量**，称为因子。

因子分析模型

一、数学模型

设有n个样品，每个样品观测p个指标，这p个指标之间有较强的相关性。如果：

（1）$X=(X_1,X_2,\cdots,X_p)'$是可观测随机变量，且均值向量$E(X)=0$，

协方差矩阵$\mathrm{cov}(X)=\Sigma$，且协方差矩阵Σ与相关阵R相等；

（2）$F=(F_1,F_2,\cdots,F_m)'$ （$m<p$）是不可观测的变量，其均值向量$E(F)=0$，

方差矩阵$\mathrm{cov}(F)=I$，即向量F的各分量是相互独立的；

$$D(F)=\begin{bmatrix}1 & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1\end{bmatrix}=I$$

即 $F_1, F_2, \cdots, F_m$ 互不相关，方差为1。

（3）$\varepsilon=(\varepsilon_1,\varepsilon_2,\cdots,\varepsilon_p)'$与$F$相互独立，且$E(\varepsilon)=0$，

ε的协方差阵Σ_ε是对角矩阵：

$$D(\varepsilon)=\begin{bmatrix}\sigma_1^2 & & & \\ & \sigma_2^2 & & \\ & & \ddots & \\ & & & \sigma_p^2\end{bmatrix}$$

即ε的各个分量也是相互独立的。

则模型

$$\begin{cases} X_1 = a_{11}F_1 + a_{12}F_2 + \cdots + a_{1m}F_m + \varepsilon_1 \\ X_2 = a_{21}F_1 + a_{22}F_2 + \cdots + a_{2m}F_m + \varepsilon_2 \\ \cdots\cdots \\ X_p = a_{p1}F_1 + a_{p2}F_2 + \cdots + a_{pm}F_m + \varepsilon_p \end{cases}$$

称为因子 模型。

其中 $F_1, F_2, \cdots, F_m$ 为公共因子，是不可观测的变量，他们的系数称为因子载荷。 ε_i 是特殊因子，是不能被前m个公共因子包含的部分。

用矩阵的表达方式

$$X = AF + \varepsilon$$

$$E(\mathbf{F}) = \mathbf{0}$$

$$E(\boldsymbol{\varepsilon}) = \mathbf{0}$$

$$Var(\mathbf{F}) = \mathbf{I}$$

$$Var(\boldsymbol{\varepsilon}) = diag(\sigma_1^2, \sigma_2^2, \cdots, \sigma_p^2)$$

$$\mathrm{cov}(\mathbf{F}, \boldsymbol{\varepsilon}) = E(\mathbf{F}\boldsymbol{\varepsilon}') = \mathbf{0}$$

比较：

- 主成分分析与因子分析：主成分分析仅仅是变量变换，而因子分析需要构造因子模型
  - 主成分分析:原始变量的线性组合表示新的综合变量，即主成分；
  - 因子分析：潜在的假设公因子和随机影响变量的线性组合表示原始变量。

例如，在企业形象或品牌形象的研究中，消费者可以通过一个有24个指标构成的评价体系，评价百货商场24个方面的优劣。现有n 家商场，请你对这n家商场的企业形家作因子分析。

因子分析可以通过24个变量，找出消费者主要关心的三个方面，即反映商店环境、商店服务水平和商品价格的三个潜在的因子，对商店进行综合评价。而这三个公共因子可以表示为：

$$x_i = \mu_i + \alpha_{i1}F_1 + \alpha_{i2}F_2 + \alpha_{i3}F_3 + \varepsilon_i \qquad i = 1, \cdots, 24$$

称 F_1、F_2、F_3 是不可观测的潜在因子。24个变量共享这三个因子，但是每个变量又有自己的个性，不被包含的部分 ε_i ，称为特殊因子。

二、因子分析模型的性质

1、原始变量X的协方差矩阵的分解

$$\because \quad \mathbf{X}-\boldsymbol{\mu}=\mathbf{AF}+\boldsymbol{\varepsilon}$$

$$\therefore \quad Var(\mathbf{X}-\boldsymbol{\mu})=\mathbf{A}Var(\mathbf{F})\mathbf{A}'+Var(\boldsymbol{\varepsilon})$$

$$\boldsymbol{\Sigma}_x=\mathbf{AA}'+\mathbf{D}$$

A是因子模型的系数

$$Var(\boldsymbol{\varepsilon})=\mathbf{D}=diag(\sigma_1^2,\sigma_2^2,\cdots,\sigma_p^2)$$

D的主对角线上的元素值越小，则公共因子共享的成分越多。

2、模型不受计量单位的影响(模型形式不变)

将原始变量X做变换$X^*=CX$,这里

$C=diag(c_1,c_2,\ldots,c_n),c_i>0$。

$$C(X-\mu)=C(AF+\varepsilon)$$

$$CX=C\mu+CAF+C\varepsilon$$

$$X^*=C\mu+CAF+C\varepsilon$$

$$X^*=\mu^*+A^*F^*+\varepsilon^* \qquad F^*=F$$

$$E(\mathbf{F}^*) = \mathbf{0}$$

$$E(\boldsymbol{\varepsilon}^*) = \mathbf{0}$$

$$Var(\mathbf{F}^*) = \mathbf{I}$$

$$Var(\boldsymbol{\varepsilon}^*) = diag(\sigma_1^2, \sigma_2^2, \cdots, \sigma_p^2)$$

$$\text{cov}(\mathbf{F}^*, \boldsymbol{\varepsilon}^*) = E(\mathbf{F}^* \boldsymbol{\varepsilon}^{*\prime}) = \mathbf{0}$$

3、因子载荷不是惟一的

设T为一个p×p的正交矩阵，令**$A^*=AT$**，**$F^*=T'F$**，则模型可以表示为

$\mathbf{X}^* = \boldsymbol{\mu} + \mathbf{A}^*\mathbf{F}^* + \boldsymbol{\varepsilon}$ 且满足条件因子模型的条件

$$E(\mathbf{T}'\mathbf{F}) = \mathbf{0} \qquad E(\boldsymbol{\varepsilon}) = \mathbf{0}$$

$$Var(\mathbf{F}^*) = Var(\mathbf{T}'\mathbf{F}) = \mathbf{T}'Var(\mathbf{F})\mathbf{T} = \mathbf{I}$$

$$Var(\boldsymbol{\varepsilon}) = diag(\sigma_1^2, \sigma_2^2, \cdots, \sigma_p^2)$$

$$\mathrm{cov}(\mathbf{F}^*, \boldsymbol{\varepsilon}) = E(\mathbf{F}^*\boldsymbol{\varepsilon}') = \mathbf{0}$$

三、 因子载荷矩阵中的几个统计特征

因子载荷 a_{ij} 是第i个变量与第j个公共因子的相关系数

模型为 $X_i = a_{i1}F_1 + \cdots + a_{im}F_m + \varepsilon_i$

在上式的左右两边乘以F_j ，再求数学期望

$$E(X_iF_j) = a_{i1}E(F_1F_j) + \cdots + \alpha_{ij}E(F_jF_j) + \cdots + a_{im}E(F_mF_j) + E(\varepsilon_iF_j)$$

根据公共因子的模型性质，有

$\gamma_{x_iF_j} = \alpha_{ij}$ （载荷矩阵中第i行，第j列的元素）反映了第i个变量与第j个公共因子的相关重要性。绝对值越大，相关的密切程度越高。

因子载荷矩阵的估计方法-主成分分析方法

设随机向量 $\mathbf{x}=(x_1,x_2,\cdots,x_p)'$ 的均值为$\boldsymbol{\mu}$，协方差为$\boldsymbol{\Sigma}$，$\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_p\geq 0$为$\Sigma$的特征根，$\mathbf{u_1},\mathbf{u_2},\cdots,\mathbf{u_p}$ 为对应的标准化特征向量，则

$$\boldsymbol{\Sigma}=\mathbf{AA}'+\mathbf{D}=\mathbf{U}\begin{bmatrix}\lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_p\end{bmatrix}\mathbf{U}'$$

$$\begin{bmatrix}\mathbf{u_1} & \mathbf{u_2} & \cdots & \mathbf{u_p}\end{bmatrix}\begin{pmatrix}\lambda_1 & & 0\\ & \ddots & \\ 0 & & \lambda_p\end{pmatrix}\begin{bmatrix}\mathbf{u}_1'\\ \mathbf{u}_2'\\ \vdots\\ \mathbf{u}_p'\end{bmatrix}$$

$$=\lambda_1\mathbf{u_1}\mathbf{u}_1'+\lambda_2\mathbf{u_2}\mathbf{u}_2'+\cdots+\lambda_p\mathbf{u_p}\mathbf{u}_p'$$

$$=\begin{bmatrix}\sqrt{\lambda_1}\mathbf{u}_1 & \sqrt{\lambda_2}\mathbf{u}_2 & \cdots & \sqrt{\lambda_p}\mathbf{u}_p\end{bmatrix}\begin{bmatrix}\sqrt{\lambda_1}\mathbf{u}_1'\\ \sqrt{\lambda_2}\mathbf{u}_2'\\ \vdots\\ \sqrt{\lambda_p}\mathbf{u}_p'\end{bmatrix}$$

上式有一个假定，模型中的特殊因子是不重要的，因而从$\mathbf{\Sigma}$的分解中忽略了特殊因子的方差。

因子分析的步骤

选择分析的变量

用定性分析和定量分析的方法选择变量，因子分析的前提条件是**观测变量间有较强的相关性**，因为如果变量之间无相关性或相关性较小的话，他们不会有共享因子, 所以原始变量间应该有较强的相关性。

计算所选原始变量的相关系数矩阵

相关系数矩阵描述了原始变量之间的相关关系。可以帮助判断原始变量之间是否存在相关关系，这对因子分析是非常重要的，因为如果所选变量之间无关系，做因子分析是不恰当的。可以说，**相关系数矩阵是估计因子结构的基础**。

提取公共因子

这一步要确定因子求解的方法和因子的个数。需要根据研究者的设计方案或有关的经验或知识事先确定。因子个数的确定可以根据因子方差的大小。只取方差大于1(或特征值大于1)的那些因子，因为方差小于1的因子其贡献可能很小；按照因子的累计方差贡献率来确定，一般认为要达到60％才能符合要求；

因子旋转

通过坐标变换使每个原始变量在尽可能少的因子之间有密切的关系，这样因子解的实际意义更容易解释,并为每个潜在因子赋予有实际意义的名字。

SVD分解

- 矩阵分解技术能从复杂数据中提取重要的特征信息。它将原始矩阵表示成两个或多个矩阵的乘积。
- 奇异值分解（Singular Value Decomposition，SVD）作为常见的矩阵分解技术，常应用于降维。

SVD算法原理

- 特征值分解：$\boldsymbol{A = Q\Sigma Q^{-1}}$。其中，Q是原矩阵A的特征向量组成的矩阵，Σ是特征值构成的对角阵， 每一个对角线上的元素就是一个特征值。利用特征值和特征向量，我们可以还原出原始矩阵。特征值分解要求矩阵必须是方阵。
- 奇异值分解是一个能适用于任意的矩阵的一种分解的方法。

SVD分解

- 假设有一个m行n列的原始矩阵$\mathbf{A}$，对它进行奇异值分解，可以得到三个矩阵，分别是：$\boldsymbol{U}$、$\boldsymbol{\Sigma}$、$\boldsymbol{V}^{\mathrm{T}}$，

$$\boldsymbol{A}_{m\times n}=\boldsymbol{U}_{m\times m}\boldsymbol{\Sigma}_{m\times n}\boldsymbol{V}^{\mathrm{T}}_{n\times n}$$

- 其中，U称为左奇异向量，V称为右奇异向量，矩阵$\boldsymbol{\Sigma}$表示奇异值矩阵。对于矩阵$\boldsymbol{\Sigma}$，除了对角元素不为0，其他元素都为0，并且对角元素是从大到小排列的。这些对角元素就是奇异值，它对应了原始矩阵$\boldsymbol{A}$的奇异值，表征着数据集的特征值。
- 如何将奇异值和特征值对应起来呢？由于方阵的特征值和特征向量的关系如下：

$$\boldsymbol{A}v=\lambda v$$

- 将矩阵$\mathbf{A}$的转置乘以$\mathbf{A}$，将会得到一个方阵，求该方阵特征值和特征向量，有：

$$\left(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\right)v_i=\lambda_i v_i$$

$$A = U\Sigma V^T \Rightarrow AV = U\Sigma V^T V \Rightarrow AV = U\Sigma \Rightarrow Av_i = \sigma_i u_i \Rightarrow \sigma_i = Av_i / u_i$$

$$A = U\Sigma V^T \Rightarrow A^T = V\Sigma U^T \Rightarrow A^T A = V\Sigma U^T U\Sigma V^T = V\Sigma^2 V^T$$

- 这里得到的$\boldsymbol{V}$，就是右奇异向量。此外，还可以得到：

$$u_i = \frac{1}{\sigma_i} \boldsymbol{A} v_i \quad , \qquad \sigma_i^2 = \lambda_i$$

- 这里的σ就是上述的奇异值，是矩阵$\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$特征值的平方根。$\boldsymbol{U}$就是上述的左奇异向量。
- $\boldsymbol{\Sigma}$中的奇异值跟特征值类似，在矩阵中按从大到小排列，并且奇异值的减少特别快。所以可以仅保留比较大的r个奇异值，也就是说数据集中保留r个重要特征，将其余特征当作噪声或冗余信息。那么，如何确定要保留的奇异值个数r呢？通常是保留到矩阵奇异值平方和总量的90%为止。

$$\boldsymbol{A}_{m\times n} \approx \boldsymbol{U}_{m\times m}\boldsymbol{\Sigma}_{m\times n}\boldsymbol{V}^{\mathrm{T}}_{n\times n}$$

- SVD降维的过程就是舍弃不重要的特征向量的过程，而剩下的特征向量组成的空间即为降维后的空间。它不仅节省存储量，更能降低计算量。因此，在隐性语义索引、图像压缩、推荐系统、金融等领域都有应用。

低维嵌入（简介）

- 传统的线性降维方法，如主成分分析PCA，在把高维数据映射到低维空间时通常不能保留原高维数据的内在非线性结构和特征。
- 非线性的方法如LLE、hessian局部线性嵌入算法HLLE等应运而生。它们的优点是具有较少的参数需要设置，而且使用非迭代方法求解从而可以避免陷入局部极小。

算法原理

- LLE算法基本思想是把整个非线性流形分为许多个小块，这些小块由各数据点与它的k个邻居点构成，每个小块可以看作是拥有局部线性结构的。每个数据点和它的k个邻居点的线性组合系数称为重构权重。LLE认为所有数据点降到低维后邻居关系不变，包括重构权重也不变。因此将小块数据线性降维后，再把它们拼接起来，即可构成高维数据的低维表示。

低维嵌入（简介）

- 低维嵌入（LLE）算法大致如下：
- （1）假设数据由n个m维样本点$\boldsymbol{x}_i(i=1,2,\cdots,n)$组成，按照$k$NN算法计算出每个样本点$\boldsymbol{x}_i$的$k$个近邻点向量$\boldsymbol{x}_{i_j}(j=1,2,\cdots,k)$，目的是利用近邻点$\boldsymbol{x}_{i_j}$线性重构$\boldsymbol{x}_i$。
- （2）基于损失函数定义出重构误差$\varepsilon(\boldsymbol{W})$，其中$w_{ij}$表示第$j$个样本点$\boldsymbol{x}_{i_j}$对于第$i$个样本点$\boldsymbol{x}_{i_i}$重构时的贡献量，即重构权重。

$$\varepsilon(\boldsymbol{W})=\sum_{i=1}^{n}\left|\boldsymbol{x}_i-\sum_{j=1}^{k}w_{ij}\boldsymbol{x}_{i\_j}\right|^2$$

- （3）将所有样本点从m维空间映射到低维的d维空间中。原始m维空间中第i个样本点的权值w_{ij}保持不变，用于低维的d维空间中重构输出样本点向量$\boldsymbol{y}_i$。同时希望在低维空间中的重构误差$\Phi(\boldsymbol{W})$最小，即

$$\Phi(\boldsymbol{W})=\sum_{i=1}^{n}\left|\boldsymbol{y}_i-\sum_{j=1}^{k}w_{ij}\boldsymbol{y}_{i\_j}\right|^2$$