矩阵计算

李宇峰

liyf@nju.edu.cn

人工智能学院

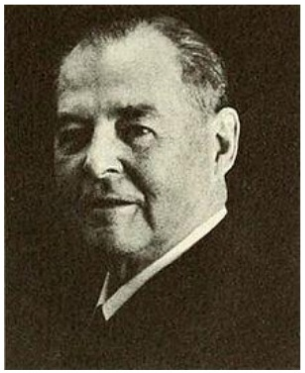
3.2 特征分析的应用—主成分分析

1. 主成分分析

对于多要素的复杂系统，多变量问题是经常会遇到的。变量太多，无疑会增加分析问题的难度与复杂性，而且在许多实际问题中，多个变量之间是具有一定的相关关系的。因此，我们就会很自然地想到，能否在各个变量之间相关关系研究的基础上，用较少的新变量代替原来较多的变量，而且使这些较少的新变量尽可能多地保留原来较多的变量所反映的信息。

- 事实上，这种想法是可以实现的，作为特征分析的一个应用，介绍的主成分分析方法(PCA)就是综合处理这种问题的一种强有力的方法。

Karl Pearson



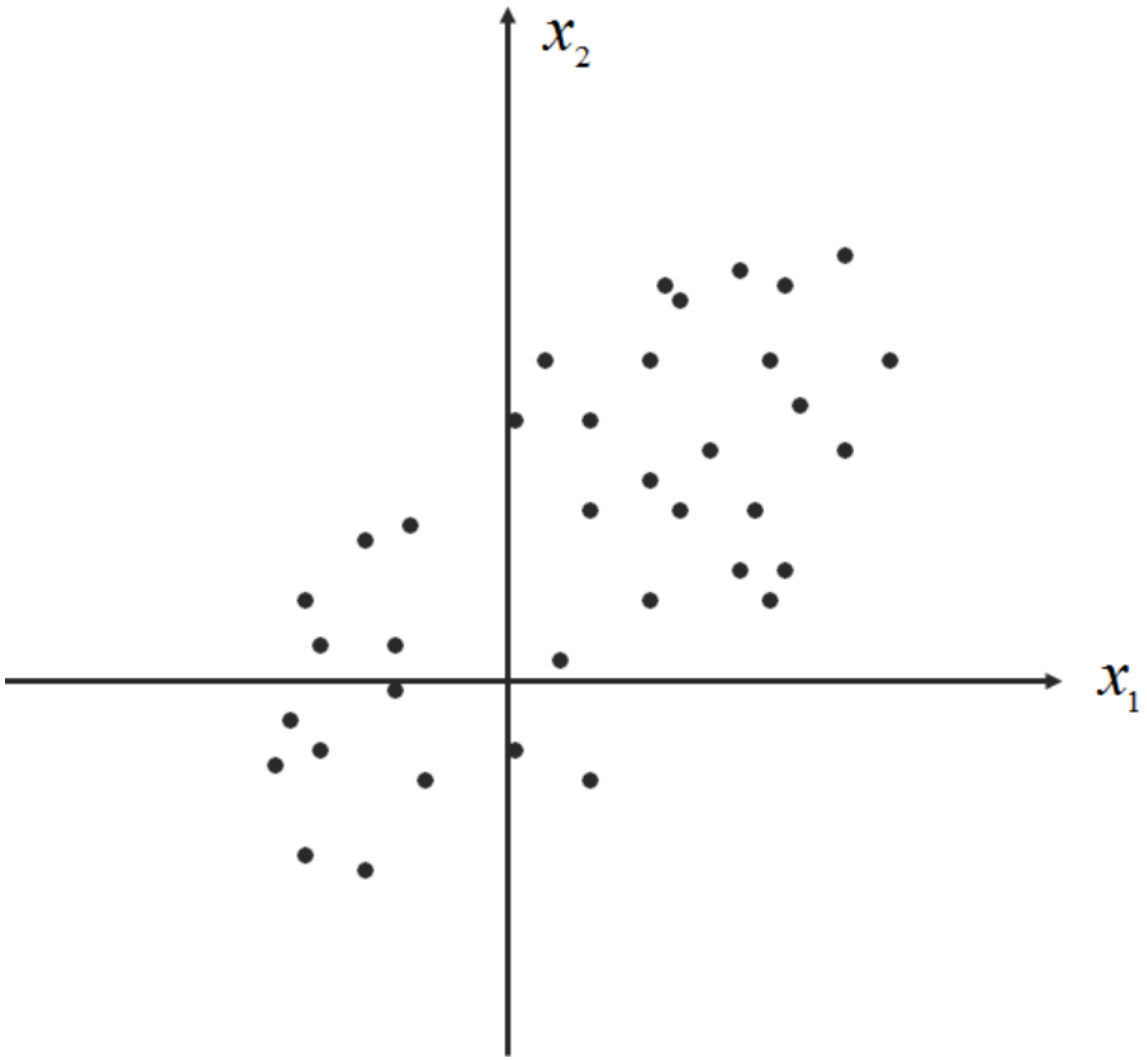
Harold Hotelling

主成分分析方法一个十分著名的例子是英国的统计学家斯通(Richard Stone)在1947年关于国民经济的研究。他曾利用美国1929—1938年各年的数据，得到了17个反映国民收入与支出的变量要素，例如雇主补贴、消费资料和生产资料、纯公共支出、净增库存、股息、利息外贸平衡等等。

在进行主成分分析后，竟以97.4%的精度，用三个新变量就取代了原17个变量。根据经济学知识，斯通发现这三个新变量与经济学中的“总收入F1”、“总收入变化率F2”和“经济发展或衰退的趋势F3”高度相关(highly correlated)。更有意思的是，这三个变量其实都是可以直接测量的。因此，这三个新变量被赋予了特殊和有意义的涵义。斯通将他得到的主成分与实际测量的总收入、总收入变化率以及时间因素做了相关分析，得到了很好的结果。

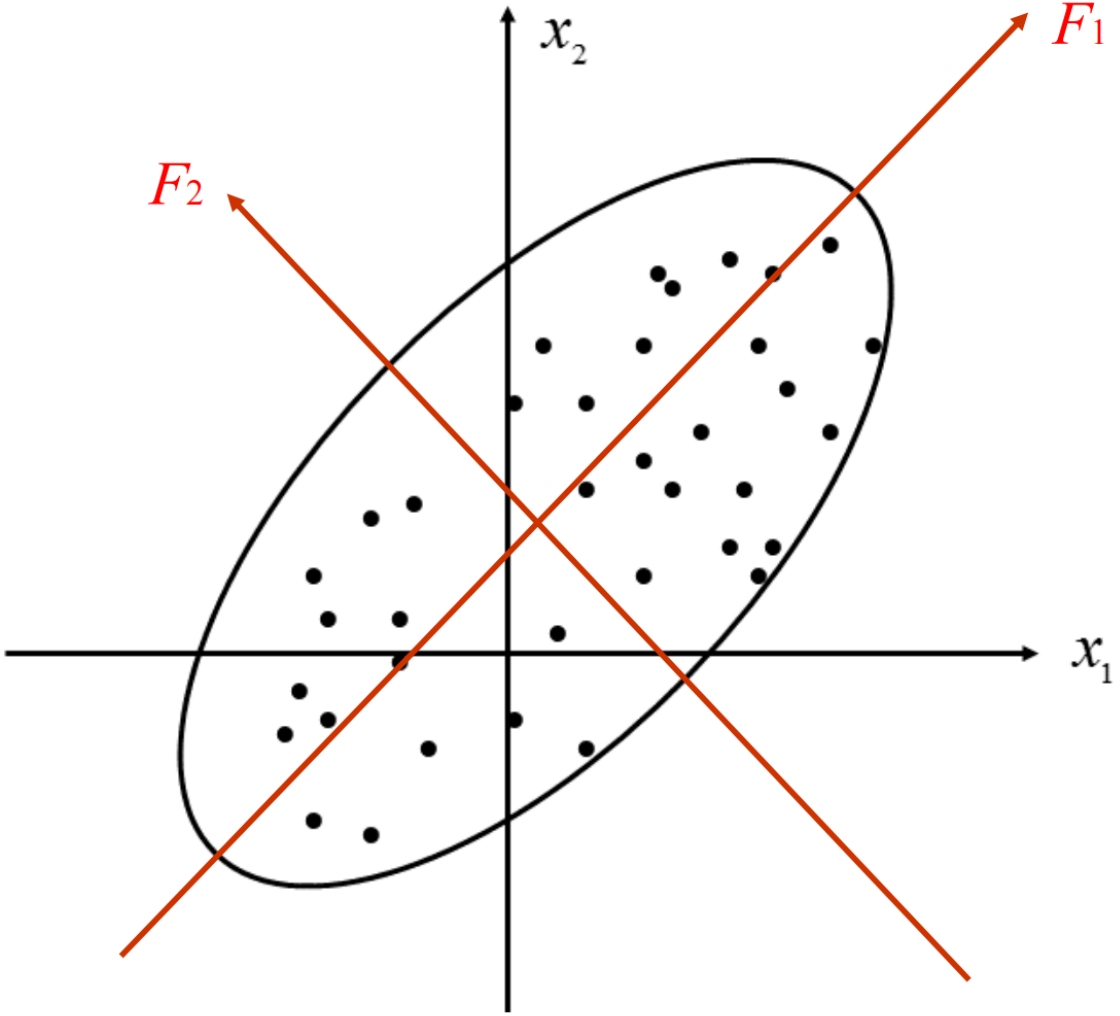
2. 主成分分析的几何解释

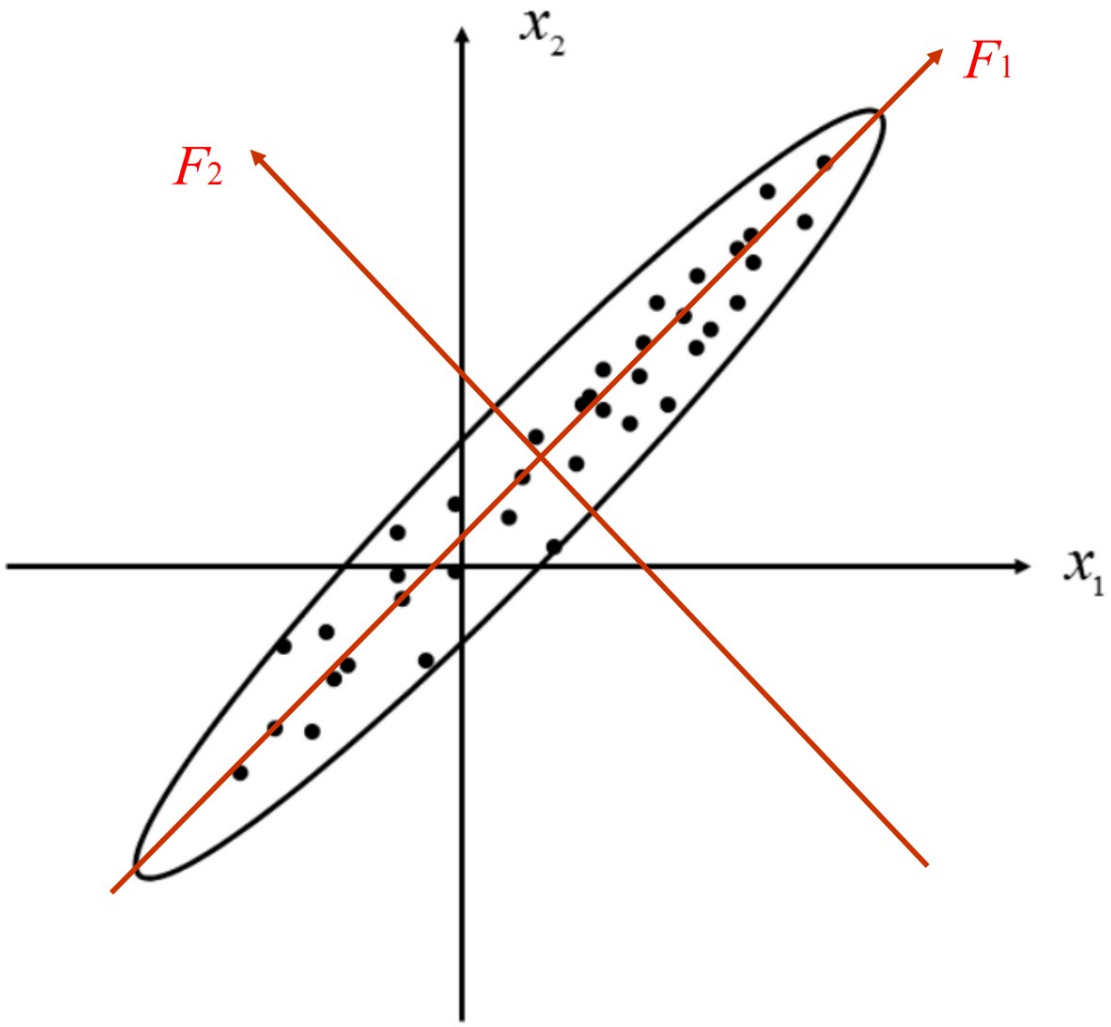
为了方便，我们在二维空间中讨论主成分的几何意义：假设有n个样品，每个样品有两个观测变量x_1和x_2，在由变量x_1和x_2所确定的二维平面中，n个样本点所散布的情况如椭圆状。如果这n个样本点无论是沿着x_1轴方向或x_2轴方向都具有较大的离散性，其离散的程度可以分别用观测变量x_1的方差和x_2的方差定量地表示。那么，如果只考虑x_1和x_2中的任何一个，包含在原始数据中的信息将会有较大的损失。


x_2
x_1

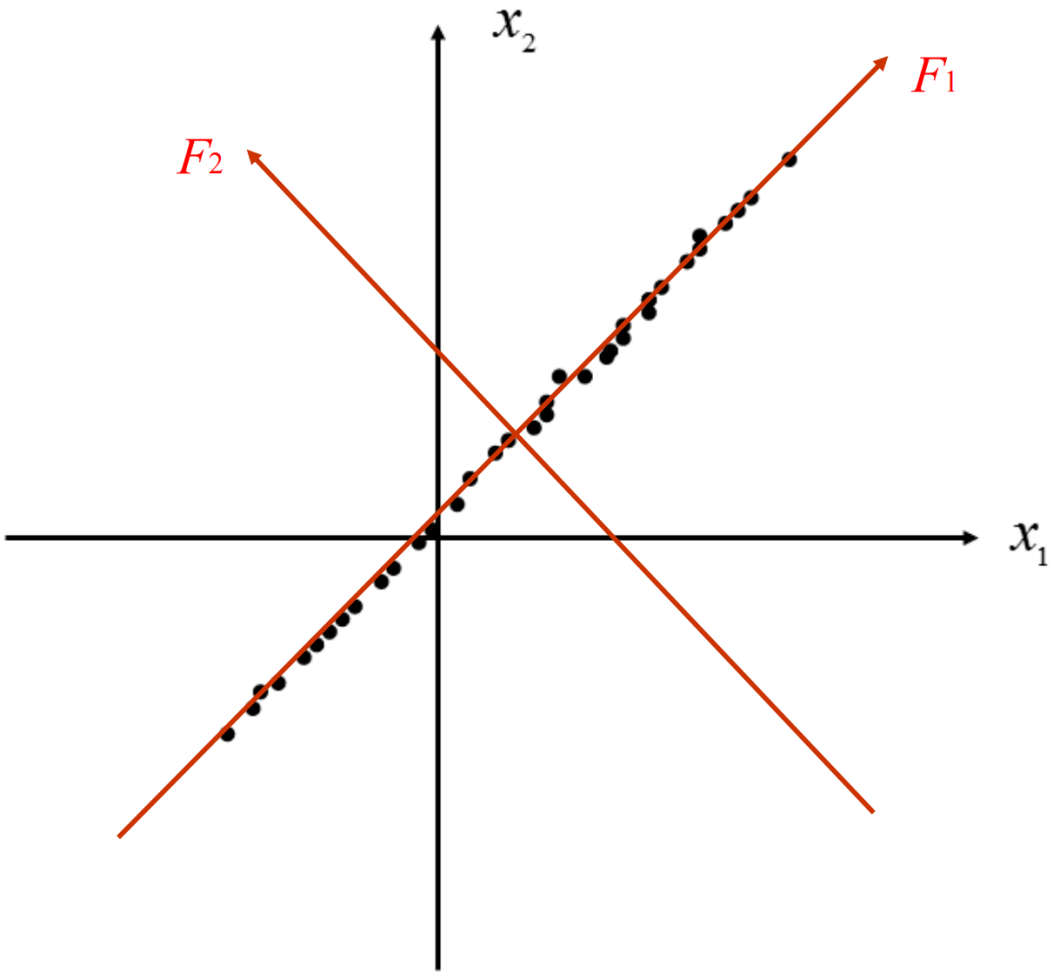
平移、旋转坐标轴

如果我们将x_1轴和x_2轴先平移，再同时按逆时针方向旋转θ角度，得到新坐标轴F_1和F_2。F_1和F_2是两个新变量。


x_2
F_1
F_2
x_1



如果数据有较大相关性？



如果数据有较大相关性？

根据旋转变换的公式：

$$\begin{cases} y_1 = \ x_1 cos\theta + x_1 sin\theta \\ y_2 = -x_1 sin\theta + x_2 cos\theta \end{cases}$$

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} cos\theta & sin\theta \\ -sin\theta & cos\theta \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \mathbf{U}^T\mathbf{x}$$

$\mathbf{U}^T$为旋转变换矩阵，它是正交矩阵，即有

$$\mathbf{U}^T = \mathbf{U}^{-1}, \qquad \mathbf{U}^T\mathbf{U} = \mathbf{I}$$

旋转变换的目的：

为了使得n个样品点在F_1轴方向上的离散程度最大，即F_1的方差最大。

(变量F_1代表了原始数据的绝大部分信息，在研究某经济问题时，即使不考虑变量F_2也无损大局)。经过上述旋转变换原始数据的大部分信息集中到F_1轴上，对数据中包含的信息起到了浓缩作用。

F_1，F_2除了可以对包含在X_1，X_2中的信息起着浓缩作用之外，还具有不相关的性质，这就使得在研究复杂的问题时避免了信息重叠所带来的虚假性。二维平面上的各点的方差大部分都归结在F_1轴上，而F_2轴上的方差很小。F_1和F_2称为原始变量x_1和x_2的综合变量。F_1简化了系统结构，抓住了主要矛盾。

由此可概括出主成分分析的几何意义：

主成分分析的过程也就是坐标旋转的过程，各主成分表达式就是新坐标系与原坐标系的转换关系，新坐标系中各坐标轴的方向就是原始数据方差最大的方向。

3.主成分分析方法的原理

主成分分析出发点是把原来多个变量化为少数几个综合指标的一种统计分析方法。

从数学角度来看，这是一种降维处理技术。假定有n个样本，每个样本共有p个零均值变量描述，这样就构成了一个$p \times n$的原始数据矩阵：

$$\mathbf{X} = [x_1, x_2, \ldots, x_n] = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{p1} & x_{p2} & \cdots & x_{pn} \end{bmatrix}_{p \times n} \quad (1)$$

如何从这p个变量的数据中抓住事物内在的规律性呢？要解决这一问题，自然要在p维空间中加以考察，这是主成分分析中的一个关键问题。

为了解决这个问题，我们采取的策略是进行降维处理，即用较少的几个综合指标来代替原来较多的变量指标，而且使这些较少的综合指标既能尽量多的反映原来较多指标所反映的信息，同时这些综合指标之间又是彼此独立的。那么，这些综合指标(即新变量)应该如何选取呢？显然，其最简单的想法是取原来变量指标的线性组合，适当调整组合系数，使新的变量指标之间相互独立且有代表性。

如果记原来的变量指标为$x_1, x_2, \ldots, x_p$，它们的综合指标(新变量)为$z_1, z_2, \ldots, z_m (m \leq p)$，则

$$\begin{cases} z_1 = w_{11}x_1 + w_{12}x_2 + \cdots + w_{1p}x_p \\ z_2 = w_{21}x_1 + w_{22}x_2 + \cdots + w_{2p}x_p \\ \cdots \\ z_m = w_{m1}x_1 + w_{m2}x_2 + \cdots + w_{mp}x_p \end{cases} \quad (2)$$

在(2)式中，系数w_{ij}(原始变量的载荷系数)由下列原则决定：

(1) z_i与z_j ($i \neq j$; i, j=1,2,…, m)相互线性无关；

(2) z_i是$x_1, x_2, \ldots, x_p$的一切线性组合中方差最大者；z_2是与z_1不相关的$x_1, x_2, \ldots, x_p$的所有线性组合中方差最大者；

…；以此类推z_m是与$z_1, z_2, \ldots, z_{m-1}$都不相关的$x_1, x_2, \ldots, x_p$的所有线性组合中方差最大者。

写成向量形式：

$$z_i = w_i^H x, \qquad \forall i = 1, \ldots, m$$

其中：要求w_i是一组标准正交基，即

$$w_i^H w_j = \begin{cases} 0 & if \;\; i \neq j \\ 1 & if \;\; i = j \end{cases}$$

令矩阵$\mathbf{W} = [\mathbf{w}_1 \;\; \mathbf{w}_2 \;\; \ldots \;\; \mathbf{w}_m]$.

新$m \times n$数据集 $\mathbf{Z} = \mathbf{W}^{\mathrm{H}}\mathbf{X}$

每一组观测数据在第i维新坐标下投影的方差为：

$$\sum_{k=1}^{n}|z_{ik}|^2 = \mathbf{w}_i^H \mathbf{X}\mathbf{X}^H \mathbf{w}_i$$

这样决定的新变量指标$z_1, z_2, \ldots, z_m$分别称为原变量指标$x_1, x_2, \ldots, x_p$的第一，第二，…第m个主成分。其中z_i在总方差中占的比例最大，$z_2, z_3, \ldots, z_m$的方差依次递减。在实际问题的分析中常挑选前几个最大的主成分，这样既减少了变量的数目，又抓住了主要矛盾，简化了变量之间的关系。

从以上分析可以看出，找主成分就是确定原来变量$x_j(j = 1,2,\cdots,p)$在诸主成分$z_i(i = 1,2,\cdots,m)$上的载荷$w_{ij}(i = 1,2,\cdots,m; j = 1,2,\cdots,p)$。

4.主成分分析的计算过程

主成分分析的计算步骤

通过上述主成分分析的基本原理的介绍，我们可以把主成分分析计算步骤归纳如下：

（1）计算相关系数矩阵。设由p个分量的随机向量组成的矩阵是

$$\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n] = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{p1} & x_{p2} & \cdots & x_{pn} \end{bmatrix}_{p \times n}$$

求出自相关矩阵：

$$R = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1p} \\ r_{21} & r_{22} & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & r_{pp} \end{bmatrix}_{p\times p} \quad (3)$$

在公式(3)中，$r_{ij}(i,j=1,2,\cdots,p)$是原来变量x_i与x_j的相关系数，计算公式是

$$r_{ij} = \frac{\sum_{k=1}^{n}(x_{ik}-\overline{x_i})(x_{jk}-\overline{x_j})}{\sqrt{\sum_{k=1}^{n}(x_{jk}-\overline{x_j})^2 \sum_{k=1}^{n}(x_{jk}-\overline{x_j})^2}} \quad (4)$$

因为R是$p\times p$的实对称矩阵（即$r_{ij}=r_{ji}$），所以只需要计算其上三角元素即可。

(2) 计算特征值和特征向量

首先求出特征方程

$$det(\lambda I - R) = 0$$

的全部特征值$\lambda_i(i = 1,2,\cdots,p)$, 并按其从大到小的顺序排列，即$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$; 然后分别求出对应于特征值$\lambda_1, \lambda_2, \cdots, \lambda_p$的标准正交特征向量$u_1, u_2, \cdots, u_p$。令$u_1, u_2, \cdots, u_p$组成的矩阵为

$$U = [u_1, u_2, \cdots, u_p] = \begin{bmatrix} u_{11} & u_{12} & \cdots & u_{1p} \\ u_{21} & u_{22} & \cdots & u_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ u_{p1} & u_{p2} & \cdots & u_{pp} \end{bmatrix}$$

这实际上完成了一个特征值分解过程：

$$R = U\Sigma_R U^T \qquad \Sigma_R = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & \vdots \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & 0 & \lambda_p \end{bmatrix}_{p\times p}$$

(3) 计算主成分贡献率及累计贡献率

主成分z_i的贡献率

$$\frac{\lambda_i}{\sum_{k=1}^{p}\lambda_k}, \qquad (i = 1,2,\cdots,p)$$

直到第m个主成分的累计贡献率$= \frac{\sum_{l=1}^{m}\lambda_l}{\sum_{k=1}^{p}\lambda_k}$.

如果到第$m(\leq p)$个主成分的累计贡献率达到85-95%的特征值$\lambda_1, \lambda_2, \cdots, \lambda_p$所对应的第一，第二，…，第$m$个主成分。

(4) 主成分（新变量）关于原始变量的关系式

$$\begin{cases} z_1 = u_{11}x_1 + u_{12}x_2 + \cdots + u_{1p}x_p \\ z_2 = u_{21}x_1 + u_{22}x_2 + \cdots + u_{2p}x_p \\ \vdots \\ z_m = u_{m1}x_1 + u_{m2}x_2 + \cdots + u_{mp}x_p \end{cases}$$

写成矩阵向量的形式就是

$$[z_1, z_2, \ldots, z_m]^T = [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_m]^T [x_1, x_2, \ldots, x_p]^T$$

这就把一个有p个特征的问题降维为仅有$m(\leq p)$的问题来分析。

例1 设$x=(x_1,x_2,x_3)^T$的协方差矩阵为

$$G=\begin{bmatrix}1 & -2 & 0\\ -2 & 5 & 0\\ 0 & 0 & 2\end{bmatrix}$$

从协方差矩阵出发，求解主成分。

（1）求协方差矩阵的特征值。由

$$|G-\lambda I|=\begin{vmatrix}1-\lambda & -2 & 0\\ -2 & 5-\lambda & 0\\ 0 & 0 & 2-\lambda\end{vmatrix}$$

$$=(1-\lambda)(5-\lambda)(2-\lambda)-(-2)(-2)(2-\lambda)=0$$

得 $\lambda_1=5.83 \quad \lambda_2=2 \quad \lambda_3=0.17$

(2) 求特征值对应的特征向量

$$u_1 = \begin{bmatrix} 0.383 \\ -0.924 \\ 0.000 \end{bmatrix} \quad u_2 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \quad u_3 = \begin{bmatrix} 0.924 \\ 0.383 \\ 0.000 \end{bmatrix}$$

(3) 计算主成分

$$z_1 = 0.383x_1 - 0.924x_2 \quad z_2 = x_3 \quad z_3 = 0.924x_1 + 0.383x_2$$

(4) 各主成分的贡献率及累计贡献率

第一主成分贡献率：5.83/(5.83+2+0.17) = 0.72875

第二主成分贡献率：2/(5.83+2+0.17) = 0.25

第三主成分贡献率：0.17/(5.83+2+0.17) = 0.02125

由此可以看出，第一和第二主成分的累计贡献率达到

$$(5.83+2) / (5.83+2+0.17) = 0.97875$$

由此可将以前三维的问题降维为两维问题，第一和第二主成分包含了以前变量的绝大部分信息97.875%.

例2 设$x=(x_1,x_2,x_3)$=的协方差矩阵为

$$G=\begin{bmatrix}16 & 2 & 30\\ 2 & 1 & 4\\ 30 & 1 & 100\end{bmatrix}$$

从协方差矩阵出发，求解主成分。

(1) 求协方差矩阵的特征值。由

$$|G-\lambda I|=\begin{vmatrix}16-\lambda & 2 & 30\\ 2 & 1-\lambda & 4\\ 30 & 1 & 100-\lambda\end{vmatrix}=0$$

得 $\lambda_1=109.793 \quad \lambda_2=6.469 \quad \lambda_3=0.738$

(2) 求特征值对应的特征向量

根据

$$(G - \lambda_i I)u_i = 0, \qquad i = 1,2,3$$

解得

$$u_1 = \begin{bmatrix} 0.305 \\ 0.041 \\ 0.951 \end{bmatrix} \quad u_2 = \begin{bmatrix} 0.944 \\ 0.120 \\ -0.308 \end{bmatrix} \quad u_3 = \begin{bmatrix} -0.127 \\ 0.992 \\ -0.0028 \end{bmatrix}$$

(3) 主成分的表达式

$$\begin{cases} z_1 = 0.305x_1 + 0.041x_2 + 0.951x_3 \\ z_2 = 0.944x_1 + 0.120x_2 - 0.308x_3 \\ z_3 = -0.127x_1 + 0.992x_2 - 0.002x_3 \end{cases}$$

(4) 各主成分的贡献率及累计贡献率

第一主成分贡献率：109.793/(109.793+6.469+0.738) =0.938

第二主成分贡献率：6.469/(109.793+6.469+0.738) =0.0553

第三主成分贡献率：0.738/(109.793+6.469+0.738) =0.0063

由从上面主成分贡献率的情况看，第一主成分的贡献率已经占到93.8%，因此这个问题可以将3维指标变量降维为1维问题 进行处理。

以上两个例子是用协方差矩阵(covariance matrix)的特征分析实现主成分分析。在实际应用中，由于变量所用量纲不同，往往先对数据进**标准化**处理用相关矩阵(correlation matrix)进行主成分分析。

例3 企业经济效益综合分析。用5个经济指标进行考核。用相关系数矩阵法求解主成分。其中计算出的相关系数矩阵为：

$$\rho = \begin{bmatrix} 1 & 0.4532 & -0.7536 & -0.3475 & 0.5621 \\ 0.4532 & 1 & -0.4545 & 0.4244 & 0.7316 \\ -0.7536 & -0.4545 & 1 & 0.3668 & -0.4168 \\ -0.3475 & 0.4244 & 0.3668 & 1 & 0.499 \\ 0.5621 & 0.7316 & -0.4168 & 0.499 & 1 \end{bmatrix}$$

(1) 计算其特征值：

$$\lambda_1 = 2.695 \quad \lambda_2 = 1.719 \quad \lambda_3 = 0.331$$

$$\lambda_4 = 0.206 \quad \lambda_5 = 0.049$$

(2) 各特征值的累计方差贡献率为：

$$\sum_{k=1}^{j} \lambda_k / p \quad 0.539 \;\; 0.883 \;\; 0.949 \;\; 0.990 \;\; 1.000$$

(3) 从以上方差贡献率看，k =2时主成分个数较为合适。

λ_1和λ_2对应的特征向量为：

$$u_1 = [0.501 \; 0.503 \; -0.470 \; 0.074 \; 0.520]^T$$

$$u_2 = [-0.348 \; 0.285 \; 0.388 \; 0.744 \; 0.305]^T$$

(4) 建立第一和第二主成分：

$$F_1 = 0.501x_1 + 0.503x_2 - 0.470x_3 + 0.074x_4 + 0.520x_5$$

$$F_2 = -0.348x_1 + 0.285x_2 + 0.388x_3 + 0.744x_4 + 0.305x_5$$