101 主成分分析（PFA）

（1）方法原理

主成分分析就是设法将原来众多具有一定相关性的指标（比如 p 个指标），重新组合成一组新的相互无关的综合指标来代替原来指标。通常数学上的处理就是将原来 p 个指标作线性组合，作为新的综合指标。在众多的线性组合中选取尽可能多地反映原来指标信息的线性组合作为第一个综合指标记为 F_1。反映信息量的多少可以用变量的方差的大小来表示，方差大的反映的信息量大。因此，在所有的线性组合中所选取的 F_1 应该是方差最大的，故称为第一主成分。如果第一主成分不足以代表原来 p 个指标的信息，在考虑选取 F_2，即选第二个线性组合。为了有效地反映原来信息，F_1 已有的信息就不需要再出现在 F_2 中，也就是说 F_1 和 F_2 是不相关的。依次类推可以找出第三，第四，……，第 p 个主成分。这些主成分之间不仅不相关，而且它们的方差依次递减。

（2）基本步骤

第一步，将原始数据标准化。

对于任何一个观测变量（指标）都进行标准化变换，变为标准化变量。标准化的变换公式为 $z=(x-\bar{x})/\sigma_x$（$\bar{x}$ 为 x 的均值，σ_x 为 x 的标准差），标准化变换并不改变变量之间的相关系数。

第二步，建立变量的相关系数矩阵。

标准化变量的相关系数矩阵和协差阵是相同的，因此可以通过建立变量的相关系数矩阵来计算相应的特征值和特征向量。

$$R=\left(r_{ij}\right)_{p\times p}$$

$$\gamma_{ij}=\frac{s_{jk}}{\sqrt{s_{jj}}-\sqrt{s_{kk}}}=\frac{\frac{1}{n-1}\sum_{i=0}^{p}(x_{ij}-\bar{x}_{0j})(x_{ij}-\bar{x}_{0k})}{\sqrt{s_{jj}}\sqrt{s_{kk}}}=\frac{1}{n-1}\sum_{i=1}^{n}Z_{ij}Z_{ik}$$

（其中 $s_{kk}=\frac{1}{n-1}\sum_{i=1}^{n}(x_{ij}-\bar{x}_{0k})$，$\bar{x}_{0k}=\frac{1}{n}\sum_{i=0}^{n}x_{ij}$，$Z_{ik}=\frac{x_{ij}-\bar{x}_{0k}}{\sqrt{s_{kk}}}$）

$$\therefore R=\left(\gamma_{ij}\right)_{p\times p}=\frac{1}{n\text{-}1}(\sum_{i=1}^{n}Z_{ij}Z_{ik})_{p\times p}=\frac{1}{n-1}(Z)^{T}Z$$

第三步，求 R 的特征根λ 1≥λ 2≥…≥λ p＞0 及相应的单位特征向量，并确定主成分：

$$a_1=\begin{bmatrix}a_{11}\\a_{21}\\\vdots\\x_{p1}\end{bmatrix},\quad a_2=\begin{bmatrix}a_{12}\\a_{22}\\\vdots\\x_{p2}\end{bmatrix},\ \ldots\ldots\ a_p=\begin{bmatrix}a_{1p}\\a_{2p}\\\vdots\\x_{pp}\end{bmatrix}$$

若 $\lambda_1\geq\lambda_2\geq\cdots\lambda_n\geq 0$，则根据因子累计方差贡献率大于等于 85%来确定主成分数目 m，与 $\lambda_k(k=1,2,\cdots,m)$ 对应的特征向量记为 $\alpha_k=(\alpha_{k1},\alpha_{k2},\cdots,\alpha_{km})$，与 λ_k 相对应的主成分称为第 k 个主成分，λ_k 是相关矩阵 R 的第 k 个特征值。

第四步，写出主成分，计算综合指标的数值。

$$F_i=a_{1i}X_1+a_{2i}X_2+\cdots+a_{pi}X_p \qquad \text{i=1，…，m}$$

第五步，计算综合得分

运用主成分分析得出综合指标的数值，以每个主成分的方差贡献率为权重来构造一个综合评价函数。

$$y=a_1F_1+a_2F_2+\cdots+a_mF_m$$

也称 y 为评估指数，依据对每个系统计算出的 y 值大小进行排序比较。

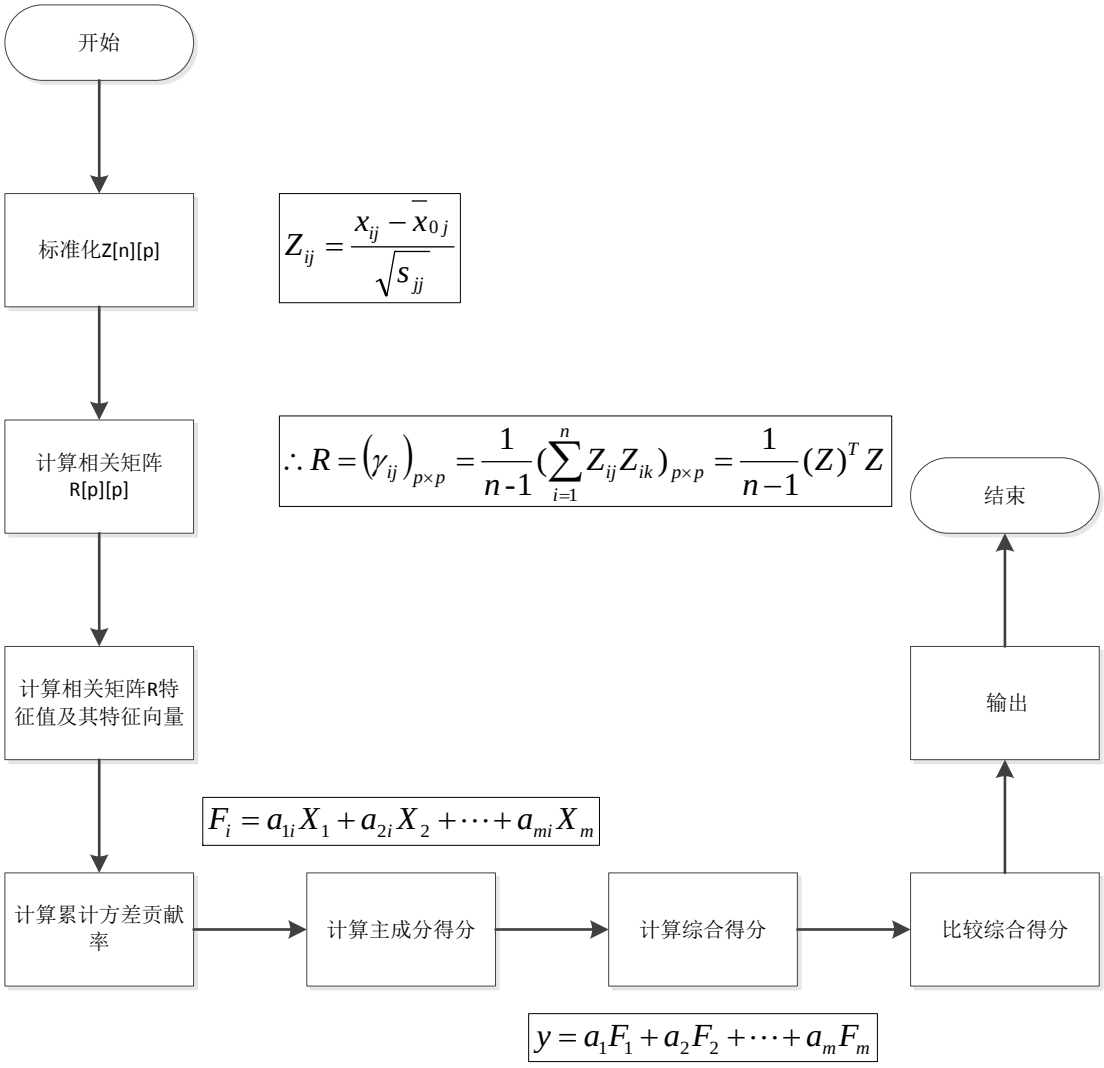
其中，a_i 为第 i 个主成分的方差贡献率

$$a_i=\lambda_i\Big/\sum_{i=1}^{n}\lambda_i$$

（3）应用及效果分析

从而可以看出利用主成分分析做综合评价是从原始数据所给定的信息直接确定权重进行评价的,所取的权重直接为对应主成分的方差贡献率。某个主成分在综合评价时所能反映的信息越多,相应的权重也越大,所能反映的信息越少,相应的权重也越少,因为其所得信息都是来源于原始数据,这也是主成分分析综合评价法最主要的优点之一。

（4）简略流程图


开始
标准化Z[n][p]
$Z_{ij}=\frac{x_{ij}-\bar{x}_{0j}}{\sqrt{s_{jj}}}$
计算相关矩阵R[p][p]
$\therefore R=(\gamma_{ij})_{p\times p}=\frac{1}{n\text{-}1}(\sum_{i=1}^{n}Z_{ij}Z_{ik})_{p\times p}=\frac{1}{n-1}(Z)^{T}Z$
结束
计算相关矩阵R特征值及其特征向量
输出
$F_i=a_{1i}X_1+a_{2i}X_2+\cdots+a_{mi}X_m$
计算累计方差贡献率
计算主成分得分
计算综合得分
比较综合得分
$y=a_1F_1+a_2F_2+\cdots+a_mF_m$