主成分分析法

Optimization and operational model

成都信息工程大学学生数学建模协会

https://blog.csdn.net/program_developer/article/details/80632779

https://blog.csdn.net/weixin_43819566/article/details/113800120

https://www.bilibili.com/video/BV1Q4411B7kS

主讲人：lsx

目录

CONTENTS

01 /主成分分析法简介

Mathematical modeling and its application

什么是主成分分析法（PCA）

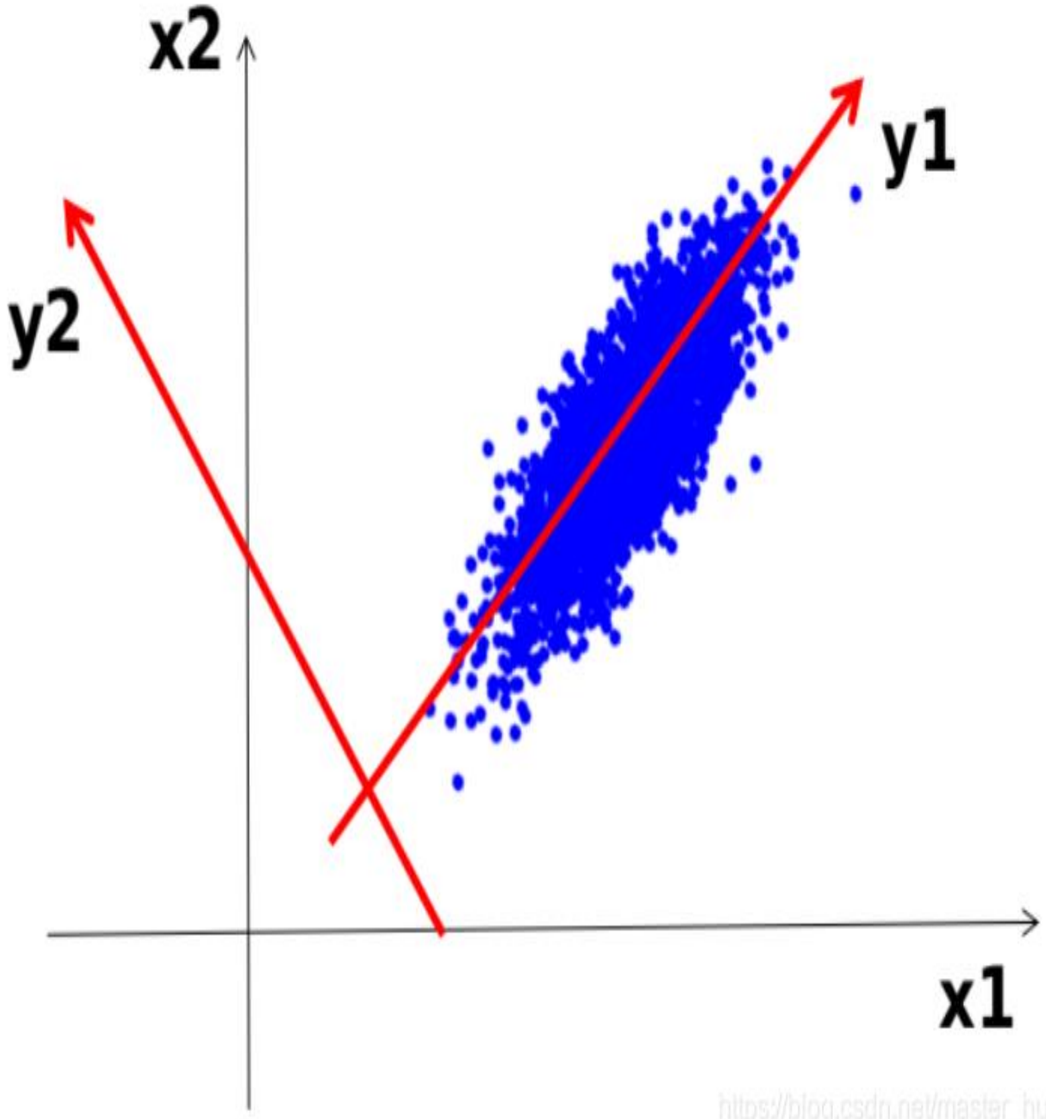
主成分分析是一种降维算法，它能将多个指标转换为少数几个主成分，这些主成分是原始变量的线性组合，且彼此之间互不相关，其能反映出原始数据的大部分信息。

一般来说，当研究的问题涉及到多变量且变量之间存在很强的相关性时，我们可考虑使用主成分分析的方法来对数据进行简化。

在数学上更简单的理解我们可以把它想象为对很多个坐标点通过映射方法到一根函数直线上。通过这种方法我们肯定会得到新的特征用来表示这个事件，新的特征剔除了原有特征的冗余信息，因此更有区分度。新的特征基于原有特征，它能够重建原有特征。主成分分析要保留最有可能重建原有特征的新特征，从而达到数据降维的作用。

02 / PCA原理及案例

Mathematical model



例如我们得到了一个二维数据集，里面一个标签仅仅只用两个（x1，x2）特征就能表示，但是我们只能用一个特征去描述这件事。在原始的坐标轴上，我们无论删掉x1或者x2是我们都不能完整的表达出这个点来。

但是我们可以通过PCA，将数据从原来的坐标系转换到了新的坐标系，新坐标系的选择是由数据本身决定的。第一个新坐标轴选择的是原始数据中方差最大的方向，第二个新坐标轴的选择和第一个坐标轴正交且具有最大方差的方向。就通过该数据集（x1，x2），我们通过该方法构建出坐标系(y1，y2)，在该坐标系中我们可以发现数据基本都集中在轴上面，而这条短轴上则差距很小。在极端的情况，短轴如果退化成一点，那只有在长轴的方向才能够解释这些点的变化了。

这样，由二维到一维的降维就自然完成了。

主成分分析法（PCA）的基本思想

假设有 n 个样本，p 个指标，则可构成大小为 $n \times p$ 的样本矩阵 X:

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} = (x_1, x_2, \cdots, x_p)$$

假设我们想找到新的一组变量 $z_1, z_2, \ldots, z_m (m \leq p)$，且它们满足:

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

系数 l_{ij} 的确定原则:

1. z_i 与 z_j ($i \neq j$; $i,j=1,2,\ldots,m$) 相互无关;
2. z_1 是 $x_1, x_2, \ldots, x_p$ 的一切线性组合中方差最大者;
3. 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;
4. 以此类推，z_m 是与 $z_1, z_2, \ldots, z_{m-1}$ 不相关的 $x_1, x_2, \ldots, x_p$ 的所有线性组合中方差最大者;
5. 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 主成分。

主成分分析法（PCA）的计算步骤

假设有 n 个样本，p 个指标，则可构成大小为 n × p 的样本矩阵 x：

$$x=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}=(x_1, x_2, \cdots, x_p)$$

主成分分析法（PCA）的计算步骤

1. 首先对其进行标准化处理：

$$\text{按列计算均值：}\bar{x}_j=\frac{1}{n}\sum_{i=1}^{n}x_{ij},\quad \text{标准差：}S_j=\sqrt{\frac{\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2}{n-1}},\quad \text{标准化数据：}X_{ij}=\frac{x_{ij}-\bar{x}_j}{S_j}$$

$$\text{原始样本矩阵经过标准化变化：}X=\begin{bmatrix}X_{11} & X_{12} & \cdots & X_{1p}\\ X_{21} & X_{22} & \cdots & X_{2p}\\ \vdots & \vdots & \ddots & \vdots\\ X_{n1} & X_{n2} & \cdots & X_{np}\end{bmatrix}=(X_1,X_2,\cdots,X_p)$$

主成分分析法（PCA）的计算步骤

2. 计算标准化样本查的协方差矩阵：

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

$$r_{ij} = \frac{1}{n-1}\sum_{k=1}^{n}(X_{ki} - \overline{X}_i)(X_{ki} - \overline{X}_j) = \frac{1}{n-1}\sum_{k=1}^{n} X_{ki}X_{kj}$$

1、2步骤可以合成一步：

$$R = \frac{\sum_{k=1}^{n}(x_{ki} - \overline{x}_i)(x_{ki} - \overline{x}_j)}{\sqrt{\sum_{k=1}^{n}(x_{ki} - \overline{x}_i)^2 \sum_{k=1}^{n}(x_{kj} - \overline{x}_j)^2}}$$

主成分分析法（PCA）的计算步骤

3. 计算 R 的特征值和特征值向量：

$$\text{特征值}\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0, \quad (\text{R是半正定矩阵，且}\mathrm{tr}(R) = \sum_{k=1}^{p} \lambda_k = p)$$

$$\text{特征向量：}\ a_1 = \begin{bmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{p1} \end{bmatrix}, a_2 = \begin{bmatrix} a_{12} \\ a_{22} \\ \vdots \\ a_{p2} \end{bmatrix}, \cdots, a_p = \begin{bmatrix} a_{1p} \\ a_{2p} \\ \vdots \\ a_{pp} \end{bmatrix}$$

主成分分析法（PCA）的计算步骤

4. 计算主成分共享率以及累计贡献率：

$$\text{贡献率} = \frac{\lambda_i}{\sum_{k=1}^{p}\lambda_k}, \quad \text{累加贡献率} = \frac{\sum_{k=1}^{i}\lambda_k}{\sum_{k=1}^{p}\lambda_k}, \quad (i = 1, 2, \cdots, p)$$

---

5. 写出主成分：

一般取累计贡献率超过 80% 的特征值所对应的第一、第二、…、第 m（$m \leq p$）个主成分。

$$\text{第}i\text{个主成分：} F_i = a_{1i}X_1 + a_{2i}X_2 + \cdots + a_{pi}X_p$$

---

6. 根据系数分析主成分代表的意义：

对于某个主成分而言，指标前面的系数越大，代表该指标对于该主成分的影响越大。

案例分析

题目来源于：《应用多元统计分析》王学民

在制定服装标准的过程中，对128名成年男子的身材进行了测量，每人测得的指标中含有这样六项：身高（x_1）、坐高（x_2）、胸围（x_3）、手臂长（x_4）、肋围（x_5）和腰围（x_6）。所得样本相关系数矩阵（对称矩阵哦）列于下表。

	x_1	x_2	x_3	x_4	x_5	x_6
x_1	1.000	0.79	0.36	0.76	0.25	0.51
x_2	0.79	1.000	0.31	0.55	0.17	0.35
x_3	0.36	0.31	1.000	0.35	0.64	0.58
x_4	0.76	0.55	0.35	1.000	0.16	0.38
x_5	0.25	0.17	0.64	0.16	1.000	0.63
x_6	0.51	0.35	0.58	0.38	0.63	1.000

案例分析

经过计算，相关系数矩阵的特征值、相应的特征向量以及贡献率列于下表：

特征向量	a1	a2	a3	a4	a5	a6
x1：身高	0.469	−0.365	0.092	-0.122	-0.080	-0.786
x2：坐高	0.404	−0.397	0.613	0.326	0.027	0.443
x3：胸围	0.394	0.397	−0.279	0.656	0.405	-0.125
x4：手臂长	0.408	−0.365	−0.705	-0.108	-0.235	0.371
x5：肋围	0.337	0.569	0.164	-0.019	-0.731	0.034
x6：腰围	0.427	0.308	0.119	-0.661	0.490	0.179
特征值	3.287	1.406	0.459	0.426	0.295	0.126
贡献率	0.548	0.234	0.077	0.071	0.049	0.021
累计贡献率	0.548	0.782	0.859	0.930	0.979	1.000

案例分析

从表中可以看到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三个主成分，它们能够很好地概括原始变量。

$$F_1 = 0.469X_1 + 0.404X_2 + 0.394X_3 + 0.408X_4 + 0.337X_5 + 0.427X_6$$
$$F_2 = -0.365X_1 - 0.397X_2 + 0.397X_3 - 0.365X_4 + 0.569X_5 + 0.308X_6$$
$$F_3 = 0.092X_1 + 0.613X_2 - 0.279X_3 - 0.705X_4 + 0.164X_5 + 0.119X_6$$

X_i 均是标准化后的指标，x_i：身高、坐高、胸围、手臂长、肋围和腰围

- 第一主成分 F_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 F_2 在 X_3、X_5、X_6。上有中等程度的正载荷，而在 X_1、X_2、X_4 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 F_3 在 X_2 上有大的正载荷，在 X_4 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。

注：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，因此我们也可以考虑只取前两个主成分进行分析。

03 / 代码实现

Mathematical Contest in Modeling

代码实现

5	6	4	5
1	1	2	2
2	2	2	3
7	7	7	6
3	3	3	3

可以看到这组4维变量也具有很强的相关性，我们把他写成矩阵的形式。$\begin{pmatrix} 5 & 6 & 4 & 5 \\ 1 & 1 & 2 & 2 \\ 2 & 2 & 2 & 3 \\ 7 & 7 & 7 & 6 \\ 3 & 3 & 3 & 3 \end{pmatrix}$他的列向量可以表示为$(X_1, X_2, X_3, X_4)$我们需要得到的是原始数据的主成分，即将这4组向量进行适当的线性组合（线性变换），并选出可以表示原始数据大部分信息的主成分。

设Y_i为第i个主成分，i=1,2,3,4,可设

$$\begin{cases} Y_1 = c_{11}X_1 + c_{12}X_2 + c_{13}X_3 + c_{14}X_4 \\ Y_2 = c_{21}X_1 + c_{22}X_2 + c_{23}X_3 + c_{24}X_4 \\ Y_3 = c_{31}X_1 + c_{32}X_2 + c_{33}X_3 + c_{34}X_4 \\ Y_4 = c_{41}X_1 + c_{42}X_2 + c_{43}X_3 + c_{44}X_4 \end{cases}$$

代码实现

Step1:标准化，对于第一列数据标准化的步骤是先算出所有元素的均值和标准差，再用每一个元素减去均值再除以标准差，用matlab程序实现为X1=(X-mean (X))/std (X)或X1=zscore(X)。对每一列都作出同样的操作，得到标准化后的矩阵。

Step2:进行标准化之后我们算出标准化数据的协方差矩阵，用matlab程序实现为R = cov(X1)。（协方差矩阵对角线就是方差，其他的是相互的协方差，即要让对角线的方差数值最大，其它协方差数值为0，这是为什么我们要算协方差矩阵。）

以上两步可以合并为一步，直接算原始数据的相关系数矩阵，matlab程序实现为R = corrcoef (x)。

（corr (x1, x2) =(cov(x1, x2)/ ($\sqrt{var(x1)}$*$\sqrt{var(x2)}$)）

代码实现

Step3:计算协方差矩阵的特征值和特征向量，特征值对应的特征向量就是理想中想取得正确的坐标轴，而特征值就等于数据在旋转之后的坐标上对应维度上的方差。matlab代码为[V,D] = eig(R); lambda=diag(D)。

得到我们这个例子的特征值和特征向量。由于lambda现在是从小到大排列，为了后面计算方便我们把他调换一下顺序，相应的对应的V也需要调换。matlab代码为lambda= lambda（end: -1:1):V=rot90(V)’ 。

```
V =

    0.8196    0.2668    0.0060    0.5070
   -0.4926    0.5510    0.4502    0.5011
   -0.2871   -0.0293   -0.8230    0.4894
   -0.0561   -0.7902    0.3465    0.5024


lambda =

    0.0030
    0.0130
    0.1044
    3.8797
```

```
V =

    0.5070    0.0060    0.2668    0.8196
    0.5011    0.4502    0.5510   -0.4926
    0.4894   -0.8230   -0.0293   -0.2871
    0.5024    0.3465   -0.7902   -0.0561


lambda =

    3.8797
    0.1044
    0.0130
    0.0030
```

代码实现

Step4:计算贡献率和累计贡献率，选择合适的主成分。matlab代码实现为:contribution_rate = lambda / sum(lambda) ; cum_contribution_rate =cumsum(contribution_rate)。

得到如下结果

```
contribution_rate =

    0.9699
    0.0261
    0.0032
    0.0007
```

```
cum_contribution_rate =

    0.9699
    0.9960
    0.9993
    1.0000
```

从这个结果我们可以看到第一个主成分的贡献率最高，且已经超过80%，我们就选择第一个主成分来反应我们数据的信息。

Y=0.5070*X1+ 0.5011*X2+0.4894*X3+ 0.5024*X4。再根据这个结果求得最后的Y，Y^T =[1.1819,-2.0173,-1.3075,2.8102,-0.6674]。

注意:虽然在前面步骤中讲到可以将标准化步骤省略但在这里求Y时参与运算的矩阵需要是标准化后的矩阵。

代码实现

1	地区	GDP	农业总产值	工业总产值	第三产业总	固定资产投	消费品零售	城乡居民储
2	济南	27611	2055.3	30195.5	12932.8	651.3	620.99	870.55
3	西安	15204	919.5	10885.5	7450.3	646.7	506.5	1432.86
4	深圳	59271	1816.1	394190.6	78754.9	1090.14	915.45	2625.39
5	南京	33050	1357.2	56288.8	14307.7	1201.88	711.44	1382.24
6	武汉	24963	1221.2	21355.6	12088	822.2	960.58	1376.12
7	成都	20777	1339	11618.7	9393.3	1085.2	875.28	1265
8	沈阳	27487	1335.6	21521.8	12242.3	971.36	808.8	1545.95
9	广州	56271	1436.3	68626	29699.2	1348.93	1675.05	4256.82
10	青岛	28150	1615.3	45598.5	11360.6	1025.4	605.5	1089.49
11	大连	34975	1312.6	35936.9	14697.2	716.21	645.22	1302.29
12	厦门	60176	825.1	112304.2	23122.5	304.65	260.31	464.69
13	杭州	38858	1658.6	63667.7	16004.7	1205.18	704.34	1732.36
14	哈尔滨	17463	1933.5	9133.9	7852.7	523.6	707.4	1261.2
15	长春	21285	1706.1	8632.5	8632.5	460	495.3	965.15
16	宁波	39174	1573	2596.3	14553	1095.67	595.63	1208.98
17	1	2	3	4	5	6	7	8

04 / 总结

Introduction to the Society for Mathematical Modeling

总结

主成分分析法（PCA）本质是一种降维算法，一种对高维度特征数据预处理方法。将高维度的数据保留下最重要的一些特征，去除噪声和不重要的特征，从而实现提升数据处理速度的目的。在实际的生产和应用中，降维在一定的信息损失范围内，可以为我们节省大量的时间和成本。降维也成为应用非常广泛的数据预处理方法。

主成分分析法（PCA）主要应用于评价、决策类问题，
主要有以下优缺点：

总结

优点：

1. 通过PCA降维之后的各个主成分之间是正交的，可以消除原始数据之间相互影响的因素。

2. PCA降维的计算过程并不复杂，因为主要就是对一个协方差矩阵做特征值分解，因此实现起来较简单容易。

3. 在保留大部分主要信息的前提下，起到了降维效果。

缺点：

1. 主成分特征维度的含义具有模糊性，解释性差。（我们最多可以理解成主成分只是由原来的坐标维度线性相加的结果，但加出来之后它到底是啥就不好说了）

2. PCA降维的标准是选取令原数据在新坐标轴上方差最大的主成分。但方差小的特征就不一定不重要，这样的唯一标准有可能会损失一些重要信息。

3. PCA毕竟是只保留特定百分比的主成分，属于“有损压缩”，难免会损失一些信息。