




# 主成分分析

- 主成分分析的基本原理
- 主成分分析的计算步骤
- 主成分分析方法应用实例

## 问题的提出：

在实际问题研究中，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，人们会很自然地想到，能否在相关分析的基础上，用较少的新变量代替原来较多的旧变量，而且使这些较少的新变量尽可能多地保留原来变量所反映的信息？



事实上，这种想法是可以实现的，主成分分析方法就是综合处理这种问题的一种强有力的工具。

主成分分析是把原来多个变量划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。

- 例如，某人要做一件上衣要测量很多尺寸，如身长、袖长、胸围、腰围、肩宽、肩厚等十几项指标，但某服装厂要生产一批新型服装绝不可能把尺寸的型号分得过多？而是从多种指标中综合成几个少数的综合指标，做为分类的型号，利用主成分分析将十几项指标综合成3项指标，一项是反映长度的指标，一项是反映胖瘦的指标，一项是反映特体的指标。

## 一、主成分分析的基本原理

假定有 $n$ 个样本，每个样本共有 $p$ 个变量，构成一个 $n \times p$ 阶的数据矩阵

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \quad (1.1)$$



当 $p$ 较大时，在 $p$ 维空间中考察问题比较麻烦。为了克服这一困难，就需要进行降维处理，即用较少的几个综合指标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地反映原来较多变量指标所反映的信息，同时它们之间又是彼此独立的。

定义：记 $x_1, x_2, \dots, x_p$ 为原变量指标， $z_1, z_2, \dots, z_m$  ( $m \leq p$ ) 为新变量指标

$$\begin{cases} z_1 = l_{11}x_1 + l_{12}x_2 + \dots + l_{1p}x_p \\ z_2 = l_{21}x_1 + l_{22}x_2 + \dots + l_{2p}x_p \\ \dots\dots\dots \\ z_m = l_{m1}x_1 + l_{m2}x_2 + \dots + l_{mp}x_p \end{cases} \quad (1.2)$$

$$l_{i1}^2 + \dots + l_{ip}^2 = 1$$

系数 $l_{ij}$ 的确定原则：

①  $z_i$ 与 $z_j$  ( $i \neq j; i, j=1, 2, \dots, m$ ) 相互无关；

②  $z_1$  是  $x_1, x_2, \dots, x_p$  的一切线性组合中方差最大者,  $z_2$  是与  $z_1$  不相关的  $x_1, x_2, \dots, x_p$  的所有线性组合中方差最大者; ...;  $z_m$  是与  $z_1, z_2, \dots, z_{m-1}$  都不相关的  $x_1, x_2, \dots, x_p$  的所有线性组合中方差最大者。

则新变量指标  $z_1, z_2, \dots, z_m$  分别称为原变量指标  $x_1, x_2, \dots, x_p$  的第1, 第2, ..., 第  $m$  主成分。



从以上的分析可以看出，主成分分析的实质就是确定原来变量 $x_j$  ( $j=1, 2, \dots, p$ ) 在诸主成分 $z_i$  ( $i=1, 2, \dots, m$ ) 上的荷载  $l_{ij}$  ( $i=1, 2, \dots, m; j=1, 2, \dots, p$ )。

从数学上可以证明，它们分别是相关矩阵 $m$ 个较大的特征值所对应的特征向量。

## 二、主成分分析的计算步骤

设有  $n$  个样品，每个样品观测  $p$  个指标，将原始数据写成矩阵

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

1. 将原始数据标准化。这里不妨设上边矩阵已标准化了。

2. 建立变量的相关系数阵：

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

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

3. 求 $\mathbf{R}$ 的特征根  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p > 0$  ;  
及相应的单位特征向量:

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

4. 写出主成分

$$F_i = a_{1i}X_1 + a_{2i}X_2 + \dots + a_{pi}X_p \quad i=1, \dots, p$$

# 计算主成分贡献率及累计贡献率

## ✓ 贡献率

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

## ✓ 累计贡献率

$$\frac{\sum_{k=1}^i \lambda_k}{\sum_{k=1}^p \lambda_k} \quad (i = 1, 2, \dots, p)$$

一般取累计贡献率达**85%~95%**的特征值  $\lambda_1, \lambda_2, \dots, \lambda_m$  所对应的第1、第2、...、第 **$m$**  ( $m \leq p$ ) 个主成分。

### 三、 实例演示

例 对全国**30**个省市自治区经济发展基本情况的八项指标作主成分分析，原始数据如下：

省份	GDP $X_1$	居民消费水平 $X_2$	固定资产投资 $X_3$	职工平均工资 $X_4$	货物周转量 $X_5$	居民消费价格指数 $X_6$	商品零售价格指数 $X_7$	工业总产值 $X_8$
北京	1394.89	2505	519.01	8144	373.9	117.3	112.6	843.43
天津	920.11	2720	345.46	6501	342.8	115.2	110.6	582.51
河北	2849.52	1258	704.87	4839	2033.3	115.2	115.8	1234.85
山西	1092.48	1250	290.9	4721	717.3	116.9	115.6	697.25
内蒙	832.88	1387	250.23	4134	781.7	117.5	116.8	419.39
辽宁	2793.37	2397	387.99	4911	1371.1	116.1	114	1840.55
吉林	1129.2	1872	320.45	4430	497.4	115.2	114.2	762.47
黑龙江	2014.53	2334	435.73	4145	824.8	116.1	114.3	1240.37
上海	2462.57	5343	996.48	9279	207.4	118.7	113	1642.95
江苏	5155.25	1926	1434.95	5943	1025.5	115.8	114.3	2026.64

续表

省份	GDP $X_1$	居民消 费水平 $X_2$	固定资 产投资 $X_3$	职工平 均工资 $X_4$	货物周 转 量 $X_5$	居民消费 价格指数 $X_6$	商品零 售价格指数 $X_7$	工业总 产 值 $X_8$
浙江	3524.79	2249	1006.39	6619	754.4	116.6	113.5	916.59
安徽	2003.58	1254	474	4609	908.3	114.8	112.7	824.14
福建	2160.52	2320	553.97	5857	609.3	115.2	114.4	433.67
江西	1205.11	1182	282.84	4211	411.7	116.9	115.9	571.84
山东	5002.34	1527	1229.55	5145	1196.6	117.6	114.2	2207.69
河南	3002.74	1034	670.35	4344	1574.4	116.5	114.9	1367.92
湖北	2391.42	1527	571.68	4685	849	120	116.6	1220.72
湖南	2195.7	1408	422.61	4797	1011.8	119	115.5	843.83
广东	5381.72	2699	1639.83	8250	656.5	114	111.6	1396.35
广西	1606.15	1314	382.59	5105	556	118.4	116.4	554.97



续表

省份	GDP $X_1$	居民消 费水平 $X_2$	固定资 产投资 $X_3$	职工平 均工资 $X_4$	货物周 转 量 $X_5$	居民消费 价格指数 $X_6$	商品零 售价格指数 $X_7$	工业总 产 值 $X_8$
海南	364.17	1814	198.35	5340	232.1	113.5	111.3	64.33
四川	3534	1261	822.54	4645	902.3	118.5	117	1431.81
贵州	630.07	942	150.84	4475	301.1	121.4	117.2	324.72
云南	1206.68	1261	334	5149	310.4	121.3	118.1	716.65
西藏	55.98	1110	17.87	7382	4.2	117.3	114.9	5.57
陕西	1000.03	1208	300.27	4396	500.9	119	117	600.98
甘肃	553.35	1007	114.81	5493	507	119.8	116.5	468.79
青海	165.31	1445	47.76	5753	61.6	118	116.3	105.8
宁夏	169.75	1355	61.98	5079	121.8	117.1	115.3	114.4
新疆	834.57	1469	376.95	5348	339	119.7	116.7	428.76

数据来源：1996年《中国统计年鉴》。



第一步 将原始数据标准化。

第二步 建立指标之间的相关系数阵 $R$ 如下

	$X_1$	$X_2$	$X_3$	$X_4$	$X_5$	$X_6$	$X_7$	$X_8$
$X_1$	1.000	.267	.951	.191	.617	-.274	-.264	.874
$X_2$	.267	1.000	.426	.718	-.151	-.234	-.593	.363
$X_3$	.951	.426	1.000	.400	.431	-.282	-.359	.792
$X_4$	.191	.718	.400	1.000	-.356	-.134	-.539	.104
$X_5$	.617	-.151	.431	-.356	1.000	-.255	.022	.659
$X_6$	-.274	-.234	-.282	-.134	-.255	1.000	.760	-.126
$X_7$	-.264	-.593	-.359	-.539	.022	.760	1.000	-.192
$X_8$	.874	.363	.792	-.104	.659	-.126	-.192	1.000

### 第三步 求R的特征值和特征向量。

主成分	特征值	方差贡献率	累计贡献率
1	3.755	46.943	46.943
2	2.195	27.443	74.386
3	1.214	15.178	89.564
4	.403	5.033	94.596
5	.213	2.660	97.256
6	.139	1.737	98.993
7	6.594E-02	.824	99.817
8	1.462E-02	.183	100.000

■ 从上表看，前3个特征值累计贡献率已达89. 564%，说明前3个主成分基本包含了全部指标具有的信息，我们取前3个特征值，并计算出相应的特征向量：

第一特征向量 $a_1$	第二特征向量 $a_2$	第三特征向量 $a_3$
0. 470641	0. 107995	0. 19241
0. 456708	0. 258512	0. 109819
0. 424712	0. 287536	0. 19241
—0. 31944	0. 400931	0. 397525
0. 312729	—0. 40431	0. 24505
0. 250802	0. 498801	—0. 24777
0. 240481	—0. 48868	0. 332179
—0. 26267	0. 167392	0. 723351

因而前三个主成分为：

第一主成分：

$$\begin{aligned} F_1 = & 0.470641X_1 + 0.456708X_2 + 0.424712X_3 \\ & - 0.31944X_4 + 0.312729X_5 + 0.250802X_6 \\ & + 0.240481X_7 - 0.26267X_8 \end{aligned}$$

第二主成分：

$$\begin{aligned} F_2 = & 0.107995X_1 + 0.258512X_2 + 0.287536X_3 \\ & + 0.400931X_4 - 0.40431X_5 + 0.498801X_6 \\ & - 0.48868X_7 + 0.167392X_8 \end{aligned}$$

### 第三主成分：

$$\begin{aligned} F_3 = & 0.19241X_1 + 0.109819X_2 + 0.19241X_3 \\ & + 0.397525X_4 + 0.24505X_5 - 0.24777X_6 \\ & + 0.332179X_7 + 0.723351X_8 \end{aligned}$$

在第一主成分的表达式中第一、二、三项指标的系数较大，这三个指标起主要作用，我们可以把第一主成分看成是由国内生产总值、固定资产投资和居民消费水平所该划的反映经济发展状况的综合指标；

在第二主成分中，第四、五、六、七项指标的影响大，且第六、七项指标的影响尤其大，可将之看成是反映物价指数、职工工资和货物周转量的综合指标；

在第三主成分中，第八项指数影响最大，远超过其它指标的影响，可单独看成是工业总产值的影响。

## 四.主成分分析的应用

### ■ 1.主成分分析用于系统评估

利用主成分  $F_1, \dots, F_p$  做线性组合, 并以每个主成分  $F_i$  的方差贡献率  $\alpha_i$  作为权数构造一个综合评价函数:

$$y = \alpha_1 F_1 + \dots + \alpha_m F_m$$

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或分类划级。

## 例1.

城市环境生态化是城市发展的必然趋势，表现为社会、经济、环境与生态全方位的现代化水平，一个符合生态规律的生态城市应该是结构合理、功能高效和关系协调的城市生态系统。所谓结构合理是指适度的人口密度，合理的土地利用，良好的环境质量，充足的绿地系统，完善的基础设施，有效的自然保护；功能高效是指资源的优化配置、物力的经济投入、人力的充分发挥、物流的畅通有序、信息流的快捷；关系协调是指人和自然协调、社会关系协调、城乡协调、资源利用和更新协调。一个城市要实现生态城市的发展目标，关键是在市场经济的体制下逐步改善城市的生态环境质量，防止生态环境质量恶化，因此，对城市的生态环境水平调查评价很有必要。



我们对江苏省十个城市的生态环境状况进行了调查，得到生态环境指标的指数值，见表1。现对生态环境水平分析和评价。

表 1 指标指数值

一级指标	结构				功能			协调度		生态环境
二级 指标	人口 结构 $x_1$	基础 设施 $x_2$	地理 结构 $x_3$	城市 绿化 $x_4$	物质 还原 $x_5$	资源 配置 $x_6$	生产 效率 $x_7$	城市 文明 $x_8$	可持 续性 $x_9$	水平排序
无锡市	0.7883	0.7633	0.4745	0.8246	0.8791	0.9538	0.8785	0.6305	0.8928	5
常州市	0.7391	0.7287	0.5126	0.7603	0.8736	0.9257	0.8542	0.6187	0.7831	7
镇江市	0.8111	0.7629	0.881	0.6888	0.8183	0.9285	0.8537	0.6313	0.5608	10
张家港市	0.6587	0.8552	0.8903	0.8977	0.9446	0.9434	0.9027	0.7415	0.8419	3
连云港市	0.6543	0.7564	0.8288	0.7926	0.9202	0.9154	0.8729	0.6398	0.8464	6
扬州市	0.8259	0.7455	0.785	0.7856	0.9263	0.8871	0.8485	0.6142	0.7616	8
泰州市	0.8486	0.78	0.8032	0.6509	0.9185	0.9357	0.8473	0.5734	0.8234	9
徐州市	0.6834	0.949	0.8862	0.8902	0.9505	0.876	0.9044	0.898	0.6384	2
南京市	0.8495	0.8918	0.3987	0.6799	0.862	0.9579	0.8866	0.6186	0.9604	4
苏州市	0.7846	0.8954	0.397	0.9877	0.8873	0.9741	0.9035	0.7382	0.8514	1



- **princomp** 函数

**功能：**进行主成分分析。

**语法：**

```
PC = princomp(X)
```

```
[PC, SCORE, latent, tsquare] = princomp(X)
```

**描述：**

`[PC, SCORE, latent, tsquare] = princomp(X)` 根据数据矩阵  $X$  返回因子成分 PC、Z 分数 SCORE、 $X$  的协方差矩阵的特征值 latent 和 Hotelling's  $T^2$  统计量 tsquare。

Z 分数是通过将原始数据转换到因子成分空间中得到的数据。latent 向量的值为 SCORE 的列的方差。Hotelling's  $T^2$  为来自数据集合中心的每一个观测量的多变量距离的度量。

利用Matlab中的princomp命令实现。具体程序如下

```
X= [0.7883 0.7391 0.8111 0.6587 0.6543 0.8259 0.8486 0.6834 0.8495 0.7846  
    0.7633 0.7287 0.7629 0.8552 0.7564 0.7455 0.7800 0.9490 0.8918 0.8954  
    0.4745 0.5126 0.8810 0.8903 0.8288 0.7850 0.8032 0.8862 0.3987 0.3970  
    0.8246 0.7603 0.6888 0.8977 0.7926 0.7856 0.6509 0.8902 0.6799 0.9877  
    0.8791 0.8736 0.8183 0.9446 0.9202 0.9263 0.9185 0.9505 0.8620 0.8873  
    0.9538 0.9257 0.9285 0.9434 0.9154 0.8871 0.9357 0.8760 0.9579 0.9741  
    0.8785 0.8542 0.8537 0.9027 0.8729 0.8485 0.8473 0.9044 0.8866 0.9035  
    0.6305 0.6187 0.6313 0.7415 0.6398 0.6142 0.5734 0.8980 0.6186 0.7382  
    0.8928 0.7831 0.5608 0.8419 0.8464 0.7616 0.8234 0.6384 0.9604 0.8514];  
  
x = x';  
  
stdr = std(x);    % 求各变量标准差  
  
[n,m] = size(x);  
  
sddata = x./stdr(ones(n,1),:);    % 标准化变换  
  
[p, princ, egenvalue] = princomp(sddata)    % 调用主成分分析程序
```

```

p3 = p(:, 1 : 3)    % 输出前三个主成分系数
sc = princ(:, 1 : 3) % 输出前三个主成分得分
egenvalue    % 输出特征根
per = 100 * egenvalue / sum(egenvalue) % 输出各个主成分贡献率

```

执行后得到所要结果，这里是前三个主成分、主成分得分、特征根。即

$$p = \begin{bmatrix} -0.3677 & 0.1442 & -0.3282 \\ 0.3702 & 0.2313 & -0.3535 \\ 0.1364 & -0.5299 & 0.0498 \\ 0.4048 & 0.1812 & 0.0582 \\ 0.3355 & -0.1601 & 0.5664 \\ -0.1318 & 0.5273 & -0.0270 \\ 0.4236 & 0.3116 & -0.0958 \\ 0.4815 & -0.0267 & -0.2804 \\ -0.0643 & 0.4589 & 0.5933 \end{bmatrix}, \quad princ = \begin{bmatrix} -0.8301 & 1.3897 & 0.3946 \\ -1.3364 & -0.2159 & 0.2729 \\ -1.8408 & -1.6267 & -2.0552 \\ 2.3754 & 0.1623 & 0.9043 \\ 0.3634 & -0.9002 & 1.4326 \\ -0.9266 & -1.6636 & 0.5837 \\ -1.8984 & -0.6907 & 0.5432 \\ 3.9332 & -1.5024 & -0.9765 \\ -1.2134 & 2.3442 & -0.4498 \\ 1.3736 & 2.7034 & -0.6469 \end{bmatrix}$$

$$egenvalue = [3.8811, 2.6407, 1.0597]', \quad per = [43.12, 29.34, 11.97]'.$$

这样，前三个主成分为

$$z_1 = -0.3677x_1 + 0.3702x_2 + 0.1364x_3 + 0.4048x_4 + 0.3355x_5 - 0.1318x_6 \\ + 0.4236x_7 + 0.4815x_8 - 0.0643x_9$$

$$z_2 = 0.1442x_1 + 0.2313x_2 - 0.5299x_3 + 0.1812x_4 - 0.1601x_5 + 0.5273x_6 \\ + 0.3116x_7 - 0.0267x_8 + 0.4589x_9$$

$$z_3 = -0.3282x_1 - 0.3535x_2 + 0.0498x_3 + 0.0582x_4 + 0.5664x_5 - 0.0270x_6 \\ - 0.0958x_7 - 0.2804x_8 + 0.5933x_9$$

第一主成分贡献率为**43.12%**，第二主成分贡献率为**29.34%**，第三主成分贡献率为**11.97%**，前三个主成分累计贡献率达**84.24%**。

如果按**80%** 以上的信息量选取新因子，则可以选取前三个新因子。第一新因子**z1** 包含的信息量最大为**43.12%**，它的主要代表变量为**X8(城市文明)**、**X7(生产效率)**、**X4 (城市绿化)**，其权重系数分别为**0.4815**、**0.4236**、**0.4048**，反映了这三个变量与生态环境水平密切相关，第二新因子**z2**

包含的信息量次之为29.34%，它的主要代表变量为X3(地理结构)、X6(资源配置)、X9 (可持续性)，其权重系数分别为0.5299、0.5273、0.4589，第三新因子 Z3包含的信息量为11.97%，代表总量为 X9(可持续性)、X5(物质还原)，权重系数分别为0.5933、0.5664。这些代表变量反映了各自对该新因子作用的大小，它们是生态环境系统中最重要影响因素。

根据前三个主成分得分，用其贡献率加权，即得十个城市各自的总得分

$$\begin{aligned} F &= 43.12\%princ(:,1) + 29.34\%princ(:,2) + 11.97\%princ(:,3) \\ &= [0.0970, -0.6069, -1.5170, 1.1801, 0.0640, -0.8178, -0.9562, 1.1383, 0.1107, 1.3077]' \end{aligned}$$

根据总得分排序，结果见表1。

## ■ 2.主成分回归

考察进口总额 $Y$ 与三个自变量：国内总产值 $x_1$ ,存储量 $x_2$ , 总消费量 $x_3$ 之间的关系，现收集数据如下，试用主成分回归分析方法求进口总额与总产值、存储量和总消费量的定量关系式.

序号	$x_1$	$x_2$	$x_3$	$Y$
1	149.3	4.2	108.1	15.9
2	161.2	4.1	114.8	16.4
3	171.5	3.1	123.2	19.0
4	175.5	3.1	126.9	19.1
5	180.8	1.1	132.1	18.8
6	190.7	2.2	137.7	20.4
7	202.1	2.1	146.0	22.7
8	212.4	5.6	154.1	26.5
9	226.1	5.0	162.3	28.1
10	231.9	5.1	164.3	27.6
11	239.0	0.7	167.6	26.3

- 分析：本题目可先尝试一般的线性回归模型，但拟合的效果一般，故可尝试主成分回归分析方法
- 解：首先对各个变量数据进行标准化处理，其次，建立指标之间的相关系数阵并求出相关阵的特征值分别为：

$$\lambda_1 = 1.999, \lambda_2 = 0.998, \lambda_3 = 0.003$$

前2个主成分的累计贡献率在**99%**以上，故取**2**个主成分（ $x_i^*$ 表示 $x_i$ 的标准化变量）：

$$Z_1 = 0.7063x_1^* + 0.0435x_2^* + 0.7065x_3^*,$$

$$Z_2 = -0.0357x_1^* + 0.9990x_2^* - 0.0258x_3^*$$

由主成分回归得到的标准化回归方程为

$$\begin{aligned} Y^* &= 0.68998Z_1 + 0.1913Z_2 \\ &= 0.4804x_1^* + 0.2211x_2^* + 0.4825x_3^* \end{aligned}$$

用原变量表示的回归方程

$$Y = -9.130 + 0.0727x_1 + 0.6091x_2 + 0.1062x_3$$