主成分分析方法

___河流水质的综合评价

思想：利用降维的思想，把多指标转化为少数几个综合指标。

在研究多变量问题时，变量太多会增大计算量和增加分析问题的复杂性，人们自然希望在进行定量分析的过程中涉及的变量较少，而得到的信息量又较多。主成分分析是解决这一问题的理想工具。（主要分析众多变量之间的相关性）

主成分分析

思想：利用降维的思想，把多指标转化为少数几个综合指标。

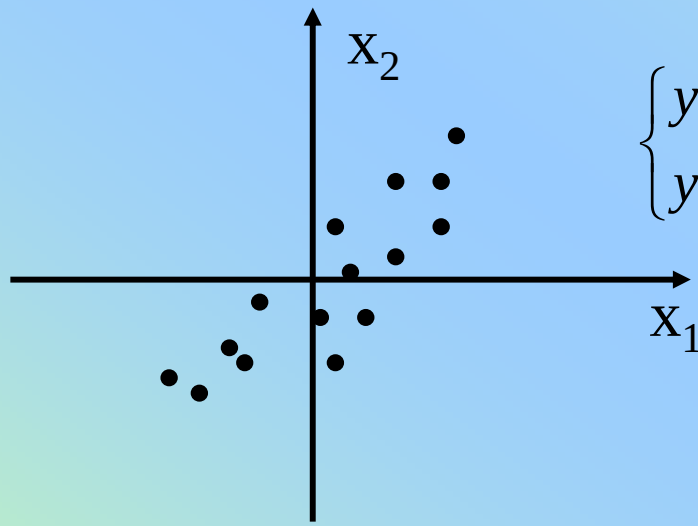
在研究多变量问题时，变量太多会增大计算量和增加分析问题的复杂性，人们自然希望在进行定量分析的过程中涉及的变量较少，而得到的信息量又较多。主成分分析是解决这一问题的理想工具。（主要分析众多变量之间的相关性）

例如：在学生学习过程中，已经修完p门课程，其成绩为$x_1, x_2, \ldots, x_p$，如何评价每个学生的综合能力？假设每门课程有权重$c_1, c_2, \ldots, c_p$，则加权之和为：

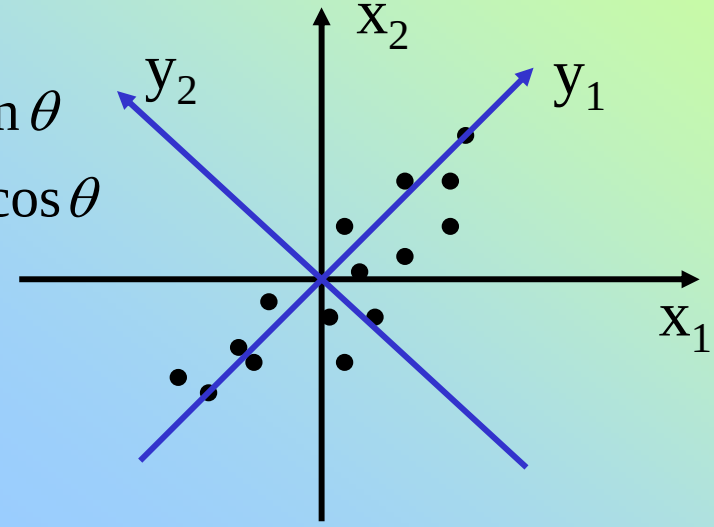
$$s = c_1x_1 + c_2x_2 + \ldots + c_px_p$$

每个学生对应这样一个成绩，假设有n个学生，其成绩分别为：$s_1, s_2, \ldots, s_n$。如果这些值很分散，表明每个人的综合能力能很好地区分。关键是如何确定权重$c_1, c_2, \ldots, c_p$，在数学上反映的问题是什么呢？

主成分的几何意义



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



变换的目的是为了使得n个样本点在y_1轴方向上的离散程度最大，既y_1的方差达最大。说明变量y_1代表了原始数据的绝大部分信息，对y_2忽略也无损大局，即由两个指标压缩成一个指标。

一、数据结构

适合用主成分分析的数据具有如下结构：

指 标

编号	X1	X2	X3	X4	……	Xm
1						
2		x_{ij}				
3						
…		$x^*_{ij} = \frac{x_{ij} - \bar{x}_j}{\sqrt{Dx_j}}, i = 1,2,\cdots,n,\ j = 1,2,\cdots,m$				
n						

样本

主成分分析最大的问题是受量纲的影响，因此，实际应用中，需要对数据进行标准化。一般使用协方差矩阵Σ或相关系数矩阵R进行分析。

二、主成分的基本思想

设$X_1,\ldots,X_p$表示以$x_1,\ldots,x_p$为样本观测值的随机变量，如果能找到$c_1,\ldots,c_p$，使得

$$\max D(c_1X_1+\ldots+c_pX_p)$$

但上述公式必须加上某种限制，否则权值可选择无穷大而没有意义，通常规定

$$c_1^2+\ldots+c_p^2=1$$

由于解$c_1,\ldots,c_p$是p维空间的一个单位向量，它代表一个“方向”，称为主成分方向。

二、主成分的基本思想

由于一个主成分不足以代表原来的p个变量的信息。因此需要寻找第二个乃至第三、四个主成分，原则上，第二个主成分不应该再包含第一个主成分的信息，统计上的描述就是让这两个主成分的协方差为零，几何上就是这两个主成分的方向正交。具体确定各个主成分的方法如下：

设Z_i表示第i个主成分，可设

$$\begin{cases} Z_1 = c_{11}X_1 + c_{12}X_2 + \cdots + c_{1p}X_p \\ Z_2 = c_{21}X_1 + c_{22}X_2 + \cdots + c_{2p}X_p \\ \vdots \\ Z_p = c_{p1}X_1 + c_{p2}X_2 + \cdots + c_{pp}X_p \end{cases}$$

二、主成分的基本思想

确定$(c_{11},\ldots,c_{1p})$, 使得$\max D(Z_1)$，并且满足

$$c_{11}^2+\ldots+c_{1p}^2=1$$

确定$(c_{21},\ldots,c_{2p})$，使得$\max D(Z_2)$，并且满足$(c_{21},\ldots,c_{2p})$与$(c_{11},\ldots,c_{1p})$垂直，和

$$c_{21}^2+\ldots+c_{2p}^2=1$$

确定$(c_{31},\ldots,c_{3p})$，使得$\max D(Z_3)$，并且满足$(c_{31},\ldots,c_{3p})$与$(c_{11},\ldots,c_{1p})$，$(c_{21},\ldots,c_{2p})$垂直，和

$$c_{31}^2+\ldots+c_{3p}^2=1$$

……

如何确定主成分的个数？

二、主成分的基本思想

在实际研究中，由于主成分的目的是为了降维，减少变量的个数，故一般选取少量的主成分（不超过5或6个），只要它们能包含原变量信息量的80%以上即可。

三、主成分分析的具体实现

设相关矩阵为$R_{p\times p}$，求特征方程$|R-\lambda I|=0$，其解为特征根λ_i将解由小到大进行排序为：

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0$$

1） $(c_{i1},\ldots,c_{ip})$实际上是对应于λ_i的特征向量。若原变量服从正态分布，则各主成分之间相互独立；

2）全部p个主成分所反映的n例样本的总信息，等于p个原变量的总信息。信息量的多少，用变量的方差来度量。

3）各主成分的作用大小是：$Z_1 \geq Z_2 \geq \ldots \geq Z_p$；

4）第i个主成分的贡献率是 $\dfrac{\lambda_i}{\sum_{j=1}^{p} \lambda_j} \times 100\%$

5）前m个主成分的累计贡献率是 ：

$$\frac{\sum_{i=1}^{m} \lambda_i}{\sum_{j=1}^{p} \lambda_j} \times 100\%$$

在应用时，一般取累计贡献率为80%以上比较好。

四、MATLAB软件实现

[pc, score, variance, t2]=princomp(X)

输入数据矩阵：

$$X=\begin{array}{c}\begin{matrix}x_1 & x_2 & & x_p\end{matrix}\\ \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}\end{array}$$

一般地，要求n > p。模型：

$$\begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_m \\ \vdots \\ z_p \end{bmatrix} = C^T \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix}$$

要求 m < p 。

输出变量：

① pc 主分量z_j的系数$(c_{i1},\ldots,c_{ip})$，也叫因子系数；注意：$pc^T pc$ = 单位阵

② score是主分量下的得分值；得分矩阵与数据矩阵X的阶数是一致的；

③ variance是score对应列的方差向量，即相关系数矩阵R的特征值；容易计算方差所占的百分比

percent-v = 100*variance/sum(variance);

④ t2表示检验的t2-统计量（主要用于方差分析）

例1：

某医学院测得20例肝病患者的4项肝功能指标：SGPT（转氨酶）X1

肝大指数X2

ZnT（硫酸锌浊度）X3

AFP（胎甲球）X4

其数据如下表ex1.xls。

OBS	X1	X2	X3	X4
1	40	2.0	5	20
2	10	1.5	5	30
3	120	3.0	13	50
4	250	4.5	18	0
5	120	3.5	9	50
6	10	1.5	12	50
7	40	1.0	19	40
8	270	4.0	13	60
9	280	3.5	11	60
10	170	3.0	9	60
11	180	3.5	14	40
12	130	2.0	30	50
13	220	1.5	17	20
14	160	1.5	35	60
15	220	2.5	14	30
16	140	2.0	20	20
17	220	2.0	14	10
18	40	1.0	10	0
19	20	1.0	12	60
20	120	2.0	20	0

zcf.m

数据未标准化的计算结果：

pc = 0.9998 0.0071 -0.0179 -0.0091
0.0082 -0.0055 -0.0469 0.9989
0.0184 -0.0249 0.9984 0.0466
0.0066 -0.9996 -0.0247 -0.0067

variance = 1.0e+003 *[7.9046 0.4786 0.0522 0.0004]；

t2 =[2.9602 3.2317 1.5008 9.6644 3.2463 2.6023 2.0675
3.9587 5.1659 2.3459 1.4789 4.9255 5.5928 8.4165
1.7758 1.0442 4.3163 4.2880 3.8299 3.5884]；

R = 1.0000 0.6950 0.2195 0.0249
0.6950 1.0000 -0.1480 0.1351
0.2195 -0.1480 1.0000 0.0713
0.0249 0.1351 0.0713 1.0000 （相关系数矩阵）

score =

-98.2666	15.0539	-7.8344	0.2006
-128.1984	4.8485	-7.5223	-0.0946
-17.9323	-14.5758	-2.0656	0.6464
111.8154	36.1911	1.7700	1.5374
-18.0019	-14.4790	-6.0827	0.9595
-127.9383	-15.3188	-1.0280	0.0971
-97.8856	-5.2821	5.6958	-0.2805
132.1077	-23.5196	-5.0403	0.2205
142.0644	-23.3965	-7.1924	-0.4626
32.0482	-24.1200	-7.2001	-0.0597
42.0113	-4.1837	-1.9154	0.7167
-7.6292	-14.9231	14.7760	0.3492
81.9100	16.0278	0.9537	-1.3687
22.5176	-24.8297	19.0082	-0.2560
81.9285	6.1005	-2.3359	-0.5769
1.9875	15.3860	5.3551	-0.0055
81.7933	26.0963	-1.8177	-0.9418
-98.3136	34.9280	-2.3006	-0.4308
-117.8791	-25.2419	-1.4306	-0.5601
-18.1390	35.2379	6.2072	0.3100

数据标准化的计算结果：　　Z=zscore(X);

		z_1	z_2	z_3	z_4
pc =	x_1	-0.7000	0.0950	0.2400	0.6659
	x_2	-0.6898	-0.2836	-0.0585	-0.6636
	x_3	-0.0879	0.9042	0.2703	-0.3189
	x_4	-0.1628	0.3050	-0.9305	0.1208

score （略）

variance =[1.7183　1.0935　0.9813　0.2069]；

t2 =[2.9602　3.2317　1.5008　9.6644　3.2463　2.6023　2.0675
　　3.9587　5.1659　2.3459　1.4789　4.9255　5.5928　8.4165
　　1.7758　1.0442　4.3163　4.2880　3.8299　3.5884]；

Mean	**138.0**	**2.325**	**15.0**	**35.5**
Std	**88.887**	**1.055**	**7.4197**	**21.8788**

$$\begin{cases} Z_1 = -0.7x_1 - 0.6898x_2 - 0.0879x_3 - 0.1628x_4 \\ Z_2 = 0.095x_1 - 0.2836x_2 + 0.9042x_3 + 0.305x_4 \\ Z_3 = 0.24x_1 - 0.0585x_2 + 0.2703x_3 - 0.9305x_4 \\ Z_4 = 0.6659x_1 - 0.6636x_2 - 0.3189x_3 + 0.1208x_4 \end{cases}$$

说明：系数的绝对值越大，该主成分受该指标的影响就越大。有如下解释：

① Z_1：x_1，x_2，指急性炎症；（由实际问题解释）

② Z_2：x_3，　　指慢性炎症；

③ Z_3：x_4，　　指向原发性肝癌可疑；

（前三项综合指标的信息量已经达到94.828%）

应用：

若测得某一个肝炎病人的4项指标分别为：$X_1=50$，$X_2=2.0$，$X_3=31$，$X_4=45$，如何判断该病人患病情况？

计算结果：(得分值)

$Z_1=-0.6452$ ， $Z_2=2.075^*$ ， $Z_3=0.0407$ ，$Z_4=1.0899$。

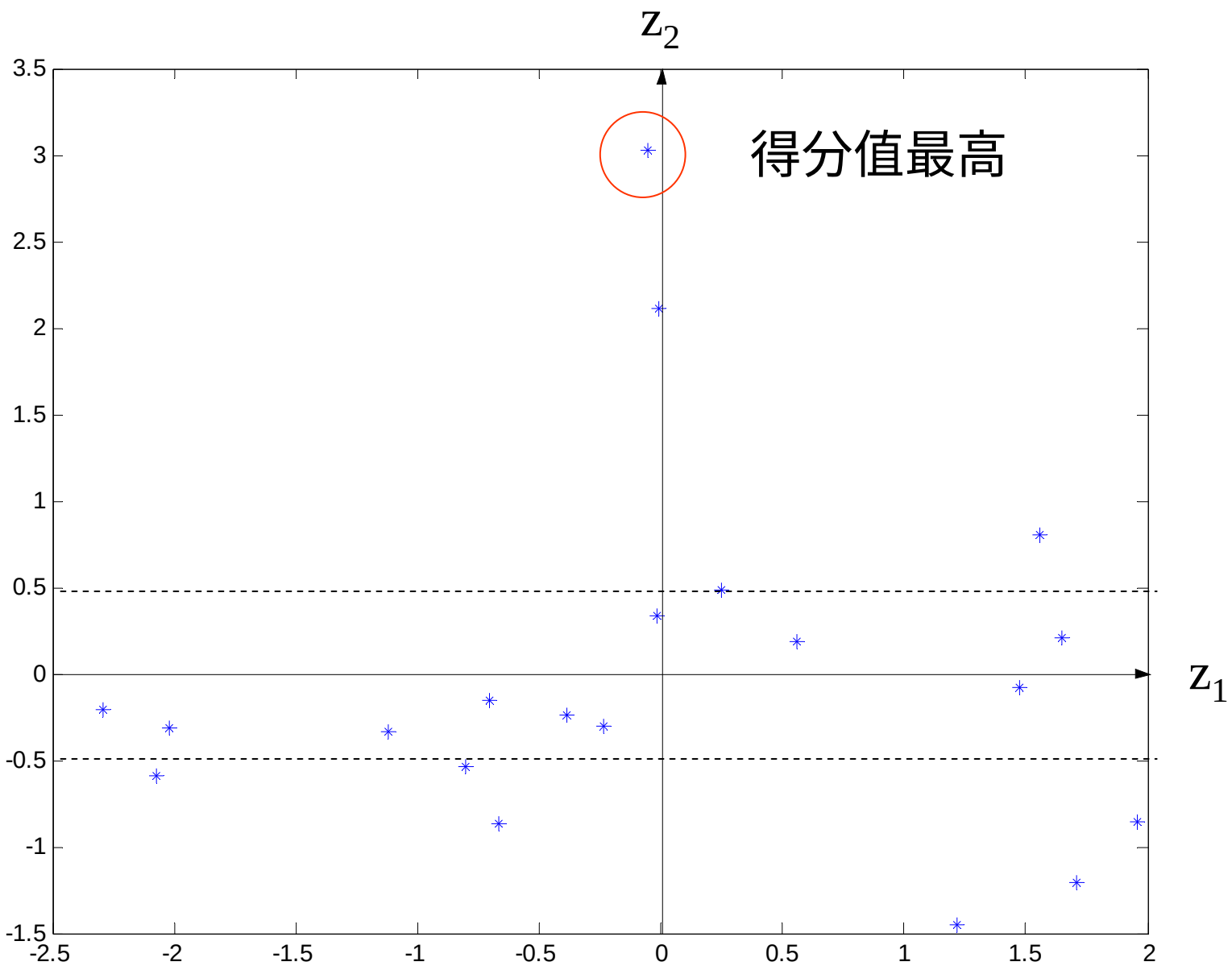
由此诊断该患者肝炎症类型很可能为慢性。

归纳：

1、主成分能降低所研究的数据空间的维数；

2、有时可通过因子负荷c_{ij}的结构，弄清x变量间的某些关系；

3、多维数据的一种图形表示方法；（选取前两个主成分或某两个主成分根据得分，画出n个样品在二维平面上的分布情况，由图形可直观地看出各样品在主分量中的地位，进而对样品进行分类处理。


z_2
得分值最高
z_1
3.5
3
2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2.5
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2

例2：河流水质综合评价

水质评价是环境质量评价的一个方面，用数学模型方法进行这方面的定量化研究，是有意义的。目前常见的方法主要有：

1）简单指数法；

2）分级加权评分法；

3）概率统计法；

4）模糊数学法等等。

表 1　湟水流域各断面及水质标准样点在主要污染指标上的标准化数据表

指标 / 断面	X_1（高锰酸盐指数）	X_2（生化需氧量）	X_3（氨氮）	X_4（亚硝酸盐氮）	X_5（硝酸盐）	X_6（挥发酚）	X_7（总氰化物）	X_8（总砷）	X_9（六价铬）	X_{10}（石油类）
1 扎马隆	-0.857	-0.454	-0.464	-0.497	-1.279	-0.029	-0.363	-0.486	-0.804	-0.596
2 西钢桥	-0.547	-0.314	-0.379	-0.439	-0.527	-0.029	0.605	-0.486	-0.633	-0.576
3 新宁桥	-0.392	-0.035	-0.114	-0.362	0.461	0.263	-0.363	-0.486	-0.804	-0.196
4 报社桥	1.547	0.943	1.673	-0.091	-0.056	0.848	0.363	-0.400	0.262	2.388
5 小峡口	0.462	0.384	0.003	-0.008	0.649	0.556	0.848	-0.187	-0.249	-0.186
6 民和桥	-0.663	0.384	-0.499	0.035	2.131	-1.199	-0.848	-0.486	-0.846	-0.576
7 峡门桥	-0.857	-0.593	-0.680	-0.439	-0.880	-1.119	-0.848	-0.315	0.731	-0.316
8 桥头桥	-0.547	-0.593	-0.457	-0.401	-0.174	-1.199	-0.848	-0.359	1.584	-0.226
9 长宁桥	-0.392	-0.454	-0.170	-0.323	-0.597	-1.199	-0.848	-0.315	1.967	-0.326
10 湟泽	-0.353	-0.593	0.283	-0.342	-0.715	-1.199	-0.848	-0.187	2.138	-0.196
11 湟泽桥	-0.120	-0.454	0.102	-0.187	-1.350	-0.029	-0.121	-0.400	-0.633	-0.556
12 塔尔桥	-1.051	-0.593	-0.681	-0.420	-0.433	-1.199	-0.848	-0.443	0.902	-0.236
13 朝阳桥	-0.081	-0.175	0.542	0.011	0.955	0.556	0.363	-0.400	-0.633	-0.376
14 老幼堡	-1.051	-0.454	-0.470	-0.434	0.367	-0.029	-0.605	-0.443	-0.931	-0.616
15 七一桥	3.215	4.015	3.695	0.253	1.166	1.726	0.121	-0.315	-0.079	2.448
16 三其桥	0.384	-0.035	-0.420	-0.299	-0.997	1.433	-0.363	-0.443	-0.676	-0.596
17 沙塘桥	0.617	-0.175	-0.383	-0.357	-1.091	1.433	0.121	-0.400	-0.676	2.198
18 第Ⅱ类	-0.547	-0.454	-0.594	-0.100	-0.386	-0.907	-0.121	1.474	-0.846	-0.636
19 第Ⅲ类	0.229	-0.134	-0.493	0.142	0.790	-0.029	1.090	1.474	-0.846	-0.636
20 第Ⅳ类	1.004	-0.035	-0.493	4.258	1.966	1.433	3.512	3.604	1.072	-0.186

原始数据为年均值，摘自《1998 年青海省环境质量监测报告书》(1995，5)

主要计算结果及分析

R=corrcoef(X)

1、计算相关系数矩阵R

1.0000	0.8674	0.8360	0.3622	0.3367	0.7520	0.4569	0.1938	-0.0122	0.7746
	1.0000	0.9042	0.1324	0.4219	0.5451	0.1631	-0.0525	-0.1275	0.6806
		1.0000	0.0136	0.2503	0.4716	0.0642	-0.1767	0.0336	0.7271
			1.0000	0.5900	0.3914	0.8557	0.8576	0.1842	0.0227
				1.0000	0.2055	0.4932	0.4270	-0.1381	0.0155
					1.0000	0.5980	0.1904	-0.3940	0.5424
						1.0000	0.7934	-0.0949	0.1019
							1.0000	0.1053	-0.1413
								1.0000	0.0537
									1.0000

显然，x_1,x_2,x_3,x_6,x_{10}具有较强的相关性，这些指标均属于有机污染指标;指标X_4、X_5、X_7、X_8之间具有较强的相关性,而这些指标属于无机污染指标。X_9为一特殊的无机污染指标,与其它指标相关性较弱。

2、计算R的特征值：

variance = 4.6031 2.9169 1.3024 0.8673 | 0.2516 0.2407 0.1097
0.0639 0.0501 0.0167

由于前四个特征值对应的累计方差贡献率已达92.97%,故前四个主成分已反映原始指标所提供的绝大部分信息,可利用它们对各断面水质污染程度进行评价。

3、计算前四个主成分

$$z_1 = 0.453x_1 + 0.391x_2 + 0.354x_3 + 0.275x_4 + 0.245x_5 + 0.38x_6 + 0.312x_7 + 0.18x_8 - 0.049x_9 + 0.324x_{10}$$

$$z_2 = -0.142x_1 - 0.266x_2 - 0.34x_3 + 0.461x_4 + 0.222x_5 - 0.02x_6 + 0.412x_7 + 0.511x_8 + 0.048x_9 - 0.318x_{10}$$

$$z_3 = 0.088x_1 + 0.026x_2 + 0.164x_3 + 0.162x_4 - 0.053x_5 - 0.355x_6 - 0.102x_7 + 0.108x_8 + 0.873x_9 + 0.158x_{10}$$

$$z_4 = -0.094x_1 + 0.31x_2 + 0.188x_3 - 0.016x_4 + 0.741x_5 - 0.399x_6 + 0.206x_7 - 0.077x_8 - 0.11x_9 - 0.303x_{10}$$

由线性表达式中系数的大小及符号，可对各主成分的实际意义作如下解释：第一主成分为除 x_9 以外的其它九项指标的综合；第二主成分则与五项无机污染指标成正相关，而与五项有机污染指标负相关；第三、四主成分又分别体现 x_9、 x_5 的信息相对多一些。

4、计算主成分得分及综合得分

最后计算出二十个样点的主成分得分及综合得分,给予各断面水质污染程度的定量化描述,得分越大,表明污染程度越严重,由此便可对样点就污染程度进行排序和分级,具体结果见下表:

注意：计算综合得分的计算公式：

$$z=\frac{1}{\sum_{i=1}^{4}\lambda_i}(\lambda_1 z_1+\lambda_2 z_2+\lambda_3 z_3+\lambda_4 z_4)$$

	z_1	z_2	z_3	z_4	综合得分z	排序	污染程度分级
	-1.5617	-0.3688	-0.9731	-0.7338	-1.0494	17	轻
	-0.8432	0.1202	-0.9083	-0.3653	-0.5192	10	轻
	-0.3593	-0.3487	-0.9751	0.4668	-0.3649	7	轻
	2.7294	-1.9184	0.6527	-0.7517	0.7396	2	重
	0.8939	0.2514	-0.5317	0.2199	0.4486	3	中
	-0.7640	0.2122	-0.6039	2.5833	-0.1490	5	中（轻）
	-2.0988	-0.2478	0.8159	-0.2530	-0.9846	16	轻
	-1.7475	-0.2013	1.6310	0.1972	-0.6539	13	轻
	-1.6472	-0.3429	2.0538	-0.0495	-0.6141	12	轻
	-1.5009	-0.4680	2.3146	-0.1659	-0.5576	11	轻
	-0.6195	-0.1772	-0.6743	-0.1004	-0.4473	8	轻
	-2.1082	-0.1929	0.9544	0.0945	-0.9228	15	轻
	0.4687	0.1046	-0.8578	0.6407	0.1962	4	中
	-1.2986	-0.0216	-1.1534	0.5675	-0.7277	14	轻
	6.1050	-3.3352	0.6344	1.0350	2.0741	1	严重
	-0.1401	-0.5177	-1.2313	-1.1071	-0.4870	9	轻
	0.9488	-1.2404	-0.8150	-2.1838	-0.2277	6	轻
I	-1.0552	1.1410	-0.4953	0.0353	-0.2212		
II	0.5226	1.7637	-0.8608	0.3481	0.6947		
III	4.0756	5.7879	1.0232	-0.4778	3.7732		

总结：

关于主成分的实际意义要结合具体问题和有关专业知识才能给出合理的解释。虽然利用主成分本身可对所研究的问题在一定程度上作分析，但主成分分析本身往往并不是最终目的，更重要的是利用主成分综合原始变量的信息，达到降维的目的，然后对数据作进一步的分析，如回归分析、聚类分析、判别分析等。