

MATLAB统计分析与应用

——主成分分析

主讲人：谢中华

SSstudy.com
科学软件学习网

主要内容

- 主成分分析的基本思想
- 主成分分析的MATLAB函数
- 主成分分析案例

第一节 主成分分析的基本思想

一、概念

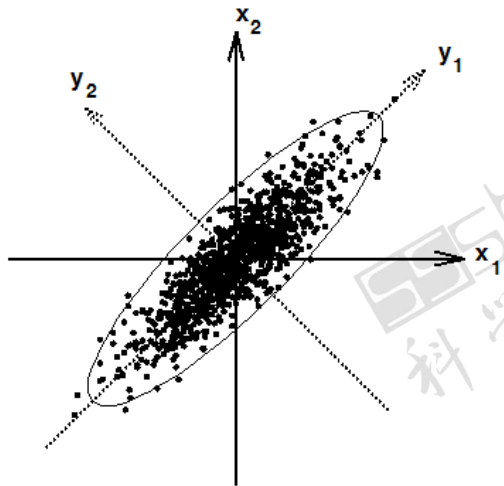
主成分分析 (Principal Component Analysis) 又称主分量分析，是由皮尔逊 (Pearson) 于1901年首先引入，后来由霍特林 (Hotelling) 于1933年进行了发展。是一种通过降维技术把多个变量化为少数几个主成分 (即综合变量) 的多元统计方法，这些主成分能够反映原始变量的大部分信息，通常表示为原始变量的线性组合，为使得这些主成分所包含的信息互不重叠，要求各主成分之间互不相关。

二、几何意义

假设从二元总体 $\mathbf{x} = (x_1, x_2)'$ 中抽取容量为 n 的样本，绘出样本观测值的散点图，如图所示。从图上可以看出，散点大致分布在一个椭圆内， x_1 与 x_2 呈现出明显的线性相关性。

进行坐标旋转

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



第二节 主成分分析的MATLAB函数

- pcacov : 根据协方差矩阵或相关系数矩阵进行主成分分析
- princomp : 根据样本观测值矩阵进行主成分分析
- pcares : 由样本的主成分得分重建（恢复）原始数据

第三节 主成分分析案例

一、从协方差矩阵或相关系数矩阵出发求解主成分

【例15.3-1】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，每人测了六项指标：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 ，样本相关系数矩阵如表所列。试根据样本相关系数矩阵进行主成分分析。

变 量	身高x1	坐高x2	胸围x3	手 臂 长 x4	肋围x5	腰围x6
身高x1	1	0.79	0.36	0.76	0.25	0.51
坐高x2	0.79	1	0.31	0.55	0.17	0.35
胸围x3	0.36	0.31	1	0.35	0.64	0.58
手臂长x4	0.76	0.55	0.35	1	0.16	0.38
肋围x5	0.25	0.17	0.64	0.16	1	0.63
腰围x6	0.51	0.35	0.58	0.38	0.63	1

MATLAB程序

```
PHO = [1    0.79  0.36  0.76  0.25  0.51  
       0.79  1    0.31  0.55  0.17  0.35  
       0.36  0.31  1    0.35  0.64  0.58  
       0.76  0.55  0.35  1    0.16  0.38  
       0.25  0.17  0.64  0.16  1    0.63  
       0.51  0.35  0.58  0.38  0.63  1    ];
```

```
[COEFF,latent,explained] = pcacov(PHO)
```

```
result1(1,:) = {'特征值', '差值', '贡献率', '累积贡献率'};
```

```
result1(2:7,1) = num2cell(latent);
```

```
result1(2:6,2) = num2cell(-diff(latent));
```

```
result1(2:7,3:4) = num2cell([explained, cumsum(explained)])
```

MATLAB程序

```
s = {'标准化变量'; 'x1:身高'; 'x2:坐高'; 'x3:胸  
围'; ... 'x4:手臂长'; 'x5:肋围'; 'x6:腰围'};  
result2(:,1) = s ;  
result2(1, 2:4) = {'Prin1', 'Prin2', 'Prin3'};  
result2(2:7, 2:4) = num2cell(COEFF(:,1:3))
```

结果分析

前3个主成分的累积贡献率达到了85.8756%，因此可以只用前3个主成分进行后续的分析

$$y_1 = 0.4689x_1^* + 0.4037x_2^* + 0.3936x_3^* + 0.4076x_4^* + 0.3375x_5^* + 0.4268x_6^*$$

$$y_2 = -0.3648x_1^* - 0.3966x_2^* + 0.3968x_3^* - 0.3648x_4^* + 0.5692x_5^* + 0.3084x_6^*$$

$$y_3 = -0.0922x_1^* - 0.6130x_2^* + 0.2789x_3^* + 0.7048x_4^* - 0.1643x_5^* - 0.1193x_6^*$$

3个主成分分别可以解释为：身材的**综合成分**（或**魁梧成分**）、高矮和胖瘦的**协调成分**、臂长成分。

二、从样本观测值矩阵出发求解主成分

【例15.3-2】 下表列出了2007年全国31个省、市和自治区的农村居民家庭平均每人全年消费性支出的8个主要变量数据，这8个变量是

x_1 = 食品， x_2 = 衣着， x_3 = 居住， x_4 = 家庭设备用品及服务， x_5 = 交通和通信， x_6 = 文教娱乐用品及服务， x_7 = 医疗保健， x_8 = 其他商品和服务，试根据这8个主要变量的观测数据，进行主成分分析。

地 区	x1	x2	x3	x4	x5	x6	x7	x8
北 京	2132.51	513.44	1023.21	340.15	778.52	870.12	629.56	111.75
天 津	1367.75	286.33	674.81	126.74	400.11	312.07	306.19	64.30
河 北	1025.72	185.68	627.98	140.45	318.19	243.30	188.06	57.40
山 西	1033.68	260.88	392.78	120.86	268.75	370.97	170.85	63.81
内蒙古	1280.05	228.40	473.98	117.64	375.58	423.75	281.46	75.29
辽 宁	1334.18	281.19	513.11	142.07	361.77	362.78	265.01	108.05
吉 林	1240.93	227.96	399.11	120.95	337.46	339.77	311.37	87.89
黑龙江	1077.34	254.01	691.02	104.99	335.28	312.32	272.49	69.98
上 海	3259.48	475.51	2097.21	451.40	883.71	857.47	571.06	249.04
江 苏	1968.88	251.29	752.73	228.51	543.97	642.52	263.85	134.41
浙 江	2430.60	405.32	1498.50	338.80	782.98	750.69	452.44	142.26
安 徽	1192.57	166.31	479.46	144.23	258.29	283.17	177.04	52.98
福 建	1870.32	235.61	660.55	184.21	465.40	356.26	174.12	107.00
江 西	1492.02	147.71	474.49	121.54	277.15	252.78	167.71	61.08
山 东	1369.20	224.18	682.13	195.99	422.36	424.89	230.84	71.98
河 南	1017.43	189.71	615.62	136.37	269.46	212.36	173.19	62.26
湖 北	1479.04	168.64	434.91	166.25	281.12	284.13	178.77	97.13
湖 南	1675.16	161.79	508.33	152.60	278.78	293.89	219.95	86.88
广 东	2087.58	162.33	763.01	163.85	443.24	254.94	199.31	128.06
广 西	1378.78	86.90	554.14	112.24	245.97	172.45	149.01	47.98
海 南	1430.31	86.26	305.90	93.26	248.08	223.98	95.55	73.23
重 庆	1376.00	136.34	263.73	138.34	208.69	195.97	168.57	39.06
四 川	1435.52	156.65	366.45	142.64	241.49	177.19	174.75	52.56
贵 州	998.39	99.44	329.64	70.93	154.52	147.31	79.31	34.16
云 南	1226.69	112.52	586.07	107.15	216.67	181.73	167.92	38.43
西 藏	1079.83	245.00	418.83	133.26	156.57	65.39	50.00	68.74
陕 西	941.81	161.08	512.40	106.80	254.74	304.54	222.51	55.71
甘 肃	944.14	112.20	295.23	91.40	186.17	208.90	149.82	29.36
青 海	1069.04	191.80	359.74	122.17	292.10	135.13	229.28	47.23
宁 夏	1019.35	184.26	450.55	109.27	265.76	192.00	239.40	68.17
新 疆	939.03	218.18	445.02	91.45	234.70	166.27	210.69	45.25

主成分分析程序

```
[X,textdata] = xlsread('Princomp.xls');  
XZ = zscore(X);  
[COEFF,SCORE,latent,tsquare] = princomp(XZ)  
explained = cumsum(100*latent/sum(latent))  
n = size(X,2);  
varname = textdata(3,2:end)';  
result2 = cell(n+1, 3);  
result2(1,:) = {'标准化变量', '特征向量t1', '特征向量t2'};  
result2(2:end, 1) = varname;  
result2(2:end, 2:end) = num2cell(COEFF(:,1:2))
```

结果分析

前两个主成分的累积贡献率达到了92.9957%，所以只用前2个主成分就可以了

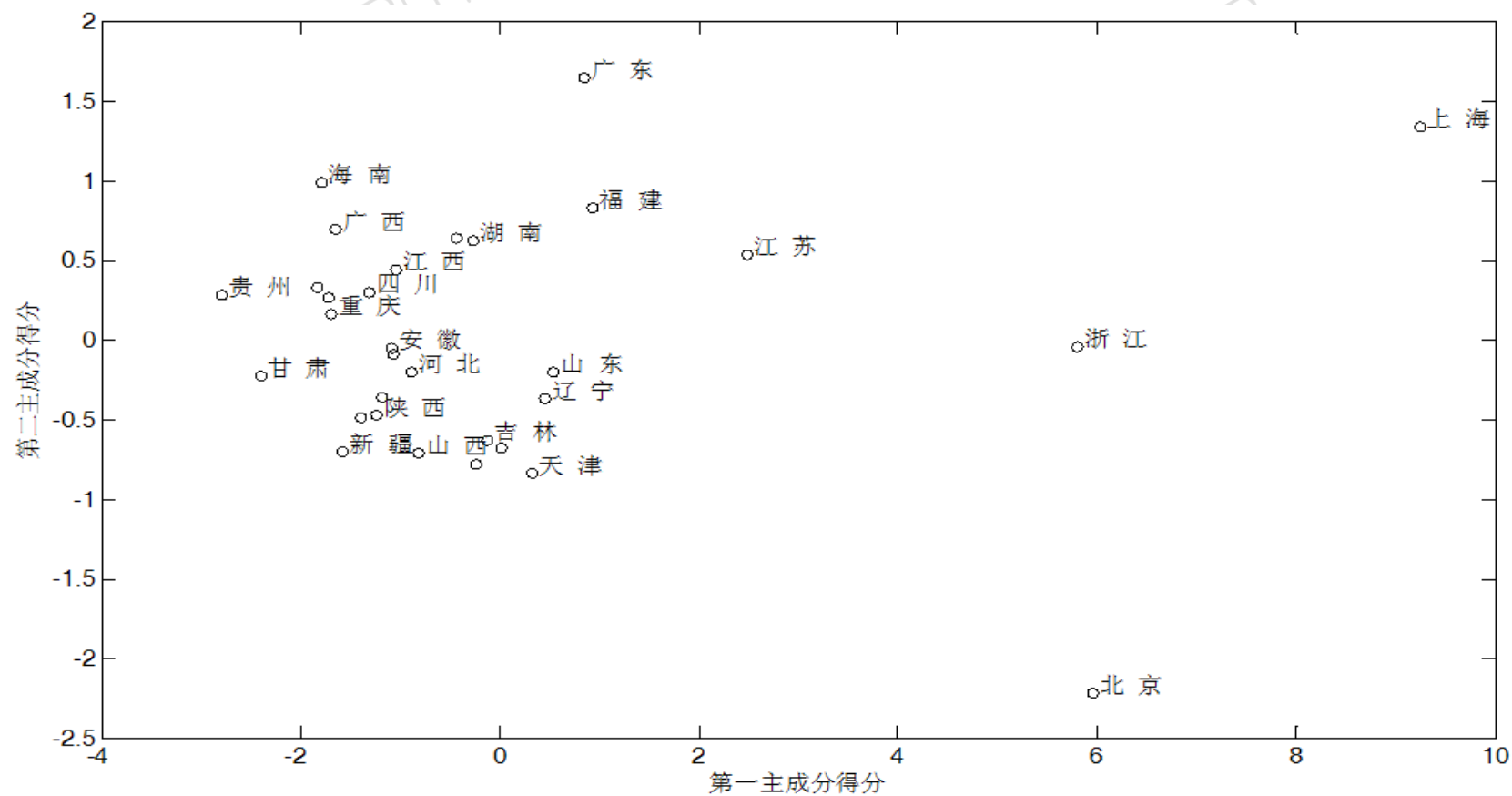
$$y_1 = 0.3431x_1^* + 0.3384x_2^* + 0.3552x_3^* + 0.3692x_4^* + 0.3752x_5^* + 0.3587x_6^* + 0.3427x_7^* + 0.3441x_8^*$$

$$y_2 = 0.5035x_1^* - 0.4866x_2^* + 0.1968x_3^* + 0.1088x_4^* - 0.0547x_5^* - 0.2208x_6^* - 0.4783x_7^* + 0.4225x_8^*$$

2个主成分分别可以解释为：综合消费性支出成分、消费倾向成分。

绘制主成分得分散点图的程序

```
cityname = textdata(4:end,1);  
plot(SCORE(:,1),SCORE(:,2),'ko');  
xlabel('第一主成分得分');  
ylabel('第二主成分得分');  
gname(cityname);
```



Thank You