聚类分析

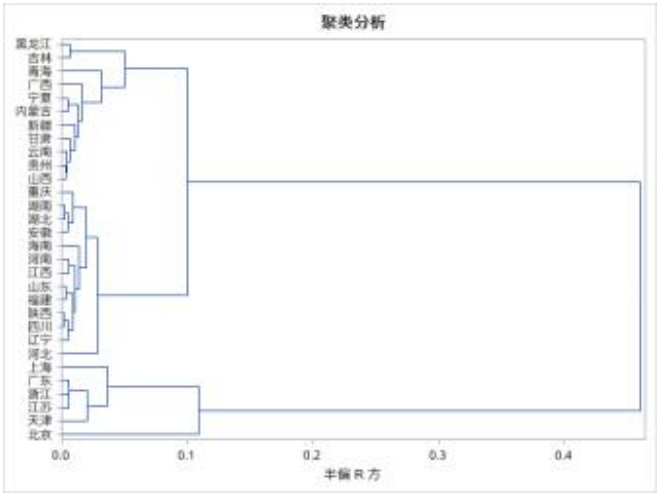
代码 1：系统聚类法

```
Proc cluster data=tmp1.exe3_1
method=ward outtree=a1
standard;#single最短距离法、
complete最长距离法、median中间距离
法、centroid重心法、average类平均
法、ward离差平方和法
id region;
run;
```

代码结果分析：

相关矩阵的特征值				
	特征值	差分	比例	累积
1	12.4123349	10.6985309	0.6206	0.6206
2	1.7138040	0.3461240	0.0857	0.7063
3	1.3676800	0.3560959	0.0684	0.7747
4	1.0115841	0.1261195	0.0506	0.8253
5	0.8854646	0.3364299	0.0443	0.8695
6	0.5490347	0.0602734	0.0275	0.8970
7	0.4887613	0.1004969	0.0244	0.9214
8	0.3882644	0.0408646	0.0194	0.9408
9	0.3473998	0.1111956	0.0174	0.9582
10	0.2362042	0.0543411	0.0118	0.9700
11	0.1818631	0.0219709	0.0091	0.9791
12	0.1598923	0.0526897	0.0080	0.9871
13	0.1072025	0.0604825	0.0054	0.9925
14	0.0467200	0.0008741	0.0023	0.9948
15	0.0458459	0.0199153	0.0023	0.9971
16	0.0259306	0.0118858	0.0013	0.9984
17	0.0140448	0.0038864	0.0007	0.9991
18	0.0101583	0.0052984	0.0005	0.9996
19	0.0048599	0.0019092	0.0002	0.9999
20	0.0029507		0.0001	1.0000

上表给出了离差平方和的聚类分析统计量，该图包含特征值，解释变异的比重和累计比重。



上图为 Ward 离差平方和树形图，给定阈值 0.75，所有地区聚类为 4 类。但北京单独分为一类分类效果并不是很好，我们最终将地区分为以下几类：

高质量发展地区：北京、天津、江苏、浙江、广东、上海。

中高质量发展地区：河北、辽宁、四川、山西、福建、山东、江西、河南、海南。

中低质量发展地区：安徽、湖北、湖南、重庆。

低质量发展地区：山西、贵州、云南、甘肃、新疆、内蒙古、宁夏、广西、青海、吉林、黑龙江。

代码 2：动态聚类法

```
Proc standard data=tmp1.exe3_1
out=st mean=0 std=1;
run;
```

```
data st1;
set st;
if n_=11 then output;
if n_=18 then output;
if n_=30 then output;
run;
proc fastclus data=st
maxclusters=3 seed=st1 out=aa
mean=m;
run;
data zz;
set aa;
keep region cluster;
run;
```

代码运行结果分析：

FASTCLUS 过程
替换=FULL 半径=0 最大聚类

聚类	x1	x2	x3	x4	x5	x6	x7	x8	
1	1.536266432	0.690931140	0.296578031	0.180202362	0.184907424	1.477051694	0.509111099	0.404371882	0.8506
2	-0.340197424	-0.041902121	0.109265590	-0.230206922	-0.314768023	-0.216160100	-0.426826458	-0.395698115	-0.3828
3	-0.810189802	-1.066473455	-2.044827479	-0.859205802	-0.387317989	-0.727233233	0.172053628	-0.489775883	-0.6680

上表给出了三个初始种子即凝聚点的各项指标数据。

基于最终种子的准则 = 0.6528

聚类汇总

聚类	频数	均方根标准差	从种子到观测的最大距离	半径超出	最近的聚类	聚类质心间的距离
1	6	1.0145	7.2477		2	6.5585
2	13	0.5249	3.9011		3	3.1199
3	11	0.6550	4.3811		2	3.1199

上表为动态聚类完成后每一类的频数、标准差以及与凝聚点的最大距离信息。

变量的统计量

变量	总标准差	标准差内	R 方	RSQ/(1-RSQ)
x1	1.00000	0.43975	0.819957	4.554236
x2	1.00000	0.59706	0.668104	2.012992
x3	1.00000	0.77845	0.435805	0.772437
x4	1.00000	0.87770	0.282766	0.394246
x5	1.00000	0.86988	0.295493	0.419433
x6	1.00000	0.52498	0.743406	2.897213
x7	1.00000	0.72776	0.506892	1.027952
x8	1.00000	0.77527	0.440407	0.787014
x9	1.00000	0.45365	0.808394	4.219034
x10	1.00000	0.69396	0.551638	1.230342
x11	1.00000	0.51878	0.749429	2.990885
x12	1.00000	0.70134	0.542039	1.183591
x13	1.00000	0.58532	0.681023	2.135019
x14	1.00000	0.47322	0.791507	3.796328
x15	1.00000	0.52731	0.741124	2.862849
x16	1.00000	0.51293	0.755050	3.082468
x17	1.00000	0.63121	0.629048	1.695766
x18	1.00000	0.95539	0.150178	0.176717
x19	1.00000	0.87406	0.288705	0.405887
x20	1.00000	0.84805	0.330415	0.493463
OVER-ALL	1.00000	0.68701	0.560569	1.275670

伪 F 统计量 = 17.22

近似期望总体 R 方 = 0.18997

立方聚类准则 = 24.427

上表为参与聚类的每一个变量以及变量整体的一些相关统计量。

聚类均值

聚类	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12	x13	x14	
1	1.748967489	1.469448248	0.874124724	0.962156650	1.067557513	1.672844886	1.395437126	1.232142202	1.749085107	1.216977338	1.598971099	1.403243060	1.461582175	1.713439717	1.66
2	-0.265447441	-0.032503163	0.274965057	-0.030010908	-0.239220566	-0.276362179	-0.406891845	-0.087019918	-0.304675921	0.110960625	-0.099425703	-0.167707676	-0.002855242	-0.246832879	-0.26
3	-0.640271655	-0.763104397	-0.801754008	-0.489345281	-0.299588883	-0.585850999	-0.280275343	-0.569235844	-0.593974879	-0.794941105	-0.754662951	-0.567205324	-0.793852263	-0.642891897	-0.59

聚类标准差

聚类	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12	x13	x14	x15
1	0.917322460	1.178040024	0.665991660	1.455879978	1.995391026	1.071100227	1.588337671	0.664741175	0.999401653	0.532792491	0.460901484	1.171829973	0.523405207	0.700287693	1.082030071
2	0.269311625	0.444088425	0.644522038	0.635920235	0.189931482	0.311253935	0.256429096	1.039946627	0.204838955	0.453773230	0.619133686	0.620852645	0.746618785	0.441133776	0.288731111
3	0.119788420	0.178751161	0.957028103	0.731375761	0.094804372	0.232892522	0.299497646	0.322632235	0.076848749	0.954580101	0.400560909	0.423019275	0.345152624	0.354836531	0.255554127

上表为聚成 3 类后，每一类中每个变量的均值和标准差。

	地区	聚类
1	北京	1
2	天津	1
3	河北	2
4	山西	3
5	内蒙古	3
6	辽宁	2
7	吉林	3
8	黑龙江	3
9	上海	1
10	江苏	1
11	浙江	1
12	安徽	2
13	福建	2
14	江西	2
15	山东	2
16	河南	2
17	湖北	2
18	湖南	2
19	广东	1
20	广西	3
21	海南	2
22	重庆	2
23	四川	2
24	贵州	3
25	云南	3
26	陕西	2
27	甘肃	3
28	青海	3
29	宁夏	3
30	新疆	3

由上表可知，通过动态聚类分析全国 30 个地区可以分为 3 类。第一类包含 6 个地区，第二类包含 13 个地区，第三类包含 11 个地区。

第一类：北京、天津、上海、江苏、浙江、广东。这 6 个省市是我国经济社会综合发展程度最高的地区，在可命名为高质量发展地区。

第二类：河北、安徽、福建、江西、山东、河南、湖北、湖南、海南、重庆、四川、陕西、甘肃。这 13 个地区经济有一定发展，但仍存在发展不平衡、发展程度不充分等问题，可命名为中质量发展地区。

第三类：陕西、内蒙古、吉林、黑龙江、贵州、云南、甘肃、青海、宁夏、新疆。这 11 地区的发展较为落后，可命名为低质量发展地区。

判别分析

距离判别代码：

```
proc discrim
data=tmp2.enterprise_classified list listerr
testdata=tmp1.enterprise_unclassified out=a1 testout=to1
outstat=os pool=yes;
class type;
var x1-x11;
#priors 'ST'=0.2 '非ST'=0.8;贝叶斯判别
run;
```

代码结果分析：

到 TYPE 的广义平方距离		
从 TYPE	ST	非ST
ST	0	6.59359
非ST	6.59359	0

以下对象的线性判别函数: TYPE			
变量	标签	ST	非ST
常数		-2.92421	-2.07285
X1	总资产收益率	-6.80562	1.51080
X2	净资产收益率	0.03062	-0.04370
X3	营业净利率	-0.56583	-0.08834
X4	应收账款周转率	0.01389	0.00481
X5	总资产周转率	0.56413	3.05703
X6	流动资产周转率	0.59758	-0.10940
X7	主营业务收入增长率	-1.25878	-0.60597
X8	RandD强度	-4.33727	7.92585
X9	流动比率	0.19914	0.42368
X10	速动比率	-0.08679	-0.23302
X11	现金比率	-0.14221	0.17651

根据上述输出可以写出具体的判别函数和判别规则。

第一类的判别函数为：

f1(x)=-2.9242-6.8056x1+0.0306x2-0.5658x3+0.0139x4+0.55641x5+0.5976x6-1.2588x7-4.3373x8+0.1991x9-0.0868x10-0.1422x11

第二类判别函数为：

f2(x)=-2.0729+1.5108x1-0.0437x2-0.0883x3+0.0048x4+3.057x5-0.1094x6-0.606x7+7.9259x8+0.4237x9-0.233x10+0.1765x11

判别规则为：若 f1(x)> f2(x)，企业属于 ST 类型。

成员的后验概率TYPE

观测	从 TYPE	分为TYPE		ST	非ST
1	非ST	非ST		0.1516	0.8484
2	非ST	非ST		0.0635	0.9365
3	非ST	非ST		0.0176	0.9824
4	非ST	非ST		0.3596	0.6404
5	非ST	非ST		0.2351	0.7649
6	非ST	非ST		0.0297	0.9703
7	非ST	ST	*	0.6538	0.3462
8	非ST	非ST		0.0439	0.9561
9	非ST	非ST		0.0002	0.9998
10	非ST	非ST		0.0015	0.9985
11	非ST	非ST		0.0603	0.9397
12	非ST	非ST		0.0022	0.9978
13	非ST	非ST		0.1167	0.8833
14	非ST	非ST		0.4419	0.5581
15	非ST	非ST		0.0088	0.9912

上述表格为回代结果（不完整），出现*表示判别结果与训练样本中的分类不一致。

分入“TYPE”的观测数和百分比

从 TYPE	ST	非ST	合计
ST	16 80.00	4 20.00	20 100.00
非ST	1 3.33	29 96.67	30 100.00
合计	17 34.00	33 66.00	50 100.00
先验	0.5	0.5	

“TYPE”的出错数估计

	ST	非ST	合计
比率	0.2000	0.0333	0.1167
先验	0.5000	0.5000	

上述输出表面：在训练样本中，20 家 ST 企业中有 16 家企业判别为 ST 类别，4 家企业判断为非 ST 类别，错判率为 20%；30 家非 ST 企业中有 1 家企业判断为 ST 类别，29 家企业判断为非 ST 类别，错判率为 3.33%；合计的错判率为 11.67%。

费歇判别代码：

```
proc candisc data=tmp1.flower
out=flower_out;
class a;
var x1-x4;
run;
proc plot data=flower_out;
plot can2*can1=A;
run;
proc discrim data=flower_out
out=out;
class a;
```

```
var can1;
run;
```

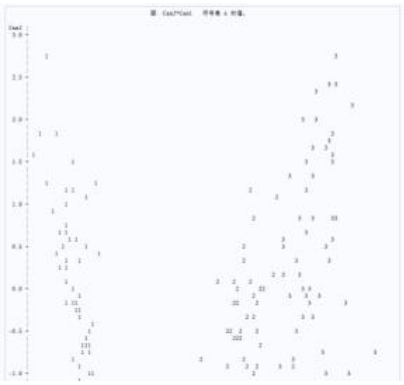
代码结果分析：

原始典型系数			
变量	标签	Can1	Can2
X1	花萼长	-.0829377642	0.0024102149
X2	花萼宽	-.1534473068	0.2164521235
X3	花瓣长	0.2201211656	-.0931921210
X4	花瓣宽	0.2810460309	0.2839187853

由上述输出得到中心化的费歇判别函数为：

y1=-0.0829(x1-58.433)+0.1524(x2-30.573)+0.2201(x3-37.580)+0.2818(x4-11.993)

y2=0.0024(x1-58.433)+0.2165(x2-30.573)-0.0932(x3-37.580)+0.2839(x4-11.993)



上图输出将 150 个样品的判别函数得分(y1,y2)做散点图得到的结果，可以看到，三组的分离效果非常好，且分离的很大程度显示在 can1 上，这与一个判别函数解释的方差贡献率相符合。因此，对于一个新的待判样本，通过计算判别得分，可以通过在坐标图中的位置判别所属类型。接下来利用 can1 对样品进行距离判别。距离判别后错判率降低至 1.33，得到了显著改善。

"A" 的出错数估计				
	1	2	3	合计
比率	0.0000	0.0400	0.0000	0.0133
先验	0.3333	0.3333	0.3333	

主成分分析

主成分分析代码：

```
Proc princomp data= tmp1.exe5_3
out=out;
run;
ods graphics on;
proc princomp data=tmp1.exe5_3
out=out n=2
plot=pattern(ncomp=2)
plot=score(ncomp=2);
var x1-x13;
id region;
run;
ods graphics off;
proc plot data=out;
plot out2*out1
$ region='*'/href=0vref=0;
```

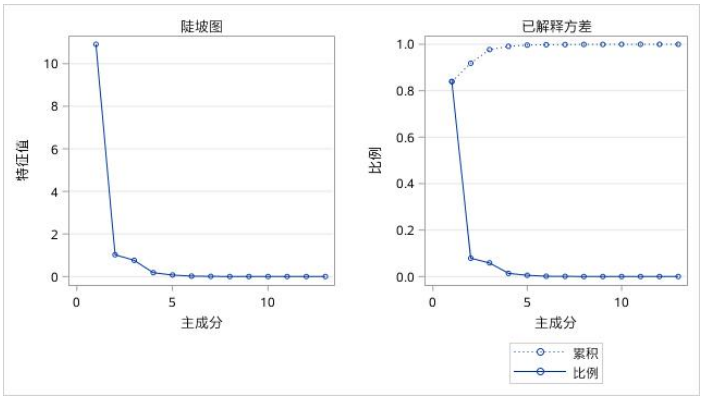
```
run;

proc sort data=out;

by descending prin1;

run;

proc print data=out;

var region prin1;

run;
```

结果分析：

相关矩阵的特征值

	特征值	差分	比例	累积
1	10.9121956	9.8856877	0.8394	0.8394
2	1.0265078	0.2640777	0.0790	0.9184
3	0.7624301	0.5823642	0.0586	0.9770
4	0.1800659	0.1060393	0.0139	0.9909
5	0.0740266	0.0550267	0.0057	0.9966
6	0.0189999	0.0091376	0.0015	0.9980
7	0.0098622	0.0043460	0.0008	0.9988
8	0.0055162	0.0014962	0.0004	0.9992
9	0.0040201	0.0010041	0.0003	0.9995
10	0.0030160	0.0008977	0.0002	0.9997
11	0.0021183	0.0012988	0.0002	0.9999
12	0.0008195	0.0003977	0.0001	1.0000
13	0.0004218		0.0000	1.0000

上表显示了原始数据相关系数矩阵特征值、方差贡献率及累计方差贡献率。可以看出前两个主成分的累积方差贡献率达到91.84%，可以解释大部分的变量信息。因此选取两个主成分进行分析。

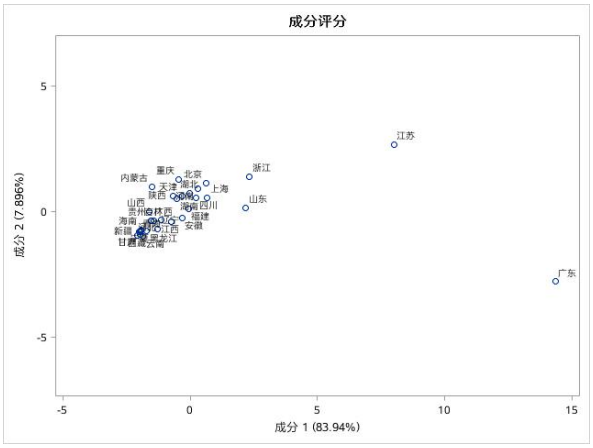


可以看到陡坡图拐点在2处，进一步肯定了选取两个主成分。

特征向量

		Prin1	Prin2
X1	人均生产总值	0.139537	0.490520
X2	企业数	0.297096	0.069616
X3	资产总计	0.301053	0.039268
X4	主营业务收入	0.296641	0.057982
X5	利润总额	0.290946	0.138545
X6	R&D经费内部支出	0.297910	-.128563
X7	新产品开发项目数	0.296850	0.030803
X8	新产品开发经费支出	0.294909	-.187961
X9	新产品销售收入	0.299869	-.054071
X10	专利申请数	0.296837	-.167732
X11	拥有发明专利数	0.275695	-.347455
X12	新增固定资产	0.146512	0.722675
X13	从业人员年平均人数	0.299391	-.059978

分析 prin1 和 prin2 在各变量上的系数可以发现:第一主成分在各变量上的系数都为正，而且数值相差不大，因而可以认为 prin1 代表地区高新技术产业综合竞争力水平，prin1 得分越高，表明地区高新技术产业综合竞争力水平实力越强；第二主成分在变量前的系数由有正有负，可以认为是地区高新技术产业经营规模与高新技术展业经营潜力的比较。当得分为正值时，表明相对于经营规模而言经营潜力较好；当得分接近于零时，表明地区高新技术产业经营规模与其潜力较为均衡；当得分为负值时，表示相对于高新技术经营规模而言潜力较差。



用图解样品的方法，可以非常直观地看出各地区高新技术产业竞争力状况。其中越往图的右上角分布表示该地区高新技术产业竞争力越强，越往左下表明该地区高新技术产业发展水平落后、潜力较低。

Obs	region	Prin1
1	广东	14.3410
2	江苏	8.0186
3	浙江	2.3288
4	山东	2.1652
5	上海	0.6606
6	北京	0.6155
7	河南	0.3145
8	四川	0.2453
9	湖北	-0.0262
10	福建	-0.0742
11	安徽	-0.3242
12	天津	-0.3295
13	重庆	-0.4639
14	湖南	-0.5309
15	陕西	-0.6645

由于第一主成分为各个地区高新技术产业综合竞争力水平，根据第一主成分对各地区进行排序，得到最终各地区高新技术产业竞争力状况排名情况。

主成分回归代码：

方法1：

```
Proc corr data=tmp1.exe5_5;
var y x1-x3;
run;
proc reg data=tmp1.exe5_5;
model y=x1 x2 x3;
run;
proc standard data=tmp1.exe5_5
out=sv mean=0 std=1;
var y x1-x3;
run;
proc princomp data=sv out=opcr;
var x1-x3;
run;
proc reg data=opcr;
model y=prin1 prin2;
run;
quit;
```

方法2：

```
Proc reg data=tmp1.exe5_5
outset=out;
model y=x1-x3/pcomit=1,2;
run;
```

```
quit;

proc print data=out;

run;
```

结果分析：

(1)相关系数矩阵。

Pearson 相关系数, N = 18 Prob > |r| under H0: Rho=0				
	y	x1	x2	x3
y 旅游人数	1.00000	0.86244 <.0001	0.99584 <.0001	0.99814 <.0001
x1 公路里程数	0.86244 <.0001	1.00000	0.89039 <.0001	0.86426 <.0001
x2 农村人均可支配收入	0.99584 <.0001	0.89039 <.0001	1.00000	0.99783 <.0001
x3 城镇人均可支配收入	0.99814 <.0001	0.86426 <.0001	0.99783 <.0001	1.00000

相关系数矩阵显示自变量与因变量高度相关，但自变量之间相关性也较高，说明可能存在多重共线性。

(2)普通最小二乘估计。

方差分析					
源	自由度	平方和	均方	F 值	Pr > F
模型	3	31009779	10336593	1252.15	<.0001
误差	14	115571	8255.08174		
校正合计	17	31125350			

均方根误差	90.85748	R 方	0.9963
因变量均值	2234.61111	调整 R 方	0.9955
变异系数	4.06592		

参数估计						
变量	标签	自由度	参数估计	标准误差	t 值	Pr > |t|
Intercept	Intercept	1	-53.85875	81.10991	-0.66	0.5175
x1	公路里程数	1	0.09547	0.71508	0.13	0.8957
x2	农村人均可支配收入	1	-0.01102	0.06330	-0.17	0.8643
x3	城镇人均可支配收入	1	0.38784	0.15130	2.56	0.0225

回归方程显著性检验结果显示 F 值为 1252.15，p<0.0001，认为该模型通过 F 检验，模型的整体拟合效果好。

回归模型调整后的 R 方值为 0.9955，接近于 1，说明该模型对数据的拟合程度高。

但回归模型的参数估计结果显示在 0.01 的显著性水平下，所有参数的显著性水平均不通过检验。

这与自变量与因变量高度相关相矛盾，且 x2 前系数为负与相关系数矩阵高度正相关结果也矛盾。

各变量之间可能存在多重共线性导致回归估计结果失真，进一步考虑采用主成分回归。

(3)主成分回归。

相关矩阵的特征值				
	特征值	差分	比例	累积
1	2.83622475	2.67312956	0.9454	0.9454
2	0.16309519	0.16241514	0.0544	0.9998
3	0.00068006		0.0002	1.0000

前两个主成分累积方差贡献率达到 99.98%，大于 95%，因此选取 2 个主成分进行建模和分析。

特征向量				
		Prin1	Prin2	Prin3
x1	公路里程数	0.559888	0.824650	0.080489
x2	农村人均可支配收入	0.588462	-.327373	-.739283
x3	城镇人均可支配收入	0.583300	-.461280	0.668567

根据上表可以得到主成分的表达式。

方差分析					
源	自由度	平方和	均方	F 值	Pr > F
模型	2	16.92991	8.46495	1811.54	<.0001
误差	15	0.07009	0.00467		
校正合计	17	17.00000			

均方根误差	0.06836	R 方	0.9959
因变量均值	-2.4672E-16	调整 R 方	0.9953
变异系数	-2.7707E16		

参数估计						
变量	标签	自由度	参数估计	标准误差	t 值	Pr > |t|
Intercept	Intercept	1	-4.9032E-16	0.01611	-0.00	1.0000
Prin1		1	0.58215	0.00984	59.13	<.0001
Prin2		1	-0.46123	0.04105	-11.23	<.0001

此时模型 F 值为 1811.54，p<0.0001,认为该模型通过 F 检验，模型的整体拟合效果显著。

调整后的 R 方值为 0.9953，接近于 1，说明该模型对数据的拟合程度高。

主成分回归模型的参数估计结果，在 1%的显著性水平下通过了 t 检验，将主成分表达式带入，可以得到主成分回归模型：

Y=-0.11157*x1+0.49357*x2+0.55232*x3

因子分析

代码：

```
Proc factor data=tmp1.exe6_2
method=prin n=4r=v out=out
outstat=stat reorder;
var x1-x8;
run;
proc plot data=out;
plot factor2*factor1
$ region='*'/herf=0verf=0;
run;
data a1;
set out;
f=0.28607345*factor1+0.277776
3375*factor2+0.2051361*factor
3+0.203346*factor4;
keep region f;
run;
proc sort data=a1;
by descending f;
run;
```

结果分析：

旋转因子模式

		Factor1	Factor2	Factor3	Factor4
x8	每万人高等学校在校生数	0.91513	0.28392	0.22067	0.10783
x7	公共图书馆藏书量	0.76347	0.26158	0.04407	0.56625
x5	人均财政收入	0.76290	0.52900	0.23243	-0.03320
x1	城镇居民人均可支配收入	0.28705	0.92893	-0.02532	0.19060
x2	农村居民人均可支配收入	0.35482	0.90008	-0.03295	-0.21596
x3	社会福利院数	0.19230	0.22113	0.87665	0.35806
x4	每万人拥有福利院床位数	0.15449	-0.25556	0.87529	-0.35058
x6	医院和卫生院数	0.13066	-0.07793	-0.00648	0.97945

每个因子已解释方差

Factor1	Factor2	Factor3	Factor4
2.2885876	2.2222107	1.6410888	1.6267680

最终的公因子方差估计: 总计 = 7.778655

x1	x2	x3	x4	x5	x6	x7	x8
0.98227234	0.98376246	0.98260670	0.97821910	0.91698637	0.98250636	0.97390264	0.97839915

经过因子旋转后，前 4 个公共因子的累计方差贡献率达到 97.23%，且各变量的变量共同度均在 0.75 以上。结合指标的实际经济意义，保留四个公共因子的分析效果最为理想。

X8、x7、x5 在第一公共因子上有较大载荷值，可以将 Factor1 命名为：经济条件和精神生活因子。

X1、x2 在第二公共因子上有较大载荷值，可以将 Factor2 命名为:可支配收入因子。

X3、x4 在第三公共因子上有较大载荷值，

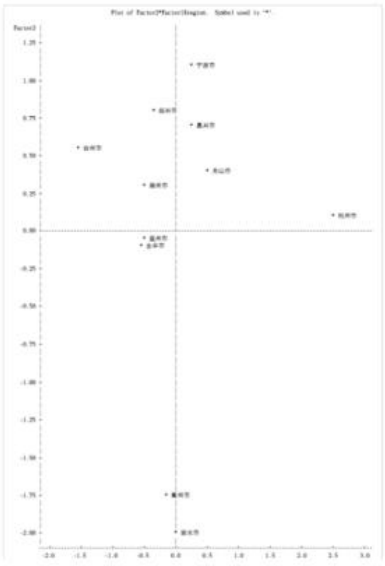
可以将 Factor3 命名为:福利资源因子。

X6 在第四公共因子上有较大载荷值，可以将 Factor4 命名为:医疗资源因子。

综合得分计算公式为：

f=0.28607345*factor1+0.2777763375*factor2+0.2051361*factor3+0.203346*factor4

对因子得分进行排序并画出散点图：



根据输出散点图可以浙江不同地区福利水平进行分析：

落在第一象限的有宁波、嘉兴、舟山、杭州地区，表明这些地区福利水平发展较为均衡，居民生活质量较高。

落在第二象限的有绍兴、台州、湖州地区。表明这些地区福利发展水平较高，但居民可支配收入较低。

落在第三象限的有温州、金华、衢州。这些地区居民可支配收入较高，但社会福利水平较低。落在第四象限的有丽水，福利水平较为落后，居民可支配收入也有提升空间。

	地区	f
1	杭州市	1.1026075486
2	宁波市	0.5442016476
3	台州市	0.1978053714
4	绍兴市	0.1918003515
5	温州市	-0.032057963
6	舟山市	-0.09742002
7	嘉兴市	-0.203225904
8	湖州市	-0.253600576
9	金华市	-0.373908693
10	丽水市	-0.511428085
11	衢州市	-0.564773678

最终各地区综合得分排序结果见上图。

对应分析

对应分析代码：

```
data aa;
set tmp1.exe7_1(obs=40);
run;
data aa1;
set aa;
select;
when(_n_<=8) G='理学';
when(9<=_n_<=14) G='医学';
when(15<=_n_<=18) G='农业科学';
when(19<=_n_<=39) G='工程与技术科学';
otherwise G='其他';
end;
run;
```

```
proc means data=aa1 sum;
var sci;
class G;
outputsum=o1;
run;
proc means data=aa1 sum;
varei CPCI_S;
class G;
outputsum=o2;
run;
proc means data=aa1 sum;
var CPCI_S;
class G;
outputsum=o3;
run;
data aam;
merge data1 data2 data3;
run;
proc corresp data=aamall;
var o1 o2 o3;
id g;
run;
#proc corresp data=exe7_4;
tables row,column;
```

run;针对不是列联表情况

结果分析：

要求在对学科进行分类的基础上对各类学科论文数量进行对应分析。首先将数据表中的空记录剔除，然后按照学科进行分类并对各类学科论文数量进行汇总。

Row Profiles

	SCI	EI	CPCI-S
	0.505039	0.370786	0.124175
工程	0.286487	0.524930	0.188583
理学	0.671139	0.289150	0.039711
农业	0.958490	0.028614	0.012896
其他	0.064584	0.038259	0.897157
医学	0.921288	0.004794	0.073918

Column Profiles

	SCI	EI	CPCI-S
	0.500000	0.500000	0.500000
工程	0.145532	0.363210	0.389627
理学	0.230469	0.135246	0.055463
农业	0.012274	0.000499	0.000672
其他	0.000317	0.000255	0.017884
医学	0.111408	0.000790	0.036355

输出行轮廓于列轮廓，是列联表中每一行数值除以行和得到的结果和每一列数值除以列和得到的结果。

SAS 系统

The CORRESP Procedure

Inertia and Chi-Square Decomposition

Singular Value	Principal Inertia	Chi-Square	Percent	Cumulative Percent	18 36 54 72 90
0.34389	0.11826	136113	87.59	87.59	**********
0.12942	0.01675	19279	12.41	100.00	***
Total	0.13501	155392	100.00		

Degrees of Freedom = 10

输出各维汇总表，其中 Singular 是奇异值，Principal inertia 是主惯量，Percent 是惯量的百分比，最后一列数据是惯量占比的累计值。从中可以看出，第一维和第二

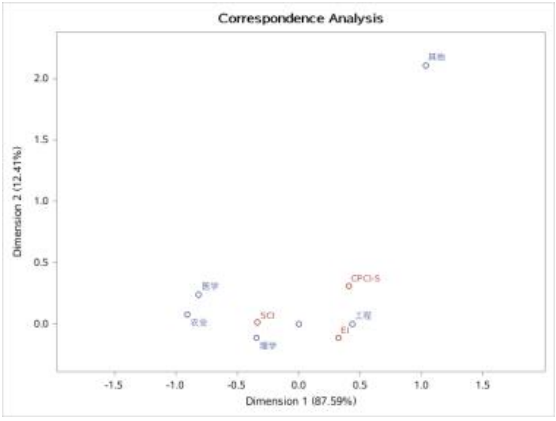
维的惯量比例占总惯量的 100%,因此前两维解释了列联表数据 100%的变异。

Row Coordinates		
	Dim1	Dim2
	-0.0000	0.0000
工程	0.4383	-0.0000
理学	-0.3414	-0.1161
农业	-0.9040	0.0732
其他	1.0345	2.1033
医学	-0.8177	0.2369

上表输出 R 型因子分析中的公因子载荷，表示”样品”投影到公共因子 dim1 和 dim2 的坐标值。

Column Coordinates		
	Dim1	Dim2
SCI	-0.3396	0.0092
EI	0.3262	-0.1156
CPCI-S	0.4071	0.3077

上表输出是 Q 型因子分析中的公因子载荷，表示变量投影到公共因子 dim1 和 dim2 的坐标值。



上图是对应分析图结果，根据对应分析的思路，通过观察邻近区域进行关联性分析。从上图可以看到，“理学”与“SCI”的距离较近，“工程与技术科学”与“EI”的距离较近。结果表明，学科类别与检索机构论文收录有一定联系。

对应分析图分析:

(1)观察临近区域进行关联性分析。

从图中可以看出各个省份和四种收入的距离，距离越近说明地区的收入来源特征与该类收入关联度越高。例如 xx 与工资性收入的关联度较高.

(2)通过向量分析进行偏好排序。

从中心向任意点连线作向量，例如从中心向“工资性收入”作向量，然后让所有地区往这条向量及其延长线作垂线，垂点越靠近向量正向的表示工资性收入比重越高。以 xx、xx、xx 三省为例，在这三个地区中，xx 的工资性收入比重最高，xx 次之，xx 的工资性收入比重最小。

(3)通过向量的夹角来分析两者之间的相关性。

可以通过向量夹角的角度大小看不同地区或不同收入来源之间的相似情况，向量夹角越小说明相似程度越高。Xx 与 xx 的夹角小于 xx 与 xx 的夹角，这可以说明前者比后者的相似度要高。运用同样的方法可以比较不同省份之间的相似程度。

(4)通过坐标点离中心的距离研究其特征的显著性。

坐标点越靠近中心，越没有特征；越远离中心，说明其特征越明显。例如：一个 xx 越靠近原点，表明其收入来源越没有特征；若一个 xx 的坐标点与原点的距离越远，说明其收入来源的特征越显著。如图所示，xxx 等地区距离 O 点的距离较近，说明这些地区的收入来源的特征不突出。Xx 等地区与原点 O 的距离较远，说明这些地区的收入来源的特征性较显著，例如 xx 的 xx 比重最高，xx 的 xx 比重最高。

典型相关分析

典型相关分析代码：

```
proc cancorr data=tmp1.tech
out=techout outstat=techvalue
all;
with y1-y5;
var x1-x4;
run;
```

运行结果分析：

	典型相关	调整典型相关	近似标准误差	典型相关平方	特征值: Inv(E)*H = CanRsq/(1-CanRsq)				H0 检验: 当前行和之后的所有行的典			
					特征值	差分	比例	累积	似然比	近似 F 值	分子自由度	分母自
1	0.960413	0.951478	0.014169	0.922393	11.8854	9.6066	0.8122	0.8122	0.01604341	9.15	20	7
2	0.833671	0.807183	0.055684	0.695007	2.2788	1.8207	0.1557	0.9679	0.20672553	4.15	12	6
3	0.560496	0.516533	0.125218	0.314155	0.4581	0.4462	0.0313	0.9992	0.67780425	1.72	6	
4	0.108274	-.032084	0.180434	0.011723	0.0119		0.0008	1.0000	0.98827673	0.15	2	

上述输出给出了典型相关系数及其检验。第一对典型变量的相关系数是 0.960413，调整后的典型相关系数是 0.951478，p<0.0001 拒绝相关系数为零的原假设，这说明第一对典型相关系数在 0.01 的显著性水平下显著。第二对典型变量的相关系数为 0.833671，调整后的典型相关系数为 0.807183，同样，第二对典型相关系数也通过检验。因此选择前两组典型变量进行解释。

VAR 变量 及其典型变量之间的相关性

		V1	V2	V3	V4
x1	人均地区生产总值	0.8326	0.4975	-0.2292	0.0826
x2	人均可支配收入	0.9093	0.3694	-0.0333	-0.1889
x3	第三产业占GDP比重	0.8243	-0.1658	-0.3982	-0.3667
x4	人均地方财政收入	0.9934	0.1030	-0.0466	-0.0175

WITH 变量 及其典型变量之间的相关性

		W1	W2	W3	W4
y1	每万从业人员有效发明专利数	0.8279	0.0810	-0.3220	-0.4487
y2	每万从业人员发明专利申请数	0.7290	0.2324	-0.0622	-0.5696
y3	每万人RD项目数	0.5494	0.5245	-0.5031	0.0184
y4	RD投入强度	0.3034	0.9232	0.1587	-0.1702
y5	每万名就业人员的RD人力投入	0.9241	0.2817	0.1623	0.1912

上述输出是典型载荷。”VAR 变量及其典型变量之间的相关性”表明了第一组各原变量与其各个典型变量之间的相关性，例如 X1 和 V1 的相关系数为 0.8326。“WITH 变量及其典型变量之间的相关性”表明了第二组各原变量与其各个典型变量之间的相关性，例如 Y1 和 W1 之间的相关系数为 0.8279。

通过以下变量解释的 VAR 变量 标准化方差					
	它们自己的典型变量			对立面典型变量	
典型变量号	比例	累积比例	典型R 方	比例	累积比例
1	0.7966	0.7966	0.9224	0.7347	0.7347
2	0.1055	0.9021	0.6950	0.0733	0.8081
3	0.0536	0.9557	0.3142	0.0168	0.8249
4	0.0443	1.0000	0.0117	0.0005	0.8254

通过以下变量解释的 WITH 变量 标准化方差					
	它们自己的典型变量			对立面典型变量	
典型变量号	比例	累积比例	典型R 方	比例	累积比例
1	0.4929	0.4929	0.9224	0.4547	0.4547
2	0.2535	0.7464	0.6950	0.1762	0.6308
3	0.0824	0.8288	0.3142	0.0259	0.6567
4	0.1183	0.9472	0.0117	0.0014	0.6581

输出是典型冗余分析。典型变量 V1 解释了 X 组变量 79.66%的信息，V2 解释了 X 组变量 10.55%的信息，V1 和 V2 累计解释了 X 组变量 90.21%的信息，典型变量 V3 和 V4 的解释力很小。典型变量 W1 解释了 Y 组变量 49.29%%的信息，W2 解释了 Y 组变量 25.35%的信息，W1 和 W2 累计解释了 Y 组变量 74.64%的信息。同样地，典型变量 W3 和 W4 的解释力很小。第二组变量的典型变量 W1 解释了 X 组变量 73.47%的信息，W2 解释了 X 组变量 7.33%的信息，W1 和 W2 累计解释了 X 组变量 80.81%的信息。第一组变量的典型变量 V1 解释了 Y 组变量 45.47%的信息，V2 解释了 Y 组变量 17.62%的信息，V1 和 V2 累计解释了 Y 组变量 63.08%的信息。

因此，可以认为各地区的科学技术发展水平和经济发展水平不但能被本组的典型变量解释，也可以被对立组的典型变量解释，说明科学技术发展水平和经济发展水平之间的相关程度比较密切。