**33. 主成分分析**

**（一）原理**

**一、基本思想**

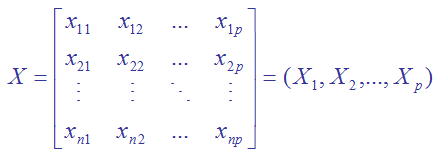
主成份分析，是数学上对数据降维的一种方法，是将多个变量转化为少数综合变量（集中了原始变量的大部分信息）的一种多元统计方法。其主要目的是将变量减少，并使其改变为少数几个相互独立的线性组合形成的新变量（主成份，其方差最大），使得原始资料在这些成份上显示最大的个别差异来。

在所有的线性组合中所选取的*F*1应该是方差最大的，称为第一主成分。如果第一主成分不足以代表原来所有指标的信息，再考虑选取第二个线性组合*F*2, 称为第二主成分。为了有效地反映原有信息，*F*1已有的信息就不需要再出现在*F*2中，用数学语言表达就是要求Cov(*F*1，*F*2)＝0. 依此类推可以构造出第三、第四、…、第*p*个主成分。

主成份分析，可以用来综合变量之间的关系，也可用来减少回归分析或聚类分析中的变量数目。

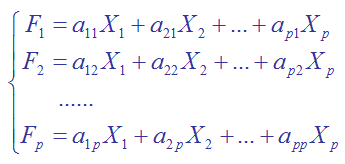
**二、基本原理**

设有*n*个样品（多元观测值），每个样品观测*p*项指标（变量）：X1，…，Xp，得到原始数据资料阵：



其中，Xi = (*x*1i，*x*2i，…，*x*ni)T，i = 1, …, p.

用数据矩阵X的*p*个列向量（即*p*个指标向量）X1，…，Xp作线性组合，得到综合指标向量：



简写成：

*Fi* = *a*1*iX*1 + *a*2*iX*2 +…+*apiXp* *i* = 1, …, *p*

限制系数*a*i = (*a*1i，*a*2i，…，*a*pi)T为单位向量，即



且由下列原则决定：

（1）*Fi*与*Fj*互不相关，即COV(*Fi*, *Fj*)= *a*iT∑*a*i=0，其中∑为X的协方差矩阵；

（2）*F1*是X1，X2，…，Xp的所有满足上述要求的线性组合中方差最大的，即

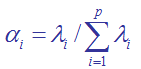


*F2*是与*F1*不相关的X1，…，Xp所有线性组合中方差最大的，…，*Fp*是与*F1*，…，*F*p-1都不相关的X1，…，Xp所有线性组合中方差最大的。

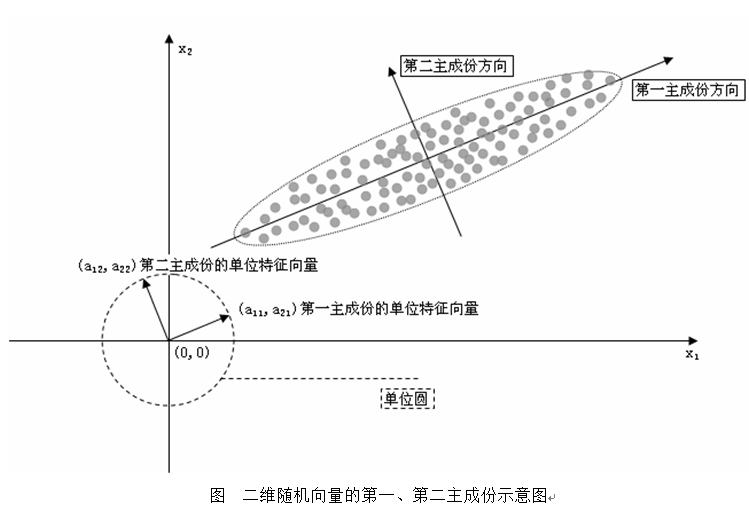
满足上述要求的综合指标向量*F1，F2，…，Fp*就是主成分，这p个主成分从原始指标所提供的信息总量中所提取的信息量依次递减，每一个主成分所提取的信息量用方差来度量，主成分方差的贡献就等于原指标相关系数矩阵相应的特征值*λ*i，每一个主成分的组合系数

*a*i = (*a*1i，*a*2i，…，*a*pi)T

就是特征值*λ*i所对应的单位特征向量。方差的贡献率为



αi越大，说明相应的主成分反映综合信息的能力越强。



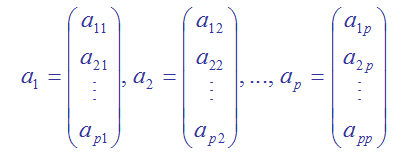
注：主成分分析是将原始变量组成的坐标系进行平移变换，使得新的坐标原点和数据群点的重心重合。新坐标第一轴与数据变化最大方向对应。*F1，F2，…，Fp*可以理解为p维空间中互相垂直的*p*个坐标轴。

**三、基本步骤**

1. 计算样品数据协方差矩阵*Σ* = (*sij*)*p*×*p*，其中



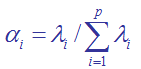
2. 求出Σ的特征值及相应的特征向量*λ*1>*λ*2>…>*λ*p>0, 及相应的正交化单位特征向量：



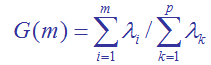
则X的第i个主成分为*Fi*= *a*iTX，i=1, …, p.

3. 选择主成分

在已确定的全部*p*个主成分中合理选择*m*个来实现最终的评价分析。一般用方差贡献率



解释主成分*Fi*所反映的信息量的大小，*m*的确定是用累计贡献率



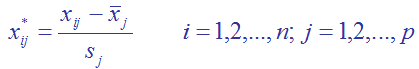
达到足够大（一般在85%以上）为原则。

4. 计算*n*个样品在*m*个主成分得分

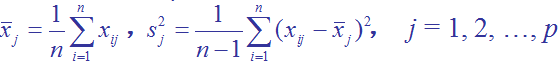


5. 数据的标准化

实际应用时，指标的量纲往往不同，所以在主成分计算之前应先消除量纲的影响。消除数据的量纲有很多方法，常用方法是将原始数据标准化



其中，



即把每列变为均值为0，方差为1.

注：标准化后变量的协方差矩阵*Σ* = (*sij*)*p*×*p*，与原变量的相关系数矩阵R= (*rij*) *p*×*p*相同，故主成分分析可以从原始变量数据的相关系数矩阵，也可以从标准化数据的协方差矩阵出发做分析。

**（二）PROC PRINCOMP过程步**

基本语法：  
**PROC PRINCOMP data = 数据集** <可选项> **;**

**<BY** 变量**;>**

**<FREQ** 变量**; >**

**<ID** 变量**;**

**<PARTIAL** 变量列表**; >**

**VAR** 变量变量**;**

**<WEIGHT** 变量**; >**

说明：

（1）可选项

type = COV/CORR

原始数据用CORR（相关系数矩阵，默认）；标准化数据用COV（协方差矩阵）；

out＝数据集——存储原始数据和各主成分得分（相应特征值）；

outstat＝数据集——存储若干统计量；

n=——设定主成分个数；

std——OUT=数据集中，将主成分得分标准化为单位方差；

noint——要求模型中不含截距。

（2）VAR语句——指定做主成分分析的变量；

（3）PARTIAL语句——指明对偏相关阵或偏协方差矩阵进行分析时，被偏出去的数值变量名。

**例1** 对全国30个省市自治区经济发展基本情况的八项指标作因子分析。

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 省份 | GDP  x1 | 居民消费水平x2 | 固定资产投资x3 | 职工平均工资x4 | 货物周转量  x5 | 居民消费价格指数  x6 | 商品零售价格指数x7 | 工业总产值  x8 |
| 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 |
| 内蒙古 | 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 |
| 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 |
| 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 |

**代码：**

**data** Economics;

infile 'C:\MyRawData\Economics.txt';

input Province $ x1 x2 x3 x4 x5 x6 x7 x8;

**run**;

**proc** **princomp** data = Economics n = **3** out = Resusts plots= score(ellipse ncomp=**2**);

id Province;

var x1-x8;

**run**;

**proc** **print** data = Resusts;

**run**;

运行结果及说明：

PRINCOMP 过程

|  |  |
| --- | --- |
| **观测** | 30 |
| **变量** | 8 |

| **简单统计量** | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **x1** | **x2** | **x3** | **x4** | **x5** | **x6** | **x7** | **x8** |
| **均值** | 1921.092667 | 1745.933333 | 511.5083333 | 5457.633333 | 666.1200000 | 117.2866667 | 114.9066667 | 862.9980000 |
| **StD** | 1474.806031 | 861.641934 | 402.8854765 | 1310.218052 | 459.9352604 | 2.0253111 | 1.8980813 | 584.5872585 |

| **相关矩阵** | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **x1** | **x2** | **x3** | **x4** | **x5** | **x6** | **x7** | **x8** |
| **x1** | 1.0000 | 0.2668 | 0.9506 | 0.1906 | 0.6172 | -.2726 | -.2636 | 0.8737 |
| **x2** | 0.2668 | 1.0000 | 0.4261 | 0.7181 | -.1510 | -.2351 | -.5927 | 0.3631 |
| **x3** | 0.9506 | 0.4261 | 1.0000 | 0.3996 | 0.4306 | -.2805 | -.3591 | 0.7919 |
| **x4** | 0.1906 | 0.7181 | 0.3996 | 1.0000 | -.3556 | -.1350 | -.5392 | 0.1044 |
| **x5** | 0.6172 | -.1510 | 0.4306 | -.3556 | 1.0000 | -.2532 | 0.0217 | 0.6586 |
| **x6** | -.2726 | -.2351 | -.2805 | -.1350 | -.2532 | 1.0000 | 0.7628 | -.1252 |
| **x7** | -.2636 | -.5927 | -.3591 | -.5392 | 0.0217 | 0.7628 | 1.0000 | -.1921 |
| **x8** | 0.8737 | 0.3631 | 0.7919 | 0.1044 | 0.6586 | -.1252 | -.1921 | 1.0000 |

相关系数矩阵可以看出，相关性较强的变量依次是

*x*1（GDP）与*x*3（固定资产投资）， *r*13=0.9506

*x*1（GDP）与*x*8（工业总产值）， *r*18=0.8737

*x*3（固定资产投资）与*x*8（工业总产值）， *r*38=0.7919

*x*6（居民消费价格指数）与*x*7（商品零售价格指数）， *r*67=0.7628

| **相关矩阵的特征值** | | | | |
| --- | --- | --- | --- | --- |
|  | **特征值** | **差值** | **比例** | **累积** |
| **1** | 3.75512808 | 1.55839208 | 0.4694 | 0.4694 |
| **2** | 2.19673600 | 0.98184422 | 0.2746 | 0.7440 |
| **3** | 1.21489177 |  | 0.1519 | 0.8958 |

给出相关系数矩阵的特征值、上下特征值之差、各主成分对方差的贡献率以及累积的贡献率。相关系数矩阵的特征值即各主成分的方差，可以看出，第一主成分对方差的贡献率为46.94%，第二主成分对方差的贡献率为27.46%，第三主成分对方差的贡献率为15.19%。前三个主成分的累积贡献率为89.58% > 85%，因此，用前三个主成分就可以很好地概括这组数据。

| **特征向量** | | | |
| --- | --- | --- | --- |
|  | **Prin1** | **Prin2** | **Prin3** |
| **x1** | 0.456650 | 0.258828 | 0.109669 |
| **x2** | 0.313091 | -.403601 | 0.246160 |
| **x3** | 0.470475 | 0.108706 | 0.192322 |
| **x4** | 0.240591 | -.487408 | 0.333810 |
| **x5** | 0.250709 | 0.498125 | -.249714 |
| **x6** | -.262440 | 0.169963 | 0.722755 |
| **x7** | -.319734 | 0.401028 | 0.396976 |
| **x8** | 0.424564 | 0.287908 | 0.191413 |

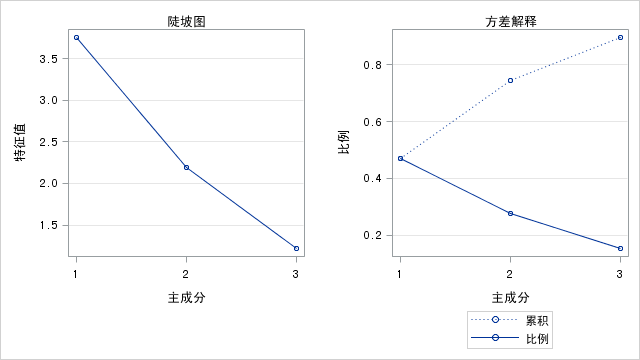
给出三个特征值（主成分）的特征向量，由此可以得到三个主成分的用原变量的线性组合表示（\*表示列向量的标准化）：

Prin1 = 0.46x1\* + 0.31x2\* + 0.47x3\* + 0.24x4\* + 0.25x5\* – 0.26x6\* – 0.32x7\* + 0.42x8\*

Prin2 = 0.26x1\* – 0.40x2\* + 0.11x3\* – 0.49x4\* + 0.50x5\* + 0.17x6\* + 0.40x7\* + 0.29x8\*

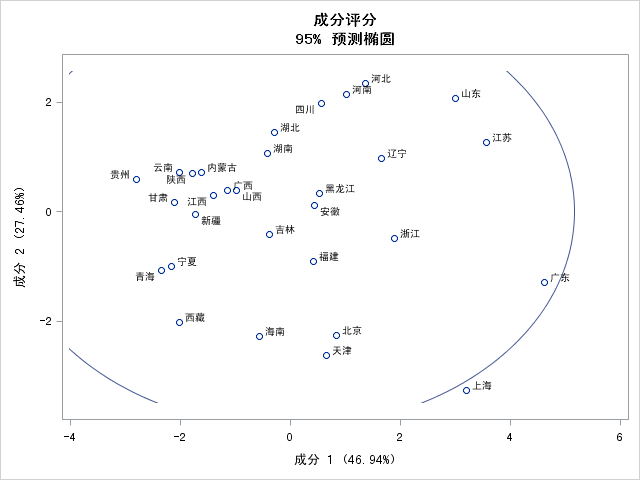
Prin3 = 0.11x1\* + 0.25x2\* + 0.19x3\* + 0.33x4\* – 0.25x5\* + 0.72x6\* + 0.40x7\* + 0.19x8\*

可见，第一主成分中x3、x1、x8的系数最大；第二主成分中x5、x7具有较大的正系数，x4、x2则具有较大的负系数；第三主成分中x6的系数最大，远远超过其他指标的影响。因此，可以把第一主成分看成是由固定资产投资（x3）、GDP（x1）、工业总产值（x8）所刻画的反映经济发展水平的综合指标；把第二主成分看成是由货物周转量（x5）、职工平均工资（x4）、居民消费水平（x2）、商品零售价格指数（x7）所刻画的与人民生活水平有关的综合指标；把第三主成分单独看成是居民消费价格指数（x6）的影响指标。



| **Obs** | **Province** | **x1** | **x2** | **x3** | **x4** | **x5** | **x6** | **x7** | **x8** | **Prin1** | **Prin2** | **Prin3** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **1** | 北京 | 1394.89 | 2505 | 519.01 | 8144 | 373.9 | 117.3 | 112.6 | 843.43 | 0.82827 | -2.25758 | 0.54030 |
| **2** | 天津 | 920.11 | 2720 | 345.46 | 6501 | 342.8 | 115.2 | 110.6 | 582.51 | 0.65759 | -2.63821 | -1.17127 |
| **3** | 河北 | 2849.52 | 1258 | 704.87 | 4839 | 2033.3 | 115.2 | 115.8 | 1234.85 | 1.35759 | 2.35127 | -1.31401 |
| **4** | 山西 | 1092.48 | 1250 | 290.90 | 4721 | 717.3 | 116.9 | 115.6 | 697.25 | -0.98882 | 0.38922 | -0.57132 |
| **5** | 内蒙古 | 832.88 | 1387 | 250.23 | 4134 | 781.7 | 117.5 | 116.8 | 419.39 | -1.62129 | 0.72368 | -0.38131 |
| **6** | 辽宁 | 2793.37 | 2397 | 387.99 | 4911 | 1371.1 | 116.1 | 114.0 | 1840.55 | 1.66278 | 0.97196 | -0.62314 |
| **7** | 吉林 | 1129.20 | 1872 | 320.45 | 4430 | 497.4 | 115.2 | 114.2 | 762.47 | -0.38675 | -0.42395 | -1.20965 |
| **8** | 黑龙江 | 2014.53 | 2334 | 435.73 | 4145 | 824.8 | 116.1 | 114.3 | 1240.37 | 0.52962 | 0.33875 | -0.70859 |
| **9** | 上海 | 2462.57 | 5343 | 996.48 | 9279 | 207.4 | 118.7 | 113.0 | 1642.95 | 3.19719 | -3.27750 | 2.88302 |
| **10** | 江苏 | 5155.25 | 1926 | 1434.95 | 5943 | 1025.5 | 115.8 | 114.3 | 2026.64 | 3.57017 | 1.26123 | 0.38489 |
| **11** | 浙江 | 3524.79 | 2249 | 1006.39 | 6619 | 754.4 | 116.6 | 113.5 | 916.59 | 1.88349 | -0.48552 | 0.22547 |
| **12** | 安徽 | 2003.58 | 1254 | 474.00 | 4609 | 908.3 | 114.8 | 112.7 | 824.14 | 0.44489 | 0.11872 | -1.86164 |
| **13** | 福建 | 2160.52 | 2320 | 553.97 | 5857 | 609.3 | 115.2 | 114.4 | 433.67 | 0.41861 | -0.91913 | -0.65652 |
| **14** | 江西 | 1205.11 | 1182 | 282.84 | 4211 | 411.7 | 116.9 | 115.9 | 571.84 | -1.38992 | 0.29904 | -0.52855 |
| **15** | 山东 | 5002.34 | 1527 | 1229.55 | 5145 | 1196.6 | 117.6 | 114.2 | 2207.69 | 2.99980 | 2.06713 | 0.54600 |
| **16** | 河南 | 3002.74 | 1034 | 670.35 | 4344 | 1574.4 | 116.5 | 114.9 | 1367.92 | 1.02208 | 2.14539 | -0.94079 |
| **17** | 湖北 | 2391.42 | 1527 | 571.68 | 4685 | 849.0 | 120.0 | 116.6 | 1220.72 | -0.28288 | 1.44846 | 1.14458 |
| **18** | 湖南 | 2195.70 | 1408 | 422.61 | 4797 | 1011.8 | 119.0 | 115.5 | 843.83 | -0.41034 | 1.06234 | 0.25469 |
| **19** | 广东 | 5381.72 | 2699 | 1639.83 | 8250 | 656.5 | 114.0 | 111.6 | 1396.35 | 4.61321 | -1.29561 | 0.09506 |
| **20** | 广西 | 1606.15 | 1314 | 382.59 | 5105 | 556.0 | 118.4 | 116.4 | 554.97 | -1.14932 | 0.38142 | 0.37036 |
| **21** | 海南 | 364.17 | 1814 | 198.35 | 5340 | 232.1 | 113.5 | 111.3 | 64.33 | -0.56304 | -2.28905 | -2.40729 |
| **22** | 四川 | 3534.00 | 1261 | 822.54 | 4645 | 902.3 | 118.5 | 117.0 | 1431.81 | 0.56919 | 1.97647 | 0.85166 |
| **23** | 贵州 | 630.07 | 942 | 150.84 | 4475 | 301.1 | 121.4 | 117.2 | 324.72 | -2.80270 | 0.58752 | 1.22127 |
| **24** | 云南 | 1206.68 | 1261 | 334.00 | 5149 | 310.4 | 121.3 | 118.1 | 716.65 | -2.01953 | 0.72284 | 1.89026 |
| **25** | 西藏 | 55.98 | 1110 | 17.87 | 7382 | 4.2 | 117.3 | 114.9 | 5.57 | -2.01580 | -2.01797 | 0.01626 |
| **26** | 陕西 | 1000.03 | 1208 | 300.27 | 4396 | 500.9 | 119.0 | 117.0 | 600.98 | -1.77727 | 0.70634 | 0.45966 |
| **27** | 甘肃 | 553.35 | 1007 | 114.81 | 5493 | 507.0 | 119.8 | 116.5 | 468.79 | -2.11587 | 0.16697 | 0.69430 |
| **28** | 青海 | 165.31 | 1445 | 47.76 | 5753 | 61.6 | 118.0 | 116.3 | 105.80 | -2.34690 | -1.07557 | 0.26359 |
| **29** | 宁夏 | 169.75 | 1355 | 61.98 | 5079 | 121.8 | 117.1 | 115.3 | 114.40 | -2.16125 | -0.99544 | -0.48691 |
| **30** | 新疆 | 834.57 | 1469 | 376.95 | 5348 | 339.0 | 119.7 | 116.7 | 428.76 | -1.72281 | -0.04321 | 1.01964 |

输出原始数据及各主成分的得分。



绘图语句plots= score(ellipse ncomp=2)，“2”要求绘制第二主成分与第一主成分的散点图，ellipse绘制95%置信椭圆。

广东、江苏、上海、山东的第一主成分得分较高，说明这些省市经济发展水平较高，其次是浙江、辽宁、河北、河南、北京、天津等。

由于在第二主成分中职工平均工资与居民消费水平具有负的载荷量，因此处于右半图中的河北、河南、山东等地的职工平均工资与居民消费水平较低，商品零售价格指数较高；而左半图中上海、天津、海南、北京等地的职工平均工资与居民消费水平较高，商品零售价格指数较低。