南京航空航天大學

NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

计算机科学与技术学院
/人工智能学院

多元统计分析大作业

课程 多元统计分析
姓名 颜劭铭
班级 *1620501*
学号 *162050127*
指导教师 李野

对全国 31 省市 2021 年十四项综合指标的多元统计分析

162050127 颜劭铭

日期：2022 年 12 月 10 日

摘　　要

本文为多元统计分析大作业报告，主要是使用《多元统计分析》课程中所学的多种多元统计方法 [3]，在自己制作的全国 31 省市 2021 年十四项综合指标的数据集上分析。选取的十四项指标分别为：地区生产总值, 年末常住人口, 城镇单位就业人员平均工资, 金融业城镇单位就业人员平均工资, 信息传输、计算机服务和软件业城镇单位就业人员平均工资, 住宅商品房平均销售价格, 经营单位所在地进出口总额, 全体居民人均可支配收入, 城市绿地面积, 旅客周转量, 规模以上工业企业专利申请数, 普通高等学校招生数, 医疗卫生机构数, 城市居民最低生活保障人数。使用的方法包括：回归分析、层次聚类分析、贝叶斯判别分析、主成分分析、因子分析。该数据集的十四项综合指标包括经济、环境、人口、教育、外贸、医疗、交通、文化等方面，该任务主要目标是检验该十四项指标是否能反映一个省市的发展水平及不同指标对于发展水平的贡献。本次作业使用 python 实现。

关键词: 多元统计分析，31 省市，python

1　数据集选择及部分可视化

本任务选择的数据来自国家统计局网站（）。

十四项指标为地区, 地区生产总值 (亿元), 年末常住人口 (万人), 城镇单位就业人员平均工资 (元), 金融业城镇单位就业人员平均工资 (元), 信息传输、计算机服务和软件业城镇单位就业人员平均工资 (元), 住宅商品房平均销售价格 (元/平方米), 经营单位所在地进出口总额 (千美元), 全体居民人均可支配收入 (元), 城市绿地面积 (万公顷), 旅客周转量 (亿人公里), 规模以上工业企业专利申请数 (件), 普通高等学校招生数 (万人), 医疗卫生机构数 (个), 城市居民最低生活保障人数 (万人)。其中包括了经济、人口、金融计算机两大热门行业工资、住宅、外贸、生活、环境、旅游、科技、教育、医疗、社会服务等可以反映一个地区发展水平的指标。有些指标的选择可能不够恰当，因为更恰当的指标在某些地区的 2021 年数据是缺失的，因此只能选择其他指标进行代替。

由于地区生产总值是最能反映一个地区发展水平的指标，因此在回归分析任务中，将地区生产总值作为因变量，剩余的十三项指标作为自变量，分析其他指标与地区发展水平的关系。

而在其他分析任务中，这十四项指标将一同作为反映各省市地区的特征，比较不同指标对于发展水平的贡献。

在表1中列出部分地区及部分指标，完整地区及指标见附录。

在图1、图2中列出部分指标的可视化图片，所有数据可视化图片见附录。

地区	地区生产总值（亿元）	年末常住人口（万人）
北京市	40269.6	2189
天津市	15695	1373
河北省	40391.3	7448
山西省	22590.2	3480

表 1: 数据集部分地区部分指标

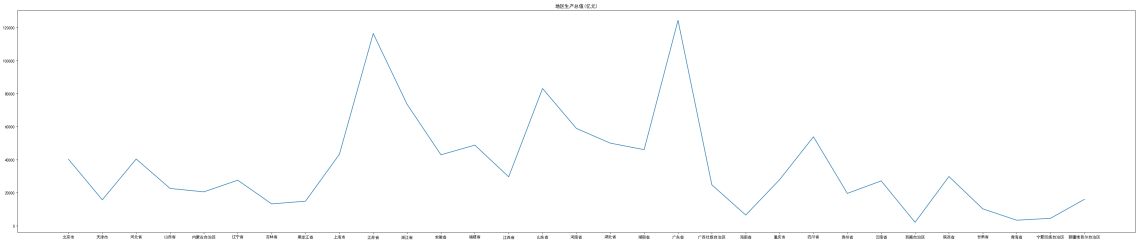

图 1: 地区生产总值

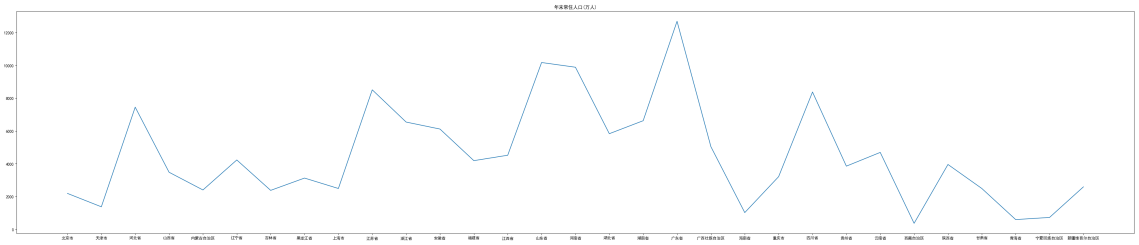

图 2: 年末常住人口

2 回归分析与显著性检验

在回归分析部分中，我们将地区生产总值作为因变量，剩余的十三项指标作为自变量，通过经典多元线性回归模型对参数进行估计，并对回归方程和回归系数进行显著性检验。

具体来说，该任务使用的是经典的一对多多元线性回归模型 [3]，利用最小二乘法建立出以下矩阵形式的回归方程:

$$\begin{cases} Y = C\beta + \varepsilon \\ \varepsilon \sim N_n\left(0, \sigma^2 I_n\right) \end{cases}$$

建立后利用方程的显著性检验:

$$H_0: \quad \beta_1 = \beta_2 = \ldots = \beta_m = 0 F\ F(m, n-m-1)$$

和回归系数的显著性检验:

$$t\ t(n-m-1)$$

得到以下指标：R^2: 判定系数，F-statistic：统计检验 F 统计量,Prob (F-statistic): F 检验的 P 值, coef：自变量和常数项的系数，$b_1, b_2, \ldots, b_m, b_0$,std err：系数估计的标准误差,t：统计检验 t 统计量,P>|t|：t 检验的 P 值。

得到的结果如表 2 所示，其中 $x_1, x_2, \ldots, x_{13}$ 分别按顺序代表除了地区生产总值以外的十三项指标，const 指的是常数项，除此以外，

由2可知, 若记地区生产总值 (亿元) 为 Y, 其余十三项指标为 $X_{\mathrm{i}}(\mathrm{i}=1,2,\cdots,13)$, 则线性回归模型可表示为:

$$\begin{aligned} Y = & -30050 + 0.8340 * X_1 + 0.3073 * X_2 - 0.0969 * X_3 + 0.0439 * X_4 - 1.0114 * X_5 + \\ & 0.00004667 * X_6 + 0.5115 * X_7 - 120.6064 * X_8 + 8.7424 * X_9 + 0.0475 * X_10 + \\ & 609.2443 * X_11 - 0.0243 * X_12 - 21.6557 * X_13 \end{aligned}$$

由于复相关系数的平方 $R^2 = 0.973$，可以说明所建立的一对多线性回归模型可以很好地解释地区生产总值的变化原因。

	coef	std err	t	P>|t|
const	-30050	17400	-1.0728	0.102
x1	0.8340	4.248	0.196	0.847
x2	0.3073	0.216	1.423	0.173
x3	-0.0969	0.071	-1.368	0.189
x4	0.0439	0.055	0.792	0.439
x5	-1.0114	0.604	-1.675	0.112
x6	0.00004667	0.0000341	1.370	0.188
x7	0.5115	0.355	1.441	0.168
x8	-120.6064	522.373	-0.231	0.820
x9	8.7424	15.366	0.569	0.577
x10	0.0475	0.131	0.362	0.722
x11	609.2443	367.755	1.657	0.116
x12	-0.0243	0.206	-0.118	0.908
x13	-21.6557	123.227	-0.176	0.863

表 2: 回归分析结果

对回归方程进行显著性检验后结果如下：F-statistic：47.12,Prob (F-statistic): 1.03e-10。说明剩余作为自变量的 13 个指标至少一项会对因变量地区生产总值产生线性影响关系。

再对各项指标的回归系数进行显著性检验，以 0.85 为阈值，发现指标医疗卫生机构数和城市居民最低生活保障人数两个指标可以通过检验，其中医疗卫生机构数的系数为-0.0243，t 统计量为-0.119，p 值为 0.908，城市居民最低生活保障人数的系数为-21.6557，t 统计量为-0.176，p 值为 0.863, 证明这两个指标都会对于因变量地区生产总值产生显著的负向线性关系。

总的来说，通过显著性检验的结果我们可以发现，回归分析建立的方程证明选取的 13 个指标在一定程度上可以很好地解释地区生产总值的变化原因，除此以外，医疗卫生机构数和城市居民最低生活保障人数作为人民生活最重要的生命健康和社会服务保障方面的两个数据，与地区生产总值有很强的线性关系，符合我们生活中的认知，当人民生活健康及社会服务得不到保障的时候，很难提升所在地区的生产总值。

3 聚类分析

在聚类分析中，我选择欧氏距离作为距离度量，采用系统聚类法 [**2**]，分别通过最小距离法、最长距离法和离差平方和法（WARD）来对三十一个省市进行聚类。

对于系统聚类法进行简要的说明：

```
(1)对于数据进行标准化
(2)计算31个省市两两间距离，得到初始的距离矩阵
(3)对于得到的矩阵，每次合并类间距离最小的两个类为新的一个类
(4)通过采用的方法来计算新类与其他类之间的距离
(5)重复(3)、(4)步，直到类的个数为1
(6)画出谱系聚类图
```

最短距离法如下所示：

$$D_{pq} = \min_{X_{(i)} \in G_p, X_{(j)} \in G_q} d_{ij}$$

$$D_{rk} = \min \{D_{pk}, D_{qk}\}$$

最长距离法如下所示：

$$D_{pq} = \max_{X_{(i)} \in G_p, X_{(j)} \in G_q} d_{ij}$$

$$D_{rk} = \max\{D_{pk}, D_{qk}\}$$

离差平方和法如下所示：

$$W = \sum_{t=1}^{k} W_t = \sum_{t=1}^{k} \sum_{i=1}^{n_t} \left(\bar{X}_{(i)}^{(t)} - \bar{X}^{(t)}\right)^T \left(\bar{X}_{(i)}^{(t)} - \bar{X}^{(t)}\right)$$

$$D_{pq}^2 = W_r - (W_p + W_q)$$

聚类结果如下图所示，其中图3为离差平方和法层次聚类结果，图4为最短距离法层次聚类结果，图5为最长距离法层次距离结果：

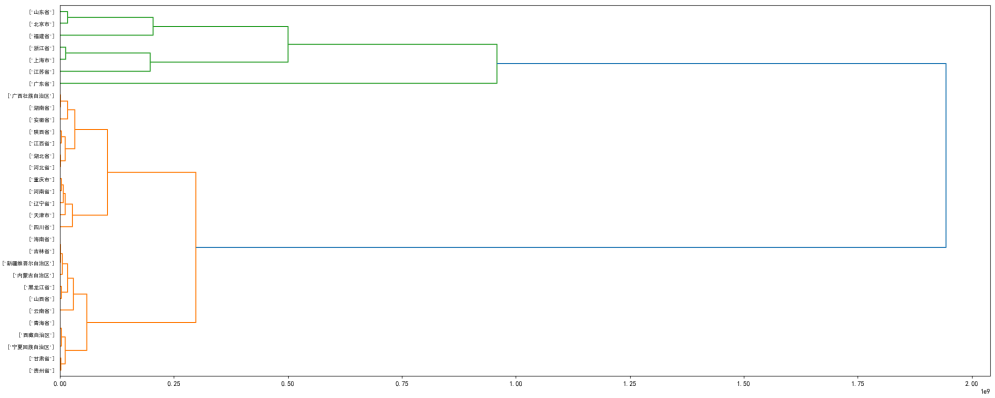

图 3: 离差平方和法层次聚类结果

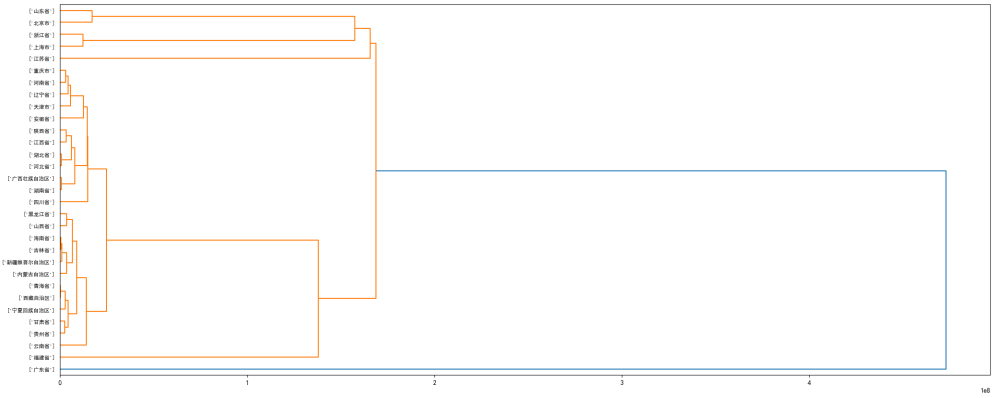

图 4: 最短距离法层次聚类结果

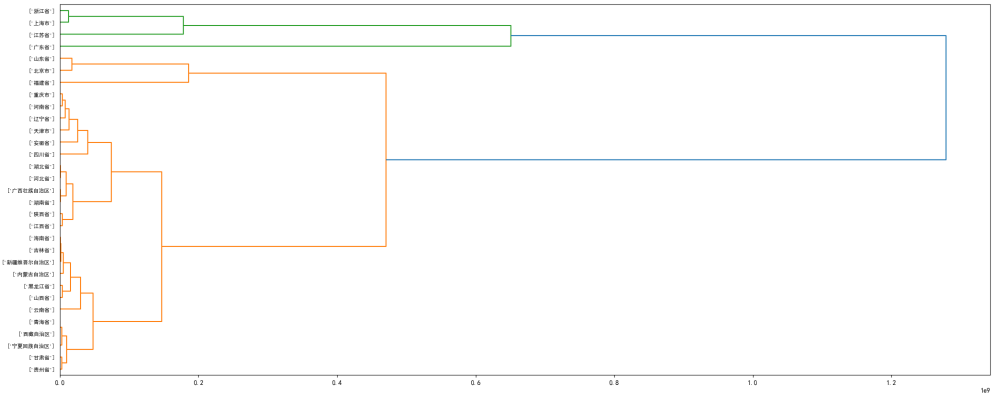

图 5: 最长距离法层次聚类结果

可以发现，三种聚类方法的结果有着较大的差别，而结合我们生活中对于各个省市发展结果的认知，离差平方和层次聚类法应该是较为准确的，这是因为不同省市的不同指标通常存在较大差异，而单纯使用最长距离和最短距离法进行判断通常会导致判别的使用过于注重某个指标，无法综合考虑多个指标的影响结果，因此离差平方和法是较为适合的。

4 贝叶斯判别

贝叶斯判别 [3] 是判别分析的一种，在这里，使用该方法来检验聚类结果是否正确，而由上节对于聚类结果的观察，决定采用离差平方和法的结果进行检验。

贝叶斯判别方法是以错分概率或风险最小为准则的判别规则，并且利用得到的先验概率、条件概率密度和贝叶斯公式计算后验概率来进行判别。

在聚类后，31 个省市分别被分为 6 个类，如表3所示：

类编号	包含省市
1	山西省、内蒙古自治区、吉林省、黑龙江省、海南省、贵州省、云南省、西藏自治区、甘肃省、青海省、宁夏回族自治区、新疆回族维吾尔自治区
2	天津市、河北省、辽宁省、安徽省、江西省、河南省、湖北省、湖南省、广西壮族自治区、重庆市、四川省、陕西省
3	上海市、江苏省、浙江省
4	北京市、上海市、山东省
5	福建省
6	广东省

表 3: 离差平方和法聚类结果

通过随机数随机选取其中 21 个省份作为训练集，其余 10 个省份作为测试集，可以得到判别结果的准确率为 90%，判别结果如表4所示：

地区	所属类别	贝叶斯判别结果
广西壮族自治区	2	2
福建省	5	4
安徽省	2	2
江西省	2	2
河南省	2	2
内蒙古自治区	1	1
新疆维吾尔自治区	1	1
宁夏回族自治区	1	1
浙江省	3	3
湖北省	2	2

表 4: 贝叶斯判别结果

可以看到，使用离差平方和聚类对全国 31 个省市进行分组的结果，与贝叶斯判别预测的分组结果基本吻合，说明聚类的分组效果较好。

5 主成分分析

在指标的选择过程中，我们会有部分指标和其他指标实际上信息会存在重复，这会增加数据的复杂度，而数据的复杂度会引起分析的难度加大，并且会影响分析的结果，通过使用主成分分析的方法，我们可以达到降维的目的，降低数据的复杂度。

主成分分析 [2] 通过将原始变量转换为原始变量的线性组合（主成分），在保留主要信息的基础上，达到简化和降维的目的，公式如下所示：

$$\begin{cases} Z_1 = a_1' X = a_{11}X_1 + a_{21}X_2 + \cdots + a_{01}X_0, \\ Z_2 = a_2' X = a_{12}X_1 + a_{22}X_2 + \cdots + a_{02}X_0, \\ \cdots\cdots\cdots\cdots\cdots\cdots \\ Z_0 = a_0' X = a_{10}X_1 + a_{20}X_2 + \cdots + a_{00}X_0, \end{cases}$$

其中 $Z_i = a_i^T X$ 被称为 X 的第 i 主成分，目标就是在限制条件 $a_i^T a_i = 1$ 的条件下，$Var(Z_i)$ 越大，包含的信息越多。

5.1 计算特征值、贡献率

首先我们可以计算出特征值，并根据每个指标的特征值计算出贡献率，公式如下所示：

$$G(m) = \frac{\sum_{i=1}^{m} \lambda_i}{\sum_{i=1}^{n} \lambda_i}$$

对于数据集进行计算，得到的特征根，贡献率，累计贡献率如表5所示，其中 $x_1, x_2, \ldots, x_{14}$ 分别按顺序代表十四项指标：

	特征值	贡献率/%	累计贡献率/%
x1	6.5864	45.52	45.52
x2	5.5595	38.42	83.94
x3	0.9000	6.22	90.16
x4	0.5699	3.93	94.09
x5	0.2468	1.70	95.79
x6	0.2094	1.44	97.23
x7	0.1918	1.32	98.55
x8	0.0608	0.42	98.97
x9	0.0459	0.31	99.28
x10	0.0393	0.27	99.55
x11	0.0271	0.21	99.76
x12	0.0154	0.14	99.90
x13	0.0084	0.07	99.97
x14	0.0055	0.03	100.00

表 5: 特征根，贡献率，累计贡献率

由表 4 结果可知，当保留 2 个主成分时，累计贡献率达到 83.94%，超过了 80%，因此在后续分析过程中选取了前两个特征值 $\lambda_1 = 6.5864, \lambda_2 = 5.5595$ 及其对应的主成分，其贡献率分别为 45.52%,38.42%.

之后，通过计算获得了各个主成分在各个指标上的因子负荷量，它反映了主成分在各个变量上的载荷情况。具体数值如表6所示：

同时由表 5 可以得到主成分与各个变量之间的线性关系（将主成分记为 $Z_i(i = 1, 2)$，各项指标记为 $X_i(i = 1, 2, \ldots, 14)$。

$$Z_1 = 0.38203176 * X_1 + 0.32133501 * X_2 + 0.12761281 * X_3 + 0.15274555 * X_4 + 0.17765656 * X_5 + 0.16256967 * X_6 + 0.36873067 * X_7 + 0.22368911 * X_8 + 0.36474319 * X_9 + 0.28527893 * X_{10} + 0.35930962 * X_{11} + 0.2819734 * X_{12} + 0.19557968 * X_{13} - 0.07683722 * X_{14}$$

$$Z_2 = -0.07992167 * X_1 - 0.24188046 * X_2 + 0.38845776 * X_3 + 0.35182274 * X_4 + 0.34621783 * X_5 + 0.36021989 * X_6 + 0.10382162 * X_7 + 0.31702445 * X_8 - 0.06158961 * X_9 - 0.25470318 * X_{10} - 0.03534915 * X_{11} - 0.2722959 * X_{12} - 0.30243376 * X_{13} - 0.25296506 * X_{14}$$

指标	主成分 1	主成分 2
地区生产总值	0.38203176	-0.07992167
年末常住人口	0.32133501	-0.24188046
城镇单位就业人员平均工资	0.12761281	0.38845776
金融业城镇单位就业人员平均工资	0.15274555	0.35182274
信息传输、计算机服务和软件业城镇单位就业人员平均工资	0.17765656	0.34621783
住宅商品房平均销售价格	0.16256967	0.36021989
经营单位所在地进出口总额	0.36873067	0.10382162
全体居民人均可支配收入	0.22368911	0.31702445
城市绿地面积	0.36474319	-0.06158961
旅客周转量	0.28527893	-0.25470318
规模以上工业企业专利申请数	0.35930962	-0.03534915
普通高等学校招生数	0.2819734	-0.2722959
医疗卫生机构数	0.19557968	-0.30243376
城市居民最低生活保障人数	-0.07683722	-0.25296506

表 6: 因子负荷量

5.2 主成分排序

因此简单利用第一主成分，将 31 省市的数据带入计算可得到主成分得分如表 6 所示，表7只展示了前十名省市，具体数据见附录:

地区	第一主成分得分
广东省	472099069
江苏省	297584094
浙江省	236529696
上海市	231951346
北京市	173843152
山东省	167500126
福建省	105293882
四川省	54479651
天津市	48977893
河南省	46942120

表 7: 第一主成分得分

由表 5 和表 6 的结果可以发现，第一主成分受地区生产总值、年末常住人口、经营单位所在地进出口总额、城市绿地面积、规模以上工业企业专利申请数这五项指标的影响最大，使得规模较大、人口众多、经济发达、科技发达的大省排名较前，如广东省、江苏省、浙江省，而北京市、上海市受占地面积和人口的制约，排名相比起我们所说的“北上广深”较为落后，这反映了省级区划和直辖市实际情况和发展模式的不同。

6 因子分析

主成分分析法和因子分析法 [3] 都寻求少数的几个变量（或因子）来综合反映全部变量（或因子）的大部分信息，变量虽然较原始变量少，但所包含的信息量却占原始信息量的 85% 以上，用这些新变量来分析问题，其可信程度仍然很高，而且这些新的变量彼此间互不相关，消除了多重共线性。这两种分析法得出的新变量，并不是原始变量筛选后剩余的变量。

但是，在主成分分析中，最终确定的新变量是原始变量的线性组合，如原始变量为 $x_1, x_2, \ldots$，经过坐标变换，将原有的 p 个相关变量 x_i 作线性变换，每个主成分都是由原有 p 个变量线性组合得到。在诸多主成分 Z_i 中 Z_1 在方差中占的比重最大，说明它综合原有变量的能力最强，越往后主成分在方差中的比重也小，综合原信息的能力越弱。

而因子分析是要利用少数几个公共因子去解释较多个要观测变量中存在的复杂关系，它不是对原始变量的重新组合，而是对原始变量进行分解，分解为公共因子与特殊因子两部分。公共因子是由所有变量共同具有的少数几个因子；特殊因子是每个原始变量独自具有的因子。

因子分析建立的模型如下所示：

$$\begin{cases} X_1 - \mu_1 = a_{11}f_1 + a_{12}f_2 + \cdots + a_{1m}f_m + \varepsilon_1 \\ X_2 - \mu_2 = a_{21}f_1 + a_{22}f_2 + \cdots + a_{2m}f_m + \varepsilon_2 \\ \vdots \\ X_p - \mu_p = a_{p1}f_1 + a_{p2}f_2 + \cdots + a_{pm}f_m + \varepsilon_p \end{cases}$$

可将因子分析一般模型写成矩阵表示形式：

$$X = \mu + AF + \varepsilon$$

其中 $F = (f_1, f_2, \cdots f_m)^T$ 为公因子向量, $\varepsilon = (\varepsilon_1, \varepsilon_2, \cdots \varepsilon_p)^T$ 为特殊因子向量 $A = (a_{ij})_{p\times m}$ 为因子载荷矩阵。

首先利用 Bartlett's 球状检验和 KMO 检验查看是否可以进行因子分析，可以得到统计量 p-value 的值为 0，表明变量的相关矩阵不是单位矩阵，即各个变量之间是存在一定的相关性，而 KMO 大于 0.6，也说明变量之间存在相关性，可以进行因子分析。

其次进行因子个数的选择，首先计算特征值和特征向量，并由特征值和因子个数的变化绘制出折线图如图6所示：

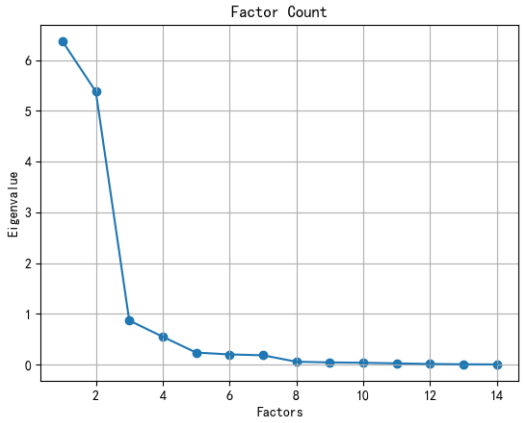

图 6: 特征值和因子个数的变化

可以发现选择 3 个因子即可进行因子分析。

因子分析的过程中利用方差最大化进行因子旋转建立因子分析模型，得到公因子方差如表8所示。

而因子贡献率得到的结果如下矩阵所示：

$$[4.92369347, \quad 4.4892255, \quad 2.81584827]$$
$$[0.35169239, \quad 0.32065896, \quad 0.20113202]$$
$$[0.35169239, \quad 0.67235135, \quad 0.87348337]$$

指标	公因子方差
地区生产总值 (亿元)	0.967505
年末常住人口 (万人)	1.004351
城镇单位就业人员平均工资 (元)	0.944354
金融业城镇单位就业人员平均工资 (元)	0.807432
信息传输、计算机服务和软件业城镇单位就业人员平均工资 (元)	0.810700
住宅商品房平均销售价格 (元/平方米)	0.951383
经营单位所在地进出口总额 (千美元)	0.991676
全体居民人均可支配收入 (元)	0.872760
城市绿地面积 (万公顷)	0.891435
旅客周转量 (亿人公里)	0.850399
规模以上工业企业专利申请数 (件)	1.002483
普通高等学校招生数 (万人)	0.984818
医疗卫生机构数 (个)	0.795543
城市居民最低生活保障人数 (万人)	0.353928

表 8: 公因子方差

旋转后的成分矩阵如表9所示，将其转化为热力图如图7所示：

指标	因子 1	因子 2	因子 3
地区生产总值 (亿元)	0.235614	0.707590	0.641333
年末常住人口 (万人)	-0.080676	0.921142	0.386444
城镇单位就业人员平均工资 (元)	0.947934	-0.198334	0.080240
金融业城镇单位就业人员平均工资 (元)	0.887230	-0.097825	0.103368
信息传输、计算机服务和软件业城镇单位就业人员平均工资 (元)	0.859965	-0.115448	0.240484
住宅商品房平均销售价格 (元/平方米)	0.974031	-0.038814	0.033777
经营单位所在地进出口总额 (千美元)	0.526409	0.362891	0.763465
全体居民人均可支配收入 (元)	0.909054	0.065369	0.205200
城市绿地面积 (万公顷)	0.201553	0.569648	0.725474
旅客周转量 (亿人公里)	-0.147638	0.844115	0.340693
规模以上工业企业专利申请数 (件)	0.185115	0.436587	0.881821
普通高等学校招生数 (万人)	-0.143324	0.954821	0.229334
医疗卫生机构数 (个)	-0.266381	0.849861	0.048176
城市居民最低生活保障人数 (万人)	-0.442678	0.280631	-0.281443

表 9: 成分矩阵

可以看出隐藏变量与城镇单位就业人员平均工资、金融业城镇单位就业人员平均工资、信息传输、计算机服务和软件业城镇单位就业人员平均工资、住宅商品房平均销售价格、全体居民人均可支配收入这几个指标关系较大，可以看出经济指标对于发展水平的影响是最大的，这几个指标都是经济类指标。

由于之前主成分分析已经进行过主成分得分的计算，因此因子分析没有进行因子得分的计算。

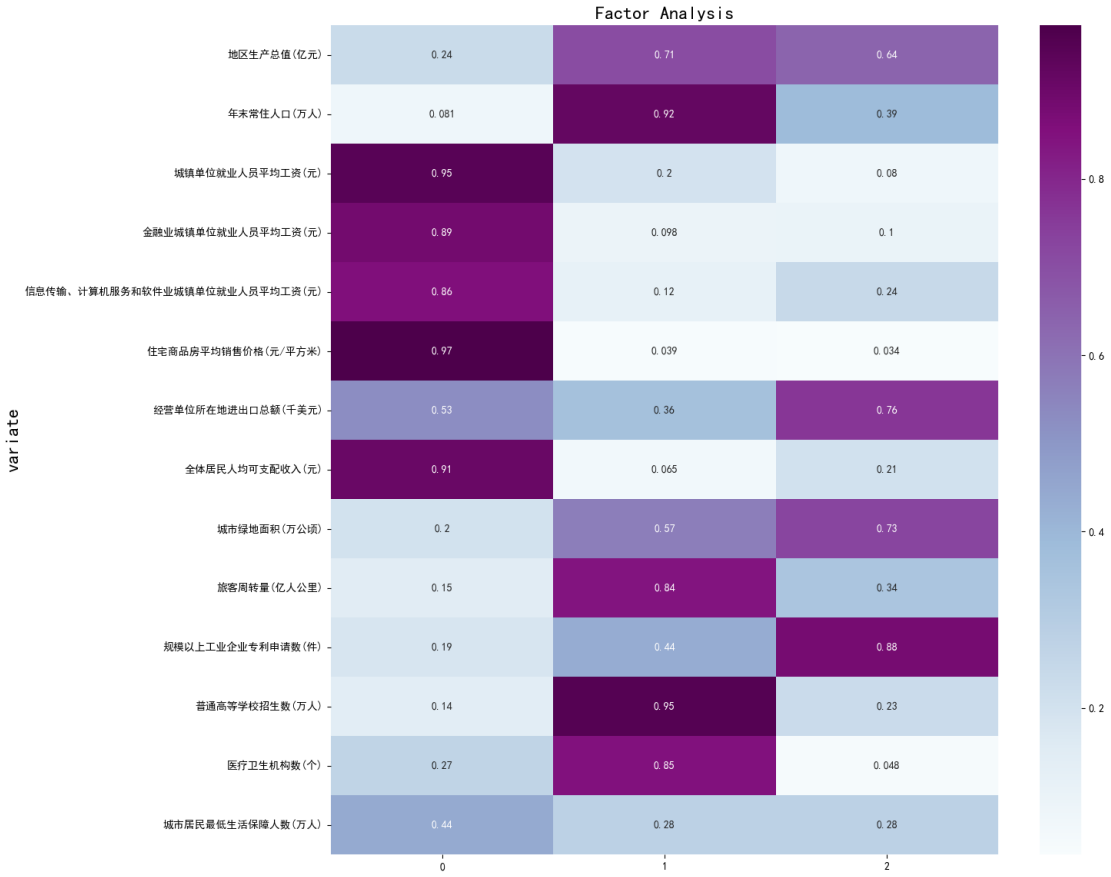

图 7: 成分矩阵可视化

7 总结

由表 2 聚类结果和附录中的第一主成分得分排序结果相比较可以发现结果还是较为一致的。广东省、江苏省、浙江省这几个沿海省份首先外贸出口多，占地面积广，高等教育，科技等方面也较为重视，并且分别有深圳、杭州、广州等城市作为支撑，是省份中的第一梯队。

上海市、北京市作为直辖市，虽然各方面都有着自己的优势，但是由于占地面积与人口的问题，距离第一的广东省有着较大的差距，但是与江苏省和浙江省的差距较小。

后续的山东省、福建省、四川省也符合我们的认知。

排在最后的主要是西部省份和东北省份，这些省份发展经济较晚，虽然占地广，但是各方面资源都较为缺乏，高等教育、外贸等实力较弱，因此得分较低，但是其中海南省是比较出乎意料的，我认为主要是因为海南省主要依靠的是旅游产业，而 2021 年的疫情冲击导致海南省发展水平较为停滞，各方面都落后很多。

参考文献

[1] George Karypis, Eui-Hong Han, and Vipin Kumar. “Chameleon: Hierarchical clustering using dynamic modeling”. In: *Computer* 32.8 (1999), pp. 68–75.

[2] Karl Pearson. “LIII. On lines and planes of closest fit to systems of points in space”. In: *The London, Edinburgh, and Dublin philosophical magazine and journal of science* 2.11 (1901), pp. 559–572.

[3] 何晓群. “应用多元统计分析 (第二版)”. In: 中国统计 10 (2015).

附录

A 数据可视化

完整地区及指标如表10、表11、表12所示。

	地区	地区生产总值 (亿元)	年末常住人口 (万人)	住宅商品房平均销售价格 (元/平方米)	经营单位所在地进出口总额 (千美元)
0	北京市	40269.6	2189	46941	470990000
1	天津市	15695.0	1373	16370	132570000
2	河北省	40391.3	7448	8330	83850000
3	山西省	22590.2	3480	6689	34530000
4	内蒙古自治区	20514.2	2400	6516	19140000
5	辽宁省	27584.1	4229	9049	119460000
6	吉林省	13235.5	2375	7050	23260000
7	黑龙江省	14879.2	3125	6242	30850000
8	上海市	43214.9	2489	40974	628520000
9	江苏省	116364.2	8505	13666	806470000
10	浙江省	73515.8	6540	20386	640930000
11	安徽省	42959.2	6113	7904	107000000
12	福建省	48810.4	4187	12653	285250000
13	江西省	29619.7	4517	7648	77020000
14	山东省	83095.9	10170	8743	453870000
15	河南省	58887.4	9883	6438	127010000
16	湖北省	50012.9	5830	9100	83080000
17	湖南省	46063.1	6622	6481	92400000
18	广东省	124369.7	12684	16453	1279570000
19	广西壮族自治区	24740.9	5037	5992	91720000
20	海南省	6475.2	1020	17550	22750000
21	重庆市	27894.0	3212	9678	123820000
22	四川省	53850.8	8372	8304	147430000
23	贵州省	19586.4	3852	5623	10130000
24	云南省	27146.8	4690	7869	48680000
25	西藏自治区	2080.2	366	8443	620000
26	陕西省	29801.0	3954	9680	73560000
27	甘肃省	10243.3	2490	5985	7610000
28	青海省	3346.6	594	7851	490000
29	宁夏回族自治区	4522.3	725	6916	3320000
30	新疆维吾尔自治区	15983.6	2589	5548	24300000

表 10: 完整地区及指标 1

地区	城镇单位就业人员平均工资 (元)	金融业城镇单位就业人员平均工资 (元)	信息传输、计算机服务和软件业城镇单位就业人员平均工资 (元)	全体居民人均可支配收入 (元)
北京市	194651	298200	290038	75002
天津市	123528	155286	157725	47449
河北省	82526	95401	132218	29383
山西省	82413	87734	99130	27426
内蒙古自治区	90426	97446	113337	34108
辽宁省	86062	102613	121947	35112
吉林省	83028	84244	93158	27770
黑龙江省	80369	74150	93945	27159
上海市	191844	397655	303573	78027
江苏省	115133	164177	180782	47498
浙江省	122309	175773	257631	57541
安徽省	93861	110117	111935	30904
福建省	98071	131573	143350	40659
江西省	83766	106042	104940	30610
山东省	94768	97701	116084	35705
河南省	74872	125279	91501	26811
湖北省	96994	127544	131663	30829
湖南省	85438	108042	117793	31993
广东省	118133	202771	213031	44993
广西壮族自治区	88170	112750	111988	26727
海南省	97471	113934	247110	30457
重庆市	101670	129860	155067	33803
四川省	96741	113147	147727	29080
贵州省	94487	142097	118727	23996
云南省	98730	144635	111243	25666
西藏自治区	140355	245574	179818	24950
陕西省	90996	112682	192699	28568
甘肃省	84500	86769	93005	22066
青海省	109346	144717	137031	25920
宁夏回族自治区	105266	112052	130972	27905
新疆维吾尔自治区	94281	132160	119526	26075

表 11: 完整地区及指标 2

地区	城市绿地面积 (万公顷)	旅客周转量 (亿人公里)	规模以上工业企业专利申请数 (件)	普通高等学校招生数 (万人)	医疗卫生机构数 (个)	城市居民最低生活保障人数 (万人)
北京市	9.31	149.59	28221	15.98	10699	7.1
天津市	4.61	171.01	18952	15.85	6076	6.7
河北省	10.15	682.13	30171	48.28	88162	15.7
山西省	5.66	221.32	10152	25.81	41007	23.8
内蒙古自治区	7.08	164.82	7722	14.49	24948	28.3
辽宁省	14.77	431.43	20104	29.01	33051	30.9
吉林省	9.45	208.72	7949	20.63	25344	35.0
黑龙江省	7.30	190.24	6691	23.90	20578	49.0
上海市	17.12	134.65	41431	14.70	6308	13.6
江苏省	31.44	991.53	207371	60.99	36448	10.0
浙江省	18.32	713.62	159920	34.75	35120	6.0
安徽省	12.76	753.60	75058	44.02	29554	31.7
福建省	8.09	313.97	51551	29.39	28693	6.5
江西省	7.96	603.91	32350	38.71	36764	31.1
山东省	27.25	706.81	98190	71.11	85715	10.9
河南省	12.82	985.32	45391	82.37	78536	35.8
湖北省	11.33	620.77	54807	47.09	36529	28.3
湖南省	9.76	857.69	40576	47.46	55677	39.0
广东省	53.29	940.86	340935	69.43	57964	15.0
广西壮族自治区	7.61	521.94	11641	40.39	34112	34.4
海南省	1.84	89.94	1613	6.63	6277	3.4
重庆市	7.34	281.17	22240	29.68	21361	23.9
四川省	13.95	596.57	41236	56.04	80249	58.9
贵州省	9.94	410.16	8372	25.52	29292	60.8
云南省	5.32	283.36	9467	31.97	26885	39.2
西藏自治区	0.64	30.16	100	1.08	6907	2.3
陕西省	7.62	435.53	16285	35.10	34971	18.6
甘肃省	3.12	335.90	4645	16.83	25759	32.6
青海省	0.87	85.01	1354	2.35	6408	5.9
宁夏回族自治区	2.71	55.92	3935	4.86	4571	7.6
新疆维吾尔自治区	8.56	260.78	5181	16.91	16970	25.8

表 12: 完整地区及指标 3

每个指标 31 省市可视化对比图如图8-图19所示。

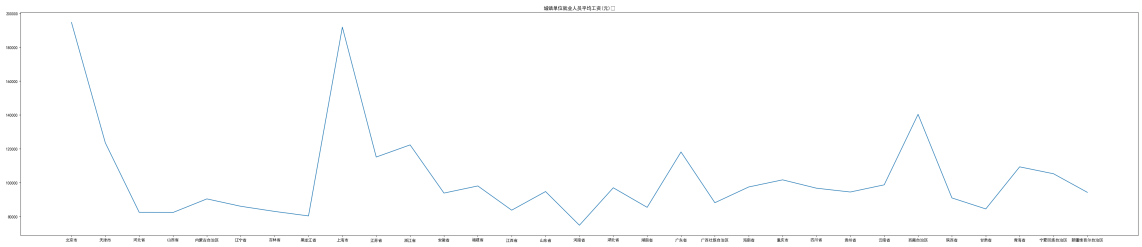

图 8: 城镇单位就业人员平均工资

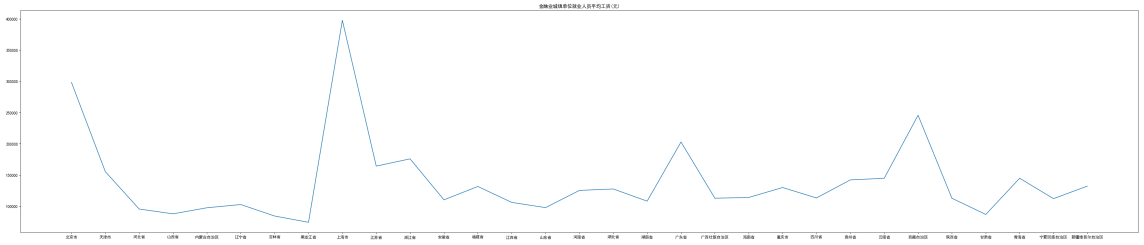

图 9: 金融业城镇单位就业人员平均工资

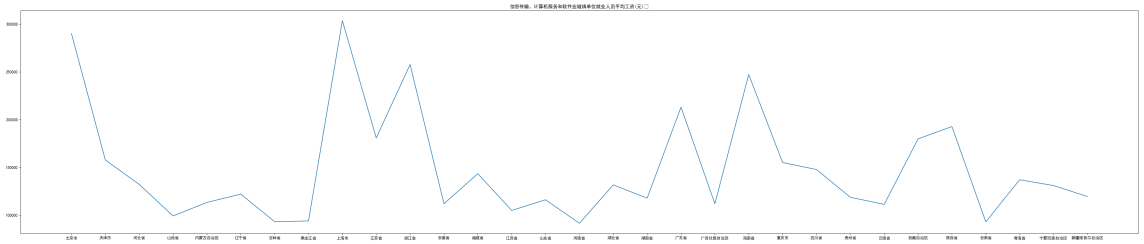

图 10: 信息传输、计算机服务和软件业城镇单位就业人员平均工资

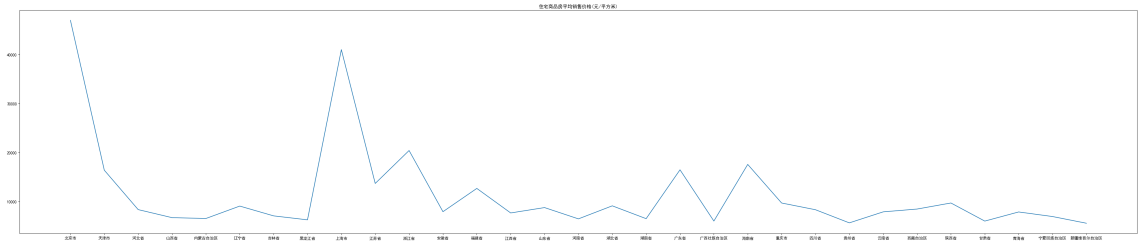

图 11: 住宅商品房平均销售价格

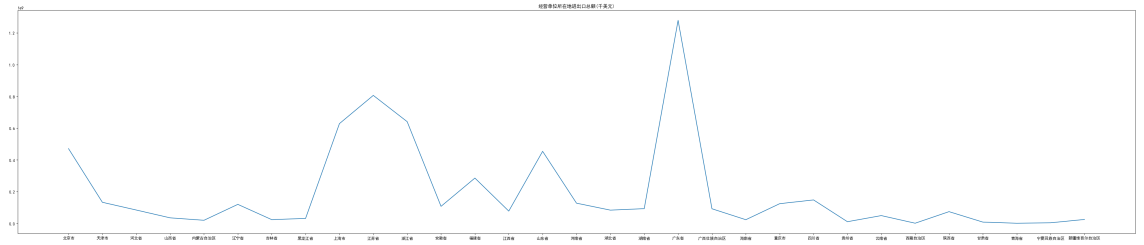

图 12: 经营单位所在地进出口总额

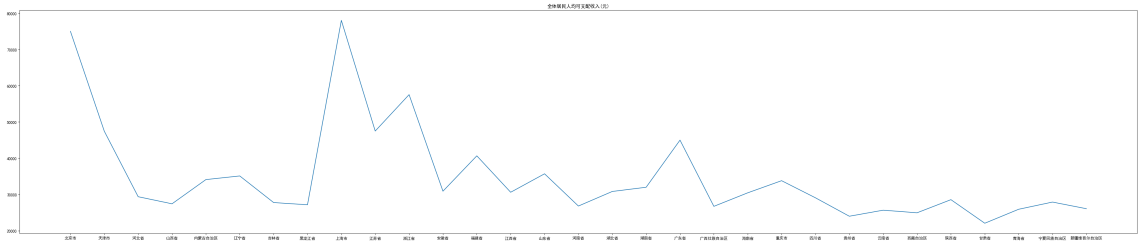

图 13: 全体居民人均可支配收入

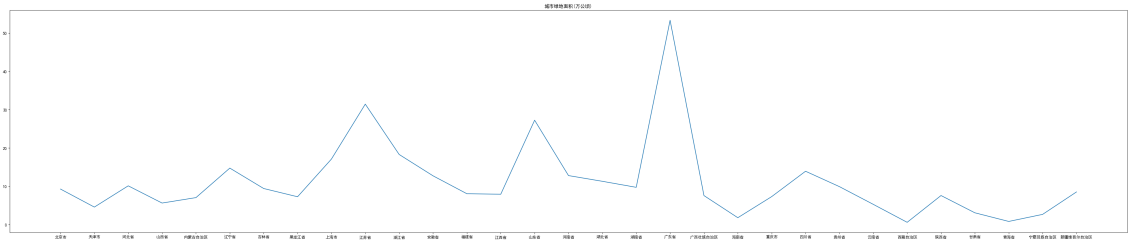

图 14: 城市绿地面积

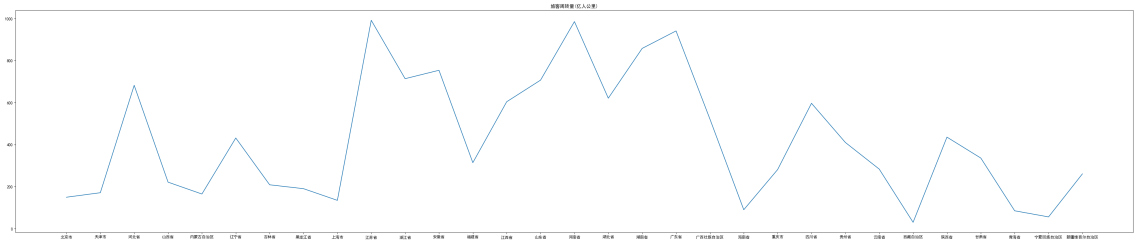

图 15: 旅客周转量

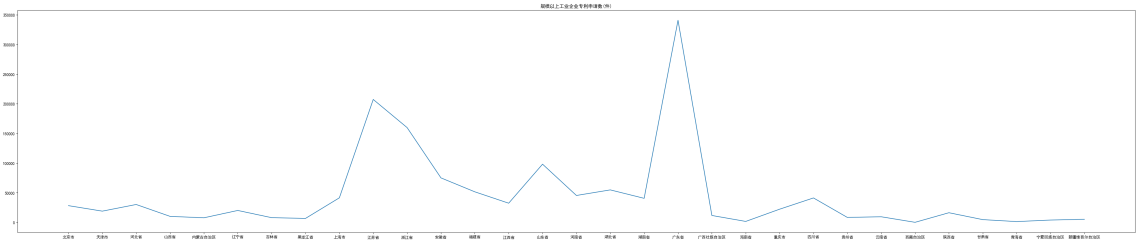

图 16: 规模以上工业企业专利申请数

图 17: 普通高等学校招生数

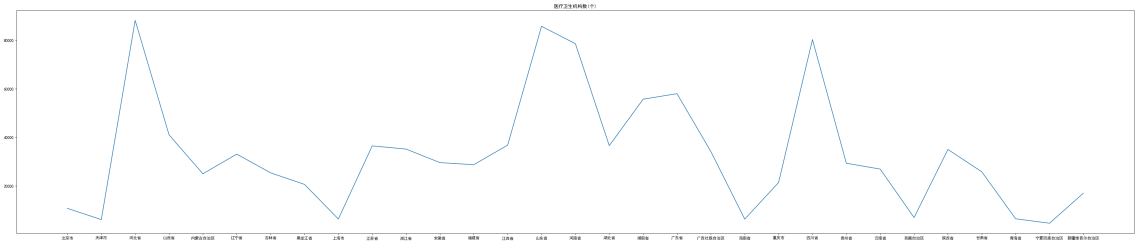

图 18: 医疗卫生机构数

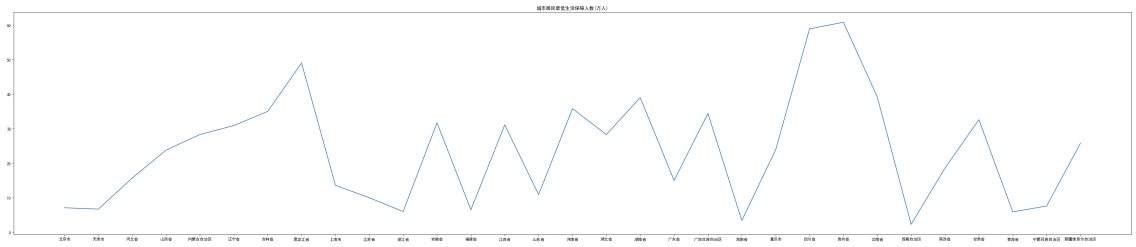

图 19: 城市居民最低生活保障人数

B 主成分

利用第一主成分，将 31 省市的数据带入计算可得到主成分得分如表13所示：

地区	第一主成分得分
广东省	472099069
江苏省	297584094
浙江省	236529696
上海市	231951346
北京市	173843152
山东省	167500126
福建省	105293882
四川省	54479651
天津市	48977893
河南省	46942120
重庆市	45749671
辽宁省	44131930
安徽省	39562418
湖南省	34172709
广西壮族自治区	33897383
河北省	31020704
湖北省	30745779
江西省	28485021
陕西省	27220332
云南省	18031910
山西省	12802508
黑龙江省	11433880
新疆维吾尔自治区	9032541
吉林省	8637745
海南省	8476667
内蒙古自治区	7129066
贵州省	3813959
甘肃省	2864028
宁夏回族自治区	1289655
西藏自治区	325241
青海省	251389

表 13: 第一主成分得分