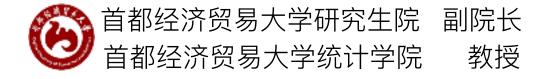
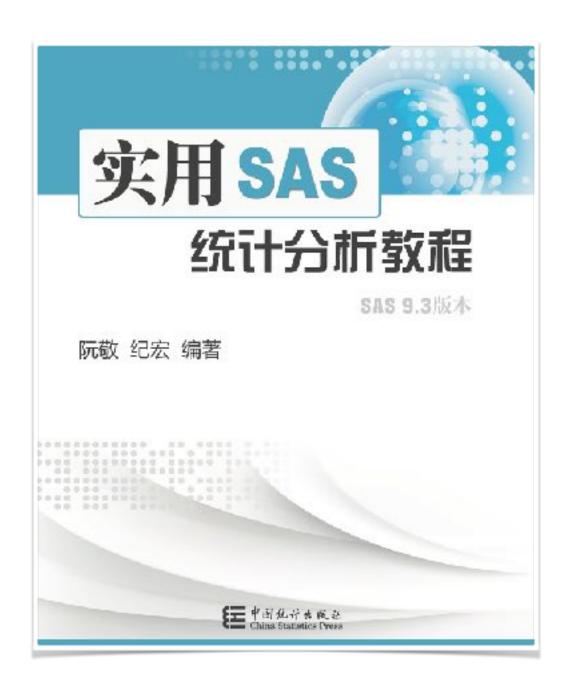
#### 阮敬 博士



© ruanjing@msn.com



#### 因子分析

- 客观世界是复杂多变的,在社会发展的过程中体现出多样性,人们的生活因此而丰富多彩。那么,人们如何简练的从若干个方面去归纳概括出事物发展的历程和特征呢,如何抓住主要矛盾,抓住矛盾的主要方面?即如何对事物发展过程中呈现出纷繁芜杂的数据进行简单明了的描述?这需要对数据进行精简和概括。
- 人们往往希望能够找出少数具有代表性的变量来对复杂事物进行描述, 这需要把反映该事物的很多变量或数据进行高度概括。本章所阐述的因子分析便是如何利用复杂多样的数据来综合描述客观事物特征的分析方法和过程。

## 数据降维

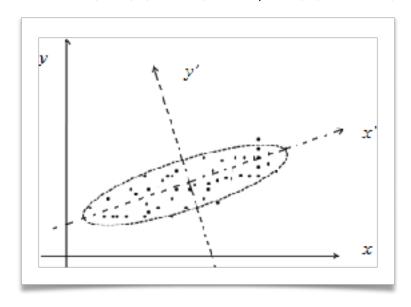
- 每个人都会遇到有很多变量的数据, 如反映全国或各省市经济、 社会发展状况的变量数据、反映一个国家总体发展状况的数据等。这些数据的共同特点是变量很多, 在如此多的变量之中, 有很多变量之间是相关的, 人们同时分析很多个变量是比较困难的, 这就会带来"维度灾难"的问题。因此, 人们希望能够找出这些变量的"代表", 来对更多的变量进行描述。
- 如在学校中进行奖学金的评定,需要考虑学生各门课程的学习成绩、与人相处的能力、尊重师长的程度、乐于助人的程度、担任学生干部努力的程度、参加社会实践活动的积极性等因素。假设有一个学生本学期考试的科目有 10 门课程,那么其参加奖学金评定按照上述的参考变量,便会有 15 个变量之多。在实际工作中,不会同时考虑这 15 个变量的数据来进行奖学金评定,通常的做法是把相互关联的变量进行综合,如把上述 15 个变量综合为学习成绩(含 10 门课程)变量、思想品德(含与人相处能力、尊重师长程度、乐于助人程度)变量、工作态度(含担任学生干部努力程度、参加实践积极性)变量等 3 个具有代表性的综合变量,然后依照这 3 个综合变量进行奖学金的评定,从而实现了化繁为简的目的。

## 数据降维的基本问题

- 把反映一个事物特征的多个变量用较少且具有代表性的变量来描述, 这个过程称之为数据降维。 不同的 变量往往是从不同的侧面或方面去描述事物特征的, 这些不同的方面称之为事物的维度,如从身高、体 重、血型 3 个方面反映一个人的特征,则具有三个维度。当反映事物方面太多的时候,过多的数据会对所 描述对象造成混乱,很难得到正确结论。因此,应当把相关的维度进行总结概括,尽量降低数据维度,简 要地对事物特征进行描述
- 为了能够简要而不遗漏的反映事物特征,数据降维过程中应当解决如下几个基本问题:
  - ◎能否把数据的多个变量用较少的综合变量来表示?
  - 较少的综合变量包含有多少原来的信息?
  - ●能否利用找到的综合变量来对事物进行较为全面的分析?
- 上述第 1 个问题具体是指在进行数据降维之前,应当考虑原始变量数据之间的关联性,即变量之间是否具有可提取综合变量所存在的必然联系; 而第 2 个问题主要考虑所提取出来的综合变量在多大程度上代表了原始数据的信息, 这是利用综合变量进行统计分析, 进而得到正确结论的理论基础; 第 3 个问题主要阐述了数据降维应当在统计分析过程中发挥的重要作用,并且在降维得到综合变量的基础之上进行进一步的统计分析活动。
- 解决好这些基本问题之后, 就可以用简化的数据对事物进行描述或判定, 从而得出统计分析的结论。

## 数据降维的基本原理

• 数据降维过程可以从最简单的二维数据降为一维数据开始。 先假设只有二维, 即只有两个变量,它们可由二维坐标轴上的横坐标(x)和纵坐标(y)来表示,因此每个观测值都有相应于这两个坐标轴的坐标值。 在正态分布的假定下, 这些数据在二维坐标轴上形成一个椭圆的分布形状,如图 14-1 所示。



• 众所周知,椭圆有一个长轴和一个短轴,且互相垂直。在短轴方向上,数据变化很少;而长轴方向,数据变化的范围较大。在极端的情况,短轴如果退化成一点,则只有在长轴的方向才能够解释这些点的变化了。因此,长轴就是要找的主要综合变量。至此,由二维到一维的降维过程就完成了。

#### 数据降维的基本原理

- 当坐标轴和椭圆的长短轴平行,那么代表长轴的变量就描述了数据的主要变化,而代表短轴的变量就描述了数据的次要变化。但是,坐标轴通常并不和椭圆的长短轴平行。因此,需要寻找椭圆的长短轴,即进行坐标平移或旋转变换,使得新变量和椭圆的长短轴平行。如果长轴变量代表了数据包含的大部分信息,就用变量在该轴上的变化代替原先的两个变量(舍去次要的另一个维度) ,降维就完成了。椭圆的长短轴相差得越大,降维效果就越好。
- 对于多维变量的降维情况和二维降维类似, 主要从高维椭球入手。首先把高维椭球的主轴找出来, 再用代表大多数数据信息的最长的几个轴作为新变量。与二维椭圆分布形状类似,高维椭球的主轴也是互相垂直的。这些互相正交的新变量是原先变量的线性组合, 可叫做主成分。正如二维椭圆有两个主轴, 三维椭球有 3 个主轴一样, 有几个变量, 就有几个主成分。
- 究竟要选择多少个主成分,是不是越少越好呢?这个问题有一定的选择标准,就是这些被选中主成分所代表的主轴的长度之和与主轴长度总和的比值,这个比值也称之为"阈值"。根据相关文献建议,所选的主轴总长度之和占所有主轴长度之和大约85%(即阈值为85%)即可。但在实际应用过程中,要依据研究目的、研究对象和所搜集的变量具体情况而定。
- 主轴越长, 表示变量在该主轴方向上的变动程度越大, 亦即方差越大。 所以一般情况下, 不去计算主轴的长度, 而是计算其主轴方向的方差, 根据所选取主轴的方差之和与所有主轴方向上方差之和的比值, 即方差贡献率的大小来判断应该取多少个主成分。

## 主成分分析—基本概念与原理

· 主成分分析 (Principal Component) 是数据降维的基本方法之一。从第 14.1 节的数据降维过程中,已经得知主成分提取的几何意义。主成分是由原始变量提取的综合变量,可以用如下式子来表示:

$$Y_{1} = \mu_{11}x_{1} + \mu_{12}x_{2} + \cdots + \mu_{1p}x_{p}$$

$$Y_{2} = \mu_{21}x_{1} + \mu_{22}x_{2} + \cdots + \mu_{2p}x_{p}$$

$$\cdots$$

$$Y_{p} = \mu_{p1}x_{1} + \mu_{p2}x_{2} + \cdots + \mu_{pp}x_{p}$$

- ・ 其中, Y 表示主成分, x 为原始变量;  $\mu_{ij}$ 为系数, 有约束条件:  $\mu_{k1}^2 + \mu_{k2}^2 + \cdots + \mu_{kp}^2 = 1$ ,  $\mu_{ij}$ 可由原始数据协方 差矩阵或相关系数矩阵确定。
- ・ 在提取出来的各个主成分当中, $Y_i$ 与 $Y_j$ 相互无关, 且第一个主成分 $Y_1$ 是 $x_1,x_2,...,x_p$ 的一切线性组合最大的; 第二个主成分 $Y_2$ 是 $x_1,x_2,...,x_p$ 的一切线性组合第二大的; 以此类推,第 n 个主成分 $Y_n$ 是 $x_1,x_2,...,x_p$ 的一切线性组合第 n 大的。

## 主成分分析—基本概念与原理

由原始数据的协方差阵或相关系数据阵,可计算出矩阵的特征值或特征根:

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$$
,

・ 其中:  $λ_1$ 对应 $Y_1$ 的方差, $λ_2$ 对应 $Y_2$ 的方差,…, $λ_p$ 对应 $Y_p$ 的方差,因此有:

阀值 = 
$$\frac{被选择的主成分长度}{主轴成分总合}$$
 =  $\frac{选择的特征根的和}{特征根总合}$  = 累积方差贡献率

- ・ λ对应的特征向量μ就是主成分分析线性模型中对应的系数,如: $\lambda_{11}$ 对应的特征向量为 $\mu_{11}$ , $\mu_{12}$ ,… $\mu_{1p}$ 为第 1 主成分的线性组合系数,即: $Y_1=\mu_{11}x_1+\mu_{12}x_2+\cdots+\mu_{1p}x_p$ 。
- 这些系数称为"主成分载荷",它表示主成分和相应原始变量的相关系数。相关系数绝对值越大,主成分对该变量的代表性也越大。根据上式计算出来的 Y 值称之为主成分得分。
- 在实际问题中,不同的变量往往有不同的量纲(即计量单位或测度规模),为了不同量纲数据之间的可比性,保证所提取主成分与原始变量意义上的一致性,在进行主成分分析之前,可按照如下 Z-Score 公式将变量标准化或无量纲化:

$$x_i^* = \frac{x_i - E(x_i)}{\sqrt{Var(x_i)}} (i = 1, 2, \dots, p)$$

• 其中, $E(x_i)$ ,表示原始变量 $x_i$ 的期望, $Var(x_i)$ 表示 $x_i$ 的方差。

- 在主成分分析的过程中,通常要先把各变量进行无量纲化(即标准化)。把变量进行标准化之后,可按照如下顺序进行主成分分析:
  - 选择协方差阵或相关阵计算特征根及对应的特征向量;
  - 计算方差贡献率,并根据方差贡献率阈值选取合适的主成分个数;
  - ●根据主成分载荷大小对选取的主成分进行命名;
  - ●根据主成分载荷计算各个主成分得分。
- SAS 系统主要使用 PRINCOMP 过程进行主成分分析。该过程自动会对所分析的变量进行标准化。因此,用户无需自行对原始变量进行标准化处理。
- PRINCOMP 过程的主要语法如下:

PROC PRINCOMP <选项>;

BY 变量列表;

FREQ 变量;

ID 变量列表;

PARTIAL 变量列表:

VAR 变量列表:

WEIGHT 变量:

- PRINCOMP 过程中的语句,最常用的是 VAR 语句,主要用于指定进行主成分分析的原始变量;而 FREQ、WEIGHT 语句与前面讲述的用法一致;PARTIAL 语句主要用于根据偏相关(或协方差)系数矩阵进行特征根及其对应特征向量的计算。
- PRINCOMP 的过程选项主要有:
  - DATA=: 指定分析数据集;
  - OUT=: 指定输出数据集;
  - OUTSTAT=: 指定含有各种统计量的输出数据集;
  - NOINT: 忽略模型中的截距项;
  - PREFIX=: 为提取出来的主成分加上指定的前缀名称;
  - PARPREFIX=: 为剩余变量加上指定的前缀名称;
  - SINGULAR=: 指定奇异标准;
  - STD: 标准化主成分得分;
  - VARDEF=: 指定计算方差或标准差时所采用的分母;
  - PLOTS=: 指定该过程输出的图形;
  - NOPRINT: 压缩所有的过程输出结果, 即不显示分析结果;
  - COV: 指定系统利用协方差矩阵计算特征根及特征向量; 缺省则表示用相关系数矩阵计算;
  - N=: 指定系统提取主成分的数目; 缺省则表示有多少个原始变量就提取多少个主成分;

- 例14-1: 为评价全国各省/直辖市/ 自治区的综合发展水平,现收集了 全国 24 个地区的人均GDP、人均 可支配收入、人均消费支出等数据 进行综合考察,如图 14-2 所示。 试利用主成分分析方法对各地区综 合发展状况进行评价。
- 在本例中,共有7个变量可供用来综合评价各地区的发展状况。但是,如果从7个方面来考察综合发展状况,不免显得过于复杂。因此,可以把7个变量进行降维,从中提取出若干个综合变量,利用综合变量所反应的主成分来对地区发展状况进行评价。

			Da	taset of SAS	USEK.LIVE			
Obs	District 地区	GDP 人均 <b>四内</b> 生产总值	Income 人均可 支配收入	Consumption 人均消 费支出	Employment 城镇 沈业军	Education 人均額 育經典	Health 医疗保 健腐用	Life 頂期寿命
- 1	北京	45444	17652 95	13244-2	0.3937	584 43	1295 76	76
2	山西	12495	8913.91	6312.63	0.2554	548.83	538.7	71.6
3	内蒙土	16331	9136.79	6928.6	0.2158	504.77	533,36	69.8
4	吉双	13348	8690 62	6794 71	0 1836	502 08	675 77	73
5	黑龙山	14434	8272.51	6178.01	0.2418	479.85	<b>6</b> 13. <b>15</b>	72.3
					····			
21	甘肃	7477	8086.82	6529.2	0.2496	505.9	492.23	67.47
22	青海	<b>100</b> 45	8057.85	6245.26	0.2002	360.52	554.11	66.00
23	宁 夢	10239	8093.64	6404.31	0.2367	388.3	535.92	/0.1
74	新疆	13108	7990 15	6207.52	0.328	456.25	499 16	67.4

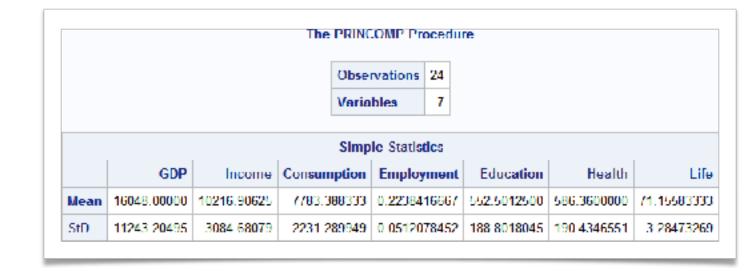
本例使用 PRINCOMP 过程进行测算的程序如下:

proc princomp data=sasuser.live out=prin\_score; /\*指定把计算结果存储于 PRIN SCORE 数据集\*/

var gdp income consumption employment education health life; /\*指定用于分析的原始变量\*/

run;

- 本段程序默认采用原始数据的相关系数矩阵进行特征根测算。程序运行后,首先可得到关于每个原始变量的均值和标准差统计量,如图 14-3 所示。
- 利用图 14-3 的输出结果,可以对原始变量进行 无量纲化或标准化(SAS 系统中会自动进行标 准化,该步骤可以省略)。



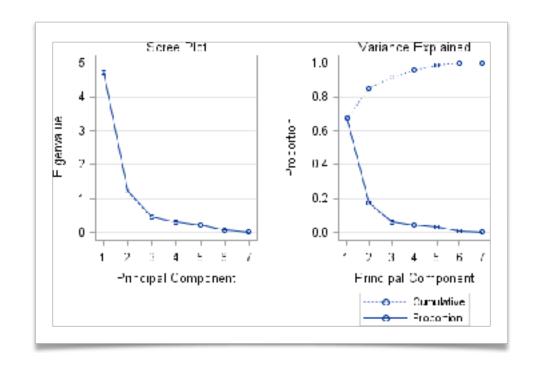
- 在主成分的分析结果中,还会输出用于 计算特征根的矩阵(协方差矩阵或相关 系数矩阵),本例输出原始变量之间的 相关系数矩阵,如图 14-4 所示。
- 相关系数矩阵可以列示出变量两两之间的简单相关系数,用于观察变量之间的联系,在一定程度上可考察数据降维是否恰当;此外,在手工计算过程中可用图 14-4 所示的相关系数矩阵计算特征根及其对应的特征向量。

			Corre	lation Matrix				
		GDP	Income	Consumption	Employment	Education	Health	Life
GDP	人物區 为生产总值	1.0000	0.9382	0.8930	0.3345	0.6867	0.7498	0.742
Income	人均可支配收入	0.9382	1.0000	0.9775	0.2111	0.7557	0.7432	0.690
Consumption	人均消费性支出	0 8930	0.9775	1 0000	0.2141	0.7628	0.7331	0.595
Employment	域領領业率	0.3345	0.2111	0.2141	1.0000	.1231	0.4341	0.014
Education	人均勒言经 <b>携</b>	0.6867	0.7557	0.7628	1231	1.0000	0.4653	0.638
llealth	医疗保健费用	0.7498	0./432	0.7331	0.4341	0.4653	1.0000	0.5259
l ife	预期寿命	0.7421	0 6905	0.5957	0 0147	0.6387	0.5259	1 0000

- 在 PRINCOMP 过程的输出结果中,会自动按照给定的矩阵数据进行 特征根计算,结算结果如图 14-5 所示。
- 图 14-5 所示结果一共输出 4 列结果,即特征根(Eigenvalue)、相邻两个特征根之间的差分(Difference)、特征根值占所有特征根值之和的比例(Proportion)和特征根累计比例(Cumulative)。依据14.2.1 节的理论分析,可知有多少个原始变量就可以提取出多少个主成分,因此图 14-5 所示的结果中可以输出 7 个特征根。
- 可以依据特征根的贡献率来决定应当取多少个主成分比较合适。 通常情况下,可参照特征根累计贡献率阈值 85%的标准来进行判定。在图 14-5 中,第 1 个特征根的值占所有特征根值之和的比例亦称之为贡献率)为 4.725/(4.725+1.234+0.449+0.306+0.214+0.061+0.011)=0.6751;第 2 个特征根的贡献率为 0.1763;由图 14-5 所示结果的最后一列可知,上述两个特征根的累积贡献率已经达到 0.8514, 说明针对本例的原始数据提取出两个主成分即可代表原始数据的大部分信息,而且前两个特征根的贡献远远大于其余特征根的贡献。因此,根据累积特征根贡献率,本例可考虑提取出两个主成分进行分析。

	Ligenva	lues of the (	orrelation N	latrix	
	Figenvalue	Difference	Proportion	Cumulative	
1	4.72549923	3.49115815	0.6751	0.6751	
2	1.23434108	0.78567923	0.1763	0.8514	
3	0.44866185	0.14254777	0.0641	0 9198	
4	0.30611408	0.09235871	0.0437	0.9592	
5	0.21375537	0.15318106	0.0305	0.9898	
6	0.06057431	0.04952023	0.0087	0.9984	
7	0.01105408		0.0016	1 0000	

- 上述分析过程还可以用如图 14-6 所示的碎石图 和带有方差贡献率的碎石图来描述。
- 碎石图可以形象的展现各特征根的大小及其贡献, 图中横轴代表所计算出来的主成分特征根,可以根据图中线条的倾斜程度来进行主成分个数的选择(越平缓表示贡献越小)。



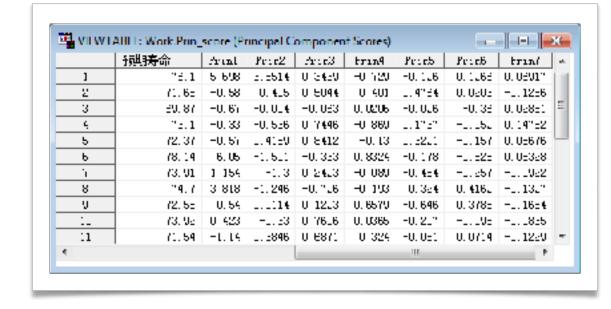
• 除考查特征根外,主成分分析中还应当考察如何提取主成分并计算主成分得分的问题,即计算特征根对应的特征向量,将其作为原始变量线性组合的系数。在PRINCOMP 过程输出结果中同样可输出特征根对应的特征向量结果,如图 14-7 所示。

			Ligenve	ectors				
		Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7
GDP	人物區 医生产总值	0.441618	0.073883	0.083499	0.153700	.325047	.796130	0.171583
Income	人均可支配收入	0.447192	029164	193227	0.036980	358861	0.213764	765499
Consumption	人均消费性支出	0.435590	- 016302	- 394963	0.005468	- 276029	0.450039	0.611568
Employment	域領領业率	0.122961	0.827743	0.098366	0.463616	0.231263	0.143879	.030374
Education	人均勒言经典	0.365034	397744	255464	0.336966	0.718524	101607	056051
llealth	医疗保健费用	0.374018	0.307351	001/32	801352	0.345642	062374	010942
l ife	預期寿命	0.356365	- 235799	0.851325	0.055691	0.011964	0.288153	0.079822

- 根据图 14-7 所示各主成分对应的特征向量即系数,可以计算出各主成分的得分。如第一个主成分得分为:
   Y<sub>1</sub> = 0.442GDP+0.447Income+0.436Consumption+0.123Employment+0.356Education+0.347Health+0.356Life
- 第二个主成分得分为:

 $Y_2 = 0.074 GDP - 0.029 Income - 0.016 Consumption + 0.828 Employment - 0.398 Education + 0.307 Health - 0.236 Life$ 

- · 可以根据主成分计算公式中的系数, 即主成分载荷绝对值的大小来判定该主成分所主要代表原始变量的含义。如第一主成分中,除变量 Enployment 的系数之外,其余变量的系数均比较大, 说明这些变量在第一主成分中发挥的影响作用比较大。 因此, 对于第一主成分Y<sub>1</sub>,可归纳除变量 Employment 之外其余变量所共同表示的含义,可把第一主成分命名为"经济生活"成分。
- · 同理,对于第二主成分Y<sub>2</sub>,变量 Employment 的系数显著 大于其余变量,说明该变量在第二主成分中发挥的影响 作用比较大。 因此,可以根据第二主成分中发挥作用 最大的变量所代表的含义,可以把第二主成分命名为 "就业成分"。
- · 把原始变量进行标准化之后, 代入上述公式, 便可计算出每个样本在对应主成分的得分,本例程序在 PRINCOMP的过程选项中使用关键字OUT=指定了将得分结果存储在WORK数据库下的 PRIN\_SCORE 数据集中,具体结果如图 14-8。



- 最后,可以把所有样本按照所提取出来的主成分得分进行排序,来考察各个样本的综合情况。此外,也可以利用特征根贡献率及其对应的主成分得分,计算出综合得分(在 SAS系统中,综合得分要通过用户手工计算才能得到),对所有样本进行综合评价,一般可采用的综合得分计算公式如下:
- 综合得分 = ∑<sub>i=1</sub>(第i主成分得分 × 第i特征根贡献率)
- 其中 n 是根据阈值确定的主成分个数。
- 本例中,根据 85%贡献率阈值,选择两个主成分进行分析。因此,每个样本的综合得分可按照如下公式进行计算:
- · 综合得分=第一主成分得分×第二主成分得分×0.1763

- 本例各样本的第一、二主成分及综合得分如表 14-1 所示。
- 由表 14-1 所列示的结果,可以分别对第一、第二主成分进行分析,也可以对各地区的发展依据综合得分进行综合分析。
- 北京、上海等大城市及沿海地区等地,由于地理位置优越及改革开放政策的优惠措施,人才普遍比较集中,其经济、教育等方面的发展比较好,人民生活水平也随之较高,受教育的程度也较高,因此体现为这些地区的第一主成分(即经济生活成分)排名靠前;而对于二主成分(即就业成分)而言,反而是经济发展相对落后的地区排名靠前,究其原因可能是由于经济发展滞后,人才竞争不激烈、流动性不大等原因造成的;从社会经济生活发展的总体情况看,北京、上海、江浙一带总体发展状况相对较好,其综合得分也名列前茅。

		表 14-1	主成分得分及綜合	1得分		
地区	第一主成分		第二主成	159	综合	
	得分	排名	得分	排名	得分	排名
北京	5.697979	2	3.551375	1	4.472813	1
di 🙉	-0.58103	12	0.404957	7	-0.32086	10
内蒙古	-0.66753	13	-0.00435	13	-0.45142	12
古 林	-0.32792	8	-0.53561	17	-0.3158	9
黑龙江	-0.57168	11	0.418921	6	-0.31209	8

江苏	1.154256	4	-1.30043	22	0.549973	4
新江	3.817791	3	-1.24587	21	2.357743	3
福建	0.540295	5	0.011427	12	0.366768	5
山东	0.42284	6	-0.3298	14	0.227315	6
24 M	-1.13698	16	0.384593	8	-0.69977	16
湖北	-0.87939	14	-4.3325	15	-0.6523	15
测图	-0.46306	9	-0.84158	19	-0.46098	13
广西	-0.93805	15	-0.8768	20	-0.78786	18
海南	-1.59391	22	-0.45978	16	-1.15711	23
重 庆	0.4128	7	-1.3373	23	0.042915	7
四川	-1.27316	18	-0.60026	18	-0.96533	20
云 南	-1.57886	21	0.373034	9	-1.00012	22
西麻	-1.63371	23	0.663251	3	-0.98599	21
陕 西	-0.4932	10	0.063999	- 11	-0.32168	- 11
甘 肃	-1.5033	20	0.600191	4	-0.90907	19
背海	-1.89646	24	0.330405	10	-1.22205	24
宁夏	-1.29782	19	0.535104	5	-0.78182	17
新福	-1.25958	17	2.027835	2	-0.49283	14

以上内容仅为本文档的试下载部分,为可阅读页数的一半内容。如要下载或阅读全文,请访问: <a href="https://d.book118.com/93712014106">https://d.book118.com/93712014106</a>
0006142