主成分分析法

主成分分析法

概念：把原来多个变量划为少数几个综合指标的一种统计分析方法，是一种降维处理技术.

一个研究对象，往往是多要素的复杂系统。变量太多无疑会增加分析问题的难度和复杂性，利用原变量之间的相关关系，用较少的新变量代替原来较多的变量，并使这些少数变量尽可能多的保留原来较多的变量所反应的信息，这样问题就简单化了。

一、基本原理

假设有n个对象，每一个对象都有x_1，x_2，…，x_p个要素构成，它们所对应的要素数据用下表给出：

研究对象	要素					
	x_1	x_2	…	x_j	…	x_p
1	x_{11}	x_{12}	…	x_{1j}	…	x_{1p}
2	x_{21}	x_{22}	…	x_{2j}	…	x_{2p}
…	…	…	…	…	…	…
i	x_{i1}	x_{i2}	…	x_{ij}	…	x_{ip}
…	…	…	…	…	…	…
n	x_{n1}	x_{n2}	…	x_{nj}	…	x_{np}

原变量为x_1，x_2，…，x_p，降维处理后，设它们的综合指标，即新变量为z_1，z_2，z_3，…，$z_m(m\leq p)$，则

$$\begin{cases} z_1 = l_{11}x_1 + l_{12}x_2 + \ldots + l_{1p}x_p \\ z_2 = l_{21}x_1 + l_{22}x_2 + \ldots + l_{2p}x_p \\ \cdots\cdots \quad \cdots\cdots \quad \cdots\cdots \quad \cdots\cdots \\ z_m = l_{m1}x_1 + l_{m2}x_2 + \ldots + l_{mp}x_p \end{cases}$$

系数l_{ij}由以下原则确定

1、z_i与z_j($i \neq j$；i，j=1，2，…，m)相互无关

2、z_1是x_1，x_2，…，x_p的一切线性组合中方差最大者；z_2是与z_1不相关的x_1，x_2，…，x_p的所有线性组合中方差最大者；……；z_m是与z_1，z_2，z_3，…，z_{m-1}都不相关的x_1，x_2，…，x_p的所有线性组合中方差最大者。

z_1称为原变量x_1，x_2，…，x_p的第一主成分

Z_2称为原变量x_1，x_2，…，x_p的第二主成分

……

z_m称为原变量x_1，x_2，…，x_p的第m主成分

找主成分z_i就是要确定系数l_{ij}。从数学上知道，它们分别是x_1，x_2，…，x_p的相关系数矩阵的m个较大的特征值所对应的特征向量。

二、 主成分分析的计算步骤

1、计算相关系数

相关系数计算公式

$$r_{ij} = \frac{\sum_{k=1}^{n} (x_{ki} - \overline{x}_i)(x_{kj} - \overline{x}_j)}{\sqrt{\sum_{k=1}^{n} (x_{ki} - \overline{x}_i)^2 \sum_{k=1}^{n} (x_{kj} - \overline{x}_j)^2}}$$

据公式得这p个变量之间的相关系数矩阵为

$$R = \begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1p} \\ r_{21} & r_{22} & \cdots & r_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ r_{p1} & r_{p2} & \cdots & r_{pp} \end{pmatrix}$$

2、计算特征值和特征向量

解特征方程

$$|\lambda E-R|=0$$

求出特征值

$$\lambda_i(i=1，2，\ldots，p)$$

将这P个特征值按大小顺序排列，即

$$\lambda_1\geq\lambda_2\geq\ldots\geq\lambda_p\geq0$$

然后按公式

$$|\lambda_i E-R|e_i=0$$

分别求出对应于λ_i的特征向量$e_i(i=1，2，\ldots，p)$

3、计算主成分贡献率及累计贡献率

主成分z_i的贡献率为

$$Q_i = \frac{\lambda_i}{\sum_{k=1}^{p} \lambda_k} \quad (i = 1, 2, \ldots, p)$$

前i个主成分的累计贡献率为

$$Q = \frac{\sum_{k=1}^{i} \lambda_k}{\sum_{k=1}^{p} \lambda_k} \quad (i = 1, 2, \ldots, p)$$

当前 i个主成分累计贡献率达到85%——95%，就取前i个主成分作为新变量。

4、计算主成分载荷

计算公式为

$$l_{ij} = \sqrt{\lambda_i} e_{ij} \left(i, \ \ j = 1,2,..., \ \ p \right)$$

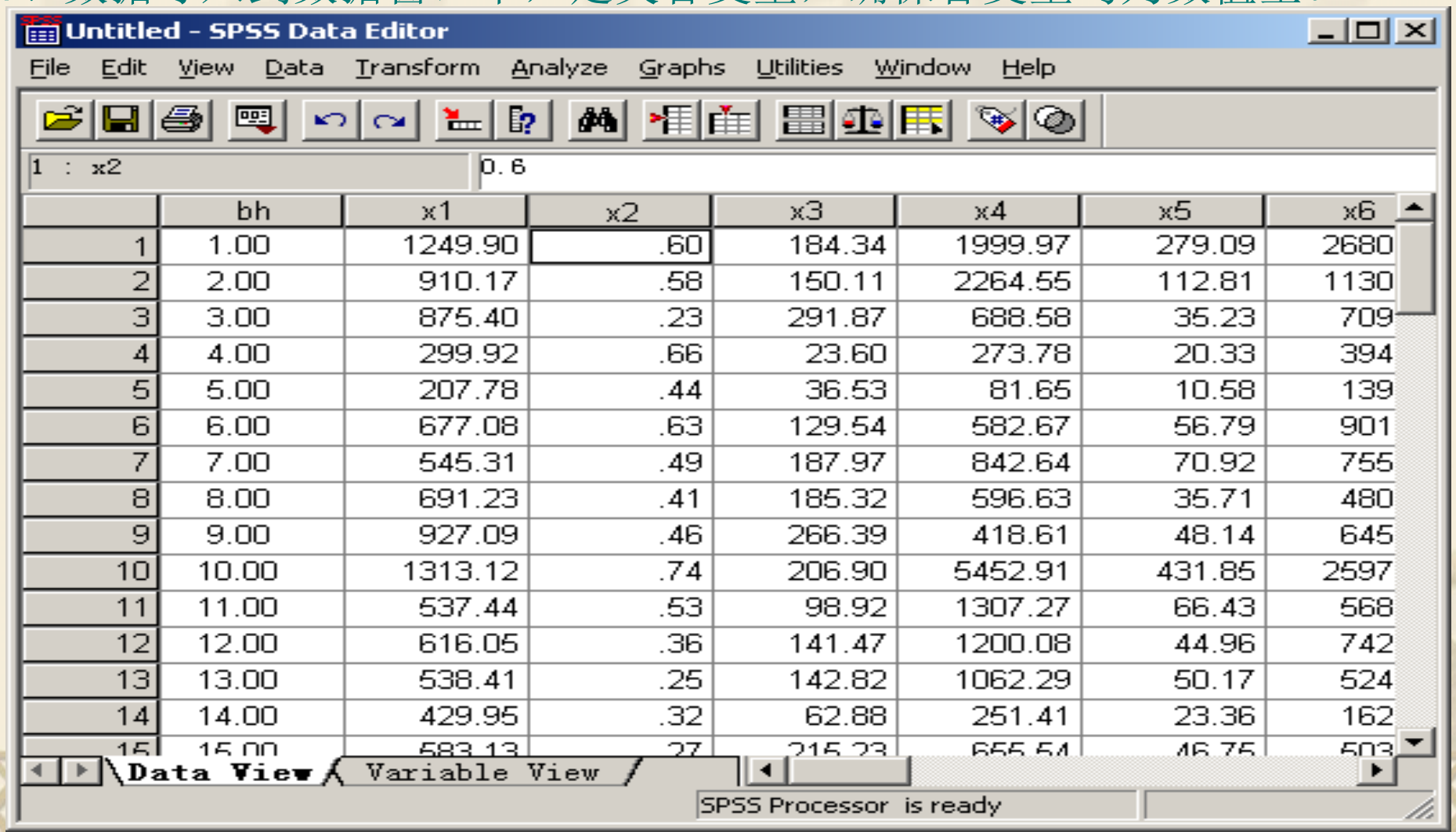
得前i个主成分在原变量上的载荷

原变量x_i	主成分			
	Z_1	Z_2	…	Z_i
x_1	l_{11}	l_{21}	…	l_{i1}
x_2	l_{12}	l_{22}	…	l_{i2}
…	…	…	…	…
x_p	l_{1p}	l_{2p}	…	l_{ip}

三、主成分分析方法的SPSS实现

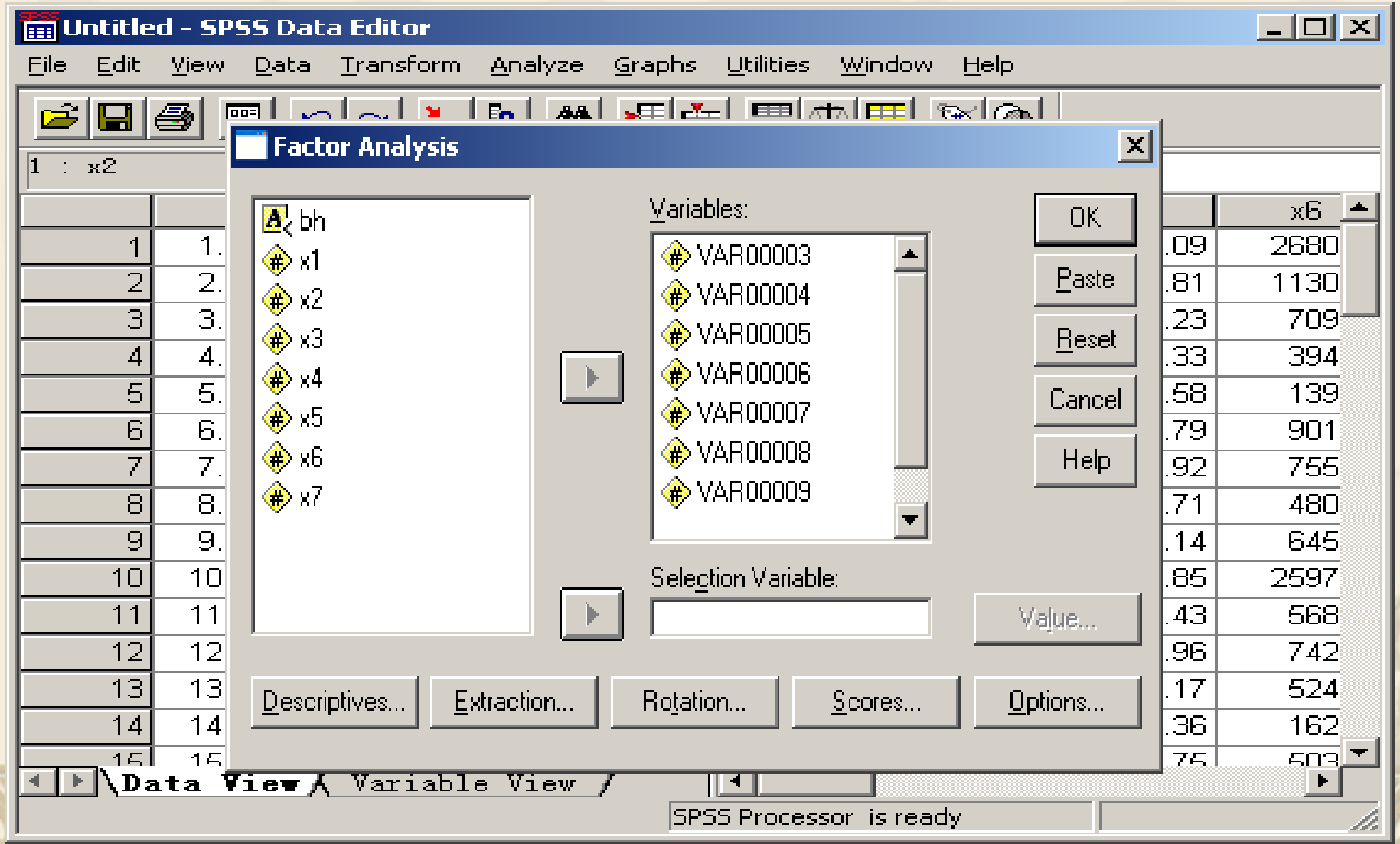
对45个城市7项经济指标进行主成分分析

1、数据导入到数据窗口中，定义各变量，确保各变量均为数值型。

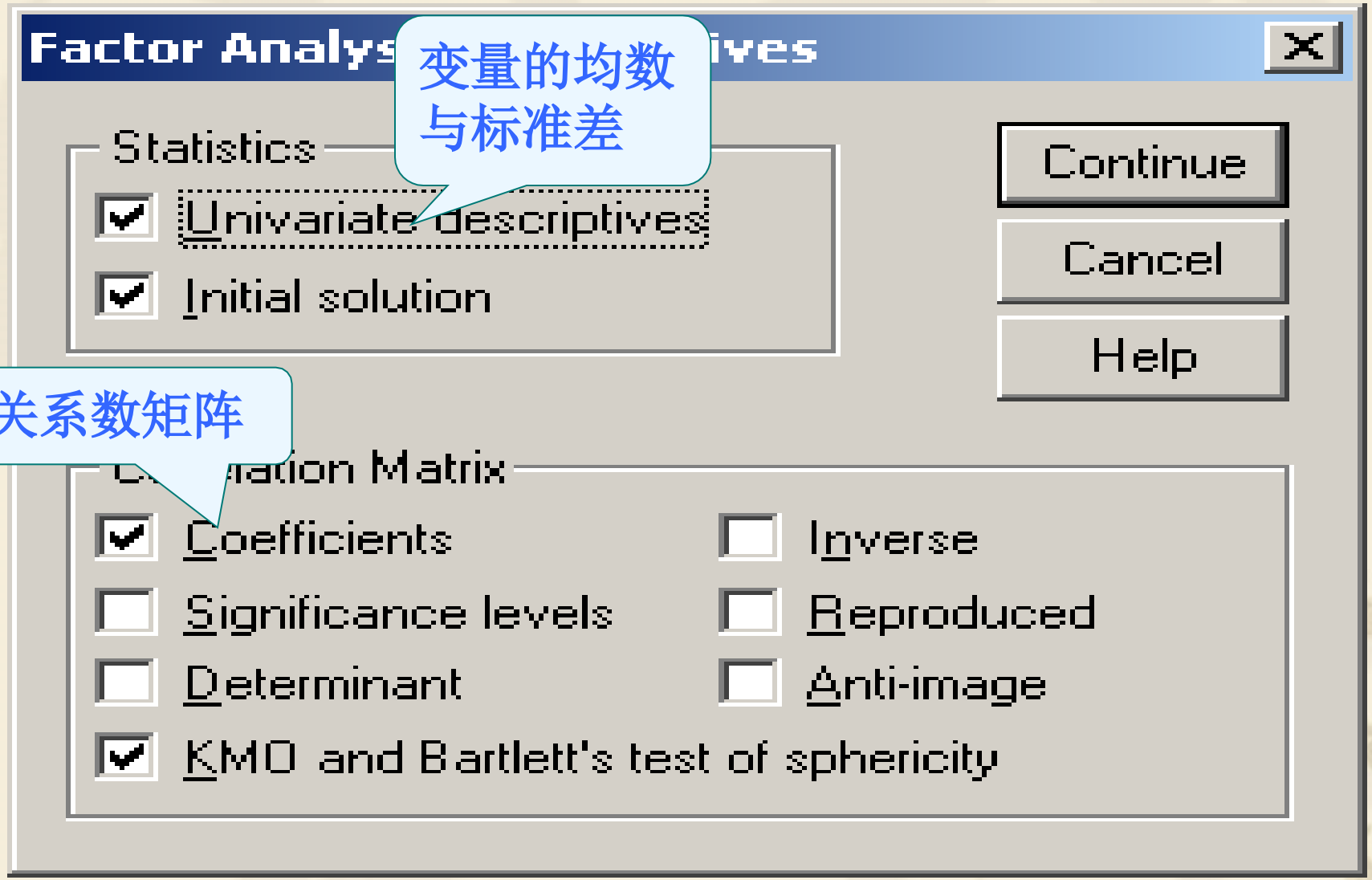


	bh	x1	x2	x3	x4	x5	x6
1	1.00	1249.90	.60	184.34	1999.97	279.09	2680
2	2.00	910.17	.58	150.11	2264.55	112.81	1130
3	3.00	875.40	.23	291.87	688.58	35.23	709
4	4.00	299.92	.66	23.60	273.78	20.33	394
5	5.00	207.78	.44	36.53	81.65	10.58	139
6	6.00	677.08	.63	129.54	582.67	56.79	901
7	7.00	545.31	.49	187.97	842.64	70.92	755
8	8.00	691.23	.41	185.32	596.63	35.71	480
9	9.00	927.09	.46	266.39	418.61	48.14	645
10	10.00	1313.12	.74	206.90	5452.91	431.85	2597
11	11.00	537.44	.53	98.92	1307.27	66.43	568
12	12.00	616.05	.36	141.47	1200.08	44.96	742
13	13.00	538.41	.25	142.82	1062.29	50.17	524
14	14.00	429.95	.32	62.88	251.41	23.36	162
15	15.00	583.13	.27	215.23	655.54	46.75	503

2、激活Analysis 菜单选Data Reduction 的Factor...命令项，弹出Factor Analysis 对话框。在对话框左侧的变量列表中选变量X1 至X7，点击钮使之进入Variables 框。



Descriptives...钮



Extraction...钮：因子提取方法