第 1 章 主成分分析与因子分析

为什么要进行数据降维？直观地好处是维度降低了，便于计算和可视化，其深层次的意义在于有效信息的提取综合及无用信息的摈弃，并且数据降维保留了原始数据的信息，我们就可以用降维的数据进行机器学习模型的训练和预测，将有效提高训练和预测的时间与效率.

降维方法分为线性和非线性降维，非线性降维又分为基于核函数和基于特征值的方法（流形学习），代表算法有

线性降维方法：PCA、ICA、LDA、LFA

基于核的非线性降维方法：KPCA、KFDA

流形学习：ISOMAP、LLE、LE、LPP

§1.1 主成分分析

1.1.1 主成分分析的基本思想

主成分分析（Principal Component Analysis　PCA）是一种常用的数据分析方法. 它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. 主成分与原始变量之间的关系：

（1）主成分是原始变量的线性组合；

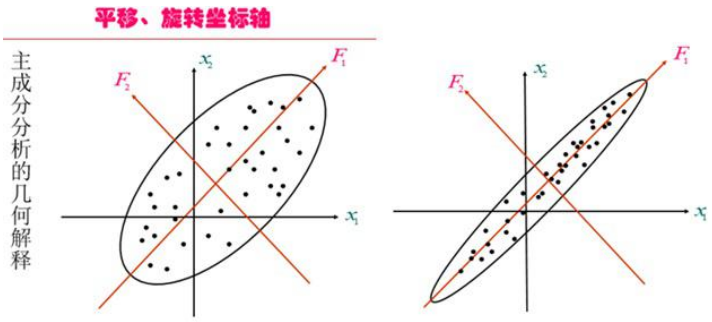
（2）主成分的数量相对于原始数量更少；

（3）主成分保留了原始变量的大部分信息；

（4）主成分之间相互独立.

1.1.2 主成分分析的几何意义

通过旋转变换，将分布在 x_1, x_2 坐标轴上的原始数据，转换到 F_1, F_2 坐标轴表示的坐标系上，使得数据在 F_1 轴上离散程度最大，此时可以忽略 F_2 轴，仅通过 F_1 轴就可以表示数据的大部分信息，从而达到降维的目的，如图所示.



1.1.3 主成分分析的基本原理

设有 n 个样本，每个样本有 p 个指标（变量）：$X_1, X_2, \cdots, X_p$，得到原始数据矩阵：

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} = (X_1, X_2, \cdots, X_p)$$

其中，$X_i = (x_{1i}, x_{2i}, \cdots, x_{ni})^T, i = 1, 2, \cdots, p$.

假定 X 的期望和协方差矩阵均存在并已知，记 $E(X) = \boldsymbol{\mu}, \mathrm{Var}(X) = \boldsymbol{\Sigma}$，考虑如下线性变换：

$$\begin{cases} F_1 = a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p \\ F_2 = a_{21}X_1 + a_{22}X_2 + \cdots + a_{2p}X_p \\ \cdots \\ F_p = a_{p1}X_1 + a_{p2}X_2 + \cdots + a_{pp}X_p \end{cases}$$

设 F_i 表示第 i 个主成分，$i = 1, 2, \cdots, p$.

不同的线性变换，得到的 F_i 统计特性不同，为得到较好的效果，我们希望主成分之间相互独立，同时方差尽可能的大. 规定：

（1）F_i 与 F_j 互不相关，即 $\mathrm{Cov}(F_i, F_j) = 0$；

（2）让系数 $a_i = (a_{i1}, a_{i2}, \ldots, a_{ip})^T$ 为单位向量，即 $a_{i1}^2 + a_{i2}^2 + \cdots + a_{ip}^2 = 1, i = 1, 2, \cdots, p$；

且 a_1 使得 $\mathrm{Var}(F_1)$ 的值达到最大；a_2 不仅垂直于 a_1，而且使 $\mathrm{Var}(F_2)$ 的值达到最大；a_3 同时垂直于 a_1 和 a_2，并使 $\mathrm{Var}(F_3)$ 的值达到最大；以此类推可得全部 p 个主成分，这项工作用手做是很繁琐的，但借助于计算机很容易完成.

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 $a_i = (a_{i1}, a_{i2}, \ldots, a_{ip})^T$ 就是特征值 λ_i 所对应的单位特征向量. 方差的贡献率为

$$\alpha_i = \frac{\lambda_i}{\sum_{i=1}^{p} \lambda_i},$$

α_i 越大，说明相应的主成分反映综合信息的能力越强.

1.1.4 主成分个数的选取

原则上如果有 p 个变量，则最多可以提取出 p 个主成分，但如果将它们全部提取出来就失去了该方法简化数据的实际意义. 主成分的个数选多少个比较合适？主要有 3 个衡量标准：

（1）保留的主成分使得方差累积贡献率达到 85%以上；

前 k 个主成分的累计贡献率指按照方差贡献率从大到小排列，前 k 个主成分累计提取了多少的原始信息. 一般来说，如果该指标达到 85%，表明这些主成分包含了全部测量指标所具有的主要信息，这样既减少了变量的个数，又便于对实际问题的分析和研究.

（2）保留的主成分的方差（特征值）大于 1；

注：特征值可以被看成是主成分影响力度的指标，表示引入该主成分后可以解释平均多少个原始变量的信息. 如果特征值小于 1，说明该主成分的解释力度还不如直接引入一个原变量的平均解释力度大. 因此一般可以用特征根大于 1 作为纳入标准.

（3）Cattell 碎石检验绘制了关于各主成分及其特征值的图形，我们只需要保留图形中变化最大之处以上的主成分即可.

1.1.5 主成分分析的优缺点

1. 优点

（1）不要求数据呈正态分布，主成分就是按数据离散程度最大的方向对基组进行旋转，这特性扩展了其应用范围，比如，用于人脸识别；

（2）通过对原始变量进行综合与简化，可以客观地确定各个指标的权重，避免主观判断的随意性；

2. 缺点

（1）主成分分析适用于变量间有较强相关性的数据，若原始数据相关性弱，则起不到很好的降维作用；

（2）降维后，存在少量信息丢失，不可能包含 100%原始数据；

（3）原始数据经过标准化处理之后，含义会发生变化，且主成分的解释含义较原

始数据比较模糊；

（4）假设标准化后的原始变量间存在多重共线性，即原始变量之间存在不可忽视的信息重叠，主成分分析不能有效剔除信息重叠；

1.1.6 主成分分析的操作步骤

（1）获取原始数据，统一量纲，将数据进行标准化处理；

（2）计算相关系数矩阵，求特征值和特征向量；

（3）确定主成分个数；

（4）提取主成分；

（5）对主成分做经济解释，主成分的经济意义由各线性组合中权重较大的几个指标来确定.

注：

相关系数矩阵：相当于消除量纲的表示变量间相关性的一个矩阵；

协方差矩阵：它是没有消除量纲的表示变量间相关性的矩阵.

1.1.7 方法用途

主成分分析往往会在大型研究中成为一个中间环节，用于解决数据信息浓缩等问题，这就可能产生各种各样的组合方法. 这里仅举最为典型的两种应用情况.

（1）主成分评价：在进行多指标综合评价时，由于要求评价结果客观、全面，就需要从各个方面用多个指标进行测量，但这样就使得观测指标间存在信息重叠，同时还会存在量纲、累加时如何确定权重系数等问题. 为此就可以使用主成分分析方法进行信息的浓缩，并解决权重的确定等问题.

（2）主成分回归：在线性回归模型中，常用最小二乘法求回归系数的估计. 但是当存在多重共线性时，最小二乘法的估计结果并不很理想，这时可考虑用主成分回归方法进行分析. 所谓主成分回归是用原自变量的主成分代替原自变量做回归分析，这些主成分既保留了原指标的绝大部分信息，又互不相关，故用主成分替代原指标后，再用最小二乘法建立的回归方程其同归系数就能避开共线性问题. 但主成分估计显然不是无偏估计.

1.1.8 案例：各省经济发展情况综合评价

现希望根据全国 30 个省、市、自治区（未包括香港、澳门、台湾地区）经济发展基本情况的 8 项指标对其进行分析和排序. 具体指标有：GDP、居民消费水平、固定资产投资、职工平均工资、货物周转量、居民消费价格指数、商品零售价格指数、工业总产值，数据文件见 development. xlsx.

这是一个综合评价问题，但各指标间存在数值关联，且各指标重要性也存在差异，

因此可以考虑首先用主成分分析法进行信息综合.

打开文件后在 SPSS 中的操作步骤如下，相应的软件界面如图 1 所示.

（1）选择“分析”→“降维”→“因子分析”菜单项.

（2）将 x1~x8 选入“变量”框.

（3）在“描述”对话框中，选中“相关系数”选项组中的“系数”复选框.

（4）单击“确定”按钮.

图 1 因子分析主对话框和描述统计对话框

SPSS 在进行分析时，首先会自动对原始变量进行标准化，因此在以后的输出结果中通常情况下都是指标准化后的变量. 在结果输出中会涉及一些因子分析中的内容，因此这里仅给出与主成分分析有关的部分. 图 2 为 8 个原始变量之间的相关系数矩阵，可见许多变量之间直接的相关性比较强，的确存在信息上的重叠. 该结果进一步确认了信息浓缩的必要性. 图 3 给出的是各成分的方差贡献率和累计贡献率，可见只有前 3 个主成分的特征根大于 1，因此 SPSS 默认只提取了前 3 个主成分. 第一主成分的方差所占所有主成分方差的 46.92%，接近一半，前 3 个主成分的累计方差贡献率达到 89.55%，因此选前三个主成分已足够描述经济发展的水平.

相关性矩阵

		GDP	居民消费水平	固定资产投资	职工平均工资	货物周转量	居民消费价格指数	商品价格指数	工业总产值
相关性	GDP	1.000	.267	.951	.187	.617	-.273	-.264	.874
	居民消费水平	.267	1.000	.426	.716	-.151	-.235	-.593	.363
	固定资产投资	.951	.426	1.000	.396	.431	-.280	-.359	.792
	职工平均工资	.187	.716	.396	1.000	-.357	-.145	-.543	.099
	货物周转量	.617	-.151	.431	-.357	1.000	-.253	.022	.659
	居民消费价格指数	-.273	-.235	-.280	-.145	-.253	1.000	.763	-.125
	商品价格指数	-.264	-.593	-.359	-.543	.022	.763	1.000	-.192
	工业总产值	.874	.363	.792	.099	.659	-.125	-.192	1.000

图 2 相关系数矩阵

总方差解释

成分	初始特征值 总计	初始特征值 方差百分比	初始特征值 累积 %	提取载荷平方和 总计	提取载荷平方和 方差百分比	提取载荷平方和 累积 %
1	3.754	46.924	46.924	3.754	46.924	46.924
2	2.203	27.532	74.456	2.203	27.532	74.456
3	1.208	15.096	89.551	1.208	15.096	89.551
4	.403	5.042	94.593			
5	.214	2.673	97.266			
6	.138	1.722	98.988			
7	.066	.829	99.817			
8	.015	.183	100.000			

提取方法：主成分分析法。

图 3 总方差解释

随后图 4 给出的是主成分系数矩阵，可以说明各主成分在各变量上的载荷，从而得出各主成分的表达式，注意在表达式中各变量已经不是原始变量，而是标准化变量：

$$F_1 = 0.884Zx_1 + 0.606Zx_2 + 0.911Zx_3 + 0.465Zx_4 + 0.486Zx_5 - 0.510Zx_6 - 0.621Zx_7 + 0.822Zx_8$$

$$F_2 = 0.385Zx_1 - 0.596Zx_2 + 0.163Zx_3 - 0.725Zx_4 + 0.737Zx_5 + 0.257Zx_6 + 0.596Zx_7 + 0.429Zx_8$$

$$F_3 = 0.120Zx_1 + 0.277Zx_2 + 0.213Zx_3 + 0.362Zx_4 - 0.279Zx_5 + 0.79\,4Zx_6 + 0.433Zx_7 + 0.210Zx_8$$

成分矩阵[a]

	成分		
	1	2	3
GDP	.884	.385	.120
居民消费水平	.606	-.596	.277
固定资产投资	.911	.163	.213
职工平均工资	.465	-.725	.362
货物周转量	.486	.737	-.279
居民消费价格指数	-.510	.257	.794
商品价格指数	-.621	.596	.433
工业总产值	.822	.429	.210

提取方法：主成分分析法。

a. 提取了 3 个成分。

图 4 成分矩阵

由于各自变量已经过标准化，因此以上 3 个主成分的均数均为 0. 可以证明，各主成分的方差应当为前述特征根 λ_i，但按上述公式计算出的数值方差均为特征根的二次方，即各主成分的原始数值还应该除以一个特征根的平方根才行.

通过上面的分析，已经可以求出用来代替 8 个原始变量的 3 个主成分，下一步就可以进一步利用这 3 个主成分来计算出综合指标，进行各地区的综合排序了. 但是，由于主成分分析本质上是一种矩阵变换过程，并不要求各主成分都具有实际意义，目前得到的各主成分含义显得并不十分明确：在第一主成分的表达式中，X_1, X_2, X_3, X_8 的系数较大，可以看成是反映 GDP、固定资产投产、居民消费水平和工业总产值的综合指标；第二主成分中，X_4 和 X_5 的系数较大，可以看成是反映职工平均工资和货物周转量方面的综合指标. 在第三主成分中，X_6 系数较大，可以看成是反映居民消费价格指数方面的综合指标. 但显然这些含义的解释不够完美，这会导致最终得到的排序结果在解释上不够清晰，下一节将进一步考虑如何使得所提取的信息含义更加清晰，随后再进行综合排序.

§1.2 因子分析

因子分析由 Charles Spear man 在 1904 年首次提出，并在其后半生一直致力于发展此理论，使之最终成为了现代统计学的重要分支，因此他被公认为因子分析之父. 因子分析在某种程度上可以被看成是主成分分析的推广和扩展，它对问题的研究更为深入，是将具有错综复杂关系的变量（或样品）综合为少数几个因子，以再现原始变量与因子之间的相互关系，探讨多个能够直接测量，并且具有一定相关性的实测指标是如何受少数几个内在的独立因子所支配，并在条件许可时借此尝试对变量进行分类.

因子分析（Factor Analysis，FA）是一种数据简化技术，通过研究众多变量之间的内部依赖关系，探求观测数据的基本结构，并用少数几个假想变量（因子）来表示原始数据. 因子能够反映众多原始变量的主要信息. 其特点为

（1）因子个数远远少于原始变量个数；

（2）因子并非原始变量的简单取舍，而是一种新的综合；

（3）因子之间没有线性关系；

（4）因子具有明确解释性，可以最大限度地发挥专业分析的作用；

作个形象的比喻. 对面来了一群女生，我们一眼就能分辨出孰美孰丑，这是判别分析；并且我们的脑海中会迅速地将这群女生聚为两类：美的一类和丑的一类，这是聚类分析. 我们之所以认为某个女孩漂亮，是因为她具有腰亮女孩所具有的一些共同特点，比如漂亮的脸蛋、高挑的身材、白皙的皮肤，等等. 其实这种从研究对象中寻找公共因子的办法就是因子分析（Factor Analysis）.

1.2.1 因子分析引例

在市场调查中我们收集了食品的五项指标（x1~x5）：味道、价格、风味、是否快餐、能量，经过因子分析，我们发现了：

x1 = 0.02 * z1 + 0.99 * z2 + e1

x2 = 0.94 * z1 - 0.01 * z2 + e2

x3 = 0.13* z1 + 0.98 * z2 + e3

x4 = 0.84 * z1 + 0.42 * z2 + e4

x5 = 0.97 * z1 - 0.02 * z2 + e5

（数字代表实际变量间的相关系数，值越大，相关性越大）

第一个公因子 z1 主要与价格、是否快餐、能量有关，代表“价格与营养”；

第二个公因子 z2 主要与味道、风味有关，代表“口味”；

e1-5 是特殊因子，是公因子中无法解释的，在分析中一般略去.

1.2.2 因子分析模型

因子分析是通过研究多个变量间相关系数矩阵（或协方差矩阵）的内部依赖关系，找出能综合所有变量主要信息的少数几个随机变量，这几个随机变量不可直接测量，通常称为因子. 各个因子间互不相关，所有变量都可以表示成公因子的线性组合. 因子分析的目的就是减少变量的数目，用少数因子代替所有变量去分析整个问题.

设有 N 个样本， p 个指标， $X=(x_1,x_2,\cdots,x_p)^T$ 为随机向量，要寻找的公因子为 $F=(F_1,F_2,\cdots,F_m)^T$ ，则模型为

$$
\begin{aligned}
X_1 &= a_{11}F_1 + a_{12}F_2 + \cdots + a_{1m}F_m + \varepsilon_1 \\
X_2 &= a_{21}F_1 + a_{22}F_2 + \cdots + a_{2m}F_m + \varepsilon_2 \\
&\vdots \\
X_p &= a_{p1}F_1 + a_{p2}F_2 + \cdots + a_{pm}F_m + \varepsilon_p
\end{aligned}
$$

被称为因子模型. 矩阵 $A=(a_{ij})$ 称为因子载荷矩阵， a_{ij} 为因子载荷（Loading），其实质就是公因子 F_j 和变量 X_i 的相关系数. ε 为特殊因子，代表公因子以外的影响因素所导致的（不能被公共因子所解释的）变量变异，实际分析时忽略不计.

对求得的公因子，需要观察它们在哪些变量上有较大的载荷，再据此说明该公因子的实际含义. 但对于分析得到的初始因子模型，其因子载荷矩阵往往比较复杂，难于对因子 F_i 给出一个合理的解释，此时可以考虑进一步做因子旋转，以求旋转后能得到更加合理的解释.

因子分析得到的模型有两个特点：其一，模型不受量纲的影响；其二，因子载荷不是唯一的，通过因子轴的旋转，可以得到新的因子载荷阵，使其意义更加明显.

求出公因子后，还可以用回归估计等方法求出因子得分的数学模型，将各公因子表示成变量的线性形式，并进一步计算出因子得分，从而对各案例进行综合评价：

$$F_i = b_{i1}X_1 + b_{i2}X_2 + \cdots + b_{in}X_n \quad (i=1,2,\cdots,m) .$$

1.2.3 各统计量的意义

除了特征根、方差贡献率、累计贡献率等主成分分析中已经解释过的统计量之外，因子分析中还新增了如下两个比较重要的统计概念:

（1）因子载荷（Loading）：因子载荷 a_{ij} 为第 i 个变量在第 j 个因子上的载荷，实际上就是 X_i 与 F_j 的相关系数，表示变量 X_i 依赖因子 F_j 的程度，或者说反映了第 i 个变量 X_i 对于第 j 个公因子 F_j 的重要性.

（2）变量共同度（Communalities）：也被称为公因子方差比，记为 h_i^2，表示全部公因子对变量 X_i 的总方差所作出的贡献，或者说变量 X_i 的信息能够被 k 个公因子所描述的程度，数值在 0～1 之间. 取值越大，说明该变量能被公共因子解释的信息比例越高.

1.2.4 适用条件

（1）样本量不能太小. 因子分析的任务是分析变量间的内在关联结构，因此要求样本量比较充足，否则结果可能不太可靠. 一般而言，样本量应至少是变量数的 5 倍，如果要想得到比较理想的结果，则应该在 10 倍以上. 其次，除了比例关系外，样本总量也不能太少，按理论要求应该在 100 例以上. 不过在实际的经济和社会问题中，很多时候样本量都达不到这个要求，这时也可以适当放宽要求，通过检验来判断结果的可靠性.

（2）各变量间应该具有相关性. 如果变量间彼此独立，则无法从中提取公因子，也就谈不上应用因子分析了. 这可以通过 Bartlett 球形检验来加以判断，如果相关阵是

单位阵，则各变量独立，因子分析法无效.

（3）KMO 检验. KMO 检验用于考察变量间的偏相关性，取值在 0～1 之间. KMO 统计量越接近于 1，变量间的偏相关性越强，因子分析的效果越好. 实际分析中，KMO 统计量在 0.7 以上时，因子分析效果一般会比较好；而当 KMO 统计量在 0.5 以下时，此时不适合应用因子分析法，应考虑重新设计变量结构或者采用其他统计分析方法.

（4）因子分析中各公因子应该具有实际意义. 在主成分分析中，各主成分实际上是矩阵变换的结果，因此意义不明显并不重要. 但是在因子分析中，提取出的各因子应该具有实际意义，否则就应该重新进行分析.

1.2.5 因子分析的操作步骤

（1）选择分析变量；

（2）计算原始变量的相关系数矩阵；

（3）提取公因子. 取方差（特征值）大于 0 的因子，因子的累积方差贡献率达到 80%；

（4）因子旋转. 因子的实际意义更容易解释；

（5）计算因子得分.

1.2.6 因子分析与主成分分析的比较

1. 区别：

（1）因子分析需要构造因子模型，着重要求新变量具有实际的意义，能解释原始变量间的内在结构.

（2）主成分分析仅仅是变量变换，是原始变量的线性组合表示新的综合变量，强调新变量贡献了多大比例的方差，不关心新变量是否有明确的实际意义.

2. 联系：

两者都是降维和信息浓缩的方法.

生成的新变量均代表了原始变量的大部分信息且互相独立，都可以用于后续的回归分析、判别分析、聚类分析等等.

1.2.7 案例：对各省经济数据的进一步分析

在前面对全国 30 个省市、自治区的经济发展状况进行了主成分分析，最终将原始信息浓缩为了数个主成分. 但如果分析目的是探讨这 8 个原始变量背后代表的是哪些地区发展内容，则显然各主成分的含义并不十分明确. 下面按照因子分析的思路对这个数据继续进行分析，在前面操作的基础上，新增的操作步骤如下.

（1）在“描述”对话框中，选中“相关系数”选项组中的“KMO 和 Bartlett 的球形度检验”复选框.

（2）在“抽取”对话框中，选中“输出”选项组中的“碎石图”复选框.

KMO 和巴特利特检验		
KMO 取样适切性量数。		.620
巴特利特球形度检验	近似卡方	231.285
	自由度	28
	显著性	.000

KMO 和 Bartlett 的球形度检验

1. 基本分析结果

这里只对比较重要的结果加以解释，对相同的输出结果不再重复说明. 图 5 为 KMO 和球形 Bartlett 检验结果. KMO 检验变量间的偏相关是否较大，Bartlett 球形检验是判断相关阵是否是单位阵. 由 Bartlett 检验可以看出，应拒绝各变量独立的假设，即变量间具有较强的相关性. 但是 KMO 统计量为 0.620，小于 0.7，说明各变量间信息的重叠程度可能不是特别高，有可能做出的因子分析模型不是很完善，但仍然值得尝试.

图 6 给出了公因子方差，它表示各变量中所含原始信息能被提取的公因子所表示的程度，可见几乎所有变量的共同度都在 80%以上，因此按照默认数量提取出的这几个公因子对各变量的解释能力是较强的.

公因子方差		
	初始	提取
GDP	1.000	.945
居民消费水平	1.000	.799
固定资产投资	1.000	.902
职工平均工资	1.000	.873
货物周转量	1.000	.857
居民消费价格指数	1.000	.957
商品价格指数	1.000	.928
工业总产值	1.000	.904
提取方法：主成分分析法。		

图 6 公因子方差

本例仍然是按照特征根大于 1 的默认标准提取了 3 个公因子，但是这个提取标准是否合适呢？还可以利用碎石图（Scree Plot）来协助判断，如图 7 所示. 碎石图用于显示各因子的重要程度，其横轴为因子序号，纵轴表示特征根大小. 它将因子按特征根从大到小依次排列，从中可以直接观察到哪些是最主要的因子. 前面的陡坡对应较大的特征根，作用明显；后面的平台对应较小的特征根，其影响较弱. 本例中可见前 3 个因子的散点位于陡坡上，而后 5 个因子散点形成了平台，且特征根均小于 1，因此至多考虑前 3 个公因子即可.

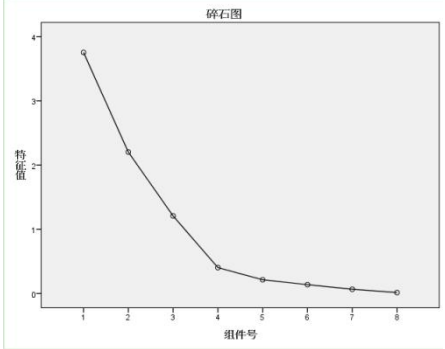


图 7 碎石图

注：Scree 一词来自于地质学，表示在岩石断层斜坡下方发现的小碎石，这些碎石可能是因风化、水流等从其他地点带来，因此其地质学价值不高，可以忽略.

最后图 8 给出的是和图 4 完全相同的“成分矩阵”表. 在第 1 节中直接按列的方向将其解释为各主成分的系数，并据此写出了公因子的计算公式. 在因子分析中，该表格则按行阅读，反映的是各因子在各变量上的载荷，即各因子对各变量的影响度：

$$Z_{x_1} = 0.884F_1 + 0.385F_2 + 0.120F_3 + \varepsilon_1$$
$$Z_{X_2} = 0.606F_1 - 0.596F_2 - 0.277F_3 + \varepsilon_2$$
$$\vdots$$
$$Z_{X_8} = 0.822F_1 + 0.429F_2 - 0.210F_3 + \varepsilon_8$$

成分矩阵[a]

	成分		
	1	2	3
GDP	.884	.385	.120
居民消费水平	.606	-.596	.277
固定资产投资	.911	.163	.213
职工平均工资	.465	-.725	.362
货物周转量	.486	.737	-.279
居民消费价格指数	-.510	.257	.794
商品价格指数	-.621	.596	.433
工业总产值	.822	.429	.210

提取方法：主成分分析法。

a. 提取了 3 个成分。

图 8 成分矩阵表

注意在表达式中的各变量不是原始变量，而是标准化变量. ε_i 表示特殊因子，是除了这 3 个公因子外影响该变量的其他因素，其对该变量的影响程度为：1-变量共同度.

2. 因子旋转

因子分析要求提取出的公因子有实际含义，但是从上面各因子和原始变量的相关系数可以看出，现在各因子的意义不是很明显，为了使因子载荷矩阵中系数更加显著，可以对初始因子载荷矩阵进行旋转，使因子和原始变量间的关系进行重新分配，相关系数的绝对值向 $(0,1)$ 区间的两极分化，从而更加容易进行解释.

“旋转”对话框用来实现因子旋转功能，其中提供了 5 种因子旋转方法，分为正交旋转和斜交旋转两大类，具体说明如下.

（1）方差最大正交旋转（Varimax）：是最常用的旋转方法，使各因子仍然保持正交的状态，但尽量使得各因子的方差差异达到最大，即相对的载荷平方和达到最大，从而方便对因子的解释.

（2）4 次方最大正交旋转（Quartimax）：该方法对各因子方差差异化的效果显然更强，同时倾向于减少和每个变量有关联的因子数，从而简化对原变量的解释.

（3）最大平衡值法（Equamax）：该方法的特点正好介于方差最大正交旋转和 4 次方最大正交旋转之间.

（4）直接 Oblimin 法：斜交旋转方法，需要首先指定一个因子映像的自相关范围.

（5）Promax：最常用的斜交旋转方法，是在方差最大正交旋转的基础上再进行斜交旋转. 旋转后允许因子间存在相关，这种旋转方式往往是在有具体的结果倾向时选用，它可以按分析者的目的将因子分解为最希望的形式. 但是在实际应用中，由于斜交旋转的结果太容易受研究者主观意愿的左右，所以建议尽量采用默认的正交旋转. 因子分析的旋转、选项和得分对话框如图 13. 10 所示.

图 9 因子分析的旋转、选项和得分对话框

对于本例可以采用方差最大旋转加以分析，结果输出中的变化如图 10 所示.

总方差解释

成分	初始特征值 总计	初始特征值 方差百分比	初始特征值 累积 %	提取载荷平方和 总计	提取载荷平方和 方差百分比	提取载荷平方和 累积 %	旋转载荷平方和 总计	旋转载荷平方和 方差百分比	旋转载荷平方和 累积 %
1	3.754	46.924	46.924	3.754	46.924	46.924	3.207	40.092	40.092
2	2.203	27.532	74.456	2.203	27.532	74.456	2.217	27.708	67.800
3	1.208	15.096	89.551	1.208	15.096	89.551	1.740	21.752	89.551
4	.403	5.042	94.593						
5	.214	2.673	97.266						
6	.138	1.722	98.988						
7	.066	.829	99.817						
8	.015	.183	100.000						

提取方法：主成分分析法。

图 10 解释的总方差

解释的总方差表格在最右侧会给出旋转后各因子的载荷情况. 由于默认只提取了前 3 个公因子，因此旋转会基于所提取的这 3 个公因子进行. 在旋转后 3 个公因子的方差贡献率均发生了变化，彼此差距有所缩小，显然信息量进行了重新分配，但仍然保持从大到小的排列顺序，且累计方差贡献率仍然是 89.55%，和旋转前完全相同.

进行方差最大旋转后，旋转后的因子载荷矩阵如图 11(a) 所示，为了便于阅读，可以利用“选项”对话框中的“系数显示格式”复选框组进行表格重排序和化简，结果如图 11(b) 所示. 可见表格按照系数大小进行了排序，而且过小的系数也被抑制输出，使得结果更清晰易读. 但内容实际上是相同的. 由表中可以看出第一公因子在 x_1, x_3, x_5 和 x_8 有较大的载荷，主要从 GDP、固定资产投资、货物周转量和工业总产值反映经济发展状况，可以命名为**总量因子**；第二公因子在 x_2, x_4 上有较大载荷，从居民消费水平和职工平均工资方面反映经济发展水平，可命名为**消费因子**；第三公因子在 x_6, x_7 上有较大载荷，表现为居民消费价格指数和水平价格指数方面，因此命名为**价格因子**. 与未旋转前相比，旋转后各公因子的意义显然更加明确合理，也更有利于对数

据的解读和应用.

旋转后的成分矩阵[a]

	成分		
	1	2	3
GDP	.955	.124	-.131
居民消费水平	.219	.841	-.209
固定资产投资	.872	.351	-.137
职工平均工资		.925	-.121
货物周转量	.751	-.507	-.192
居民消费价格指数	-.135		.969
商品价格指数	-.104	-.496	.819
工业总产值	.944	.109	

提取方法：主成分分析法。
旋转方法：凯撒正态化最大方差法。
a. 旋转在 5 次迭代后已收敛。

（a）

旋转后的成分矩阵[a]

	成分		
	1	2	3
GDP	.955	.124	-.131
工业总产值	.944	.109	
固定资产投资	.872	.351	-.137
货物周转量	.751	-.507	-.192
职工平均工资		.925	-.121
居民消费水平	.219	.841	-.209
居民消费价格指数	-.135		.969
商品价格指数	-.104	-.496	.819

提取方法：主成分分析法。
旋转方法：凯撒正态化最大方差法。
a. 旋转在 5 次迭代后已收敛。

（b）

图 11 重排序并化简前后的旋转成分矩阵

在旋转成分矩阵后还会输出成分转换矩阵，给出旋转前后各公因子间的相关系数. 如图 12 所示.

成分转换矩阵

成分	1	2	3
1	.817	.407	-.408
2	.548	-.769	.331
3	.179	.494	.851

提取方法：主成分分析法。
旋转方法：凯撒正态化最大方差法。

图 12 成分旋转矩阵

3. 因子表达式

前面得到的因子结构表达式可以将各变量表示为公因子的线性形式，但是更多的时候需要将公因子表达为各变量的线性形式，这也称为因子得分函数.

在 SPSS 中可以利用“得分”对话框中的“显示因子得分系数阵”复选框在结果中直接输出因子系数矩阵，本例结果如图 13 所示，据此可以直接写出各公因子的表达式：

成分得分系数矩阵

	成分		
	1	2	3
GDP	.306	.011	.047
居民消费水平	.025	.387	.040
固定资产投资	.270	.129	.075
职工平均工资	-.025	.451	.096
货物周转量	.248	-.319	-.139
居民消费价格指数	.070	.180	.653
商品价格指数	.077	-.098	.462
工业总产值	.317	.026	.123

提取方法：主成分分析法。
旋转方法：凯撒正态化最大方差法。
组件得分。

图 13 成分得分系数矩阵

$$F_1 = 0.306Zx_1 + 0.025Zx_2 + 0.270Zx_3 - 0.025Zx_4 + 0.248Zx_5 + 0.070Zx_6 + 0.077Zx_7 + 0.317Zx_8$$
$$F_2 = 0.011Zx_1 + 0.387Zx_2 + 0.129Zx_3 + 0.451Zx_4 - 0.319Zx_5 + 0.180Zx_6 - 0.098Zx_7 + 0.026Zx_8$$
$$F_3 = 0.047Zx_1 + 0.040Zx_2 + 0.075Zx_3 + 0.096Zx_4 - 0.139Zx_5 + 0.65\,3Zx_6 + 0.462Zx_7 + 0.123Zx_8$$

4. 保存公因子得分进行综合评价

在对公共因子做出合理的解释之后，有时还需要求出各观测所对应的各个公共因子的得分，就比如我们知道某个女孩是一美女，可能很多人更关心该给她的脸蛋、身林等各打多少分.

上文已经得到了公因子表达式，虽然可以据此人工计算出因子得分，但是这需要先将变量标准化，再输入公式计算，实际上分析者可以直接使用“得分”对话框中的“保存为变量”复选框，直接保存各因子得分值为新变量，默认变量名称为 FAC1_1 ~ FAC3_1.

由于上述 3 个公因子是分别从不同方面反映了当地经济发展状况的总体水平，单独使用某一公因子很难全面做出综合评价，因此考虑按各公因子对应的方差贡献率比例为权数计算如下综合得分，即

score=40.09/89.55*FAC1_1+27.71/89.55*FAC2_1+21.75/89.55*FAC3_1.

按照计算出的综合因子得分 score，得分最高的前 5 个地区如表所示.

序号	地区	FAC1_1	FAC2_1	FAC3_1	Score
8	上海	0. 615	3. 662	0. 847	1. 61
9	江苏	2. 034	0. 271	-0. 171	0. 95
15	山东	2. 117	-0. 194	0. 252	0. 95
19	广东	1. 485	1. 684	-1. 182	0. 90
22	四川	1. 107	-0. 520	0. 978	0. 57

可见上海综合得分最高，其 3 个公因子表现均不错，但消费因子的表现尤其突出；排名第二的江苏和第三的山东则都是靠总量因子在拉动，另两个因子得分仅在平均水平上下；广东虽然总量因子和消费因子表现都不错，但过高的物价则拉低了其综合得分；四川则恰恰相反，虽然其消费因子表现较差，但相对的低物价则拉高了其综合得分，最终排名第五. 显然，由于各个因子有明确的含义，上述综合比较结果可以很好地解释各地区整体经济发展上的优点和劣势，更有利于各区域针对各自特点确定其经济发展方向和重点.

需要指出的是，对于相同的数据，是否进行公因子旋转所得到的综合评分结果会存在明显差异，从方法原理上不旋转的原始主成分提取结果或许更严谨，但含义清听、应用价值明确的分析结果在实际分析项目中总是最受欢迎的.

基于上述结果，该数据还可以做进一步的分析，比如利用聚类分析将所有地区分为若干类，总结出各类的经济发展特征.

§1.3 综合案例分析

1.3.1 我国上市公司赢利能力与资本结构的实证分析

已知上市公司的数据见表所示。

公司	销售净利率 x_1	资产净利率 x_2	净资产收益率 x_3	销售毛利率 x_4	资产负利率 x
歌华有线	43.31	7.39	8.73	54.89	15.35
五粮液	17.11	12.13	17.29	44.25	29.69
用友软件	21.11	6.03	7	89.37	13.82
太太药业	29.55	8.62	10.13	73	14.88
浙江阳光	11	8.41	11.83	25.22	25.49
烟台万华	17.63	13.86	15.41	36.44	10.03
方正科技	2.73	4.22	17.16	9.96	74.12
红河光明	29.11	5.44	6.09	56.26	9.85
贵州茅台	20.29	9.48	12.97	82.23	26.73
中铁二局	3.99	4.64	9.35	13.04	50.19
红星发展	22.65	11.13	14.3	50.51	21.59
伊利股份	4.43	7.3	14.36	29.04	44.74
青岛海尔	5.4	8.9	12.53	65.5	23.27
湖北宜化	7.06	2.79	5.24	19.79	40.68
雅戈尔	19.82	10.53	18.55	42.04	37.19
福建南纸	7.26	2.99	6.99	22.72	56.58

试用主成分分析和因子分析对上述企业进行综合评价.

1. 主成分分析

相关性矩阵

		销售净利率x1	资产净利率x2	净资产收益率x3	销售毛利率x4
相关性	销售净利率x1	1.000	.319	-.171	.606
	资产净利率x2	.319	1.000	.674	.344
	净资产收益率x3	-.171	.674	1.000	-.139
	销售毛利率x4	.606	.344	-.139	1.000

总方差解释

	初始特征值			提取载荷平方和		
成分	总计	方差百分比	累积 %	总计	方差百分比	累积 %
1	1.897	47.429	47.429	1.897	47.429	47.429
2	1.550	38.740	86.169	1.550	38.740	86.169
3	.393	9.826	95.995			
4	.160	4.005	100.000			

提取方法：主成分分析法。

成分矩阵[a]

	成分 1	成分 2
销售净利率x1	.731	-.513
资产净利率x2	.818	.503
净资产收益率x3	.359	.897
销售毛利率x4	.752	-.477

提取方法：主成分分析法。

a. 提取了 2 个成分。

2. 因子分析

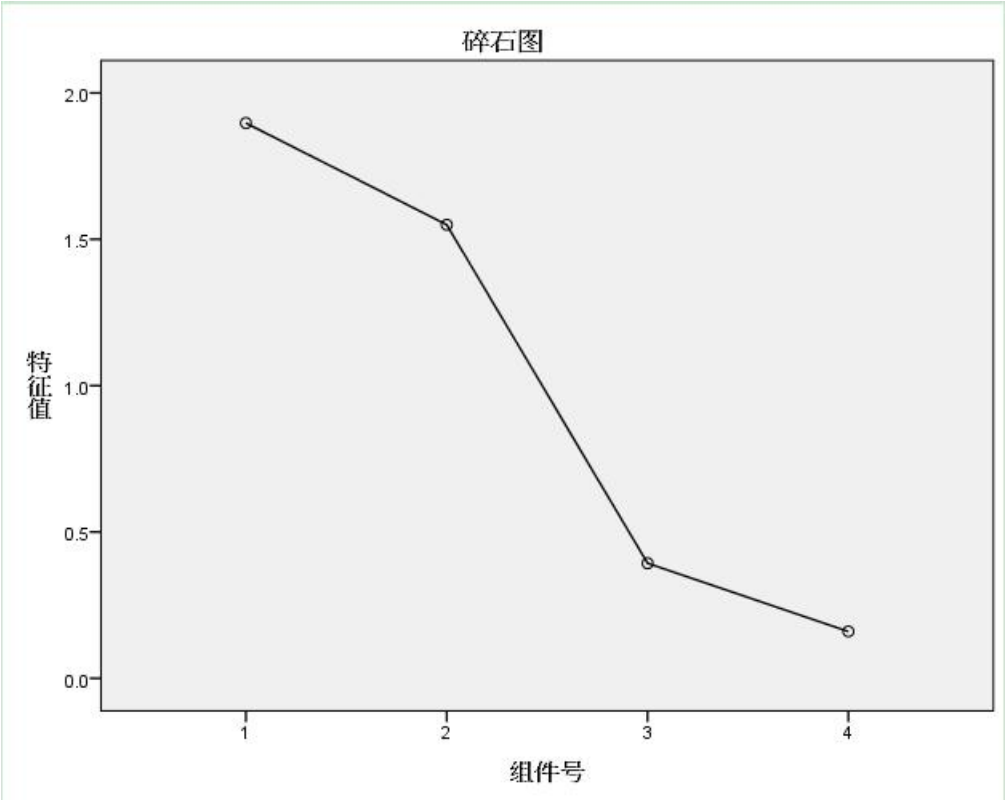
KMO 和巴特利特检验

KMO 取样适切性量数。		.455
巴特利特球形度检验	近似卡方	21.647
	自由度	6
	显著性	.001

公因子方差

	初始	提取
销售净利率x1	1.000	.797
资产净利率x2	1.000	.922
净资产收益率x3	1.000	.934
销售毛利率x4	1.000	.793

提取方法：主成分分析法。



成分矩阵[a]

	成分 1	成分 2
销售净利率x1	.731	-.513
资产净利率x2	.818	.503
净资产收益率x3	.359	.897
销售毛利率x4	.752	-.477

提取方法：主成分分析法。

a. 提取了 2 个成分。

因子旋转

总方差解释

成分	初始特征值 总计	初始特征值 方差百分比	初始特征值 累积 %	提取载荷平方和 总计	提取载荷平方和 方差百分比	提取载荷平方和 累积 %	旋转载荷平方和 总计	旋转载荷平方和 方差百分比	旋转载荷平方和 累积 %
1	1.897	47.429	47.429	1.897	47.429	47.429	1.779	44.486	44.486
2	1.550	38.740	86.169	1.550	38.740	86.169	1.667	41.684	86.169
3	.393	9.826	95.995						
4	.160	4.005	100.000						

提取方法：主成分分析法。

旋转后的成分矩阵[a]

	成分 1	成分 2
销售净利率x1	.893	.008
资产净利率x2	.372	.885
净资产收益率x3	-.230	.939
销售毛利率x4	.889	.049

提取方法：主成分分析法。
旋转方法：凯撒正态化最大方差法。

a. 旋转在 3 次迭代后已收敛。

本例中，我们选取两个主因子

第一公共因子 F_1 为销售利润因子

第一公共因子 F_2 为资产收益因子

成分得分系数矩阵

	成分 1	成分 2
销售净利率x1	.506	-.045
资产净利率x2	.161	.515
净资产收益率x3	-.183	.581
销售毛利率x4	.502	-.020

提取方法：主成分分析法。
旋转方法：凯撒正态化最大方差法。
组件得分。

计算得各个因子得分函数

$$F_1 = 0.531\tilde{x}_1 + 0.1615\tilde{x}_2 - 0.1831\tilde{x}_3 + 0.5015\tilde{x}_4$$

$$F_2 = -0.045\tilde{x}_1 + 0.5151\tilde{x}_2 + 0.581\tilde{x}_3 - 0.0199\tilde{x}_4$$

利用综合因子得分公式

$$F=\frac{44.49F_1+41.68F_2}{86.17}$$

计算出 16 家上市公司赢利能力的综合得分见表.

排名	1	2	3	4	5	6	7	8
F_1	0.0315	0.0025	0.9789	0.4558	-0.0563	1.2791	1.5159	1.2477
F_2	1.4691	1.4477	0.3959	0.8548	1.3577	-0.1564	-0.5814	-0.9729
F	0.7269	0.7016	0.6969	0.6488	0.6277	0.5847	0.5014	0.1735
公司	烟台万华	五粮液	贵州茅台	红星发展	雅戈尔	太太药业	歌华有线	用友软件
排名	9	10	11	12	13	14	15	16
F_1	-0.0351	0.9313	-0.6094	-0.9859	-1.7266	-1.2509	-0.8872	-0.891
F_2	0.3166	-1.1949	0.1544	0.3468	0.2639	-0.7424	-1.1091	-1.2403
F	0.135	-0.0972	-0.2399	-0.3412	-0.7637	-1.0049	-1.1091	-1.2403
公司	青岛海尔	红河光明	浙江阳光	伊利股份	方正科技	中铁二局	福建南纸	湖北宜化

1.3.2 我国各地区普通高等教育的发展水平实证分析

本案例运用主成分分析与因子分析法对我国各地区普通高等教育的发展水平进行综合评价.

1. 案例研究背景

近年来，我国普通高等教育得到了迅速发展，为国家培养了大批人才. 但由于我国各地区经济发展水平不均衡，加之高等院校原有布局使各地区高等教育发展的起点不一致，因而各地区普通高等教育的发展水平存在一定的差异，不同的地区具有不同的特点. 对我国各地区普通高等教育的发展状况进行聚类分析，明确各类地区普通高等教育发展 状况的差异与特点,有利于管理和决策部门从宏观上把握我国普通高等教育的整体发展现状，分类制定相关政策，更好的指导和规划我国高教事业的整体健康发展.

2. 数据资料

指标的原始数据取自《中国统计年鉴，1995》和《中国教育统计年鉴，1995》除以各地区相应的人口数得到十项指标值见表. 其中：

x_1：每百万人口高等院校数；

x_2：每十万人口高等院校毕业生数；

x_3：每十万人口高等院校招生数；

x_4：每十万人口高等院校在校生数；

x_5：每十万人口高等院校教职工数；

x_6：每十万人口高等院校专职教师数；

x_7：高级职称占专职教师的比例；

x_8：平均每所高等院校的在校生数；

x_9：国家财政预算内普通高教经费占国内生产总值的比重；

x_{10}：生均教育经费.

表 2 我国各地区普通高等教育发展状况数据

地区	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
北京	5.96	310	461	1557	931	319	44.36	2615	2.20	13631
上海	3.39	234	308	1035	498	161	35.02	3052	.90	12665
天津	2.35	157	229	713	295	109	38.40	3031	.86	9385
陕西	1.35	81	111	364	150	58	30.45	2699	1.22	7881
辽宁	1.50	88	128	421	144	58	34.30	2808	.54	7733
吉林	1.67	86	120	370	153	58	33.53	2215	.76	7480
黑龙江	1.17	63	93	296	117	44	35.22	2528	.58	8570
湖北	1.05	67	92	297	115	43	32.89	2835	.66	7262
江苏	.95	64	94	287	102	39	31.54	3008	.39	7786
广东	.69	39	71	205	61	24	34.50	2988	.37	11355
四川	.56	40	57	177	61	23	32.62	3149	.55	7693
山东	.57	58	64	181	57	22	32.95	3202	.28	6805
甘肃	.71	42	62	190	66	26	28.13	2657	.73	7282
湖南	.74	42	61	194	61	24	33.06	2618	.47	6477
浙江	.86	42	71	204	66	26	29.94	2363	.25	7704
新疆	1.29	47	73	265	114	46	25.93	2060	.37	5719
福建	1.04	53	71	218	63	26	29.01	2099	.29	7106
山西	.85	53	65	218	76	30	25.63	2555	.43	5580
河北	.81	43	66	188	61	23	29.82	2313	.31	5704
安徽	.59	35	47	146	46	20	32.83	2488	.33	5628
云南	.66	36	40	130	44	19	28.55	1974	.48	9106
江西	.77	43	63	194	67	23	28.81	2515	.34	4085
海南	.70	33	51	165	47	18	27.34	2344	.28	7928
内蒙古	.84	43	48	171	65	29	27.65	2032	.32	5581
西藏	1.69	26	45	137	75	33	12.10	810	1.00	14199
河南	.55	32	46	130	44	17	28.41	2341	.30	5714
广西	.60	28	43	129	39	17	31.93	2146	.24	5139
宁夏	1.39	48	62	208	77	34	22.70	1500	.42	5377
贵州	.64	23	32	93	37	16	28.12	1469	.34	5415
青海	1.48	38	46	151	63	30	17.87	1024	.38	7368

1. 主成分分析

相关性矩阵

		V1	V2	V3	V4	V5	V6	V7	V8	V9	V10
相关性	V1	1.000	.943	.953	.959	.975	.980	.407	.066	.868	.661
	V2	.943	1.000	.995	.995	.974	.970	.614	.350	.804	.600
	V3	.953	.995	1.000	.999	.983	.981	.626	.344	.823	.617
	V4	.959	.995	.999	1.000	.988	.986	.610	.326	.828	.612
	V5	.975	.974	.983	.988	1.000	.999	.560	.241	.859	.617
	V6	.980	.970	.981	.986	.999	1.000	.550	.222	.869	.616
	V7	.407	.614	.626	.610	.560	.550	1.000	.779	.366	.151
	V8	.066	.350	.344	.326	.241	.222	.779	1.000	.112	.048
	V9	.868	.804	.823	.828	.859	.869	.366	.112	1.000	.683
	V10	.661	.600	.617	.612	.617	.616	.151	.048	.683	1.000

总方差解释

成分	初始特征值 总计	初始特征值 方差百分比	初始特征值 累积 %	提取载荷平方和 总计	提取载荷平方和 方差百分比	提取载荷平方和 累积 %
1	7.502	75.022	75.022	7.502	75.022	75.022
2	1.577	15.770	90.791	1.577	15.770	90.791
3	.536	5.362	96.154			
4	.206	2.064	98.217			
5	.145	1.450	99.667			
6	.022	.222	99.889			
7	.007	.071	99.960			
8	.003	.027	99.987			
9	.001	.007	99.994			
10	.001	.006	100.000			

提取方法：主成分分析法。

成分矩阵[a]

	成分 1	成分 2
V1	.958	-.248
V2	.983	.043
V3	.992	.037
V4	.992	.017
V5	.987	-.064
V6	.986	-.081
V7	.614	.732
V8	.329	.882
V9	.874	-.244
V10	.672	-.360

提取方法：主成分分析法。

a. 提取了 2 个成分。

2. 因子分析

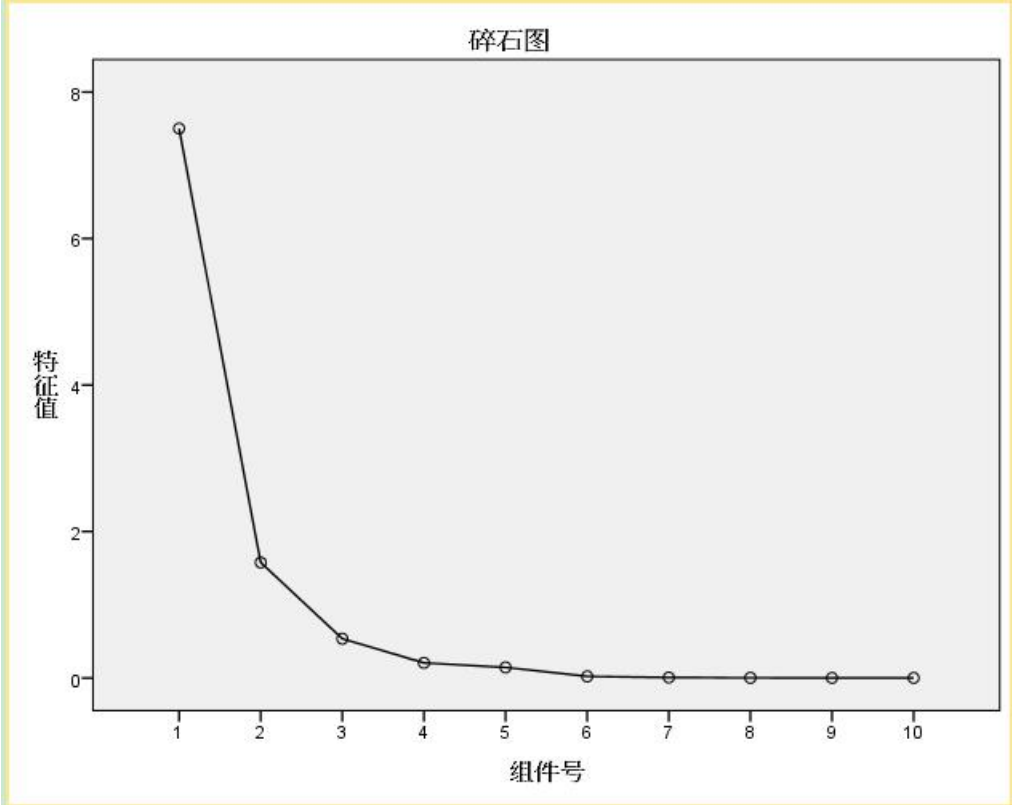
KMO 和巴特利特检验

KMO 取样适切性量数。		.834
巴特利特球形度检验	近似卡方	770.928
	自由度	45
	显著性	.000

公因子方差

	初始	提取
V1	1.000	.979
V2	1.000	.969
V3	1.000	.986
V4	1.000	.985
V5	1.000	.979
V6	1.000	.980
V7	1.000	.912
V8	1.000	.885
V9	1.000	.824
V10	1.000	.580

提取方法：主成分分析法。



成分矩阵[a]

	成分	
	1	2
V1	.958	-.248
V2	.983	.043
V3	.992	.037
V4	.992	.017
V5	.987	-.064
V6	.986	-.081
V7	.614	.732
V8	.329	.882
V9	.874	-.244
V10	.672	-.360

提取方法：主成分分析法。

a. 提取了 2 个成分。

因子旋转

总方差解释

成分	初始特征值			提取载荷平方和			旋转载荷平方和		
	总计	方差百分比	累积 %	总计	方差百分比	累积 %	总计	方差百分比	累积 %
1	7.502	75.022	75.022	7.502	75.022	75.022	6.817	68.166	68.166
2	1.577	15.770	90.791	1.577	15.770	90.791	2.263	22.625	90.791
3	.536	5.362	96.154						
4	.206	2.064	98.217						
5	.145	1.450	99.667						
6	.022	.222	99.889						
7	.007	.071	99.960						
8	.003	.027	99.987						
9	.001	.007	99.994						
10	.001	.006	100.000						

提取方法：主成分分析法。

旋转后的成分矩阵[a]

	成分	
	1	2
V1	.985	.093
V2	.910	.375
V3	.921	.372
V4	.927	.354
V5	.950	.276
V6	.955	.259
V7	.328	.897
V8	.009	.941
V9	.905	.068
V10	.754	-.110

提取方法：主成分分析法。
旋转方法：凯撒正态化最大方差法。

a. 旋转在 3 次迭代后已收敛。

成分转换矩阵

成分	1	2
1	.940	.340
2	-.340	.940

提取方法：主成分分析法。
旋转方法：凯撒正态化最大方差法。

本例中，我们选取两个主因子

成分得分系数矩阵

	成分 1	成分 2
V1	.173	-.104
V2	.114	.070
V3	.116	.067
V4	.121	.055
V5	.138	.007
V6	.141	-.004
V7	-.081	.464
V8	-.149	.541
V9	.162	-.106
V10	.162	-.184

提取方法：主成分分析法。
旋转方法：凯撒正态化最大方差法。
组件得分。

计算得各个因子得分函数

利用综合因子得分公式

$$F=\frac{68.17F_1+22.63F_2}{90.80}$$

计算出 30 省份高等教育发展水平的综合得分.