




降维技术

主成分分析与因子分析

- 
- 主成分的基本思想
 - 主成分的导出
 - 相关问题的讨论
 - 因子分析的思想
 - 因子分析三步曲

Part I 主成分分析

【引例】 根据下面的数据，对全国重点水泥企业经济效益进行综合评价，并提出改进方案。原始数据（数据来自1984年中国统计年鉴）见下表

厂家编号及指标	固定资产 利税率	资金利 税率	销售收入 利税率	资金利润 率	固定资产 产值率	流动资 金周转 天数	万元产 值能耗	全员劳动生产 率
1 琉璃河	16.68	26.75	31.84	18.4	53.25	55	28.83	1.75
2 邯郸	19.7	27.56	32.94	19.2	59.82	55	32.92	2.87
3 大同	15.2	23.4	32.98	16.24	46.78	65	41.69	1.53
4 哈尔滨	7.29	8.97	21.3	4.76	34.39	62	39.28	1.63
5 华新	29.45	56.49	40.74	43.68	75.32	69	26.68	2.14
6 湘乡	32.93	42.78	47.98	33.87	66.46	50	32.87	2.6
7 柳州	25.39	37.82	36.76	27.56	68.18	63	35.79	2.43
8 峨嵋	15.05	19.49	27.21	14.21	6.13	76	35.76	1.75
9 耀县	19.82	28.78	33.41	20.17	59.25	71	39.13	1.83
10 永登	21.13	35.2	39.16	26.52	52.47	62	35.08	1.73
11 工源	16.75	28.72	29.62	19.23	55.76	58	30.08	1.52
12 抚顺	15.83	28.03	26.4	17.43	61.19	61	32.75	1.6
13 大连	16.53	29.73	32.49	20.63	50.41	69	37.57	1.31
14 江南	22.24	54.59	31.05	37	67.95	63	32.33	1.57
15 江油	12.92	20.82	25.12	12.54	51.07	66	39.18	1.83

2020-7-24

- 主成分分析(Principal Components Analysis)也称主分量分析，是由霍特林（Hotelling）于1933年首先提出的。
- 主成分分析是利用降维的思想，在损失很少信息的前提下把多个指标转化为几个综合指标的多元统计方法。
- 通常把转化生成的综合指标称之为主成分，其中每个主成分都是原始变量的线性组合，且各个主成分之间互不相关。
- 这样在研究复杂问题时就可以只考虑少数几个主成分而不至于损失太多信息，从而更容易抓住主要矛盾，同时使问题得到简化，提高分析效率。

既然研究某一问题涉及的众多变量之间有一定的相关性，就必然存在着起支配作用的共同因素，根据这一点，通过对原始变量**相关矩阵或协方差矩阵内部结构**关系的研究，利用原始变量的线性组合形成几个综合指标（主成分），在保留原始变量主要信息的前提下起到降维与简化问题的作用，使得在研究复杂问题时更容易抓住主要矛盾。

利用主成分分析得到的主成分与原始变量之间有如下基本关系：

1. 每一个主成分都是各原始变量的线性组合
2. 主成分的数目大大少于原始变量的数目
3. 主成分保留了原始变量绝大多数信息
4. 各主成分之间互不相关

1 主成分分析的基本理论

设对某一事物的研究涉及个 p 指标，分别用 X_1, X_2, \dots, X_p 表示，这个 p 指标构成的 p 维随机向量为 $\mathbf{X} = (X_1, X_2, \dots, X_p)'$ 。设随机向量 \mathbf{X} 的均值为 μ ，协方差矩阵为 Σ 。

对 \mathbf{X} 进行线性变换，可以形成新的综合变量，用 \mathbf{Y} 表示，也就是说，新的综合变量可以由原来的变量线性表示，即满足下式：

$$\begin{cases} Y_1 = u_{11}X_1 + u_{12}X_2 + \dots + u_{1p}X_p \\ Y_2 = u_{21}X_1 + u_{22}X_2 + \dots + u_{2p}X_p \\ \dots\dots\dots \\ Y_p = u_{p1}X_1 + u_{p2}X_2 + \dots + u_{pp}X_p \end{cases} \quad (1.1)$$

上述线性变换需要约束在下面的原则之下：

1. $\mathbf{u}_i' \mathbf{u}_i = 1$, 即: $u_{i1}^2 + u_{i2}^2 + \cdots + u_{ip}^2 = 1$ ($i = 1, 2, \dots, p$)。

2. Y_i 与 Y_j 相互无关 ($i \neq j; i, j = 1, 2, \dots, p$)。

3. Y_1 是 X_1, X_2, \dots, X_p 的一切满足原则1的线性组合中方差最大者; Y_2 是与 Y_1 不相关的 X_1, X_2, \dots, X_p 所有线性组合中方差最大者; ..., Y_p 是与 Y_1, Y_2, \dots, Y_{p-1} 都不相关的 X_1, X_2, \dots, X_p 的所有线性组合中方差最大者。

基于以上三条原则决定的综合变量 Y_1, Y_2, \dots, Y_p 分别称为原始变量的第一、第二、...、第 p 个主成分。其中，各综合变量在总方差中占的比重依次递减，在实际研究工作中，通常只挑选前几个方差最大的主成分，从而达到简化系统结构，抓住问题实质的目的。

- 主成分分析的基本思想就是在保留原始变量尽可能多的信息的前提下达到降维的目的，从而简化问题的复杂性并抓住问题的主要矛盾。
- 而这里对于随机变量 X_1, X_2, \dots, X_p 而言，其协方差矩阵或相关矩阵正是对各变量离散程度与变量之间的相关程度的信息的反应，而相关矩阵不过是将原始变量标准化后的协方差矩阵。
- 因此在实际求解主成分的时候，总是从原始变量的**协方差矩阵或相关矩阵**的结构分析入手。
- 一般地说，从原始变量的协方差矩阵出发求得的主成分与从原始变量的相关矩阵出发求得的主成分是不同的。

2 主成分的导出

2.1 从协方差阵出发求解主成分

结论： 设随机向量 $\mathbf{X} = (X_1, X_2, \dots, X_p)'$ 的协方差矩阵为 Σ ，

$\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p$ 为 Σ 的特征值， $\gamma_1, \gamma_2, \dots, \gamma_p$ 为矩阵 Σ 各特征值对应的标准正交特征向量，则第 i 个主成分为：

$$Y_i = \gamma_{i1}X_1 + \gamma_{i2}X_2 + \dots + \gamma_{ip}X_p \quad (i = 1, 2, \dots, p)$$

此时：

$(i \neq j)$

$$\text{var}(Y_i) = \gamma_i' \Sigma \gamma_i = \lambda_i \quad \text{cov}(Y_i, Y_j) = \gamma_i' \Sigma \gamma_j = 0$$

由以上结论，我们把 X_1, X_2, \dots, X_p 的协方差矩阵 Σ 的非零特征值 $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p > 0$ 对应的标准化特征向量 $\gamma_1, \gamma_2, \dots, \gamma_p$ 分别作为系数向量， $Y_1 = \gamma_1' X, Y_2 = \gamma_2' X, \dots, Y_p = \gamma_p' X$ 分别称为随机向量 X 的第一主成分、第二主成分、...、第 p 主成分。 Y 的分量 Y_1, Y_2, \dots, Y_p 依次是 X 的第一主成分、第二主成分、...、第 p 主成分的充分必要条件是：

- (1) $Y = u' X, u' u = I$ ，即 u 为 p 阶正交阵；
- (2) Y 的分量之间互不相关；
- (3) Y 的 p 个分量是按方差由大到小排列。

注：无论 Σ 的各特征根是否存在相等的情况，对应的标准化特征向量 $\gamma_1, \gamma_2, \dots, \gamma_p$ 总是存在的，我们总可以找到对应各特征根的彼此正交的特征向量。这样，**求主成分的问题就变成了求特征根与特征向量的问题。**

定义 2.1 称 $\alpha_k = \frac{\lambda_k}{\lambda_1 + \lambda_2 + \cdots + \lambda_p}$ $k = 1, 2, \dots, p$ 为第 k 个主成分 Y_k 的方差贡献率，称 $\frac{\sum_{i=1}^m \lambda_i}{\sum_{i=1}^p \lambda_i}$ 为主成分 Y_1, Y_2, \dots, Y_m 的累积贡献率。

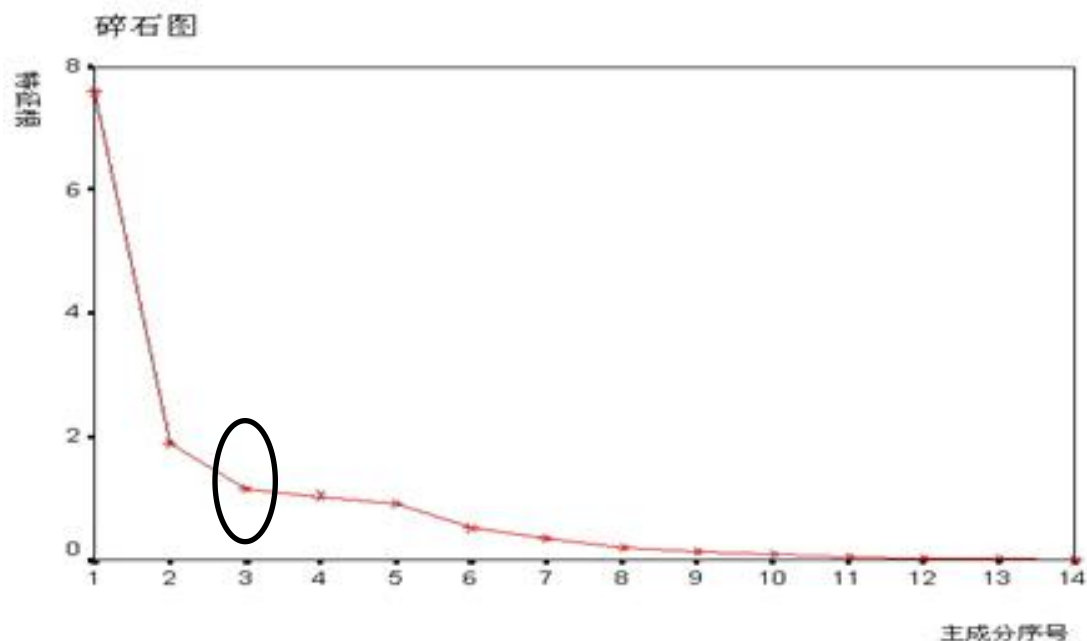
表明了主成分方差在全部方差中的比值，称 $\alpha_1 = \frac{\lambda_1}{\sum \lambda_i}$ 为第一主成分的贡献率。这个值越大，表明 $Y_1 = \gamma_1' X$ 这个变量综合 X_1, X_2, \dots, X_p 信息的能力越强，也即由 $\gamma_1' X$ 的差异来解释随机向量 X 的差异的能力越强。

正因如此，才把 $Y_1 = \gamma_1'X$ 称为 X 的主成分。进而我们就更清楚为什么主成分的名次是按特征根 $\lambda_1, \lambda_2, \dots, \lambda_p$ 取值的大小排序的。

进行主成分分析的目的之一是为了减少变量的个数，所以一般不会取 p 个主成分，而是取 $m < p$ 个主成分， m 取多少比较合适，这是一个很实际的问题，通常以所取 m 使得累积贡献率达到85%以上为宜，即

$$\frac{\sum_{i=1}^m \lambda_i}{\sum_{i=1}^p \lambda_i} \geq 85\%$$

这样，既能使损失信息不太多，又达到减少变量，简化问题的目的。另外，选取主成分还可根据特征值的变化来确定。图2-1为SPSS统计软件生成的碎石图。



碎石图

由图可知，第二个及第三个特征值变化的趋势已经开始趋于平稳，所以，取前两个或是前三个主成分是比较合适的。这种方法确定的主成分个数与按累积贡献率确定的主成分个数往往是一致的。在实际应用中有些研究工作者习惯于保留特征值大于1的那些主成分，但这种方法缺乏完善的理论支持。在大多数情况下，当 $m=3$ 时即可使所选主成分保持信息总量的比重达到85%以上。

(二) 从相关阵出发求解主成分

考虑如下的数学变换：

$$\text{令： } Z_i = \frac{X_i - \mu_i}{\sqrt{\sigma_{ii}}} \quad i = 1, 2, \dots, p$$

其中, μ_i 与 σ_{ii} 分别表示变量 X_i 的期望与方差。于是有

$$E(Z_i) = 0 \quad \text{var}(Z_i) = 1$$

$$\text{令： } \Sigma^{1/2} = \begin{pmatrix} \sqrt{\sigma_{11}} & 0 & \cdots & 0 \\ 0 & \sqrt{\sigma_{22}} & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \sqrt{\sigma_{pp}} \end{pmatrix}$$

于是，对原始变量 X 进行标准化： $Z = (\Sigma^{1/2})^{-1}(X - \mu)$

经过上述标准化后，显然有 $E(\mathbf{Z}) = \mathbf{0}$

$$\text{cov}(\mathbf{Z}) = (\sum^{1/2})^{-1} \sum (\sum^{1/2})^{-1} = \begin{pmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{12} & 1 & \cdots & \rho_{2p} \\ \vdots & \vdots & & \vdots \\ \rho_{1p} & \rho_{2p} & \cdots & 1 \end{pmatrix} = R$$

由于上面的变换过程，原始变量 X_1, X_2, \dots, X_p 的相关阵实际上就是对原始变量标准化后的协方差矩阵，因此，由相关矩阵求主成分的过程与主成分个数的确定准则实际上是与由协方差矩阵出发求主成分的过程与主成分个数的确定准则是相一致的，在此不再赘述。仍用 λ_i, γ_i 分别表示相关阵 R 的特征值与对应的标准正交特征向量，此时，求得的主成分与原始变量的关系式为：

$$Y_i = \gamma_i' \mathbf{Z} = \gamma_i' (\sum^{1/2})^{-1} (\mathbf{X} - \mu), \quad i = 1, 2, \dots, p \quad (2.3)$$

3 相关问题的讨论

3.1 关于由协方差矩阵或相关矩阵出发求解主成分

- 求解主成分的过程实际就是对矩阵结构进行分析的过程，也就是求解特征值的过程。
- 在实际分析过程中，我们可以从原始数据的协方差矩阵出发，也可以从原始数据的相关矩阵出发，其求主成分的过程是一致的。
- 从协方差阵出发和从相关阵出发所求得的主成分一般来说是有差别的，而且这种差别有时候还很大。

【例】 假定我们研究某一经济问题共涉及两个指标：产值和利税。其中产值以百万元计，利税以万元计，得原始资料矩阵如下：

$$\mathbf{X} = \begin{pmatrix} 12.5 & 586 \\ 24 & 754 \\ 15.3 & 850 \\ 18 & 667 \\ 31.2 & 750 \end{pmatrix}$$

可以得到，原始变量的协方差阵与相关阵分别为：

$$\Sigma = \begin{pmatrix} 55.90 & 242.65 \\ 242.65 & 9927.80 \end{pmatrix} \quad \mathbf{R} = \begin{pmatrix} 1 & 0.3257 \\ 0.3257 & 1 \end{pmatrix}$$

由协方差阵出发求解主成分，得到结果见表：

表

特征值	解释方差比例	累积比例
9933.7607	0.9950	0.9950
49.9343	0.0050	1.0000

对应两特征值的标准正交特征向量为：

特征值 1	特征值 2
0.0246	0.9997
0.9997	-0.0246

因此，所得的主成分的表达式为：

$$Y_1 = 0.0246X_1 + 0.9997X_2$$

$$Y_2 = 0.9997X_1 - 0.0246X_2$$

其中，第一主成分保留了原始变量99.50%的信息，我们在分析中就可以把第二主成分舍掉，这样达到简化问题的目的。

由相关矩阵求解主成分的结果见表：

表

特征值	解释方差比例	累积比例
1.3257	0.6629	0.6629
0.6742	0.3371	1.0000

对应两特征值的标准正交特征向量为：

特征值 1	特征值 2
0.7071	0.7071
0.7071	-0.7071

此时，所得主成分的表达式为：

$$Y_1 = 0.7071Z_1 + 0.7071Z_2 = 0.7071\left(\frac{X_1 - \bar{X}_1}{\sqrt{55.9}}\right) + 0.707\left(\frac{X_2 - \bar{X}_2}{\sqrt{9927.80}}\right)$$

$$Y_2 = 0.7071Z_1 - 0.7071Z_2 = 0.7071\left(\frac{X_1 - \bar{X}_1}{\sqrt{55.9}}\right) - 0.707\left(\frac{X_2 - \bar{X}_2}{\sqrt{9927.80}}\right)$$

由从相关矩阵出发求解主成分的结果可知，第一主成分保留了原始变量66.29%的信息。

我们对原始变量转换成同一度量单位再求主成分。对产值与
利税均以万元计，原始数据资料阵变为以下形式：

$$\mathbf{X}_1 = \begin{pmatrix} 1250 & 586 \\ 2400 & 754 \\ 1530 & 850 \\ 1800 & 667 \\ 3120 & 750 \end{pmatrix} \quad \text{相关矩阵没有变化，协方差矩阵变为: } \Sigma_1 = \begin{pmatrix} 558950 & 24265 \\ 24265 & 9927.8 \end{pmatrix}$$

由此协方差矩阵出发重新求主成分，结果见表：
表

特征值	解释方差比例	累积比例
560020.348	0.9844	0.9844
8857.452	0.0156	1.0000

对应两特征值的标准正交特征向量见下表：

特征值 1	特征值 2
0.999029	-0.044068
0.044068	0.999029

此时所得主成分的表达式为：

$$Y_1 = 0.999029 X_1 - 0.044068 X_2$$

$$Y_2 = 0.044068 X_1 + 0.999029 X_2$$

其中，第一主成分保留了原始变量98.44%的信息

可见，第一主成分保留原始变量的信息与主成分与原始变量的关系式均与上两种情况有很大差别，那么，究竟哪种方法得到的结果更为可信呢，在实际研究中我们应该作何选择呢？

一般而言，对于度量单位不同的指标或是取值范围彼此差异非常大的指标，我们不直接由其协方差矩阵出发进行主成分分析，而应该考虑将数据标准化。比如，在对上市公司的财务状况进行分析时，常常会涉及到利润总额、市盈率、每股净利率等指标，其中利润总额取值常常从几十万到上百万，市盈率取值一般从五到六、七十之间，而每股净利率在1以下，不同指标取值范围相差很大，这时若是直接从协方差矩阵入手进行主成分分析，明显利润总额的作用将起到重要支配作用，而其它两个指标的作用很难在主成分中体现出来，此时应该考虑对数据进行标准化处理。

但是，对原始数据进行标准化处理后倾向于各个指标的作用在主成分的构成中相等。对于取值范围相差不大或是度量相同的指标进行标准化处理后，其主成分分析的结果仍与由协方差阵出发求得的结果有较大区别。其原因是**由于对数据进行标准化的过程实际上也就是抹杀原始变量离散程度差异的过程，标准化后的各变量方差相等均为1**，而实际上方差也是对数据信息的重要概括形式，也就是说，对原始数据进行标准化后抹杀了一部分重要信息，因此才使得标准化后各变量在对主成分构成中的作用趋于相等。由此看来，对同度量或是取值范围在同量级的数据，还是直接从协方差矩阵求解主成分为宜。

3.2 主成分分析不要求数据来自于正态总体

由上面的讨论可知，无论是从原始变量协方差矩阵出发求解主成分，还是从相关矩阵出发求解主成分，均没有涉及到总体分布的问题。也就是说，与很多多元统计方法不同，主成分分析不要求数据来自于正态总体。实际上，主成分分析就是对矩阵结构的分析，其中主要用到的技术是矩阵运算的技术及矩阵对角化和矩阵的谱分解技术。

主成分分析的这一特性大大扩展了其应用范围，对多维数据，只要是涉及降维的处理，我们都可以尝试用主成分分析，而不用花太多精力考虑其分布情况。

3.3 主成分分析与重叠信息

➤首先应当认识到主成分分析方法适用于变量之间存在较强相关性的数据，如果原始数据相关性较弱，运用主成分分析后不能起到很好的降维作用，即所得的各个主成分浓缩原始变量信息的能力差别不大。一般认为当原始数据大部分变量的相关系数都小于0.3时，运用主成分分析不会取得很好的效果。

➤主成分分析对重叠信息的剔除是无能为力的，同时主成分分析还损失了一部分信息。因此在选取初始变量进入分析时应该小心，对原始变量存在多重共线性的问题，在应用主成分分析方法时一定要慎重。

如果所得到的样本协方差矩阵（或是相关阵）最小的特征值接近于0，那么就有

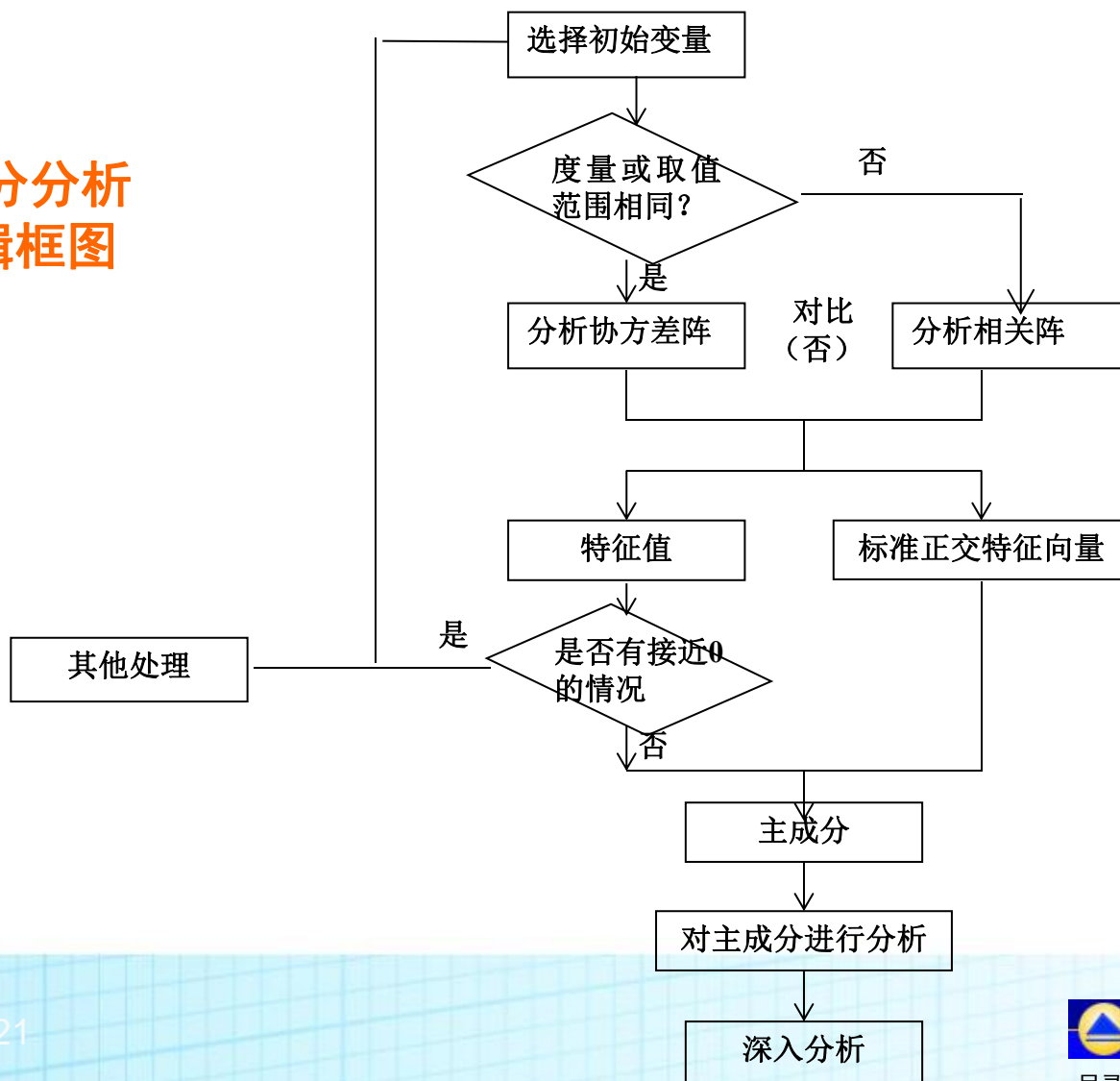
$$\sum \gamma_p = (\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})' \gamma_p = \lambda_p \gamma_p \approx 0 \quad (4.1)$$

进而推出 $(\mathbf{X} - \boldsymbol{\mu})' \gamma_p \approx 0 \quad (4.2)$

这就意味着，中心化以后的原始变量之间存在着多重共线性，即原始变量存在着不可忽视的重叠信息。**因此，在进行主成分分析得出协方差阵或是相关阵发现最小特征根接近于零时，应该注意对主成分的解释，或者考虑对最初纳入分析的指标进行筛选，**由此可以看出，虽然主成分分析不能有效地剔除重叠信息，但它至少可以发现原始变量是否存在着重叠信息，这对我们减少分析中的失误是有帮助的。

4 主成分分析SPSS实现

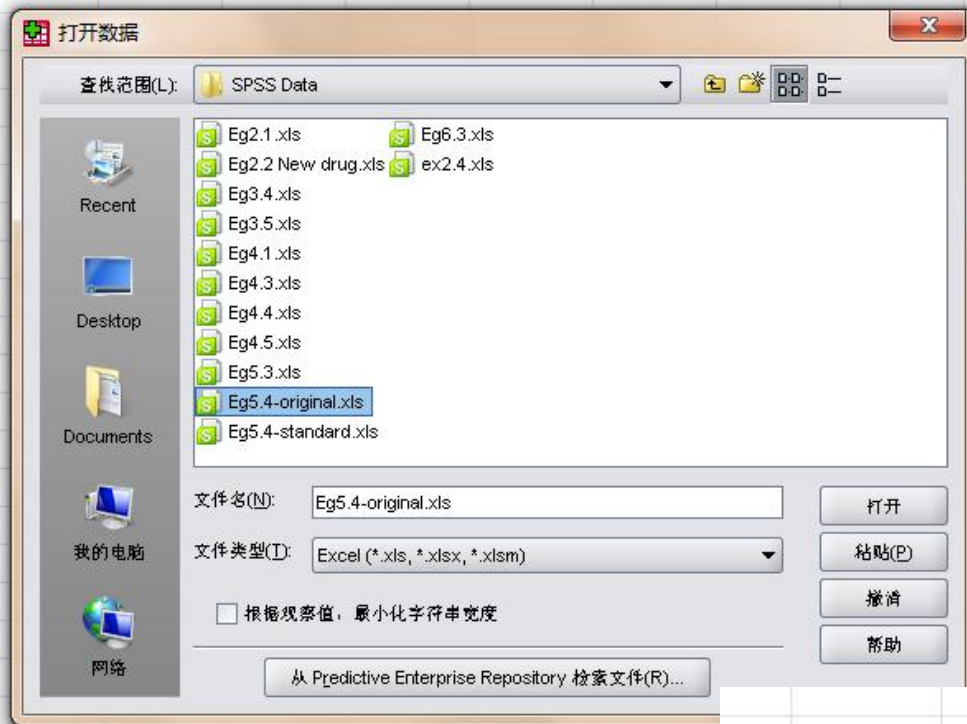
主成分分析的逻辑框图



【例】 根据下面的数据，对全国重点水泥企业经济效益进行综合评价，并提出改进方案。

厂家编号及指标	固定资产 利税率	资金利 税率	销售收入 利税率	资金利润 率	固定资产 产值率	流动资 金周转 天数	万元产 值能耗	全员劳动生产 率
1 琉璃河	16.68	26.75	31.84	18.4	53.25	55	28.83	1.75
2 邯郸	19.7	27.56	32.94	19.2	59.82	55	32.92	2.87
3 大同	15.2	23.4	32.98	16.24	46.78	65	41.69	1.53
4 哈尔滨	7.29	8.97	21.3	4.76	34.39	62	39.28	1.63
5 华新	29.45	56.49	40.74	43.68	75.32	69	26.68	2.14
6 湘乡	32.93	42.78	47.98	33.87	66.46	50	32.87	2.6
7 柳州	25.39	37.82	36.76	27.56	68.18	63	35.79	2.43
8 峨嵋	15.05	19.49	27.21	14.21	6.13	76	35.76	1.75
9 耀县	19.82	28.78	33.41	20.17	59.25	71	39.13	1.83
10 永登	21.13	35.2	39.16	26.52	52.47	62	35.08	1.73
11 工源	16.75	28.72	29.62	19.23	55.76	58	30.08	1.52
12 抚顺	15.83	28.03	26.4	17.43	61.19	61	32.75	1.6
13 大连	16.53	29.73	32.49	20.63	50.41	69	37.57	1.31
14 江南	22.24	54.59	31.05	37	67.95	63	32.33	1.57
15 江油	12.92	20.82	25.12	12.54	51.07	66	39.18	1.83

从SPSS读取 外部数据



导入spss中计算出其相关阵R如下，见下表：

相关矩阵

		固定资产利税率	资金利税率	销售收入利税率	资金利润率	固定资产产值率	流动资金周转天数	万元产值能耗	全员劳动生产率
相关	固定资产利税率	1.000	.849	.923	.902	.651	-.265	-.502	.598
	资金利税率	.849	1.000	.690	.988	.723	-.103	-.595	.265
	销售收入利税率	.923	.690	1.000	.774	.544	-.317	-.344	.531
	资金利润率	.902	.988	.774	1.000	.688	-.106	-.592	.329
	固定资产产值率	.651	.723	.544	.688	1.000	-.444	-.433	.359
	流动资金周转天数	-.265	-.103	-.317	-.106	-.444	1.000	.375	-.434
	万元产值能耗	-.502	-.595	-.344	-.592	-.433	.375	1.000	-.254
	全员劳动生产率	.598	.265	.531	.329	.359	-.434	-.254	1.000

在确定主成分个数之前，与前例相同的SPSS操作，得出软件输出结果如下：

输出结果（1）

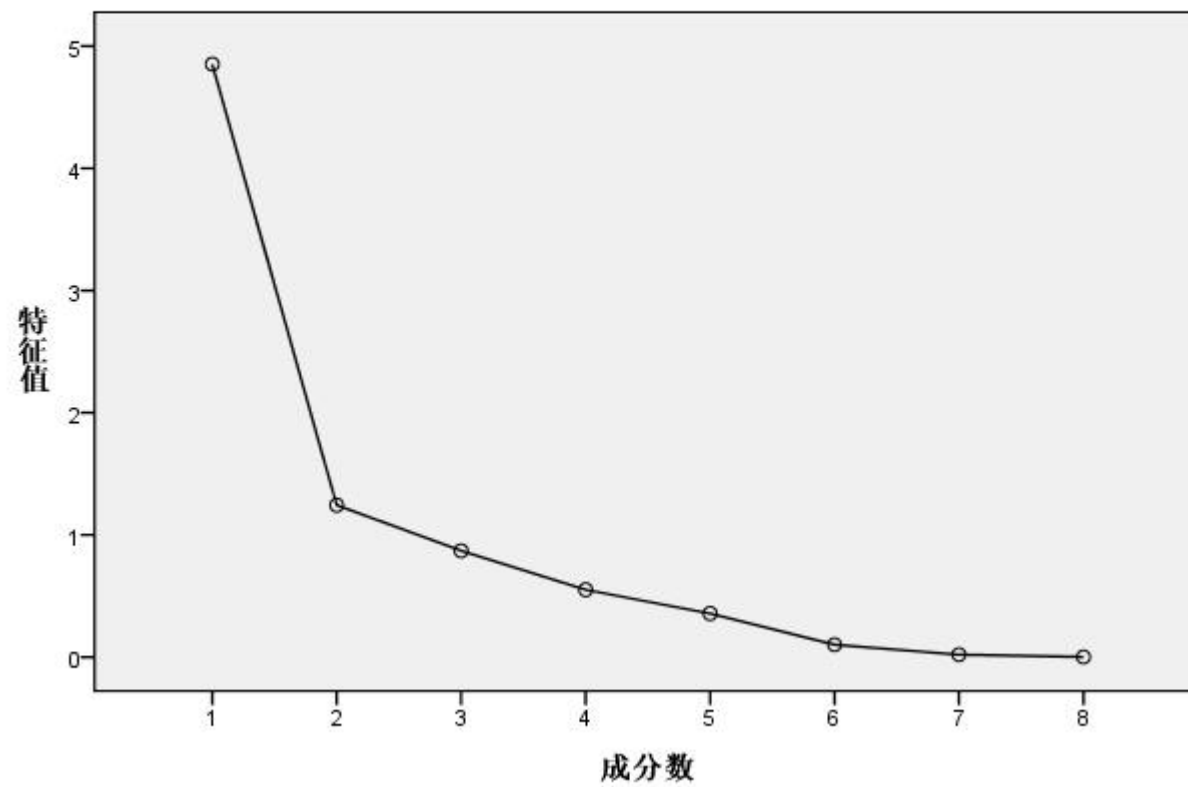
解释的总方差

成份	初始特征值			提取平方和载入		
	合计	方差的 %	累积 %	合计	方差的 %	累积 %
1	4.853	60.660	60.660	4.853	60.660	60.660
2	1.244	15.549	76.209	1.244	15.549	76.209
3	.870	10.878	87.087	.870	10.878	87.087
4	.552	6.898	93.984			
5	.357	4.463	98.447			
6	.102	1.275	99.722			
7	.021	.259	99.981			
8	.002	.019	100.000			

提取方法：主成份分析。

输出结果 (2)

碎石图



元件矩阵^a

	元件		
	1	2	3
固定资产利税率	.955	.069	.241
资金利税率	.901	.381	-.110
销售收入利税率	.858	-.013	.354
资金利润率	.928	.354	-.018
固定资产产值率	.790	-.073	-.207
流动资金周转天数	-.404	.809	.297
万元产值能耗	-.648	.040	.593
全员劳动生产率	.570	-.555	.438

撷取方法：主體元件分析。

a. 撷取 3 個元件。

从上表及上图可看出，前3个主成分解释了全部方差的87.085%，也即包含了原始数据的信息总量达到了87.085%，这说明前三个主成分代表原来的8个指标评价企业的经济效益已经有足够的把握。设这3个主成分分别用 y_1, y_2, y_3 来表示，在点击“抽取”按钮时，在“因子个数”中填写3，即可得到相关矩阵的前三个特征根的特征向量，见右表

由上表，三个主成分的线性组合如下：

$$y_1 = \frac{1}{\sqrt{4.853}} (0.955x_1^* + 0.901x_2^* + 0.858x_3^* + 0.928x_4^* + 0.790x_5^* - 0.404x_6^* - 0.648x_7^* + 0.570x_8^*)$$

$$y_2 = \frac{1}{\sqrt{1.244}} (0.069x_1^* + 0.381x_2^* - 0.013x_3^* + 0.354x_4^* - 0.073x_5^* + 0.809x_6^* + 0.040x_7^* - 0.555x_8^*)$$

$$y_3 = \frac{1}{\sqrt{0.87}} (0.241x_1^* - 0.110x_2^* + 0.354x_3^* - 0.018x_4^* - 0.207x_5^* + 0.297x_6^* + 0.593x_7^* + 0.438x_8^*)$$

主成分的经济意义由各线性组合中权数较大的几个指标的综合意义来确定。综合因子 y_1 中 x_1, x_2, x_3, x_4, x_5 的系数远大于其他变量的系数，所以， y_1 主要是固定资产利税率、资金利税率、销售收入利税率、资金利润率等五个指标的综合反映，它代表着经济效益的盈利方面，刻画了企业的盈利能力。因为由 y_1 来评价企业的经济效益已有60.76%的把握，所以这四项指标是反映企业经济效益的主要指标。

同时，从 y_1 的线性组合中可以看到前四个单项指标在综合因子 y_1 中所占的比重相当，这进而说明这四项目标用于考核评价企业经济效益每一项都是必不可少的。 y_2 主要是流动资金周转天数和全员劳动生产率的综合反映，它标志着企业的资金和人力的利用水平，以资金和个人的利用率作用于企业的经济效益。资金和人力利用得好，劳动生产率就提高，资金周转就加快，从而提高企业经济效益。 y_3 主要反映万元产值能耗，从改进生产工艺、勤俭节约方面作用于企业经济效益。这三个综合因子从三个影响企业经济效益的主要方面刻画企业经济效益，用它们来考核企业经济效益具有87.085%的可靠性。

表 主成分得分

	\hat{y}_1	\hat{y}_2	\hat{y}_3
琉璃河	0.02243	0.71623	1.83638
邯 郸	0.38122	1.97497	-0.09623
大 同	-0.71186	-0.15006	-0.88207
哈尔滨	-1.6961	0.76569	0.39536
华 新	1.79484	-1.52973	0.40186
湘 乡	1.76418	1.63618	-0.85405
柳 州	0.73074	0.28641	-1.12474
峨 嵋	-1.2721	-0.63683	-0.95769
耀 县	-0.21511	-0.53695	-1.18281
永 登	0.30076	-0.27676	-0.52736
工 源	-0.12225	0.07397	1.68818
抚 顺	-0.34114	-0.04624	1.1871
大 连	-0.48376	-1.02203	-0.34845
江 南	0.78171	-1.42385	0.70415
江 油	-0.93356	0.16899	-0.23962

最后，按照式

$$F = 0.60758 \times \hat{y}_1 + 0.15865 \times \hat{y}_2 + 0.10463 \times \hat{y}_3$$

计算各企业经济效益的综合得分，由综合得分可排出企业经济效益的名次。各主成分得分、综合得分及排名见下表。

表 主成排名

水泥厂名	盈利能力方面		资金和人力利用方面		产值能耗方面		综合效益评价	
	\hat{y}_1	名次	\hat{y}_2	名次	\hat{y}_3	名次	F	名次
琉璃河	0.02243	7	0.71623	4	1.83638	1	0.319398	6
邯 郸	0.38122	5	1.97497	1	-0.09623	7	0.534882	3
大 同	-0.71186	12	-0.15006	9	-0.88207	12	-0.54861	12
哈尔滨	-1.6961	15	0.76569	3	0.39536	6	-0.86767	14
华 新	1.79484	1	-1.52973	15	0.40186	5	0.889864	2
湘 乡	1.76418	2	1.63618	2	-0.85405	11	1.242101	1
柳 州	0.73074	4	0.28641	5	-1.12474	14	0.37174	4
峨 嵋	-1.2721	14	-0.63683	12	-0.95769	13	-0.97414	15
耀 县	-0.21511	9	-0.53695	11	-1.18281	15	-0.33964	10
永 登	0.30076	6	-0.27676	10	-0.52736	10	0.08365	8
工 源	-0.12225	8	0.07397	7	1.68818	2	0.114093	7
抚 顺	-0.34114	10	-0.04624	8	1.1871	3	-0.0904	9
大 连	-0.48376	11	-1.02203	13	-0.34845	9	-0.49253	11
江 南	0.78171	3	-1.42385	14	0.70415	4	0.322733	5
江 油	-0.93356	13	0.16899	6	-0.23962	8	-0.56547	13

在经济效益得分中，有许多企业的得分是负数，但并不是企业的经济效益就为负。这里的正负仅表示该企业与平均水平的位置关系，企业的经济效益的平均水平算作零点，这是我们在整个过程中将数据标准化的结果。

可看到，湘乡水泥厂的综合经济效益最好，是第一名；华新水泥厂的综合经济效益为第二名；……，峨嵋水泥厂的综合经济效益最差。从影响企业经济效益的三个主要因子的得分看，峨嵋水泥厂不管在企业盈利能力、资金和人力利用及产能消耗方面，都处于最差地位，因此，他们反映出峨嵋水泥厂在盈利能力方面缺乏活力，资金和人力利用率也不高，产值能耗也相对较高。企业要改变落后的状况，只能改进各项工作，提高经济效益。华新水泥厂的盈利能力最强，但这个厂的资金和人力利用效率最差，这似乎是个矛盾。

有的管理者认为只要企业盈利就一好百好，因而忽视企业的资金周转，不注重提高劳动生产率。然而，这种经济效益好、盈利能力强可能是由于企业具有得天独厚的优越条件。华新水泥厂若能正视自己，努力加快资金周转，进一步提高劳动生产率，保持自己强有力的盈利能力，该厂的经济效益从而会更好，将会立足于全国重点水泥厂的最前列。

虽然此处可以根据各上市公司的主成分得分对各公司运营情况进行一些比较分析或分类研究，但因此处主成分的意义不十分明朗，我们把更深入的分析放到下一章，以期得到更合理，更容易解释的结果。

Part II 因子分析

1 因子分析的基本原理

因子分析（factor analysis）模型是主成分分析的推广。它也是利用降维的思想，由研究原始变量相关矩阵内部的依赖关系出发，把一些具有错综复杂关系的变量归结为少数几个综合因子的一种多变量统计分析方法。相对于主成分分析，因子分析更倾向于描述原始变量之间的相关关系；因此，因子分析的出发点是原始变量的相关矩阵。

(一) Charles Spearman提出因子分析时用到的例子

为了对因子分析的基本理论有一个完整的认识，我们先给出Charles Spearman 1904年用到的例子。在该例中Spearman研究了33名学生在古典语（C）、法语（F）、英语（E）、数学（M）、判别（D）和音乐（Mu）六门考试成绩之间的相关性并得到如下相关阵：

	C	F	E	M	D	Mu
C	1.00	0.83	0.78	0.70	0.66	0.63
F	0.83	1.00	0.67	0.67	0.65	0.57
E	0.78	0.67	1.00	0.64	0.54	0.51
M	0.70	0.67	0.64	1.00	0.45	0.51
D	0.66	0.65	0.54	0.45	1.00	0.40
Mu	0.63	0.57	0.51	0.51	0.40	1.00

Spearman注意到上面相关阵中一个有趣的规律，这就是如果不考虑对角元素的话，任意两列的元素大致成比例，对C列和E列有：

$$\frac{0.83}{0.67} \approx \frac{0.70}{0.64} \approx \frac{0.66}{0.54} \approx \frac{0.63}{0.51} \approx 1.2$$

于是Spearman指出每一科目的考试成绩都遵从以下形式：

$$X_i = a_i F + e_i \quad (1.1)$$

式中， X_i 为第 i 门科目标标准化后的考试成绩，均值为0，方差为1。 F 为公共因子，对各科考试成绩均有影响，是均值为0，方差为1。 e_i 为仅对第 i 门科目考试成绩有影响的特殊因子， F 与 e_i 相互独立。也就是说，每一门科目的考试成绩都可以看作是由一个公共因子（可以认为是一般智力）与一个特殊因子的和。

对Spearman的例子进行推广，假定每一门科目的考试成绩都受到 m 个公共因子的影响及一个特殊因子的影响，于是 (1.1) 就变成了如下因子分析模型的一般形式：

$$X_i = a_{i1}F_1 + a_{i2}F_2 + \cdots + a_{im}F_m + e_i \quad (1.2)$$

式中， X_i 为标准化后的第 i 门科目的考试成绩，均值为0，方差为1。 F_1, F_2, \dots, F_m 是彼此独立的公共因子，都满足均值为0，方差为1。 e_i 为特殊因子，与每一个公共因子均不相关且均值为0。则 $a_{i1}, a_{i2}, \dots, a_{im}$ 为对第 i 门科目考试成绩的因子载荷。对该模型，有：

$$\text{var}(X_i) = a_{i1}^2 + a_{i2}^2 + \cdots + a_{im}^2 + \text{var}(e_i) = 1 \quad (1.3)$$

式中， $a_{i1}^2 + a_{i2}^2 + \cdots + a_{im}^2$ 表示公共因子解释 X_i 方差的比例，称为 X_i 的共同度，相对的 $\text{var}(e_i)$ 可称为 X_i 的特殊度或剩余方差，表示 X_i 的方差中与公共因子无关的部分。

1.1 一般因子分析模型

一般的因子分析模型可以写成：

[illegible]

称为因子模型，模型(6.7)式的矩阵形式为：

$$\mathbf{X} = \mathbf{A}\mathbf{F} + \boldsymbol{\varepsilon} \quad (1.5)$$

其中

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ a_{p1} & a_{p2} & \cdots & a_{pm} \end{bmatrix}$$

A: 因子载荷阵

因子载荷阵 \mathbf{A} 的统计意义及公共因子和原始变量的关系

1.2. 因子载荷 a_{ij} 的统计意义

由模型

$$\begin{aligned}\text{cov}(X_i, F_j) &= \text{cov}\left(\sum_{j=1}^m a_{ij} F_j + \varepsilon_i, F_j\right) \\ &= \text{cov}\left(\sum_{j=1}^m a_{ij} F_j, F_j\right) + \text{cov}(\varepsilon_i, F_j) \\ &= a_{ij}\end{aligned}$$

即 a_{ij} 是 X_i 与 F_j 的协方差，而注意到， X_i 与 F_j ($i = 1, 2, \dots, p; j = 1, 2, \dots, m$) 都是均值为0，方差为1的变量，因此， a_{ij} 同时也是 X_i 与 F_j 的相关系数。

1.3 变量共同度与剩余方差

称 $a_{i1}^2 + a_{i2}^2 + \cdots + a_{im}^2$ 为变量 x_i 的共同度，记为 h_i^2 ($i = 1, 2, \dots, p$)。由因子分析模型的假设前提，易得：

$$\text{var}(X_i) = 1 = h_i^2 + \text{var}(\varepsilon_i) \quad (1.6)$$

$$\text{记 } \text{var}(\varepsilon_i) = \sigma_i^2, \text{ 则 } \text{var}(x_i) = 1 = h_i^2 + \sigma_i^2 \quad (1.7)$$

上式表明共同度 h_i^2 与剩余方差 σ_i^2 有互补的关系， h_i^2 越大表明 X_i 对公共因子的依赖程度越大，公共因子能解释 X_i 方差的比例越大，因子分析的效果也就越好。

2 因子分析三步曲

2.1 求解因子载荷

2.2 因子旋转

2.3 因子得分

2.1 因子载荷的求解

因子分析可以分为确定因子载荷，因子旋转及计算因子得分三个步骤。首要的步骤即为确定因子载荷或是根据样本数据确定出因子载荷矩阵 \mathbf{A} 。有很多方法可以完成这项工作，如主成分法，主轴因子法，最小二乘法，极大似然法， α 因子提取法等。

我们将结合主成分分析，介绍主成分法求解因子载荷。

2.1.1 主成分法

- 用主成分法确定因子载荷是在进行因子分析之前先对数据进行一次主成分分析，然后把前面几个主成分作为未旋转的公因子。
- 这种方法所得的特殊因子 $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_p$ 之间并不相互独立，因此，用主成分法确定因子载荷不完全符合因子模型的假设前提，也就是说所得的因子载荷并不完全正确。
- 当共同度较大时，特殊因子所起的作用较小，因而特殊因子之间的相关性所带来的影响就几乎可以忽略。

用主成分法寻找公因子的方法如下：假定从相关阵出发求解主成分，设有 p 个变量，则我们可以找出 p 个主成分。将所得的 p 个主成分按由大到小的顺序排列，记为 Y_1, Y_2, \dots, Y_p ，则主成分与原始变量之间存在如下关系式：

[illegible]

式中， γ_{ij} 为随机向量 \mathbf{X} 的相关矩阵的特征值所对应的特征向量的分量，因为特征向量之间彼此正交，从 \mathbf{X} 到 \mathbf{Y} 的转换关系是可逆的，很容易得出由 \mathbf{Y} 到 \mathbf{X} 的转换关系为：

$$(2.2)$$

我们对上面每一等式只保留前 m 个主成分而把后面的部分用 ε_i 代替，则（2.2）式变为：

(2.3)

式子 (2.3) 在形式上已经与因子模型相一致, 且 Y_i ($i = 1, 2, \dots, m$) 之间相互独立, 且 Y_i 与 ε_i 之间相互独立, 为了把 Y_i 转化成合适的公因子, 现在要做的工作只是把主成分 Y_i 变为方差为 1 的变量。为完成此变换, 必须将 Y_i 除以其标准差, 由上一章主成分分析的知识知其标准差即为特征根的平方根 $\sqrt{\lambda_i}$ 。于是, 令 $F_i = Y_i / \sqrt{\lambda_i}$, $a_{ij} = \sqrt{\lambda_j} \gamma_{ji}$, 则 (2.3) 式变为:

[illegible]

上述模型与因子模型完全一致，这样，就得到了载荷 Λ 矩阵和一组初始公因子（未旋转）。

2.2 因子旋转

不管用何种方法确定初始因子载荷矩阵A，它们都不是唯一的。设 F_1, F_2, \dots, F_m 是初始公共因子，则可以建立如下它们的线性组合得到新的一组公共因子 F'_1, F'_2, \dots, F'_m ，使得， F'_1, F'_2, \dots, F'_m ，彼此相互独立同时也能很好地解释原始变量之间的相关关系。

$$F'_1 = d_{11}F_1 + d_{12}F_2 + \dots + d_{1m}F_m$$

$$F'_2 = d_{21}F_1 + d_{22}F_2 + \dots + d_{2m}F_m$$

.....

$$F'_m = d_{m1}F_1 + d_{m2}F_2 + \dots + d_{mm}F_m$$

这样的线性组合可以找到无数组，由此便引出了因子分析的第二个步骤——因子旋转。建立因子分析模型的目的不仅在于要找公共因子，更重要的是知道每一个公共因子的意义，以便对实际问题进行分析。

对初始公因子进行线性组合，即进行因子旋转，以期找到意义更为明确，实际意义更明显的公因子。

旋转以后，应当使新的因子载荷系数要么尽可能地接近于0，要么尽可能的远离0。因为一个接近于0的载荷 a_{ij} 表明 X_i 与 F_j 的相关性很弱；而一个绝对值比较大的载荷 a_{ij} 则表明公因子 F_j 在很大程度上解释了 X_i 的变化。这样，如果任一原始变量都与某些公共因子存在较强的相关关系，而与另外的公因子之间几乎不相关的话，公共因子的实际意义就会比较容易确定。

2.3 因子得分

- 当因子模型建立起来之后，我们往往需要反过来考察每一个样品的性质及样品之间的相互关系。
- 这需要进行因子分析的第三步骤的分析，即因子得分。因子得分就是公共因子 F_1, F_2, \dots, F_m 在每一个样品点上的得分。
- 这需要我们给出公共因子用原始变量表示的线性表达式，这样的表达式一旦能够得到，就可以很方便的把原始变量的取值代入到表达式中求出各因子的得分值。

即建立如下以公因子为因变量，原始变量为自变量的回归方程：

$$F_j = \beta_{j1}X_1 + \beta_{j2}X_2 + \cdots + \beta_{jp}X_p \quad (j = 1, 2, \cdots, m) \quad (2.4)$$

此处因为原始变量与公因子变量均为标准化变量，因此回归模型中不存在常数项。在最小二乘意义下，可以得到 F 的估计值：

$$\hat{F} = A'R^{-1}X \quad (2.5)$$

式中， A 为因子载荷矩阵， R 为原始变量的相关阵， X 为原始变量向量。

进行因子分析应包括如下几步：

1. 根据研究问题选取原始变量；
2. 对原始变量进行标准化并求其相关阵，分析变量之间的相关性；
3. 求解初始公共因子及因子载荷矩阵；
4. 因子旋转；
5. 因子得分；
6. 根据因子得分值进行进一步分析。

3 因子分析SPSS实现

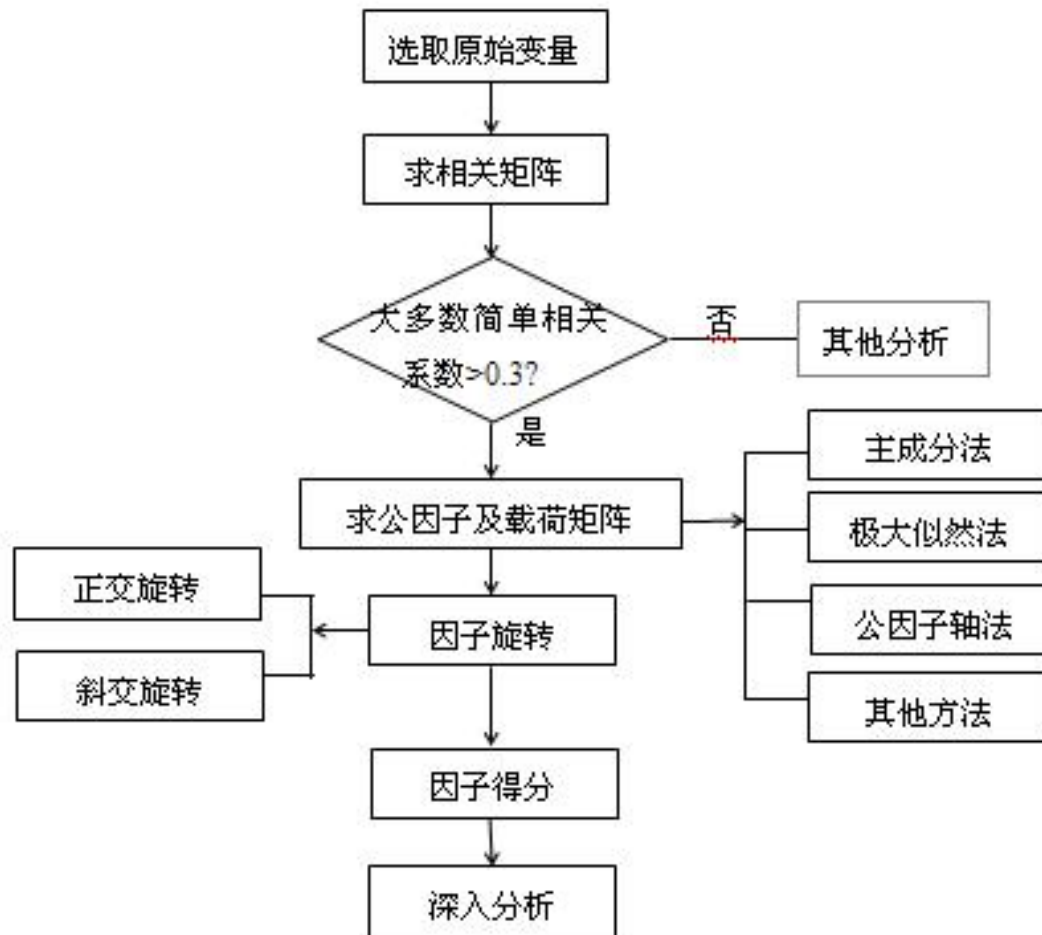


图 因子分析逻辑框图

§ 3.1 因子分析的上机实现

【例】 中心城市的综合发展是带动周边地区经济发展的重要动力。在我国经济发展进程中，各个中心城市一直是该地区经济和社会发展的“引路者”。因而，分析评价全国35个中心城市的综合发展水平，无论是对城市自身的发展，还是对周边地区的进步，都具有十分重要的意义。下面应用因子分析模型，选取反映城市综合发展水平的12个指标作为原始变量([数据](#))，运用SPSS软件，对全国35个中心城市的综合发展水平作分析评价。

1.原始数据及指标解释。我们选取了反映城市综合发展水平的12个指标，其中包括8个社会经济指标，分别为：

x_1 —非农业人口数（万人） x_2 —工业总产值（万元）
 x_3 —货运总量（万吨） x_4 —批发零售住宿餐饮业从业
人数（万人） x_5 —地方政府预算内收入（万元）
 x_6 —城乡居民年底储蓄余额（万元） x_7 —在岗职工人数
（万人） x_8 —在岗职工工资总额（万元）

4个城市公共设施水平的指标：

x_9 —人均居住面积（平方米） x_{10} —每万人拥有公共汽车
数（辆） x_{11} —人均拥有铺装道路面积（平方米）
 x_{12} —人均
公共绿地面积（平方米）

2. 计算运行结果

将标准化后的数据导入到SPSS软件，依次点选**分析-降维**进入**因子分析**对话框。把12个指标变量选入**变量**中，点击抽取按钮，在**方法**选项中选择**主成分分析**(这时，因子分析等同于主成分分析，如果是主成分分析，则只能选择此项)，点击**继续**按钮，回到主对话框点击**确定**。

按照特征根大于1的原则，选入3个公共因子，其累计方差贡献率为84.831%，特征根及累计贡献率、因子载荷矩阵如下。见输出下面的输出结果。

输出结果（1）

解释的总方差

成份	初始特征值			提取平方和载入			旋转平方和载入		
	合计	方差的 %	累积 %	合计	方差的 %	累积 %	合计	方差的 %	累积 %
1	6.388	53.233	53.233	6.388	53.233	53.233	6.258	52.149	52.149
2	2.679	22.324	75.557	2.679	22.324	75.557	2.636	21.969	74.118
3	1.113	9.274	84.831	1.113	9.274	84.831	1.286	10.714	84.831
4	.664	5.534	90.365						
5	.434	3.618	93.983						
6	.305	2.545	96.528						
7	.221	1.840	98.368						
8	.073	.610	98.978						
9	.065	.543	99.521						
10	.031	.262	99.783						
11	.020	.165	99.948						
12	.006	.052	100.000						

提取方法：主成份分析。



输出结果3-5（2）

成份矩阵^a

	成份		
	1	2	3
x1非农业人口数	.879	-.317	.146
x2工业总产值	.852	.261	.264
x3货运总量	.838	-.204	.323
x4批发零售住宿餐饮业从业人数	.780	-.199	-.391
x5地方政府预算内收入	.953	.069	.132
x6城乡居民年底储蓄余额	.981	-.025	-.054
x7在岗职工人数	.930	-.201	-.137
x8在岗职工工资总额	.829	-.076	-.213
x9人均居住面积	.053	.469	.727
x10每万人拥有公共汽车数	.200	.899	-.134
x11人均拥有铺装道路面积	.241	.927	-.063
x12人均公共绿地面积	.234	.699	-.359

提取方法:主成分分析法。

a. 已提取了 3 个成份。

此时得到的未旋转的公共因子的实际意义不好解释，因此，对公共因子进行方差最大化正交旋转。在因子分析对话框中，点击旋转按钮，进入旋转对话框，选中最大方差旋转进行方差最大化正交旋转（若是主成分分析就选择无）。得输出结果

旋转成份矩阵^a

	成份		
	1	2	3
x1非农业人口数	.927	-.180	.049
x2工业总产值	.802	.312	.351
x3货运总量	.875	-.141	.253
x4批发零售住宿餐饮业从业人数	.785	.087	-.421
x5地方政府预算内收入	.930	.196	.164
x6城乡居民年底储蓄余额	.966	.175	-.042
x7在岗职工人数	.944	.030	-.179
x8在岗职工工资总额	.819	.152	-.211
x9人均居住面积	.000	.206	.842
x10每万人拥有公共汽车数	.029	.914	.174
x11人均拥有铺装道路面积	.067	.924	.252
x12人均公共绿地面积	.089	.808	-.104

提取方法 :主成分分析法。
旋转法 :具有 Kaiser 标准化的正交旋转法。

a. 旋转在 4 次迭代后收敛。



由上表结果，原变量 x_1 可由各因子表示为：

$$x_1 = 0.929 \times F_1 - 0.183 \times F_2 + 0.039 \times F_3$$

原变量 x_2 可由各因子表示为：

$$x_2 = 0.806 \times F_1 + 0.308 \times F_2 + 0.345 \times F_3$$

其余依次类推。

为便于得出结论，在**因子分析**主对话框中点击**选项**按钮进入**选项**对话框，在**系数显示格式**中框中选中**按大小排序**使输出的载荷矩阵中各列按载荷系数大小排列，使在同一个公因子上具有较高载荷的变量排在一起。

旋转成份矩阵^a

	成份		
	1	2	3
x6城乡居民年底储蓄余额	.966	.175	-.042
x7在岗职工人数	.944	.030	-.179
x5地方政府预算内收入	.930	.196	.164
x1非农业人口数	.927	-.180	.049
x3货运总量	.875	-.141	.253
x8在岗职工工资总额	.819	.152	-.211
x2工业总产值	.802	.312	.351
x4批发零售住宿餐饮业从业人数	.785	.087	-.421
x11人均拥有铺装道路面积	.067	.924	.252
x10每万人拥有公共汽车数	.029	.914	.174
x12人均公共绿地面积	.089	.808	-.104
x9人均居住面积	.000	.206	.842

提取方法 :主成分分析法。

旋转法 :具有 **Kaiser** 标准化的正交旋转法。

a. 旋转在 4 次迭代后收敛。

最后，计算因子得分，以各因子的方差贡献率占三个因子总方差贡献率的比重作为权重进行加权汇总，得出各城市的综合得分F，即

$$F = (52.149 \cdot F_1 + 21.969 \cdot F_2 + 10.714 \cdot F_3) / 84.831$$

在因子分析主对话框中点击按钮得分进入因子得分对话框，选中变量另存为，在方法中选择回归计算因子得分，如图6-4所示：



FAC1_1	FAC2_1	FAC3_1
3.34852	0.46621	-2.97317
0.94575	-0.64965	0.97571
-0.22699	-0.33713	0.37127
0.02470	-0.18790	-0.87306
-0.81163	-0.20560	0.27423
0.00471	-0.33410	-0.65332
-0.23843	-0.24834	-0.62606
0.14846	-0.23396	-1.54615
3.58806	-0.42785	2.45858
-0.00653	0.85049	-0.59695
0.08209	-0.31025	0.37558
-0.73714	0.34316	0.00082
-0.39445	-0.14409	0.34461
-0.63304	-0.22924	-0.09257
-0.21974	-0.28866	0.89480
-0.33399	-0.28033	0.61394
0.20400	-0.73492	0.03360
-0.45519	-0.13953	0.31229
1.11453	1.25188	-0.56509

得到运行结果并计算综合得分，结果见下表：

表：

城市名	F1	F2	F3	F
北 京	3.37378	0.49045	-3.04248	1.957402
天 津	0.95462	-0.65391	0.96244	0.543583
石 家 庄	-0.2163	-0.33051	0.34879	-0.18579
太 原	-0.38828	-0.38651	-0.25036	-0.37317
呼和浩特	-0.79215	-0.19351	0.25901	-0.52774
沈 阳	0.0078	-0.3284	-0.67948	-0.15083
长 春	-0.23157	-0.24054	-0.64978	-0.27799
哈 尔 滨	0.15122	-0.22201	-1.58673	-0.1279
上 海	3.58255	-0.45691	2.45561	2.428944
南 京	-0.00443	0.85387	-0.61528	0.150699

续表:

杭 州	0.09375	-0.30626	0.35858	0.019334
合 肥	-0.72469	0.35099	-0.009	-0.37365
福 州	-0.37922	-0.13533	0.32624	-0.2422
南 昌	-0.61726	-0.21752	-0.11274	-0.46153
济 南	-0.20362	-0.28378	0.88265	-0.10935
郑 州	-0.29528	-0.28855	0.60459	-0.19844
武 汉	0.21338	-0.72838	0.00467	-0.04972
长 沙	-0.43915	-0.13069	0.29771	-0.28228
广 州	1.12851	1.25536	-0.57767	0.980497

续表:

南 宁	-0.63822	-0.02424	0.17272	-0.39529
海 口	-0.8129	-0.37573	0.96842	-0.51268
成 都	0.2126	-0.36439	0.25407	0.069272
贵 阳	-0.66464	0.31205	-0.96072	-0.44581
昆 明	-0.38986	-0.23118	-0.77627	-0.38998
西 安	-0.13289	-0.49568	-0.92188	-0.30904
兰 州	-0.61179	-0.2757	-0.75932	-0.54122
西 宁	-0.85973	-0.29906	-0.09704	-0.63549

续表:

银 川	-0.89242	0.2069	-0.91362	-0.61312
乌鲁木齐	-0.5715	-0.11384	0.73564	-0.31618
大 连	-0.0403	-0.12453	-0.88242	-0.1508
宁 波	-0.17068	-0.2716	0.34519	-0.14198
厦 门	-0.61332	0.01069	0.99742	-0.28335
青 岛	0.16015	-0.03801	-0.15229	0.0764
深 圳	-0.11722	5.19498	1.26664	1.388438
重 庆	0.92906	-1.1585	1.74667	0.480974

结果分析。由旋转后的因子载荷矩阵可以看出，公共因子 F_1 在 x_1 （非农业人口数）、 x_2 （工业总产值）、 x_3 （货运总量）、 x_4 （批发零售住宿餐饮业从业人数）、 x_5 （地方政府预算内收入）、 x_6 （城乡居民年底储蓄余额）、 x_7 （在岗职工人数）、 x_8 （在岗职工工资总额）上的载荷值都很大， x_1 ， x_7 ， x_8 是反映城市规模的指标， x_2 ， x_3 反映城市工业发展规模， x_4 反映城市第三产业的发展规模， x_5 是政府作为国家的管理者和国有资产的所有者而获得的收入， x_6 则在一定程度上反映了居民的收入水平，而在我国现今的收入分配格局下，政府和居民是再分配收入的获得大户，

有了各个公共因子合理的解释，结合各个城市在三个公共因子上的得分和综合得分，就可对各中心城市的发展水平进行评价了。在城市经济规模因子 F_1 上得分最高的前五个城市依次是上海、北京、广州、天津和重庆，其中，上海的得分为3.58，北京的为3.37，远高于其他城市，这就是说就城市经济发展规模而言，上海、北京是我国最大的城市，且其规模远大于其他城市。城市规模较小，经济发展相对较慢的城市有西宁和银川，而海口由于城市规模小，在 F_1 上的得分也较低。深圳、广州和南京在 F_2 上的得分较高，而重庆、武汉得分较低，说明深圳、广州、南京的城市基础设施在全国是较好的，而重庆等城市的基础设施相对较差，还需要下大力气进行改善。上海、重庆、深圳等城市在 F_3 上的得分比较高，说明居民在居住条件上面较别的城市好，北京、哈尔滨等则需要进行改善。

因而 x_5 , x_6 , 则在一定程度上反映了城市的国民收入水平, 因而 F_1 为反映城市规模及经济发展水平的公共因子, 在这个因子上的得分越高, 城市经济发展水平越高, 城市规模越大; 公共因子 F_2 由于在 x_{10} (每万人拥有公共汽车数)、 x_{11} (人均拥有铺装道路面积)、 x_{12} (人均公共绿地面积) 上的载荷较大, 是反映城市的基础设施水平的公共因子, 在此因子上的得分则反映了一个城市的基础设施水平; 公共因子 F_3 仅在 x_9 (人均居住面积) 上有较大的载荷, 是反映城市居民住房条件的公共因子。

将各城市在三个因子上的得分进行加权综合，就得到了综合得分。根据综合得分就可综合评价城市的发展水平。综合得分前五名的城市依次是上海、北京、深圳、广州和天津；综合得分最低的五个城市依次是西宁、银川、兰州、呼和浩特和海口。再结合各因子得分进行分析，北京在城市规模及经济发展水平，基础设施建设方面均位于前列，但是在居民住房面积上的得分较低，因此，需在这方面加大改善力度。上海在城市规模及经济发展水平及居民住房上得分最高，在基础设施方面得分不太理想，这可能是因为上海人口较多所致。而综合得分较低的城市在经济发展水平上的得分都较低，在城市发展战略上应把经济的发展放在首位，只有经济发展了，城市设施水平及其他方面才能搞上去。

另外，因子得分图分析表明，就城市规模而言，历史悠久的城市大于新兴城市；就城市设施水平而言，南方城市普遍好于北方城市，新兴城市好于老城市；综合来讲，东部地区城市发展水平高于西部地区城市。上海、北京、深圳三城市综合发展水平较接近，上海规模大，但基础设施水平较低；北京规模大、基础设施水平较高，但是居民人均住房较小；深圳规模不大，但是基础设施水平较高、人均住房面积较大。此外，综合得分值大于零的城市还有广州、天津、重庆、南京、青岛、成都、杭州等。但是这些城市与上海、北京及深圳有一定的差距。其他城市综合得分都小于零，在因子得分图中大概位于原点附近，城市综合发展水平都还较低，发展格局也较相近，其中有18个城市位于因子得分图的第三象限，而这些城市多位于中西部地区。因而，如何加快这些城市的发展以带动周边地区的进步，是影响我国整体经济发展的重要课题。

2020-7-21



目录 上页 下页 返回 结束