1. 因子分析

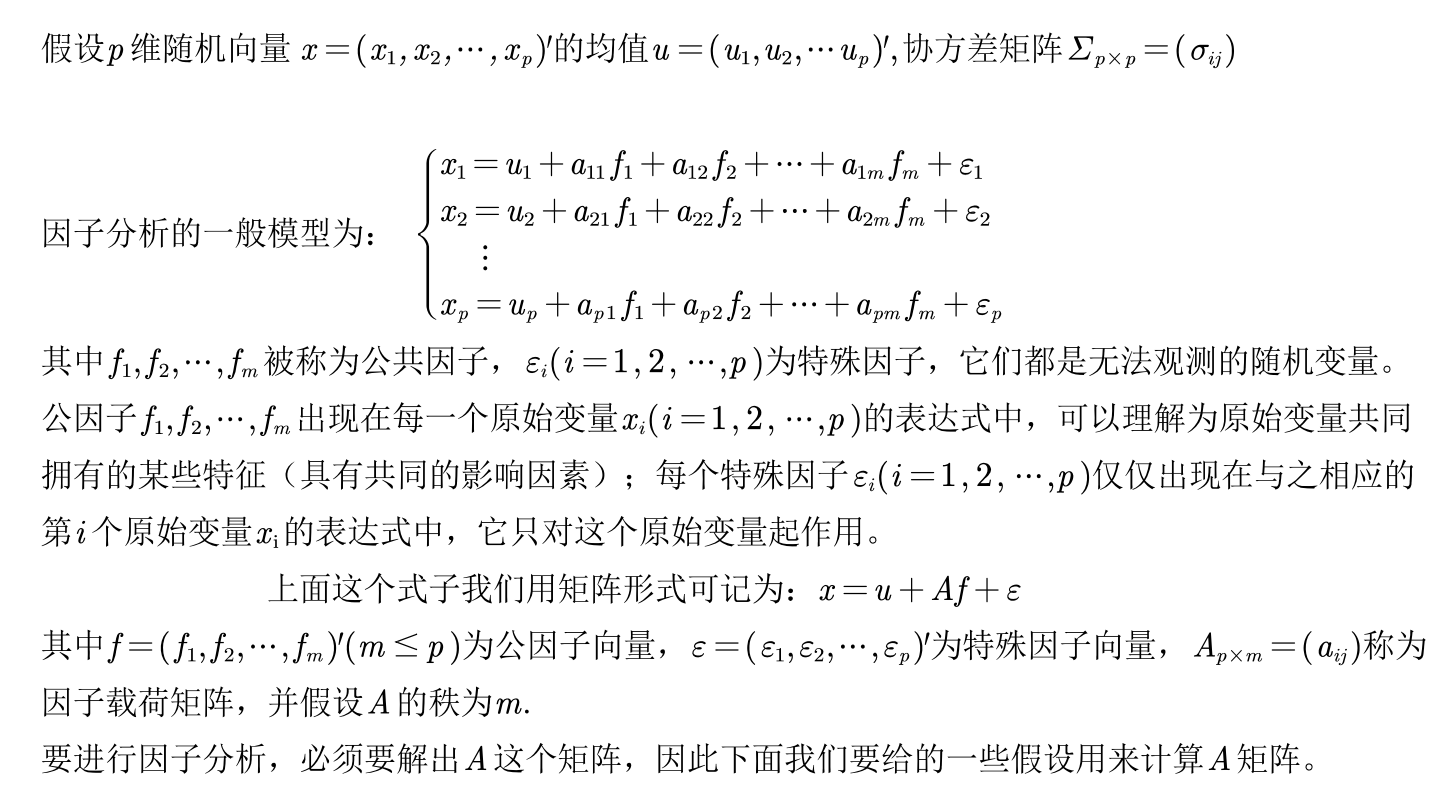
因子分析是一种常用的统计分析方法，用于探究多个变量之间的关系，识别其中存在的共性因素，并将它们组合成更少的维度（因子）来解释变量变异。

因子分析和主成分分析都是用于数据降维的方法，它们有一定的相似性，但是它们的目标和实现方式不同。

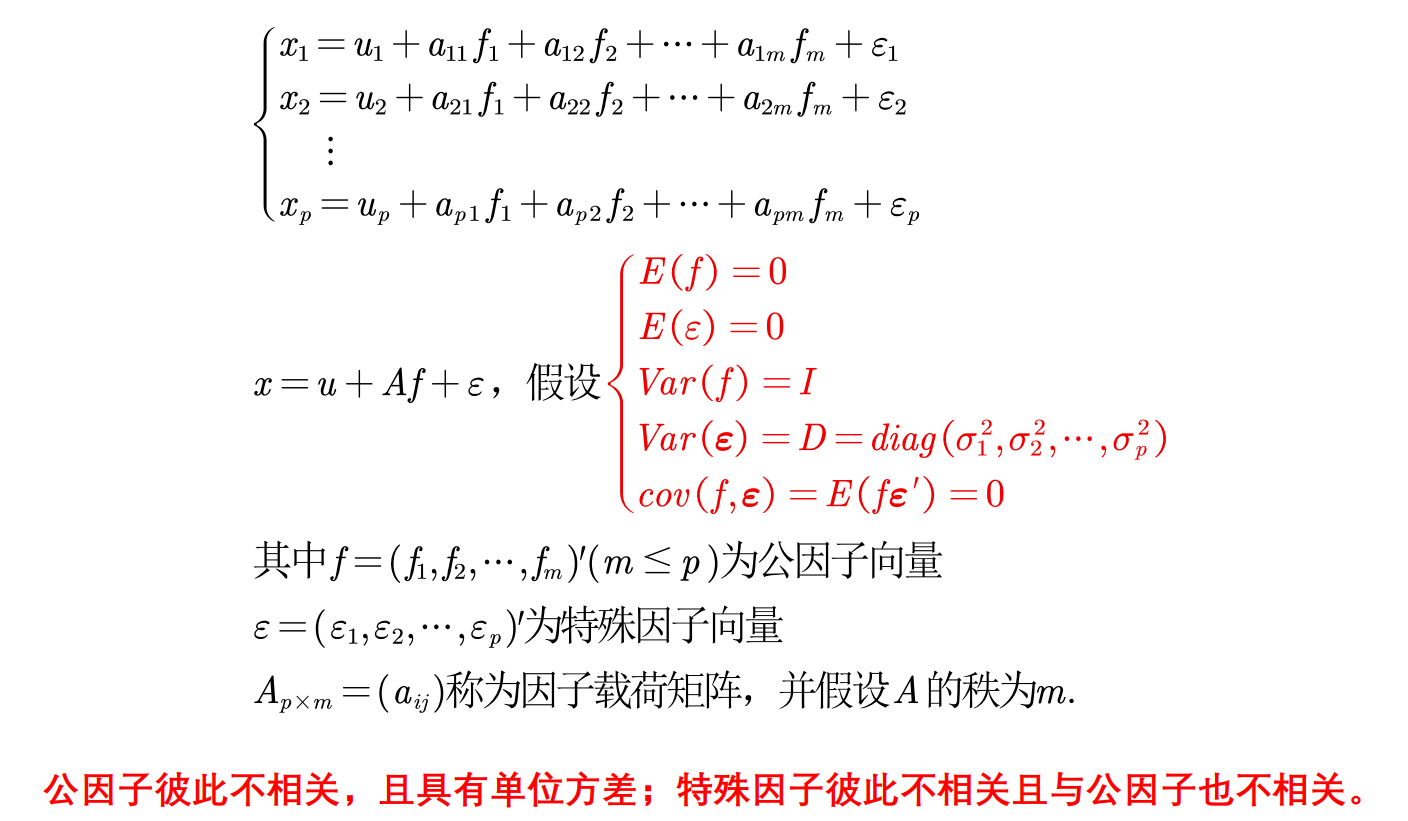
主成分分析是线性变换方法，主要目标是将高维度数据降维到低维度空间，并尽量保留原始数据信息，即通过找到能解释数据方差最多的几个主成分来代表原始数据。通常适用于数值型数据，能识别出变量之间的线性关系，但无法处理非线性关系。

因子分析更侧重于探究多个变量之间的共性因素，以便更好地理解变量之间的关系，更好地解释原始数据，但无法保留所有原始数据信息。

（1）基本原理：



（2）模型假设：



（3）对比主成分分析：

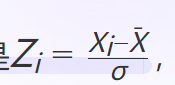
1.主成分分析只是简单的数值计算，不需要构造一个模型，几乎没什么假定;而因子分析需要构造一个因子模型，并伴随几个关键性的假定。

2.主成分的解是唯一的，而因子可有许多解。因子载荷矩阵A 不是唯一的，k可据此通 过因子变换，使新的因子具有更容易解释的实际意义，因此因子解释成功的可能性要远大于主成分解释成功的可能性。

2.数据处理

（1）剔除常数列Net Income Flag：防止KMO检验、球状检验出现无效数据

（2）标准化

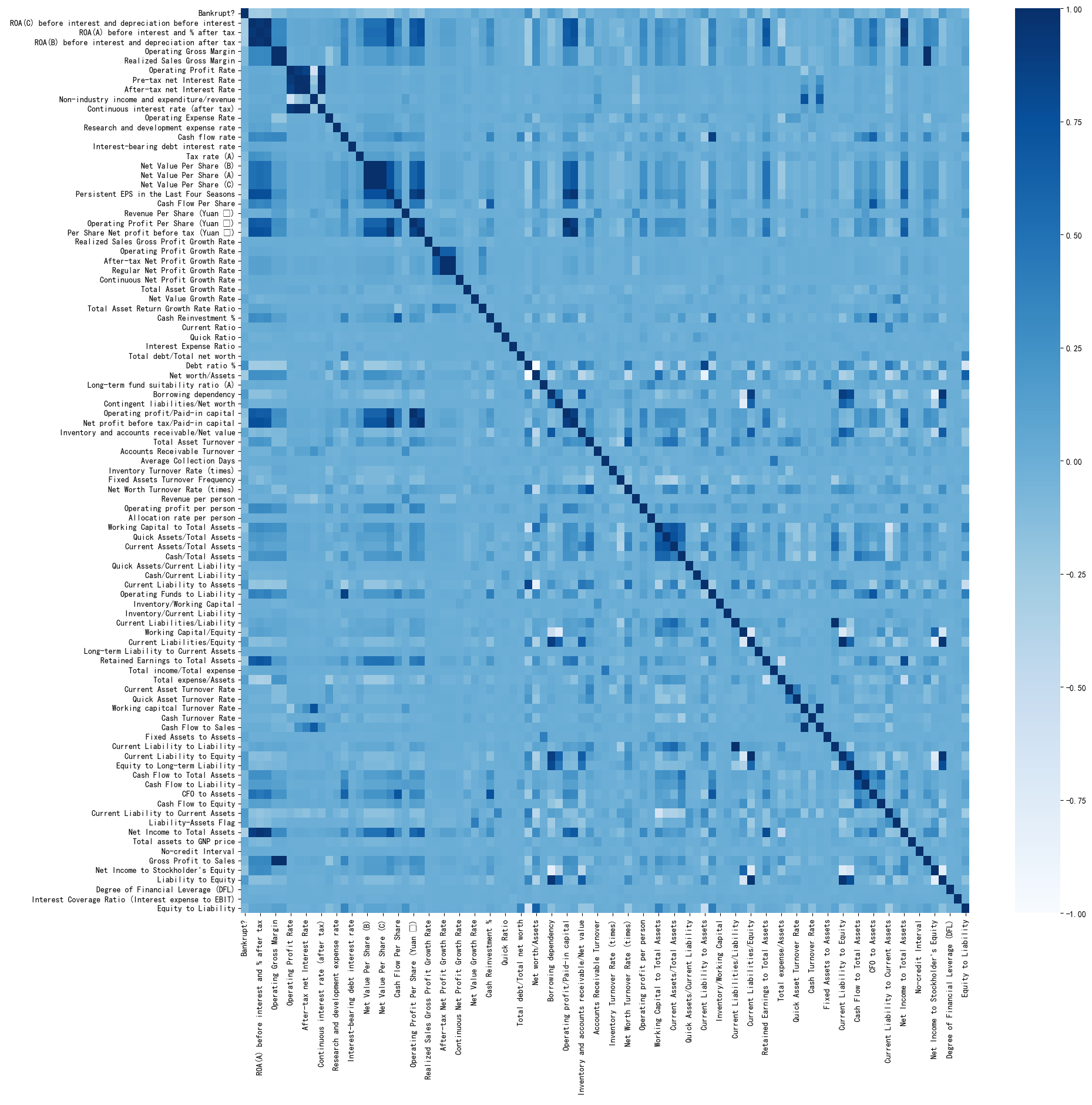


其中*Zi*​是标准化后的指标值，*Xi*​为原始数据，*X*ˉ为原始数据样本均值，*σ*为方差。用于将原始数据转换为均值为0、标准差为1的数据，使得不同量纲的指标能够在因子分析中统一处理。

标准化处理有利于理解因子载荷和因子得分。因子分析中，因子载荷表示原始变量与公共因子的相关性，因子得分是根据因子载荷计算得出的每个样本在公共因子上的得分。通过标准化处理可以得到更稳定和可比较的因子载荷和因子得分，从而更好地解释和分析结果。

（3）无缺失值，异常值

3.计算相关矩阵，热力图



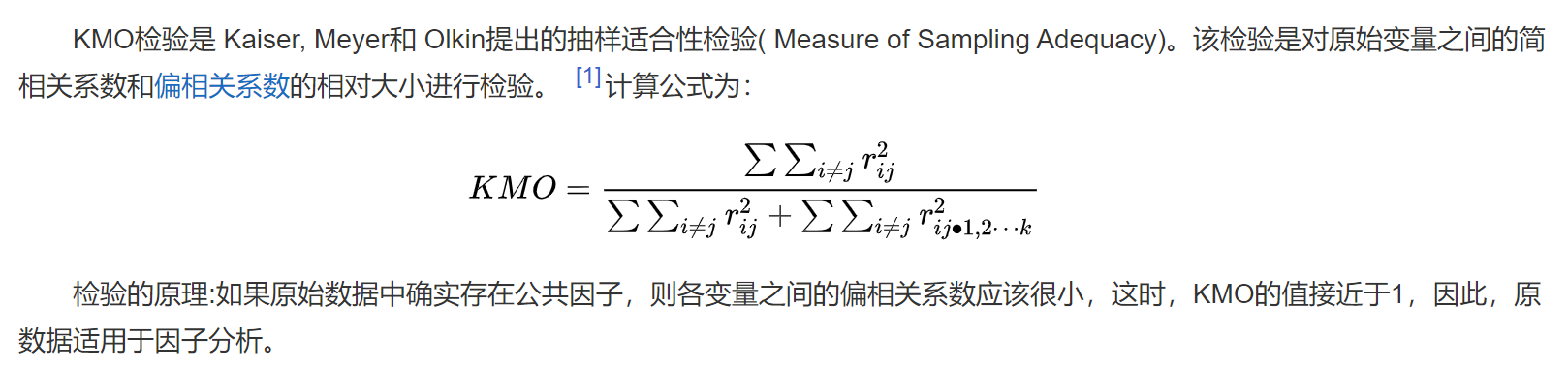
4. 充分性检验

检验总体变量的相关矩阵是否是单位阵，即检验各个变量是否各自独立。

如果不是单位矩阵，说明原变量之间存在相关性，可以进行因子分子；反之，原变量之间不存在相关性，数据不适合进行主成分分析或因子分析

（1）KMO检验

检查变量间的相关性和偏相关性，取值在0-1之间；KOM统计量越接近1，变量间的相关性越强，偏相关性越弱，因子分析的效果越好。

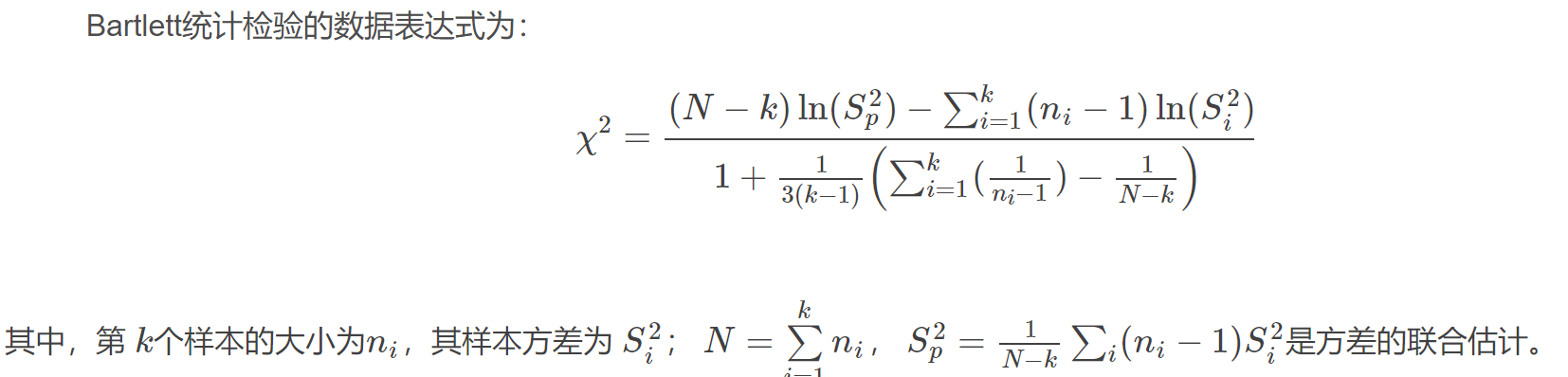


通常取值从0.6开始进行因子分析

KMO检验结果：0.6881235476321951，可以使用因子分析

（2）Bartlett's球状检验

用来对虚假设进行检验，以变量的相关系数矩阵为出发点。它的零假设相关系数矩阵是一个单位阵。巴特利特球形检验的统计量根据相关系数矩阵的行列式得到。如果该值较大，且对应的相伴概率值小于用户心中的显著性水平，那么拒绝零假设，认为相关系数不可能是单位阵，即原始变量之间存在相关性，适合于作因子分析；相反，则不适合作因子分析。



检验结果：(2161528.7119740197, 0.0)（0.0是因为小于机器表示最小值）

结果显示适用于因子分析

5. 选择因子个数

方法：计算相关矩阵的特征值，进行降序排列

特征值大于1的因子有29个，因子分析降维为29个特征

特征值

又称第j个公共因子的方差贡献。即每个变量与某一公共因子负荷量的平方总和（因子载荷矩阵中某一公共因子列所有因子负荷量的平方和）。此处降序排列

[ 1.26390058e+01 6.78742313e+00 4.73449429e+00 4.44105256e+00

3.98063606e+00 3.04214298e+00 2.86649579e+00 2.75924943e+00

2.63574669e+00 2.06502443e+00 1.94427832e+00 1.86704041e+00

1.68164280e+00 1.50795328e+00 1.48694376e+00 1.44309565e+00

1.42581491e+00 1.31365505e+00 1.23599311e+00 1.21194062e+00

1.17323408e+00 1.14536517e+00 1.09137038e+00 1.05436320e+00

1.04508630e+00 1.03439030e+00 1.01273108e+00 1.00961858e+00

1.00704235e+00 9.85209400e-01 9.77814252e-01 9.65642155e-01

9.58251120e-01 9.52762178e-01 9.39505040e-01 9.27654011e-01

9.11729774e-01 9.09149456e-01 8.90574612e-01 8.58505877e-01

8.50031700e-01 8.31141753e-01 8.16716398e-01 7.97501955e-01

7.77654266e-01 7.56408250e-01 7.15270259e-01 6.85932818e-01

6.57164817e-01 5.97929619e-01 5.64786359e-01 5.40224054e-01

5.38350271e-01 4.98351742e-01 4.64763947e-01 4.23002800e-01

4.11679978e-01 3.84199563e-01 3.23981517e-01 3.23660075e-01

2.49643491e-01 2.43240870e-01 2.15963165e-01 2.14418380e-01

1.68731552e-01 1.65497671e-01 1.54628885e-01 1.40238271e-01

1.12206600e-01 8.46964484e-02 8.20962145e-02 6.35570157e-02

5.34183653e-02 3.72401468e-02 2.90725128e-02 2.45372850e-02

2.19562537e-02 1.92898608e-02 1.51698740e-02 9.50742574e-03

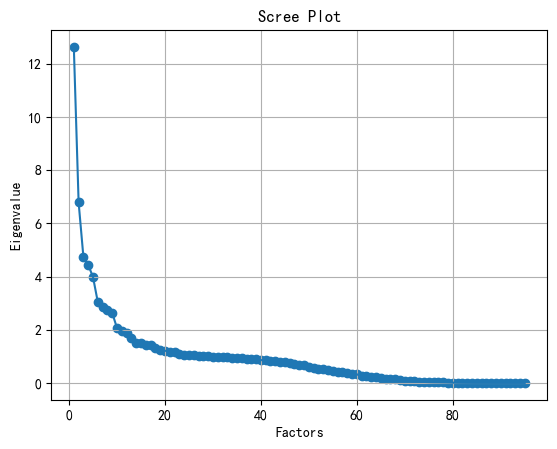
5.60704438e-03 3.76732201e-03 2.65216651e-03 1.61350363e-03

9.17762679e-04 7.80370156e-04 6.45806648e-04 3.76850153e-04

1.48399003e-04 7.41407087e-09 1.05451267e-11 1.25425650e-16

-7.02976658e-17 -7.09438348e-17 -2.75740000e-16]

特征值和因子个数的变化关系图：



6.因子旋转

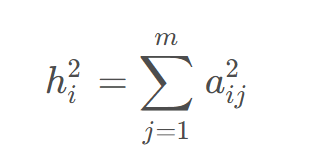
（1）建立因子分析模型

选择最大方差化因子旋转方式

是一种正交旋转，方差最大法从简化因子载荷矩阵的每一列出发，使和 每个因子有关的载荷的平方的方差最大。当只有少数几个变量在某个因子上有较高的载荷时，对因子的解释最 简单。方差最大的直观意义是希望通过因子旋转后，使每个因子上的载荷尽量拉开距离，一部分的载荷趋于1，另一部分趋于0。

（2）公因子方差：只列举了前11个属性，未提取公因子前每个属性的方差为1（Z-SCORE标准化过）

变量xi的公因子方差是因子载荷矩阵的第i行的元素的平方和。记为



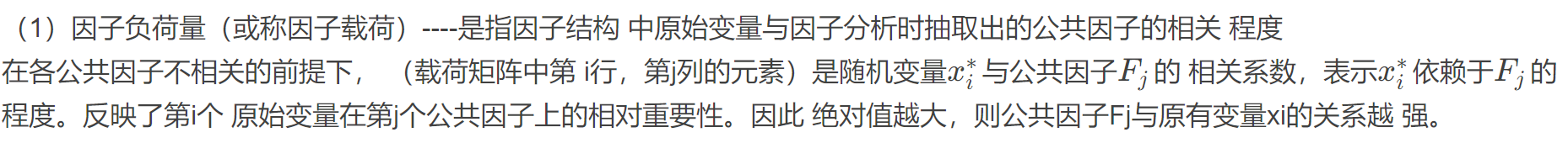
|  |  |
| --- | --- |
| Bankrupt? | 0.164974 |
| ROA(C) before interest and depreciation before interest | 0.949975 |
| ROA(A) before interest and % after tax | 0.961902 |
| ROA(B) before interest and depreciation after tax | 0.970047 |
| Operating Gross Margin | 0.998166 |
| Realized Sales Gross Margin | 0.997908 |
| Operating Profit Rate | 1.000869 |
| Pre-tax net Interest Rate | 0.997136 |
| After-tax net Interest Rate | 1.000018 |
| Non-industry income and expenditure/revenue | 0.900753 |
| Continuous interest rate (after tax) | 0.990357 |

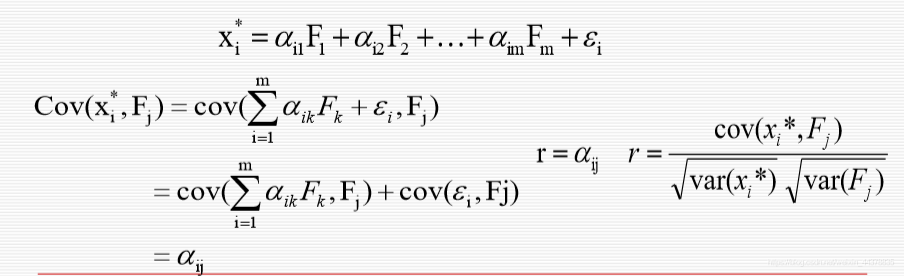
（3）因子贡献：

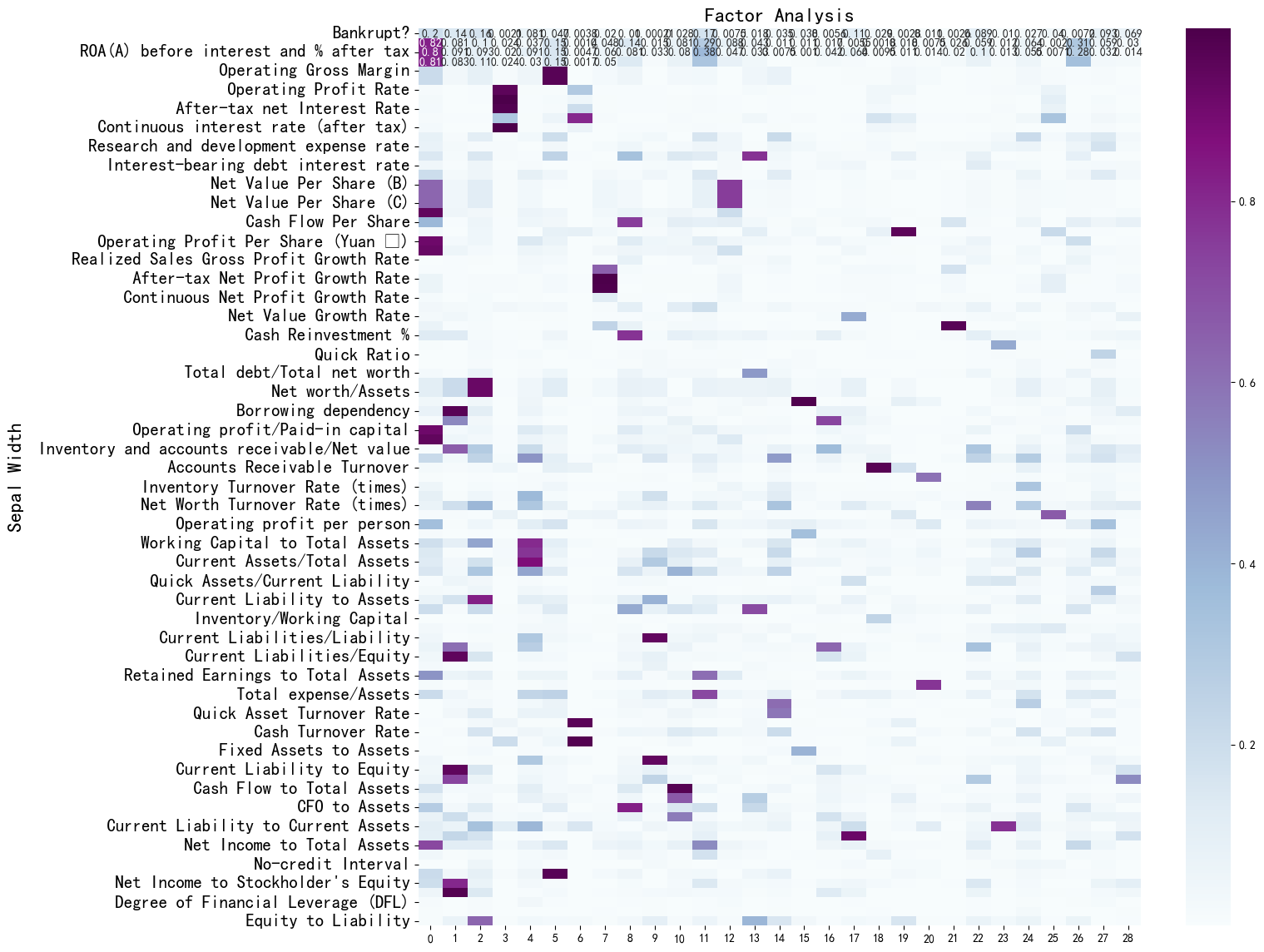
|  |  |  |  |
| --- | --- | --- | --- |
|  | 总方差贡献 | 方差贡献率 | 累计方差贡献率 |
| 0 | 9.274147 | 0.097623 | 0.097623 |
| 1 | 6.324736 | 0.066576 | 0.164199 |
| 2 | 3.991342 | 0.042014 | 0.206213 |
| 3 | 3.987802 | 0.041977 | 0.248190 |
| 4 | 3.382054 | 0.035601 | 0.283790 |
| 5 | 3.267475 | 0.034394 | 0.318185 |
| 6 | 2.691230 | 0.028329 | 0.346514 |
| 7 | 2.491041 | 0.026221 | 0.372735 |
| 8 | 2.438056 | 0.025664 | 0.398399 |
| 9 | 2.294360 | 0.024151 | 0.422550 |
| 10 | 2.135606 | 0.022480 | 0.445030 |
| 11 | 1.971304 | 0.020751 | 0.465781 |
| 12 | 1.898728 | 0.019987 | 0.485767 |
| 13 | 1.703080 | 0.017927 | 0.503694 |
| 14 | 1.445568 | 0.015217 | 0.518911 |
| 15 | 1.279122 | 0.013464 | 0.532375 |
| 16 | 1.238106 | 0.013033 | 0.545408 |
| 17 | 1.232648 | 0.012975 | 0.558383 |
| 18 | 1.075449 | 0.011321 | 0.569704 |
| 19 | 1.054267 | 0.011098 | 0.580801 |
| 20 | 0.998361 | 0.010509 | 0.591310 |
| 21 | 0.988807 | 0.010408 | 0.601719 |
| 22 | 0.879964 | 0.009263 | 0.610982 |
| 23 | 0.867906 | 0.009136 | 0.620117 |
| 24 | 0.733595 | 0.007722 | 0.627840 |
| 25 | 0.687491 | 0.007237 | 0.635076 |
| 26 | 0.633559 | 0.006669 | 0.641745 |
| 27 | 0.612371 | 0.006446 | 0.648191 |
| 28 | 0.488845 | 0.005146 | 0.653337 |

（4）旋转前后的负载矩阵

（5）载荷矩阵（旋转后）可视化







因子载荷是变量与公共因子的相关系数，某变量在某公共因子中的载荷绝对值越大，表明该变量与该公共因子更密切，即该公共因子更能代表该变量。由此可知，上图中与某因子对应变量的颜色越深（因子载荷越接近1），公共因子对该变量的代表性越强。

（6）用29个公共因子转化原始数据

7.各公司评分

以2个公因子得分为基础，再以每个因子的方差贡献率为权数进行线性加权平均，最后得到一个综合得分

经营状况得分前20的公司:

5257 48.057062

4023 43.556678

2364 43.485218

5609 43.390362

2248 43.277768

2055 43.116630

4037 43.064975

4185 43.021474

2388 43.019346

5808 42.783499

730 42.780950

2971 42.748179

2984 42.685824

3988 42.640005

3239 42.560465

4681 42.375385

4202 42.339077

4910 42.332457

756 42.217555

801 42.164587