南開大學
Nankai University

重抽样方法在因子模型选择中的应用

——Application of the Resampling Method in Factor Model Selection ——

答辩人：陶成雄　指导教师：周永道

目录

Contents

南開大學
Nankai University

PART

01

选题

南開大學 Nankai University

1.正交因子模型： $X-\mu=\Lambda^{T}F+U$

正交因子模型：现实中虽然有很多变量能被我们观测到，但在它们的背后存在着一些我们无法直接观测到的因素。这些因素在数量上少于观测到的变量，但却或多或少地同时影响着这些变量，这也是变量之间线性相关性的来源。

2.模型的求解： $X^{T}X=\Lambda^{T}\Lambda+\Phi$

模型求解的问题等价于标题方程的求解。目前流行的求解方法有：迭代主因子法、最小残差法、极大似然法和改进的主成分解法等。这些解法有的过于注重数学形式而忽略了模型的统计学意义，或是增加限制条件方便求解，很少关注模型求解的准确性：和变量背后的**真实的因子模型**相符合。这也是本文的出发点：寻找一个能够衡量求解是否准确的指标。

南開大學
Nankai University

PART

02

寻找评价指标

02 寻找评价指标

New Criterion

南開大學 Nankai University

1.分解样本阵

根据正交因子模型的定义分解样本阵X。

独特因子与公因子在现实中都是不可观测的，但它们仍然属于变量的范畴，也就各自拥有特定的分布并且可以进行抽样。在这个前提下，一个符合因子模型的样本阵X可以进行如下的分解：

$$\begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix} = \begin{pmatrix} f_{11} & \cdots & f_{1k} \\ \vdots & \ddots & \vdots \\ f_{n1} & \cdots & f_{nk} \end{pmatrix} \begin{pmatrix} \lambda_{11} & \cdots & \lambda_{1p} \\ \vdots & \ddots & \vdots \\ \lambda_{k1} & \cdots & \lambda_{kp} \end{pmatrix} + \begin{pmatrix} u_{11} & \cdots & u_{1p} \\ \vdots & \ddots & \vdots \\ u_{n1} & \cdots & u_{np} \end{pmatrix}$$

2.假设检验

H_0：通过某种方法估计得到的一组解Λ和Φ能正确反映真实的k-因子结构。

原假设等价于估计得到的载荷阵Λ能正确反映真实的载荷矩阵，当两者差别很大时拒绝原假设。由于在实际应用中我们并不知道真实的载荷阵到底是什么，因此无法直接进行比较，为此我们引入重抽样方法。

02 寻找评价指标

New Criterion

南開大學 Nankai University

3.重抽样方法

对独特样本阵进行重抽样获得新的样本阵X_1。

在原假设成立的条件下，不改变公因子样本阵F，只对估计得到的独特因子样本阵U的每一列分别进行重抽样，这样得到的新的独特因子样本阵U1仍然是合理的。从而以得到一个新的样本阵X1，且X1应当与X具有相同的k-因子结构。

4.评价指标

$$T = \| \Lambda^T \Lambda - \Lambda_1^T \Lambda_1 \| / \| \Lambda^T \Lambda \|$$

如果原假设成立，那么对第一次估计求解得到的模型进行重抽样得到新样本阵X1，再次对这个新样本阵进行估计求解，得到的结果Λ_1应当与第一次估计的结果Λ相差不多,也即原假设成立时统计量T的值应当趋近于零。通过对假设检验的分析提出的统计量T有可能正是我们寻找的评价指标。

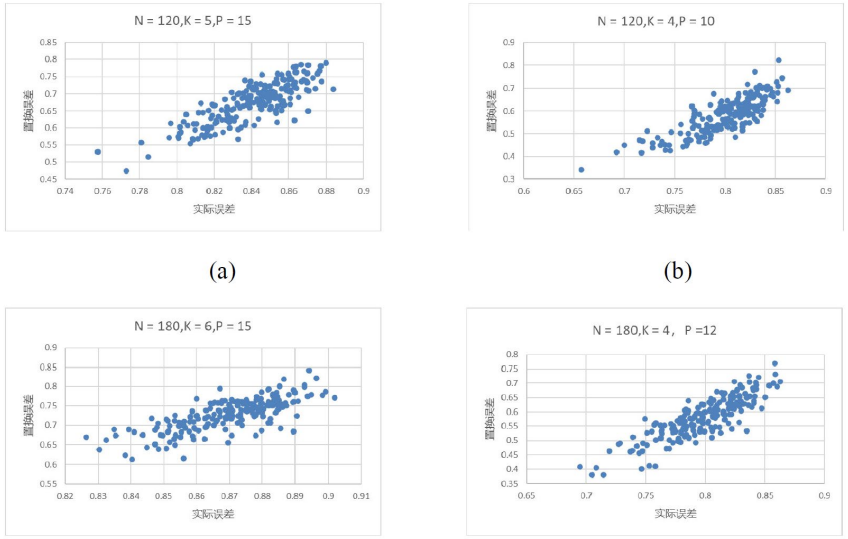
02 寻找评价指标

New Criterion

5.数值仿真

在不同的因子模型规模下检验统计量T的合理性验证。

图 2.1 检验统计量与实际误差之间的关系



(a) (b)

(c) (d)

表 2.1 统计量 T 与实际误差相关系数

编号	样本数 n	观测数 p	因子数 k	相关系数
1	120	10	4	0.816
2	120	12	6	0.824
3	120	15	5	0.684
4	150	10	6	0.800
5	150	12	5	0.773
6	150	15	4	0.671
7	180	10	5	0.859
8	180	12	4	0.780
9	180	15	6	0.759

除第三组和第六组的结果稍稍低于0.7外，其余七组的结果都大于0.75。当相关系数高于0.7时就可以认为两个变量高度相关，因此我们有理由认为这个统计量T可以被采用作为衡量估计得到的模型是否符合真实模型的一个指标。

PART

03

解的变换

03 解的变换

Transformation of Solution

南開大學 Nankai University

1.解的性质

① 如果因子模型有一组解Λ_1和Φ，那么对Λ_1左乘任意一个正交矩阵Q得到的新矩阵：$\Lambda_1 = Q\,\Lambda_2$ 任然是因子模型的解。

② 对于给定的k与同一个对角阵Φ，如果有两个载荷阵Λ_1和Λ_2同时满足因子模型方程，则存在一个k阶的正交阵Q使得： $\Lambda_1 = Q\,\Lambda_2$

③ 模型方程至少有一组满足k≤p-1的解：$\Phi = \lambda_p I_P$

④ 如果存在一组满足公因子的数量k≤p-2的解，那么在公因子数量小于等于k+1时，方程有无穷多组解。

从独特变差角度来研究解的变换

南開大學 Nankai University

2.解的变换

假设对于给定的样本阵X和确定的公因子数量k，我们已经有一组解Λ和Φ。从这组解出发，建立如下的p个方程组：

$$(X^T X - \Phi) y_i = e_i$$

当其中某一个方程有解时，在 $X^T X - \Phi$ 对应的对角元上加上一个正数t满足：$u_i - \mathrm{t} > 0$

则通过谱分解可以得到一组新的解。当同时有多个方程有解时，通过调整对应的对角元值可以得到更多的解。经过这种方法得到的解能够保持原有公因子水平k不变。

PART

04

实证分析

1. FIFA球员数据

我们感兴趣的数据是一份能够衡量FIFA球员能力的各项指标的统计结果。从中选取前六个指标以及前200位球员的数据进行分析,这些指标分别为：传中、射门、短传、凌空抽射、盘带和反击。首先对数据进行中心化标准化处理，然后再进行因子分析。选取的部分数据如下表所示：

表 4.1 球员各指标数据

Crossing	Finishing	ShortPassing	Volleys	Dribbling	Reactions
84	95	90	86	97	95
84	94	81	87	88	96
79	87	84	84	96	94
17	13	50	13	18	90
93	82	92	82	86	91
81	84	89	80	95	90
86	72	93	76	90	90
77	93	82	88	87	92
66	60	78	66	63	85

2. 变换方法的实际运用

① 迭代法得到精确解：Λ_1和Φ_1

对$\Lambda_1^T\Lambda_1$进行谱分解，通过不为零的特征向量个数分析得到Λ_1的秩。当Λ_1行满秩时，利用本文提出的方法进行变换可以保持公因子水平不变。

01

② 判断超定方程组是否有解

方程组有精确解的充要条件为系数矩阵$(X^TX-\Phi)_{-i}$与增广矩阵$((X^TX-\Phi)_{-i}, e_i)$的秩相等。其中系数矩阵不是方阵，因此我们对它的转置乘以它自身得到的矩阵进行谱分解，根据特征值大小来判断它的秩。

02

③ 调整矩阵对角元得到：Λ_2和Φ_2

如果第i个方程组有解，则在矩阵$(X^TX-\Phi)$对应的对角元上加上一个满足条件的正值u_p，变化后的$(X^TX-\Phi)$仍为非负定阵且秩不变，通过谱分解即可得到Λ_2，相应地可以得到Φ_2。

03

④ 通过极大似然法得到：Λ_3和Φ_3

04

3. 模型选择的结果

对上述三组解构造我们提出的检验统计量，经过多次重抽样我们得出第一组解置换检验前后误差相对最大，其余两组解表现相近，第三组要稍好一些。因此采用第三组解 和 作为此次因子分析模型的解。对于提取出的第一个公因子：传中、短传和盘带这三个指标的载荷在0.8及以上。由于这三个指标都与球员带球传球能力有关，因此第一个公因子可以理解为球员的基本功因子。对于提取出的第二个公因子：射门和凌空抽射这两个指标的载荷最大。由于这两个指标都与射门得分相关，因此可以理解成球员的进攻能力因子。对于提取出的第三个公因子：防守反击指标的载荷最大，其余载荷基本可以忽略不计，因此可以理解成球员的防守能力因子。

表 4.3 公因子旋转结果

	Factor1	Factor2	Factor3
Crossing	0.796	0.485	0
Finishing	0.537	0.820	0.183
ShortPassing	0.881	0.336	0
Volleys	0.576	0.746	0.148
Dribbling	0.820	0.535	0
Reactions	0	0.121	0.988

表 4.5 部分球员因子得分

球员	F1	F2	F3	F
1	0.675	0.528	2.877	1.031
2	0.089	0.808	3.139	0.891
3	0.662	0.222	2.655	0.881
4	-1.689	-1.681	2.014	-1.005
5	0.828	0.031	1.896	0.758
6	0.854	0.171	1.618	0.766

PART

05

结论与局限

南開大學 Nankai University

1.优点：

传统的因子模型求解方式大都是拟合因子模型的等价方程，只利用到了样本的协方差阵或是相关阵，样本阵自身的信息并没有被利用完全。本文将重点放在了模型背后的统计学意义上，对样本阵的分解充分地利用了样本阵背后的信息来协助我们求得模型更优的解。

2.不足：

通过分析第二章提出的假设检验问题找到了统计量T，但对其分布了解不够，只能通过数值模拟来考察统计量T的性质。

本文提出的解的变换方法能够帮助我们找到多组精确解，但这一方法并没来得及写成一个一步到位的程序，因此在实证分析部分只能逐步分析求解。

感谢各位老师的聆听！

——Thanks for Listening!——

答辩人：陶成雄