Advances in Social Sciences 社会科学前沿, 2019, 8(11), 1873-1881
Published Online November 2019 in Hans. http://www.hanspub.org/journal/ass
https://doi.org/10.12677/ass.2019.811255

Hans汉斯

Research on Competitiveness of Some Countries Based on Principal Component Analysis and K-Means Clustering

Zewei Zhang, Xiang Yuan

School of Economics & Management, Shanghai Maritime University, Shanghai
Email: yuanx2109@163.com

Received: Oct. 25th, 2019; accepted: Nov. 8th, 2019; published: Nov. 15th, 2019

Abstract

In this paper, 20 evaluation indicators from 42 countries are selected. Firstly, the information entropy reflecting the degree of information disorder is used to preliminarily screen the indicators, and 12 indicators with high contribution rate are retained, thus achieving the effect of dimensionality reduction. Then the principal components representing most information are obtained by the principal components analysis and the results are clustered by K-means method. Finally, the ranking and classification of national competitiveness are obtained.

Keywords

Information Entropy, Principal Components Analysis, K-Means Clustering, National Competitiveness

基于主成分分析的国家竞争力研究

张泽伟，袁　象

上海海事大学经济管理学院，上海
Email: yuanx2109@163.com

收稿日期：2019年10月25日；录用日期：2019年11月8日；发布日期：2019年11月15日

摘　要

本文选取42个个国家的20个评价指标，首先用反映信息无序度的信息熵对指标初步筛选，保留贡献率较高的12个指标，达到降维的效果。再用主成法求得能代表大多数信息的主成分，用K-means法对结果进

文章引用: 张泽伟, 袁象. 基于主成分分析的国家竞争力研究[J]. 社会科学前沿, 2019, 8(11): 1873-1881.
DOI: 10.12677/ass.2019.811255

行聚类。最后得到国家竞争力的排名与分类。

关键词

信息熵，主成分分析，K-Means聚类，国家竞争力

1. 引言

国家竞争力是指国家借助经营原有资产，形成自有的经济、社会模式来增加财富。国家竞争力不是一个单一的概念，而是由多种因素相互作用构成。瑞士洛桑国际管理学院认为经济实力、国际化程度、政府管理、金融体系、基础设施、企业管理、国民素质等八大因素是影响国家竞争力的主要因素。

迈克尔·波特[1]在《国家竞争优势》中提到，国家竞争力主要是由社会、经济结构、文化、制度等多个因素综合作用下形成并不断发展。Jay van Wyk (2010) [2]在钻石模型的基础上构建了“双钻石”模型，它和“单钻石”模型的不同之处是，它纳入了影响国家竞争力的国际和国内决定性因素。易顺、韩江波[3]通过研究 2010~2012 年国外学者关于国家竞争力的文献，阐述竞争力影响因素，并进行实证剖析。魏海燕[4]对《世界竞争力年鉴》的评价指标体系的构成、演变和其中所包含的科技指标状况进行分析。

本文依据上述研究现状，建立国家竞争力评价指标体系。同时参考基于主成分分析的国内城市竞争力研究方法[5] [6]进行国家竞争力综合评价。由于原始指标数据可能存在冗余或不相关属性，先利用信息熵对数据预处理和降维。

2. 研究方法

建立国家竞争力指标的数学模型，假设 X 是已知的评价矩阵，其中元素 x_{ij} 表示第 i 个国家的第 j 个指标。对于国家竞争力的评价问题，其评价数据包括多种类型，例如人均国内生产总值、国际互联网用户等评价指标的数据为绝对数值，而全球化指数、人文发展指数为相对数值。因此要先消除不同数据间量纲上的差异性。

1) 对评价矩阵 X 消除量纲并做归一化处理后得到计算矩阵 Y：

$$y_{ij} = \left(x_{ij} - \overline{x}_{\bullet j}\right) \Big/ \left(\max x_{\bullet j} - \min x_{\bullet j}\right)$$

其中 $\max x_{\bullet j}$， $\min x_{\bullet j}$， $\overline{x}_{\bullet j}$ 分别表示矩阵 X 的第 j 列最大值，最小值和平均值。

2) 然后计算每个指标的熵值，其中第 j 个指标的熵值为：

$$H_j = -k\sum_{i=1}^{n} a_{ij} \ln a_{ij} \quad \text{(式 1)}$$

取负号是要保证熵值为正，归一化系数定义为 $k = 1/\ln n$。

3) 计算评价指标权重为：

$$w_j = \left(1 - H_j\right) \Big/ \sum_{j=1}^{n} H_j \quad \text{(式 2)}$$

2.1. 指标体系构建

根据科学性、综合性等原则，结合国家竞争力评价的实际情况，本文构建的评价指标体系，见表 1。

Table 1. Index system of national competitiveness evaluation
表 1. 国家竞争力评价指标体系

指标项	指标项
居民消费率(var1)	国内生产总值(var11)
三产对国内生产总值增长的贡献率(var2)	高等教育入学率(var12)
劳动参与率(var3)	全球化指数(var13)
移动电话(部/千人) (var4)	国内生产总值增长率(var14)
国际互联网用户(个/千人) (var5)	知识经济指数(var15)
人文发展指数(var6)	人均国内生产总值(var16)
出生时预期寿命(var7)	城市人口比重(var17)
人均国民总收入(var8)	企业经营合同执行手续数(var18)
全球创新指数(var9)	每个就业者创造的国内生产总值(var19)
人均医疗支出(var10)	万美元国内生产总值能耗(var20)

2.2. 信息熵

在信息论中，信息熵用来刻画信息的无序度，熵越大表示信息的无序化程度越高，能够提供更多的信息就越多；相反，信息熵越小，则说明集合内的元素较为单一，所提供的信息较少。

2.3. 主成分分析

主成分分析法的降维思想和多标准评价指标的要求十分接近，因而近年来被大量的应用于社会学，公共管理和经济学等领域的评价体系中，成为一种独具特色的多指标评价方法。英国统计学家斯格特在对英国城镇发展水平做研究时，得到了 57 个综合指标。通过主成分法发现，仅需要 5 个由原始变量线性组合而成的新变量，就能以 95%的精度来表示原始数据的差异性，数据维度得到了大幅下降。

2.4. K-Means 聚类

聚类分析是通过距离来衡量数据之间的相似度从而实现类的划分。它是把 n 个对象根据它们的属性不同分为 k 个聚类，且使获得的聚类满足以下要求：同一聚类中的对象相似度最高；不同聚类中的对象相似度较低。

3. 实证分析

3.1. 数据来源

本文从国际统计年鉴(2016)获取 42 个国家竞争力综合测量指标数据，将 20 个指标变量的名称依次记为 $\mathrm{var}1, \mathrm{var}2, \cdots, \mathrm{var}20$。

3.2. 利用信息熵对指标变量进行初步筛选

依据公式(2-1)计算 42 个国家 20 项指标的熵值分别为：

$$H=(2.574\ \ 1.350\ \ 2.784\ \ 2.702\ \ 3.370\ \ 3.172\ \ 2.813\ \ 3.142\ \ 3.546\ \ 3.071$$
$$1.914\ \ 3.448\ \ 3.010\ \ 2.702\ \ 3.366\ \ 3.495\ \ 3.250\ \ 3.132\ \ 3.308\ \ 2.478)$$

再由公式(2-2)计算权重得：

$$W=(0.041\ \ 0.009\ \ 0.046\ \ 0.044\ \ 0.061\ \ 0.056\ \ 0.047\ \ 0.055\ \ 0.066\ \ 0.054$$
$$0.024\ \ 0.063\ \ 0.052\ \ 0.044\ \ 0.061\ \ 0.065\ \ 0.058\ \ 0.055\ \ 0.060\ \ 0.038)$$

由计算结果我们可以得到以下结论：

根据所求得得信息熵可以对维数进行初步的筛选，在总共 20 个评价指标中，第 1，2，3，4，7，11，14，20 这 8 个指标对评价的贡献率较低，因此在后面的数据分析只保留剩下的 12 个指标。

结合原始变量发现：第 2 项，即第三产业对国内生产总值增长的贡献率所占的权重最低，对整个评价体系的影响可以忽略；第 11 项，即国内生产总值的贡献率所占权重也较低，可能受庞大的人口基数影响。同时，我们看到第 16 项评价指标，人均国内生产总值，它所占的权重达到了 0.065，仅此于第 9 项全球创新指数，可见用人均国内生产总值对国家竞争力的影响很大。另外，居民消费率和万美元国内生产总值能耗对评价结果的影响也很低；劳动参与率，移动电话数，人均寿命以及国内生产总值增长率对评价结果的影响较低，说明其有一定的参考价值，但特征不是很明显。例如，从移动电话数的原数据可以看出，虽然整体服从发达国家高于发展中国家，但也存在较多个例，如柬埔寨、越南、哈萨克斯坦等发展中国家的数据高于加拿大、美国、法国等发达国家。这与通常所认为的事实恰好相反。

通过对实例的计算分析，发现对评价体系的信息熵计算权重，得到的结果客观有效，可以较好的排除评价体系中部分对结果影响较小的评价指标。

本文中由于维数较低，用信息熵筛选时，维数减少不是很明显，如果处理一个 1000 维的数据集，设定信息熵阈值进行筛选，同时调节阈值来选取主元，例如设定阈值为 0.52，满足条件的仅有 43 个变量，这使得变量个数大大降低。

3.3. 利用主成分分析对指标变量进行降维

利用 R 软件对筛选后的 12 个变量进行主成分分析，从相关矩阵求解，得到特征值和特征向量，分析方差贡献率，选取解释足够方差的主成分，并列出主成分得分，得到主成分线性组合并计算主成分得分。

3.3.1. 求出 12 个变量的相关系数矩阵，特征值和特征向量

将筛选后的数据定义为矩阵 Z，求其相关系数矩阵，特征值和所对应的特征向量。求得特征值的平方为 $\lambda_i^2=(9.214,0.819,0.540,0.416,0.359,0.189,0.144,0.128,0.085,0.060,0.030,0.016)$，相关系数矩阵所表 2 示。

3.3.2. 主成分分析

由表 3 可以看出，第 1 个特征值的方差贡献率为 76.8%，第 2 个特征值的方差贡献率为 6.8%，前三个主成分已经贡献了 88.1%，因此可以保留三个主成分。

由主成分系数矩阵可得 3 个主成分的线性组合如下：

$$y_1=-0.311\,\mathrm{var}_5^*-0.312\,\mathrm{var}_6^*-0.284\,\mathrm{var}_8^*-0.308\,\mathrm{var}_9^*-0.278\,\mathrm{var}_{10}^*-0.256\,\mathrm{var}_{12}^*$$
$$-0.288\,\mathrm{var}_{13}^*-0.321\,\mathrm{var}_{15}^*-0.307\,\mathrm{var}_{16}^*-0.271\,\mathrm{var}_{17}^*+0.237\,\mathrm{var}_{18}^*-0.279\,\mathrm{var}_{19}^*$$

$$y_2=-0.335\,\mathrm{var}_8^*-0.335\,\mathrm{var}_{10}^*+0.409\,\mathrm{var}_{12}^*+0.132\,\mathrm{var}_{13}^*$$
$$-0.311\,\mathrm{var}_{16}^*+0.292\,\mathrm{var}_{17}^*-0.517\,\mathrm{var}_{18}^*-0.364\,\mathrm{var}_{19}^*$$

$$y_3 = 0.138\,\mathrm{var}_6^* - 0.284\,\mathrm{var}_9^* + 0.111\,\mathrm{var}_{10}^* + 0.583\,\mathrm{var}_{12}^* - 0.402\,\mathrm{var}_{13}^* + 0.376\,\mathrm{var}_{17}^* + 0.483\,\mathrm{var}_{19}^*$$

其中，var_5^*，var_6^*，var_8^*，var_9^*，var_{10}^*，var_{12}^*，var_{13}^*，var_{15}^*，var_{16}^*，var_{19}^*，var_{18}^*，var_{19}^*表示对原始变量标准化后的变量。

Table 2. Correlation matrix
表 2. 相关系数矩阵

	var5	var6	var8	var9	var10	var12	var13	var15	var16	var17	var18	var19
var5	1	0.902	0.781	0.884	0.768	0.717	0.839	0.943	0.853	0.772	−0.657	0.755
var6		1	0.822	0.857	0.745	0.786	0.807	0.952	0.857	0.779	−0.632	0.782
var8			1	0.774	0.695	0.502	0.733	0.802	0.900	0.702	−0.490	0.802
var9				1	0.808	0.652	0.867	0.927	0.861	0.673	−0.703	0.761
var10					1	0.626	0.647	0.791	0.893	0.568	−0.487	0.791
var12						1	0.617	0.781	0.624	0.753	−0.576	0.559
var13							1	0.900	0.767	0.658	−0.684	0.640
var15								1	0.873	0.770	−0.684	0.772
var16									1	0.705	−0.565	0.851
var17										1	−0.631	0.638
var18											1	−0.512
var19												1

Table 3. Contributing rate of principal component
表 3. 主成分贡献率

成分	特征根	贡献率(%)	累积贡献率(%)
1	3.035	0.768	0.768
2	0.905	0.068	0.836
3	0.735	0.045	0.881
4	0.645	0.035	0.916
5	0.599	0.030	0.946
6	0.435	0.016	0.961
7	0.381	0.012	0.973
8	0.358	0.011	0.984
9	0.291	0.007	0.991
10	0.245	0.005	0.996
11	0.174	0.003	0.999
12	0.129	0.001	1.000

主成分的意义可由线性组合中系数较大，即权重较大的几个指标的综合意义来解释，所以y_1主要是知识经济水平，人文发展指数，国际互联网用户，全球创新指数，人均国内生产总值这 6 个指标的综合

反映，它更多反映的是一个国家或地区的软实力。评价国家的综合实力除了单一经济实力外，还有很多涉及文化，教育等方面的因素。这些因素也是构成国家综合实力的重要表现。软实力通常都是由经济做保障的，软实力强的国家其经济实力也通常都比较强。用 y_1 来评价国家竞争力已经有 76.8%的把握，所以这6个指标是反映国家竞争力的主要指标，每一项都必不可少。y_2 和 y_3 主要是企业经营合同手续个数，高等教育粗入学率的综合反映，它标志着国家的企业经济水平和人才教育水平。

3.3.3. 主成分得分

各国家竞争力评价得分及排序见表 4 所示。

Table 4. Scoring and ranking of national competitiveness evaluation
表 4. 国家竞争力评价得分及排序

国家或地区	综合得分	排序	国家或地区	综合得分	排序
美国	−3.967	1	土耳其	0.505	22
荷兰	−3.622	2	乌克兰	0.793	23
新加坡	−3.614	3	哈萨克斯坦	0.629	24
瑞士	−3.614	4	巴西	0.827	25
加拿大	−3.219	5	墨西哥	0.946	26
英国	−3.185	6	南非	1.024	27
德国	−2.801	7	泰国	1.274	28
法国	−2.764	8	委内瑞拉	1.355	29
新西兰	−2.710	9	蒙古	1.532	30
日本	−2.533	10	伊朗	1.801	31
澳大利亚	−2.297	11	菲律宾	2.069	32
西班牙	−1.760	12	埃及	2.078	33
比利时	−1.567	13	印度尼西亚	2.252	34
意大利	−1.539	14	越南	2.378	35
韩国	−1.338	15	斯里兰卡	2.429	36
捷克	−0.507	16	尼日利亚	2.884	37
波兰	−0.667	17	印度	2.933	38
文莱	−0.297	18	老挝	3.118	39
马来西亚	0.011	19	巴基斯坦	3.455	40
阿根廷	0.173	20	柬埔寨	3.506	41
俄罗斯	0.385	21	孟加拉国	3.641	42

在主成分得分中，正负不代表大小，仅表示该国家的竞争力与平均水平的位置关系，以国家竞争力的平均水平算作零点。在这个例子中，应该定义为：得分为正的国家其竞争力在平均水平以下；而得分为负的国家，其竞争力在平均水平以上。

为了更加客观的表示各国主成分得分情况，作出其双投影图，如图 1 所示。图中的数字表示对应编号的国家，矢量在坐标上的投影则是该变量对主成分的载荷，它解释了原始变量和主成分相关性强弱的

问题：红线在横坐标上的投影是各变量对第一主成分的载荷；在纵坐标上的投影是各变量对第二主成分的载荷。

结合排序表和双投影图对整体进行分析可以看出，各样本已经得到了粗略的分类结果。空间上相距越近的变量，正相关程度越高，国家竞争力水平越相近；距离原点越远说明这个变量被这两个主成分解释的越充分。可以看出，在第一主成分上，各个变量在横坐标上的投影相差不大，即载荷相差不大，各变量解释第一主成分的权重相似。相对而言，第 15 个变量的载荷最大，第 18 个变量的载荷系数与其它变量相反。对于第二主成分，第 12，18，19 个变量的投影较大，其载荷大，也就是说这三个变量描述了更多的第二主成分。

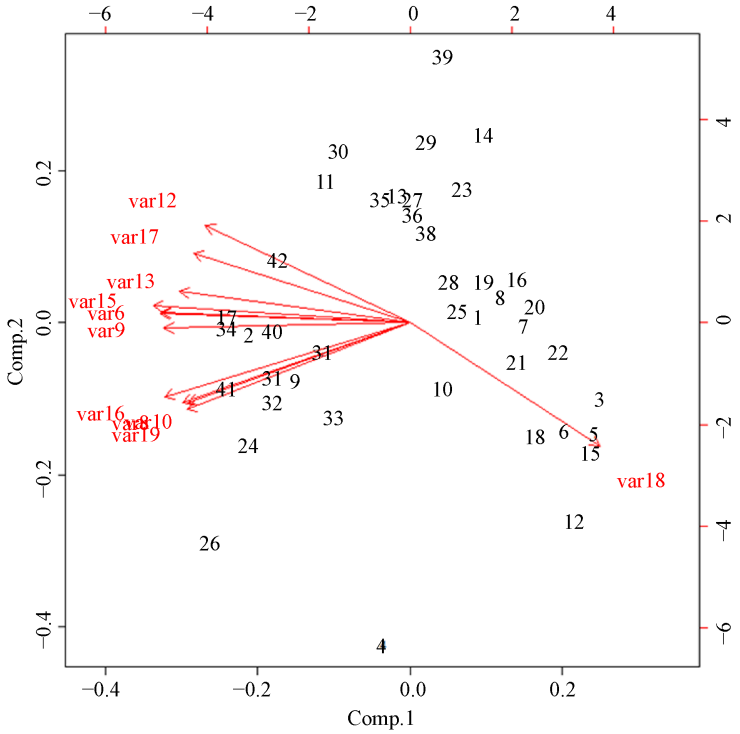

Figure 1. Double projection
图 1. 双投影图

分布在第三象限的第一主成分较好的是{26，41，34，17，24，2，40，32，31，42，9}对应国家是{美国，澳大利亚，荷兰，新加坡，加拿大，瑞士，英国，德国，法国，新西兰，日本}，这些国家的竞争力强，人均国内生产总值高，人类发展指数高，人均寿命高，城市人口比重高，全球化指数高。而第一主成分得分较差的，即位于第四象限的样本{18，22，6，12，15，5，3}对应国家{斯里兰卡，尼日利亚，印度，老挝，巴基斯坦，柬埔寨，孟加拉国}，这些国家的竞争力最弱。这些国家由于底子薄弱，工业基础差，商业不够发达，经济发展缓慢，部分地区还存在政局不稳定。

3.3.4. K-Means 聚类

取聚类数为 4，根据主成分得分的 K-means 聚类结果画出散点图，如图 2 所示，数据为聚类结果集的列“Comp.1”和“Comp.2”，颜色为用 1，2，3，4 表示缺省颜色，并用直线加以划分，从左往右一次是第一至第四区域，分别表示第一至第四聚类。由此将样本中的 42 个国家和地区分为以下四类：

黑点表示的是第一聚类，包括{美国，荷兰，新加坡，瑞士，加拿大，英国，德国，法国，新西兰，

日本，澳大利亚，西班牙，比利时，意大利}；国家竞争力强，是高度发达的资本主义国家和地区。主要分布在欧洲中西部，北美，澳洲及亚洲日本。

蓝点表示的是第二聚类，包括{文莱，韩国，马来西亚，阿根廷，巴西，委内瑞拉，捷克，波兰，俄罗斯，土耳其，乌克兰}；国家竞争力较强，主要分布在南美洲，马来群岛，东欧等地。

红点表示的是第三聚类，包括{印度尼西亚，伊朗，哈萨克斯坦，蒙古，菲律宾，泰国，越南，埃及，南非，墨西哥}；国家竞争力一般，分布范围广泛。

绿点表示的是第四聚类，包括{孟加拉国，柬埔寨，印度，老挝，巴基斯坦，斯里兰卡，尼日利亚}，国家竞争力弱主要分布在非洲，南亚，中南半岛等地。

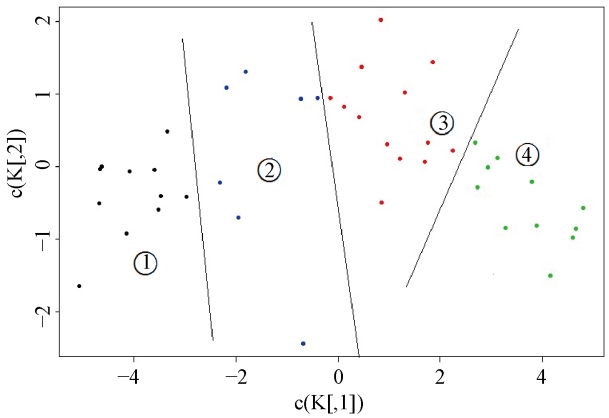

Figure 2. K-means clustering scatter
图 2. K-means 聚类散点图

4. 对实例的相关思考

事实上，权威的国家竞争力评价报告 IMD 是基于经济学理论，应用统计指标和问卷调查结果构建系统的综合评价指标体系，对国家和地区在国际上综合竞争能力进行测度。然而IMD的评价指标极其复杂，分为多个层次的评价指标，共采用了超过 300 个的竞争力指标。在实际的数据获取中会比较困难，本文只列出了 20 个竞争力指标，但最后的分析结果与当前的权威排名拟合的较好，发达国家和发展中国家得以区分。对于 300 个指标的分析，同样可以采取基于信息熵的主成分法。对于 K-means 聚类的类数选取也是一个值得思考的问题，可以通过计算轮廓系数确定，但是有时需要根据应用场景进行调整，而不能完全的依据评估参数选取。

国家竞争力评价带来的是一个横向比较的国际化视野，IMD 指标体系是从一个国家的竞争力得分情况来评价该国的竞争力，体现的是基于竞争结果的静态分析。这个评价体系反映的是已经形成的竞争力情况，而事实上，只研究一年度的评价结果并不能充分的反映一个国家竞争力，分析它可能潜在的能力。因此，在实际评价指标体系对国家竞争力进行测度和比较时，应该结合多年的评价结果和具体的国家国情，对其中的部分指标做适当修正和调整，以便更好的反映国家实力。

此外，也需要考虑客观存在的因素对评价结果的影响：例如(1) 人口因素对评价结果的影响，我国是个人口大国，人均指标结果往往被拉的很低；(2) 评价指标体系中数据采样方法对结果的影响。样本量过小或被调查者的个人偏好给整个评价体系带来了不可避免的系统误差，软指标越多，系统误差就可能越大，这也会造成评价结果在不同年度产生较大的波动。(3) 评价指标体系往往是以发达国家经验为基础制定的，对发展中国家参与国际比较有一定负面影响。

参考文献

[1] 迈克尔·波特. 国家竞争优势[M]. 北京: 华夏出版社, 2002.

[2] Van Wyk, J. (2010) Double Diamonds, Real Diamonds: Botswana's National Competitiveness Academy of Marketing Studies Journal, **14**, 55-76.

[3] 易顺, 韩江波. 国外国家竞争力研究: 理论架构与展望[J]. 兰州商学院学报, 2014(5): 93-102.

[4] 魏海燕.《世界竞争力年鉴》评价体系研究及其思考[J]. 科技管理研究, 2013(5): 65-68+77.

[5] 田美玲, 方世明. 汉江流域中心城市竞争力的评价及时空演变[J]. 统计与决策, 2016(9): 103-106.

[6] 于丽英, 郭洪晶. 长江三角洲地区城市综合竞争力的评价研究[J]. 上海大学学报(社会科学版), 2011, 18(1): 79-90.