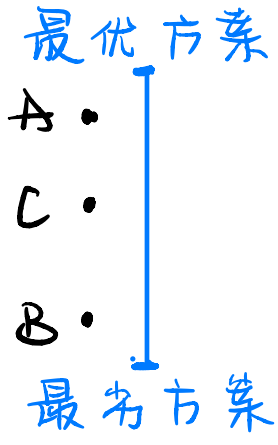
TOPSIS方法

C.L.Hwang 和 K.Yoon 于1981年首次提出 **TOPSIS (全称：Technique for Order Preference by Similarity to an Ideal Solution)**。
TOPSIS 法是一种常用的组内综合评价方法，能充分利用原始数据的信息，其结果能精确地反映各评价方案之间的差距。

基本过程为基于归一化后的原始数据矩阵，采用余弦法找出有限方案中的**最优方案和最劣方案**，然后分别计算各评价对象与最优方案和最劣方案间的距离，获得各评价对象与最优方案的相对接近程度，以此作为评价优劣的依据。

该方法对数据分布及样本含量没有严格限制，数据计算简单易行。



TOPSIS方法

为了客观地评价我国研究生教育的实际状况和各研究生院的教学质量，国务院学位委员会办公室组织过一次研究生院的评估。为了取得经验，先选5所研究生院，收集有关数据资料进行了试评估，下表是所给出的部分数据：

i \ j	人均专著 x_1 (本/人)	生师比 x_2	科研经费 x_3 (万元/年)	逾期毕业率 x_4 (%)
A	0.1	5	5000	4.7
B	0.2	6	6000	5.6
C	0.4	7	7000	6.7
D	0.9	10	10000	2.3
E	1.2	2	400	1.8

已知经过专家论证，四个指标的权重系数为 $\omega = [0.2, 0.3, 0.4, 0.1]$。

TOPSIS方法

Step1: 指标的筛选，经过观察，该数据不需要进行指标的筛选。
Step2: 指标的一致化处理。通过分析，我们知道：

- 人均专著，越多越好（极大型指标）
- 科研经费，越多越好（极大型指标）
- 逾期毕业率，越小越好（极小型指标）
- 生师比，过大过小都不好（区间型指标）

我们把两个极大型指标保持不变，对极小型指标采用取倒数操作。

对区间型指标，设研究生院的生师比最佳区间为$[5,6]$，在最佳区间内生师比得分为1，如果生师比小于2或者大于12都是0分，在其他的区间都按照线性关系进行变换。

TOPSIS方法

```
import numpy as np ## 区间型变换代码
import matplotlib.pyplot as plt
x_list = np.linspace(0,14,100)
y_list = []
for x in x_list:
    if x <= 2:
        y_list.append(0)
    elif x>2 and x<=5:
        y_list.append( (x-2)*1/3 )
    elif x>5 and x<=6:
        y_list.append(1)
    elif x>6 and x<=12:
        y_list.append( 1 - (x-6)*1/6 )
    elif x>12:
        y_list.append(0)
plt.plot(x_list,y_list)
```

TOPSIS方法

j / i	人均专著 x_1 (本/人)	生师比 x_2	科研经费 x_3 (万元/年)	逾期毕业率 x_4 (%)
A	0.1	5	5000	4.7
B	0.2	6	6000	5.6
C	0.4	7	7000	6.7
D	0.9	10	10000	2.3
E	1.2	2	400	1.8

一致化处理　　变换后 ⇓

j / i	人均专著 x_1 (本/人)	生师比 x_2	科研经费 x_3 (万元/年)	逾期毕业率 x_4 (%)
A	0.1	1.000000	5000	0.212766
B	0.2	1.000000	6000	0.178571
C	0.4	0.833333	7000	0.149254
D	0.9	0.333333	10000	0.434783
E	1.2	0.000000	400	0.555556

向量归一化法

对于极大型指标，令

$$a_{ij}^{*}=\frac{a_{ij}}{\sqrt{\sum_{i=1}^{n}a_{ij}^{2}}}(i=1,2,\cdots,n,1\leq j\leq m)$$

Step3:无量纲处理。以”人均专著”属性为例，我们使用向量归一化方法：

$$\begin{aligned}
0.1/\sqrt{0.1^2+0.2^2+0.4^2+0.9^2+1.2^2}&=0.0637576713063384\\
0.2/\sqrt{0.1^2+0.2^2+0.4^2+0.9^2+1.2^2}&=0.1275153426126767\\
0.4/\sqrt{0.1^2+0.2^2+0.4^2+0.9^2+1.2^2}&=0.2550306852253334\\
0.9/\sqrt{0.1^2+0.2^2+0.4^2+0.9^2+1.2^2}&=0.5738190417570045\\
1.2/\sqrt{0.1^2+0.2^2+0.4^2+0.9^2+1.2^2}&=0.7650920556760059
\end{aligned}\tag{27}$$

TOPSIS方法

使用同样的向量归一化方法，我们可以对其他三个指标也进行无量纲化处理，得到如下表所示的结果

j i	人均专著 x_1 (本/人)	生师比 x_2	科研经费 x_3 (万元/年)	逾期毕业率 x_4 (%)
院校 A	0.063758	0.597022	0.344901	0.275343
院校B	0.127515	0.597022	0.413882	0.231092
院校 C	0.255031	0.497519	0.482862	0.193151
院校 D	0.573819	0.199007	0.689803	0.562658
院校 E	0.765092	0.000000	0.027592	0.718952

TOPSIS方法

Step4: 选出其中的最优方案A^+和最劣方案A^-。

每一列数据取最大值，组合成最优方案

	人均专著	生师比	科研经费	逾期毕业率
最优方案 A^+	0.765092	0.597022	0.689803	0.718952
最劣方案 A^-	0.063758	0	0.027592	0.193151

最小值，最劣方案

Step5: 计算每一个学校，与最优方案以及最劣方案之间的距离

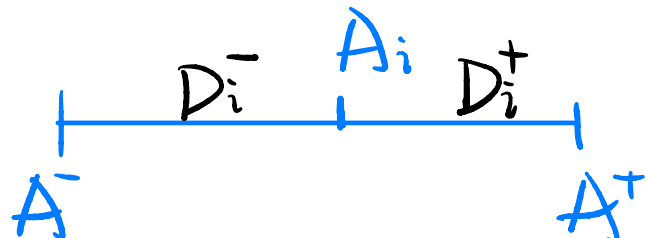
平行对象

$$D_i^+ = \sqrt{\sum_{j=1}^m w_j \left(A_j^+ - a_{ij}\right)^2}$$

$$D_i^- = \sqrt{\sum_{j=1}^m w_j \left(A_j^- - a_{ij}\right)^2}$$

然后使用如下的评价函数将其综合起来

$$C_i = \frac{D_i^-}{D_i^+ + D_i^-}$$



TOPSIS方法

Step6: 最终评价结果如下

	人均专著 x_1 (本/人)	生师比 x_2	科研经费 x_3 (万元/年)	逾期毕业率 x_4 (%)	最终得分
院校 A	0.1	5	5000	4.7	0.485830
院校 B	0.2	6	6000	5.6	0.526483
院校 C	0.4	7	7000	6.7	0.562158
院校 D	0.9	10	10000	2.3	0.677571
院校 E	1.2	2	400	1.8	0.400512

思考

- 如果理想中最好的大学是真实存在的，其得分C_i应该等于几，为什么？如果是理想中最差的大学真实存在呢?
- 上面的TOPSIS方法的权重已经给出，请思考，如果实际建模的时候没有给出权重，应该如何选取?

TOPSIS方法

当然，直接看结果可能不够直观，我们来通过一张雷达图解释这个评价的结果。

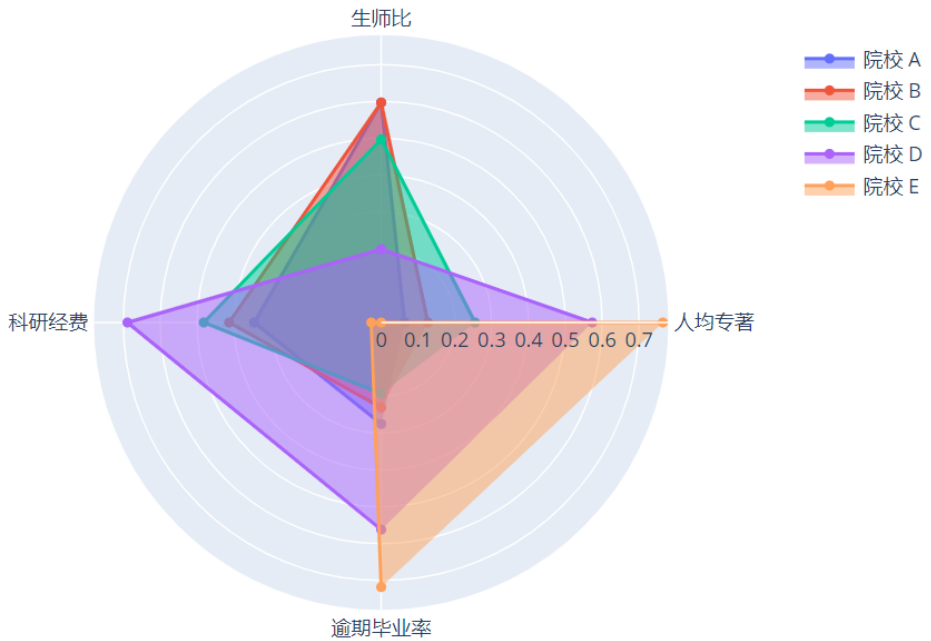


Figure:

上面是通过举例来进行TOPSIS方法的实现，关于TOPSIS方法的更详细理论推导，参考这里。在实际论文撰写的过程中，需要理论性强一些，而不是像我们上面一样简单地代入数据计算。

熵权法

下面我们来学习一种客观赋权的方法：熵权法（Entropy Weight Method），它是一种**突出局部差异的客观赋权方法**。因为它的权重选取仅依赖于数据本身的离散性。

某医院为了提高自身的护理水平，对拥有的11个科室进行了考核，考核标准包括9项整体护理，并对护理水平较好的科室进行奖励。下表是对各个科室指标考核后的评分结果。

评价指标

评价对象

科室	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_3	X_9
A	100	90	100	84	90	100	100	100	100
B	100	100	78.6	100	90	100	100	100	100
C	75	100	85.7	100	90	100	100	100	100
D	100	100	78.6	100	90	100	94.4	100	100
E	100	90	100	100	100	90	100	100	80
F	100	100	100	100	90	100	100	85.7	100
G	100	100	78.6	100	90	100	55.6	100	100
H	87.5	100	85.7	100	100	100	100	100	100
I	100	100	92.9	100	80	100	100	100	100
J	100	90	100	100	100	100	100	100	100
K	100	100	92.9	100	90	100	100	100	100

熵权法

由于各项护理的难易程度不同，因此需要对9项护理进行赋权，以便能够更加合理的对各个科室的护理水平进行评价。根据原始评分表，对数据进行归一化（最大值映射到1，最小值映射到0，其他值线性变化）后可以得到下列数据归一化表

极差变换法

科室	X_1	X_2	X_3	X_4	X_5	x_6	X_7	X_8	X_9
A	1.00	0.00	1.00	0.00	0.50	1.00	1.00	1.00	1.00
B	1.00	1.00	0.00	1.00	0.50	1.00	1.00	1.00	1.00
C	0.00	1.00	0.33	1.00	0.50	1.00	1.00	1.00	1.00
D	1.00	1.00	0.00	1.00	0.50	1.00	0.87	1.00	1.00
E	1.00	0.00	1.00	1.00	1.00	0.00	1.00	1.00	0.00
F	1.00	1.00	1.00	1.00	0.50	1.00	1.00	0.00	1.00
G	1.00	1.00	0.00	1.00	0.50	1.00	0.00	1.00	1.00
H	0.50	1.00	0.33	1.00	1.00	1.00	1.00	1.00	1.00
I	1.00	1.00	0.67	1.00	0.00	1.00	1.00	1.00	1.00
J	1.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
K	1.00	1.00	0.67	1.00	0.50	1.00	1.00	1.00	1.00

熵权法

计算第j项指标下第i个样本值所占比重

$$p_{ij} = \frac{x_{ij}}{\sum_{i=1}^{n} x_{ij}}, \quad i = 1, \cdots, n, j = 1, \cdots, m$$

计算第j个指标的熵值(熵值的计算方法是信息论中的定义，这里我们直接采用)

信息熵 $H = -\sum P_{ij} \log_2(P_{ij})$

$$e_j = -k \sum_{i=1}^{n} p_{ij} \ln(p_{ij}), \quad j = 1, \cdots, m$$

数据越分散，则数据中的信息越不可靠，所以对应的信息熵越小

其中，

$k = 1/\ln(n) > 0$

评价对象的数量.

熵权法

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9
信息熵	0.95	0.87	0.84	0.96	0.94	0.96	0.96	0.96	0.96

可以发现，熵值越小的变量，离散程度越大。接下来，我们计算信息熵冗余度d_j，并将其归一化得到权重

我们希望，离散程度越大，权重越大。

$$d_j = 1 - e_j, \quad j = 1, \cdots, m$$

$$w_j = \frac{d_j}{\sum_{j=1}^{m} d_j}, \quad j = 1, \cdots, m$$

	w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8	w_9
权重	0.08	0.22	0.27	0.07	0.11	0.07	0.07	0.07	0.07

熵权法

加权求和计算指标综合评分

$$s_i = \sum_{j=1}^{m} w_j x_{ij}, \quad i = 1, \cdots, n$$

最终得分为

科室	A	B	C	D	E	F
得分	95.71	93.14	93.17	92.77	95.84	98.01
科室	G	H	I	J	K	
得分	90.21	95.17	95.97	97.81	97.02	

F科室获得了第一名。你可以解释吗?