更新：灰色关联分析

本算法对应的视频讲解和代码可在我的微店获取：
https://weidian.com/?userid=1372657210
B 站数学建模试看课程：
https://www.bilibili.com/video/av20238704
清风数学建模视频课程的课表：
http://note.youdao.com/noteshare?id=2e1cc342928466bb632b33304a56e368
关注微信公众号：《数学建模学习交流》可获得更多数学建模技巧和比赛信息。

更新：灰色关联分析

一. 灰色关联分析概述

一般的抽象系统，如社会系统、经济系统、农业系统、生态系统、教育系统等都包含有许多种因素，多种因素共同作用的结果决定了该系统的发展态势。人们常常希望知道在众多的因素中，哪些是主要因素，哪些是次要因素；哪些因素对系统发展影响大，哪些因素对系统发展影响小；哪些因素对系统发展起推动作用需强化发展，哪些因素对系统发展起阻碍作用需加以抑制；……这些都是系统分析中人们普遍关心的问题。例如，粮食生产系统，人们希望提高粮食总产量，而影响粮食总产量的因素是多方面的，有播种面积以及水利、化肥、土壤、种子、劳力、气候、耕作技术和政策环境等。为了实现少投入多产出，并取得良好的经济效益、社会效益和生态效益，就必须进行系统分析。

数理统计中的回归分析、方差分析、主成分分析等都是用来进行系统分析的方法。这些方法都有下述不足之处：

（1）要求有大量数据，数据量少就难以找出统计规律；

（2）要求样本服从某个典型的概率分布，要求各因素数据与系统特征数据之间呈线性关系且各因素之间彼此无关，这种要求往往难以满足；

（3）计算量大，一般要靠计算机帮助；

（4）可能出现量化结果与定性分析结果不符的现象，导致系统的关系和规律遭到歪曲和颠倒。

尤其是我国统计数据十分有限，而且现有数据灰度较大，再加上人为的原因，许多数据都出现几次大起大落，没有典型的分布规律。因此，采用数理统计方法往往难以奏效。

灰色关联分析方法弥补了采用数理统计方法作系统分析所导致的缺憾。它对样本量的多少和样本有无规律都同样适用，而且计算量小，十分方便，更不会出现量化结果与定性分析结果不符的情况。

灰色关联分析的基本思想是根据序列曲线几何形状的相似程度来判断其联系是否紧密。曲线越接近，相应序列之间的关联度就越大，反之就越小。

对一个抽象的系统或现象进行分析，首先要选准反映系统行为特征的数据序列，称为找系统行为的映射量，用映射量来间接地表征系统行为。例如，用国民平均接受教育的年数来反映教育发达程度，用刑事案件的发案率来反映社会治安面貌和社会秩序，用医院挂号次数来反映国民的健康水平等。有了系统行为特征数据和相关因素的数据，即可作出各个序列的图形，从直观上进行分析。

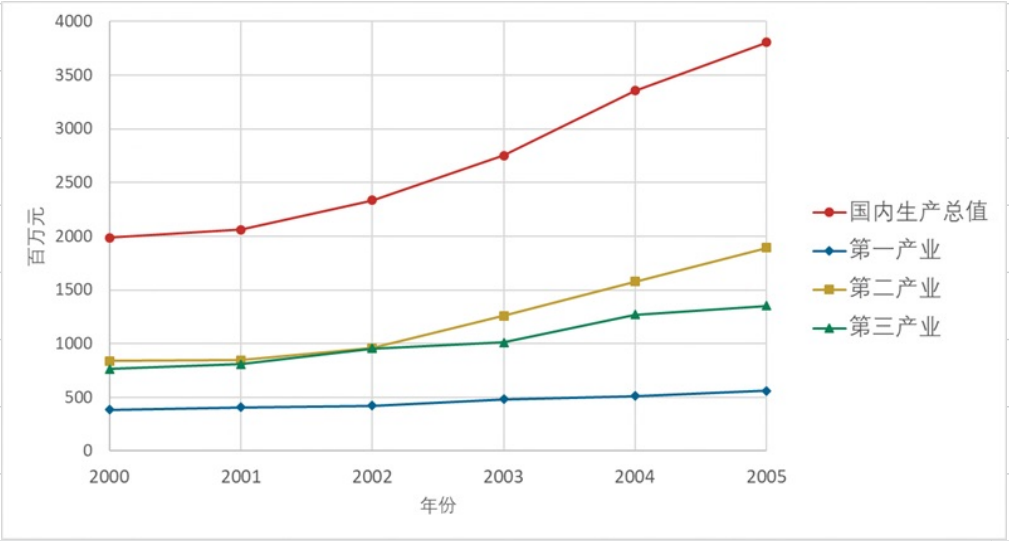
参考:刘思峰. 灰色系统理论及其应用(第五版)[M]. 北京: 科学出版社, 2010: 62.

二. 应用一：进行系统分析

下表为某地区国内生产总值的统计数据（以百万元计），问该地区从 2000 年到 2005 年之间哪一种产业对 GDP 总量影响最大。

年份	国内生产总值	第一产业	第二产业	第三产业
2000	1988	386	839	763
2001	2061	408	846	808
2002	2335	422	960	953
2003	2750	482	1258	1010
2004	3356	511	1577	1268
2005	3806	561	1893	1352

① 画统计图（Excel画的）



画图后得配上简单的分析：

① 四个变量均呈上升的趋势

② 第二产业的增幅较为明显

③ 第二产业和第三产业的差距在后两年相差更大

② 确定分析数列

(1) 母序列（又称参考数列、母指标）：能反映系统行为特征的数据序列。→类似于因变量 Y，此处记为 X_0。

(2) 子序列（又称比较数列、子指标）：影响系统行为的因素组成的数据序列。→类似于自变量 X，此处记为 $(X_1, X_2, \cdots X_m)$

在本例中：国内生产总值就是母序列（X_0），第一、第二和第三产业就是子序列（X_1 X_2 X_3）。

③ 对变量进行预处理（两个目的：去量纲、缩小变量范围简化计算）

对母序列和子序列中的每个指标进行预处理：先求出每个指标的均值，再用该指标中的每个元素都除以其均值。

年份	国内生产总值	第一产业	第二产业	第三产业
2000	0.7320	0.8361	0.6828	0.7439
2001	0.7588	0.8838	0.6885	0.7878
2002	0.8597	0.9141	0.7812	0.9292
2003	1.0125	1.0440	1.0237	0.9847
2004	1.2356	1.1069	1.2833	1.2363
2005	1.4013	1.2152	1.5405	1.3182
均值	**2716**	**461.67**	**1228.83**	**1025.67**

④ 计算子序列中各个指标与母序列的关联系数

年份	国内生产总值 X_0	第一产业 X_1	第二产业 X_2	第三产业 X_3
2000	0.7320 $X_0(1)$	0.8361 $X_1(1)$	0.6828 $X_2(1)$	0.7439 $X_3(1)$
2001	0.7588 $X_0(2)$	0.8838	0.6885	0.7878
2002	0.8597 $X_0(3)$	0.9141	0.7812	0.9292
2003	1.0125 $X_0(4)$	1.0440	1.0237	0.9847
2004	1.2356 $X_0(5)$	1.1069	1.2833	1.2363
2005	1.4013 $X_0(6)$	1.2152 $X_1(6)$	1.5405 $X_2(6)$	1.3182 $X_3(6)$

$$X_0 = (X_0(1),\ X_0(2),\ \cdots,\ X_0(n))^T \rightarrow \text{母序列}$$

$$X_1 = (X_1(1),\ X_1(2),\ \cdots,\ X_1(n))^T$$

$$\cdots$$

$$X_m = (X_m(1),\ X_m(2),\ \cdots,\ X_m(n))^T$$

$\rightarrow$ 子序列

记 $a = \min_i \min_k |X_0(k) - X_i(k)|$，$b = \max_i \max_k |X_0(k) - X_i(k)|$

a：两极最小差　　b：两极最大差

(数据保留四位小数)

$\lvert X_0 - X_1 \rvert$	$\lvert X_0 - X_2 \rvert$	$\lvert X_0 - X_3 \rvert$
0.1041	0.0492	0.0119
0.1249	0.0704	0.0289
0.0544	0.0785	0.0694
0.0315	0.0112	0.0278
0.1288	0.0477	0.0006 $\rightarrow a$
0.1862 $\rightarrow b$	0.1392	0.0832

定义：$\underset{(\text{gamma})}{\gamma}(x_0(k), x_i(k)) = \dfrac{a + \rho b}{|x_0(k) - x_i(k)| + \rho b}$

ρ：分辨系数（一般取 0.5）

$i = 1, 2, \cdots, m$

$k = 1, 2, \cdots n$

$0.4751 = \frac{0.0006 + \frac{1}{2}\times 0.1862}{0.1041 + \frac{1}{2}\times 0.1862}$	0.6586	0.8922
0.4299	0.5733	0.7680
0.6356	0.5462	0.5766
0.7520	0.8985	0.7753
0.4224	0.6657	1.0000
0.3356	0.4035	0.5317

⑤ 定义 $\gamma(X_0, X_i) = \frac{1}{n}\sum_{k=1}^{n}\gamma(x_0(k), x_i(k))$ 为 X_0 和 X_i 的灰色关联度。

求平均值 $= \gamma(X_0, X_1) = 0.5084$，$\gamma(X_0, X_2) = 0.6243$，$\gamma(X_0, X_3) = 0.7573$

⑥ 通过比较三个子序列和母序列的关联度可以得到结论：

该地区在 2000 年至 2005 年间的国内生产总值受到第三产业的影响最大。(其灰色关联度最大)

讨论：

① 什么时候用标准化回归，什么时候用灰色关联分析？

当样本个数 n 较大时，一般使用标准化回归；当样本个数 n 较少时，才使用灰色关联分析。

② 如果母序列中有多个指标，应该怎么分析？

例如：Y_1 和 Y_2 是母序列，X_1、X_2、…、X_m 是子序列

那么我们首先计算 Y_1 和 X_1、X_2、…、X_m 的灰色关联度进行分析；再计算 Y_2 和 $X_1, X_2, \cdots, X_m$ 的灰色关联度进行分析。

三. 用于综合评价问题

Topsis 结合熵权法可解决这题

题目：评价下表中20条河流的水质情况。

注：含氧量越高越好；PH值越接近7越好；细菌总数越少越好；植物性营养物量介于10-20之间最佳，超过20或低于10均不好。

$X_{n\times m}$，$n=20$：评价对象个数，$m=4$：评价的指标个数

河流	含氧量（ppm）	PH值	细菌总数(个/mL)	植物性营养物量（ppm）
A	4.69	6.59	51	11.94
B	2.03	7.86	19	6.46
C	9.11	6.31	46	8.91
D	8.61	7.05	46	26.43
E	7.13	6.5	50	23.57
F	2.39	6.77	38	24.62
G	7.69	6.79	38	6.01
H	9.3	6.81	27	31.57
I	5.45	7.62	5	18.46
J	6.19	7.27	17	7.51
K	7.93	7.53	9	6.52
L	4.4	7.28	17	25.3
M	7.46	8.24	23	14.42
N	2.01	5.55	47	26.31
O	2.04	6.4	23	17.91
P	7.73	6.14	52	15.72
Q	6.35	7.58	25	29.46
R	8.29	8.41	39	12.02
S	3.54	7.27	54	3.16
T	7.44	6.26	8	28.41

注：数据是我随手编的，仅用于讲解相应的算法，可能有不合理之处，请见谅。

① 对指标进行正向化 （第二讲：Topsis）

② 对正向化后的矩阵进行预处理（本讲上个例题），得到矩阵 $Z_{n\times m}=(z_{ij})_{n\times m}$

③ 将预处理后的矩阵每一行取出最大值构成母序列（虚构的）

④ 计算各个指标与母序列的灰色关联度：r_1、r_2、…、r_m

⑤ 计算各个指标的权重：$w_1=r_1/(r_1+r_2+\cdots+r_m)$、$w_2=r_2/(r_1+r_2+\cdots+r_m)$、…、$w_m=r_m/(r_1+r_2+\cdots+r_m)$

⑥ 第k个评价对象的得分：$S_k=\sum_{i=1}^{m} z_{ki}\cdot r_i$ $(k=1,2,\ldots,n)$

⑦ 对得分进行归一化：$S_1'=S_1/(S_1+S_2+\cdots+S_n)$、$S_2'=S_2/(S_1+S_2+\cdots+S_n)$、$S_n'=S_n/(S_1+S_2+\cdots+S_n)$