Open Journal of Natural Science 自然科学, 2025, 13(2), 231-242
Published Online March 2025 in Hans. https://www.hanspub.org/journal/ojns
https://doi.org/10.12677/ojns.2025.132024

Hans汉斯

基于趋势性改进的灰色主成分聚类模型

李美妮

重庆对外经贸学院数学与计算机学院，重庆

收稿日期：2025年1月4日；录用日期：2025年2月27日；发布日期：2025年3月6日

摘　要

在灰色主成分分析方法的研究中，针对原有方法中用于主成分分析的关联度矩阵的伪相关性和波动型序列量化不准确的问题。在两方面进行改进，一是改进灰色相对关联度的取值范围，将其从0.5~1拓展到0~1；二是基于斜率思路引入了趋势概率关联度指标，度量序列在趋势变化方向上的表征。结合两种关联度，提出了可以替换相关系数或协方差矩阵的新关联度矩阵。模拟与实证结果显示，改进的灰色关联度矩阵能够更好的度量波动型序列特征，将其用于主成分分析聚类，可以发现更加丰富和合理的结果，与一般模型相比效果更优。

关键词

主成分聚类，灰色关联度，趋势概率，波动型序列

Grey Principal Component Clustering Model Based on Trend Improvement

Meini Li

School of Mathematics and Computer Science, Chongqing College of International Business and Economics, Chongqing

Received: Jan. 4th, 2025; accepted: Feb. 27th, 2025; published: Mar. 6th, 2025

Abstract

In the research of the grey principal component analysis method, the issues of pseudo-correlation and inaccurate quantification of fluctuating sequences in the correlation matrix for principal component analysis were tackled. Two improvements were made in two respects. Firstly, the range of grey relative correlation was expanded from 0.5~1 to 0~1. Secondly, based on the slope concept, a trend probability correlation index was introduced to measure the representation of sequences in the trend change direction. By combining the two correlation indices, a new correlation matrix that

文章引用: 李美妮. 基于趋势性改进的灰色主成分聚类模型[J]. 自然科学, 2025, 13(2): 231-242.
DOI: 10.12677/ojns.2025.132024

can replace the correlation coefficient or covariance matrix was proposed. Simulation and empirical results indicate that the improved grey correlation matrix can better measure the features of fluctuating sequences and be used for principal component analysis clustering to yield more diverse and reasonable results, outperforming general models.

Keywords

Principal Component Clustering, Grey Correlation, Trend Probability, Wave Sequence

1. 引言

主成分聚类[1]方法因为基于降维的思想，在提取大量信息量的基础上进行聚类，剔除冗余信息对数据分析的影响，使数据分析结果更加优化。因此在社会、经济、自然等领域，主成分聚类方法得到了广泛的应用。它是基于变量之间的相关性进行分析的，与之相反，灰色关联度方法[2]是基于序列趋势的几何特征来度量各序列之间的关联性，由于与传统关联性度量方法的不同，其越来越多的应用于社会各方面。

那么如何结合主成分分析思想与灰色关联度，国内外学者在灰色理论与主成分分析结合的研究上进行了许多有益的探索和应用。一方面是从理论方面进行探索，蔚雪波，张辉[3]基于灰色绝对关联度矩阵提出了灰方差，将其作为度量相关性的方法引入到主成分分析，提出了基于灰方差的主成分分析方法。Tung C T，Lee Y J [4]通过计算序列点之间的绝对距离构造了新的关联度矩阵，来代替相关系数或协方差矩阵，形成不一样的度量方法，提出了基于灰色关联度的主成分分析方法。Zhao N，Shao H Y [5]基于pearson、kendall、spearman 和邓氏灰色关联度矩阵讨论了四种方法在主成分分析中的稳健性，发现了结合非参数邓氏灰色关联度能够对主成分分析起到稳健作用。袁周，方志耕[6]直接用灰色相对关联度矩阵代替传统主成分中的相关系数矩阵，然后对实证数据进行主成分评价，进一步讨论了评价中重复加权问题以及因子负荷量作为主成分权重合成的理由。另一方面着眼于方法应用，傅为忠，代露露和潘群群[7]总结归纳了主成分聚类与灰色聚类的优缺点，构建了灰色主成分聚类方法，将方法应用于对安徽省主导产业的评价与决策选择。陈宝平[8]结合主成分分析和灰色聚类，对我国居民收入差距进行实证分析。

从以上的分析，发现对于模型的改进基本上是基于灰色关联度开展的。其是为了改善对波动型序列的适用性，基于此主要有以下一些改进：崔立志，刘思峰[9]等考虑序列的斜率特征代替相对关联特征，提出了基于斜率特征的灰色相似关联度模型，并且引入了符号函数，将灰色关联度范围推广到了−1 到 1，具有了正、负相关的性质。黄山松和曾波[10]基于数据增量的变化率，提出了通过序列之间数据增量的变化率来代替原模型中通过距离来表征关联度，增强了趋势上的表征。通过这种方法对邓氏灰色关联度进行改进，获得了不错的结果。刘小妹，柯林和于俊杰[11]基于波动型序列的特征，提出了新的灰色绝对关联度模型，在保持单调型序列之间度量的优点的基础上，弥补了在波动型序列之间度量的缺点。

现有的灰色主成分分析研究主要存在以下两个问题：一是关联度矩阵考虑的是序列某一方面的特征，但是忽略了序列在趋势变化量和趋势变化方向的特征。尤其是在波动型序列上现有方法的度量不够，无法与单调型序列进行区分。二是主成分分析所需的变量之间的关联矩阵，需要有较为准确的序列之间关系度量，而现有的方法关联度度量范围偏向于 0.5~1，容易导致序列之间的伪相关性，使分析结果不合理。因此，本文对相对关联度的范围进行改进，使其分布在 0~1 上。此外，加入数据增量变化，即序列的差

分象，表征序列趋势，用序列之间的共同趋势概率度量趋势变化方向上的差异，且此度量范围在 0~1 上。结合新提出的趋势概率度量和新改进的相对关联度，在保证原有方法在单调型序列上的优越性能，还体现了序列在趋势变化方向上的特征。快捷高效的解决了相对关联度矩阵用于主成分分析的弊端，并且尝试推广至灰色主成分聚类，探究其合理性。

2. 基于趋势性的灰色主成分聚类模型

2.1. 灰色相对关联度

主成分聚类的核心在于主成分的提取，在现有的研究中，袁周和方志耕[6]在灰色主成分评价模型的构建中提出了用灰色理论中的灰色关联度矩阵代替协方差或相关系数矩阵。灰色关联度表示相邻序列之间的变化速率的关系，主要分为灰色绝对关联度和相对关联度。主要思想就是通过计算离散积分来表征序列与 y=0 所围成的面积，用面积的大小来度量序列之间的关联度。

设 $X_i=\left(x_i(1),x_i(2),\cdots,x_i(n)\right)$ 和 $X_j=\left(x_j(1),x_j(2),\cdots,x_j(n)\right)$ 是初值不为零的序列。则两组序列的初值象为：

$$X_i'=\left(x_i'(1),x_i'(2),\cdots,x_i'(n)\right)=\left(\frac{x_i(1)}{x_i(1)},\frac{x_i(2)}{x_i(1)},\cdots,\frac{x_i(n)}{x_i(1)}\right) \tag{1}$$

$$X_j'=\left(x_j'(1),x_j'(2),\cdots,x_j'(n)\right)=\left(\frac{x_j(1)}{x_j(1)},\frac{x_j(2)}{x_j(1)},\cdots,\frac{x_j(n)}{x_j(1)}\right) \tag{2}$$

继续计算序列的初始零化象：

$$X_i'^0=\left(x_i'^0(1),x_i'^0(2),\cdots,x_i'^0(n)\right)=\left(x_i'(1)-x_i'(1),x_i'(2)-x_i'(1),\cdots,x_i'(n)-x_i'(1)\right) \tag{3}$$

$$X_j'^0=\left(x_j'^0(1),x_j'^0(2),\cdots,x_j'^0(n)\right)=\left(x_j'(1)-x_j'(1),x_j'(2)-x_j'(1),\cdots,x_j'(n)-x_j'(1)\right) \tag{4}$$

计算序列的 $\left|S_i'\right|$、$\left|S_j'\right|$ 和 $\left|S_j'-S_i'\right|$

$$\left|S_i'\right|=\left|\sum_{k=2}^{n-1}x_i'^0(k)+\frac{1}{2}x_i'^0(n)\right| \tag{5}$$

$$\left|S_j'\right|=\left|\sum_{k=2}^{n-1}x_j'^0(k)+\frac{1}{2}x_j'^0(n)\right| \tag{6}$$

$$\left|S_j'-S_i'\right|=\left|\sum_{k=2}^{n-1}\left(x_j'^0(k)-x_i'^0(k)\right)+\frac{1}{2}\left(x_j'^0(n)-x_i'^0(n)\right)\right| \tag{7}$$

则可以定义序列 i 与 j 之间的灰色相对关联度 r_{ij}：

$$r_{ij}=\frac{1+\left|S_i'\right|+\left|S_j'\right|}{1+\left|S_i'\right|+\left|S_j'\right|+\left|S_j'-S_i'\right|} \tag{8}$$

但是，对 $\left|S_j'-S_i'\right|$ 进行讨论，当 S_i' 与 S_j' 同号时，$0\le\left|S_j'-S_i'\right|<\left|S_j'\right|+\left|S_i'\right|$，当 S_i' 与 S_j' 异号时，$\left|S_j'-S_i'\right|\le\left|S_j'\right|+\left|S_i'\right|$，因而，$r_{ij}$ 的取值范围：

$$r_{ij}=\frac{1+\left|S_i'\right|+\left|S_j'\right|}{1+\left|S_i'\right|+\left|S_j'\right|+\left|S_j'-S_i'\right|}=1-\frac{\left|S_j'-S_i'\right|}{1+\left|S_i'\right|+\left|S_j'\right|+\left|S_j'-S_i'\right|}>1-\frac{1}{2}=\frac{1}{2} \tag{9}$$

$$r_{ij}=\frac{1+\left|S_i'\right|+\left|S_j'\right|}{1+\left|S_i'\right|+\left|S_j'\right|+\left|S_j'-S_i'\right|}=1-\frac{\left|S_j'-S_i'\right|}{1+\left|S_i'\right|+\left|S_j'\right|+\left|S_j'-S_i'\right|}\le 1 \tag{10}$$

因此，$r_{ij}\in(0.5,1]$，可见，这样的相对关联度取值范围，并不符合直观判断，所以根据归一化方法，将其转化到$(0,1]$上，使之更符合定义范围。

2.2. 趋势性概率关联度

设$X_i=(x_i(1),x_i(2),\cdots,x_i(n))$和$X_j=(x_j(1),x_j(2),\cdots,x_j(n))$是初值不为零的序列。对于趋势性选用斜率进行度量，也就是所说的序列差分，定义其为差分象。则两组序列的差分象为：

$$X_i^-=(x_i^-(1),x_i^-(2),\cdots,x_i^-(n-1))=(x_i(2)-x_i(1),x_i(3)-x_i(2),\cdots,x_i(n)-x_i(n-1)) \tag{11}$$

$$X_j^-=(x_j^-(1),x_j^-(2),\cdots,x_j^-(n-1))=(x_j(2)-x_j(1),x_j(3)-x_j(2),\cdots,x_j(n)-x_j(n-1)) \tag{12}$$

接下来标注趋势性特征向量ε_i和ε_j：

$$\varepsilon_i=(\varepsilon_i(1),\varepsilon_i(2),\cdots,\varepsilon_i(n-1))=\begin{cases}\varepsilon_i(v)=1, & \varepsilon_i(v)\ge 0\\ \varepsilon_i(v)=0, & \varepsilon_i(v)<0\end{cases},v=1,2,\cdots,n-1 \tag{13}$$

同理，ε_j也可用同样的方法计算得到。最后，根据趋势性特征向量ε_i和ε_j计算趋势概率关联度：

$$p_{ij}=\sum_{v=1}^{n-1}I(\varepsilon_i(v),\varepsilon_j(v))/(n-1) \tag{14}$$

其中$I(\cdot)$为指示函数，定义如下：

$$I(\varepsilon_i(v),\varepsilon_j(v))=\begin{cases}1, & \varepsilon_i(v)=\varepsilon_j(v)\\ 0, & \varepsilon_i(v)\ne\varepsilon_j(v)\end{cases} \tag{15}$$

可发现趋势概率关联度计算的是相同变化趋势在整体变化趋势当中发生的频率，用来度量两个序列之间的变化趋势是较为准确的方法。

2.3. 基于趋势性方法的计算步骤

基于上文中改进了范围的灰色相对关联度，以及提出的趋势性概率关联度。分别可以从离散积分面积以及序列发展趋势两个方面来体现序列之间关联程度，既有变化趋势量上的比较又存在变化趋势方向上的特征提取。因此结合这两种方法，可以提出适用于主成分分析的关联性矩阵R_{ij}，如下所示：

$$R_{ij}=(1-\delta)r_{ij}+\delta p_{ij} \tag{16}$$

δ为调节系数，δ越大，关联性中变化趋势方向的特征体现越充分；δ越小，关联性中变化趋势量的特征体现越充分。一般而言，将δ设置为0.5，能够同时满足两种变化趋势的体现。除此之外，将关联性矩阵的取值范围转换为$(0,1]$，这样对主成分分析中的度量矩阵进行代替，不会出现关联性的误判行为。

由上述讨论，本文进一步拓展了灰色主成分分析，提出了基于趋势性改进的灰色主成分聚类方法。结合趋势方向和趋势量变化，构建新的主成分聚类方法。具体的步骤如下：

(1) 对原始数据进行初值像，始点零化象，差分象处理，得到$X_j'^0$、$X_i'^0$和X_j^-、X_i^-；

(2) 计算改进范围的灰色相对关联度矩阵r和趋势性概率关联度矩阵p；

(3) 根据调节系数，计算所提出的关联性矩阵R；

(4) 计算关联性矩阵R的特征根、特征向量和累计方差贡献率。根据累计方差贡献率大于或等于85%

的特征根数量作为主成分的数量，计算主成分综合得分。

(5) 根据以上计算出的综合得分，计算样本综合得分的距离进行主成分聚类。

3. 数值模拟

根据相关文献，灰色关联度对于单调型数据的关联量化表现性能优越，对于波动型序列的关联量化表现较弱。因此，在这一部分，针对所提出的改进方法、灰色相对关联度和传统的相关系数，对单调型序列、波动型序列和混合型序列进行模拟。比较三种方法在不同情况的性能表现，从而验证改进方法的性能表现。

3.1. 单调型序列

考虑单调型序列的模拟，构建 4 条序列，分别考虑受序列总体斜率，受序列截距项的影响，见图 1。其中序列 1 为斜率为 1，截距项为 0 的序列，序列 2 为斜率为 2，截距项为 0 的序列，序列 3 为斜率为 1，截距项为 30 的序列，序列 4 为斜率为 2，截距项为 30 的序列。

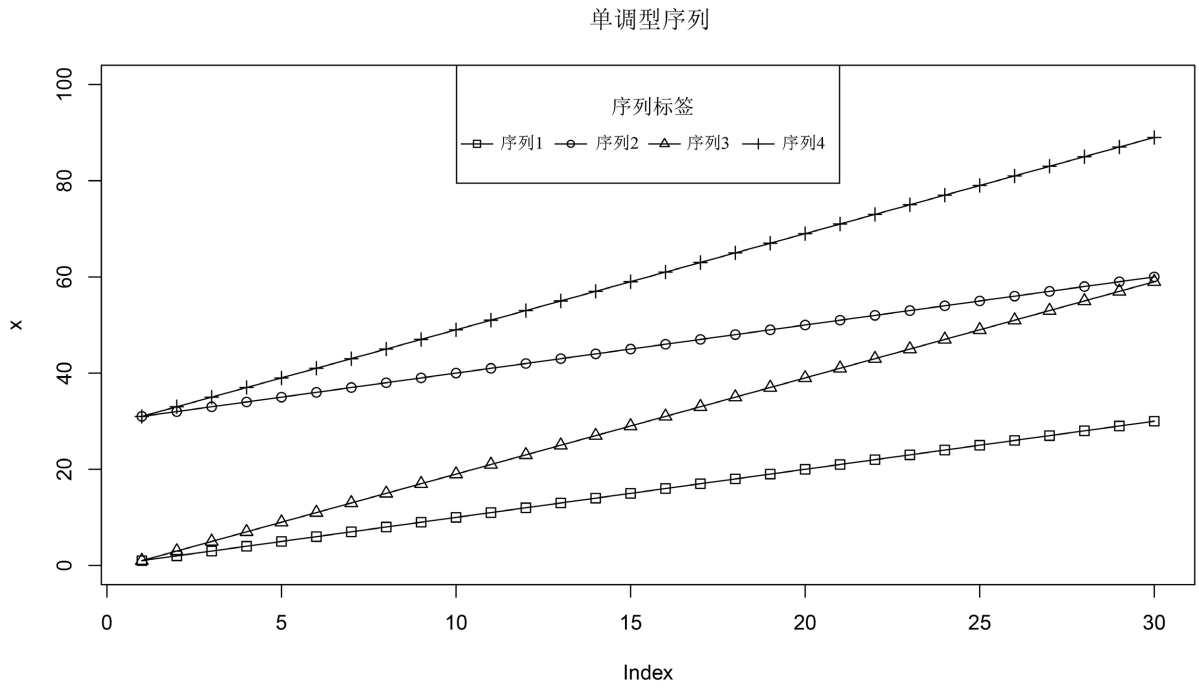

Figure 1. Sequences of monotone type
图 1. 单调型序列

趋势性概率关联度与传统统计的相关系数高度相似均为 1，与图 1 中所示的情况有所差异，见表 1。这时灰色相对关联度所表现出的关联更加准确，体现了在趋势量上的差异，改进的方法与之相同。说明了改进灰色相对关联度在吸收灰色相对关联度的优点时，在单调性情况下变化不大，同样能够很好的表征变化趋势量上的差异。而且，当截距项相同时，斜率不同时，改进灰色相对关联度系数类似，即 x1-x2，x3-x4，他们的关联系数均为 0.51 左右；当斜率相同，截距项相同时，改进灰色相对关联系数类似，即序列 x1-x3，x2-x4，他们的关联系数均为 0.75 左右。而剩余的序列 x1-x4，x2-x3 的关联系数主要体现了在变化趋势量上的差异，x2-x3 的关联系数稍小于 x1-x4。总的来看，引进灰色关联度来体现序列之间的关联，在单调性序列上的表现优于传统的相关系数。

Table 1. Correlation of monotone sequence
表 1. 单调型序列的关联度

序列对	相关系数	灰色相对关联度	趋势性概率关联度	改进灰色相对关联度
x1-x2	1	0.5167	1	0.5167
x1-x3	1	0.7501	1	0.7501
x1-x4	1	0.5328	1	0.5328
x2-x3	1	0.5084	1	0.5084
x2-x4	1	0.7545	1	0.7545
x3-x4	1	0.5164	1	0.5164

3.2. 波动型序列

考虑波动型序列的模拟，构建 4 条序列，分别考虑受序列总体斜率，受序列截距项的影响，见图 2。其中序列 1 为斜率为 1，截距项为 0 的序列，序列 2 为斜率为 2，截距项为 0 的序列，序列 3 为斜率为 −1，截距项为 30 的序列，序列 4 为斜率为−2，截距项为 60 的序列。

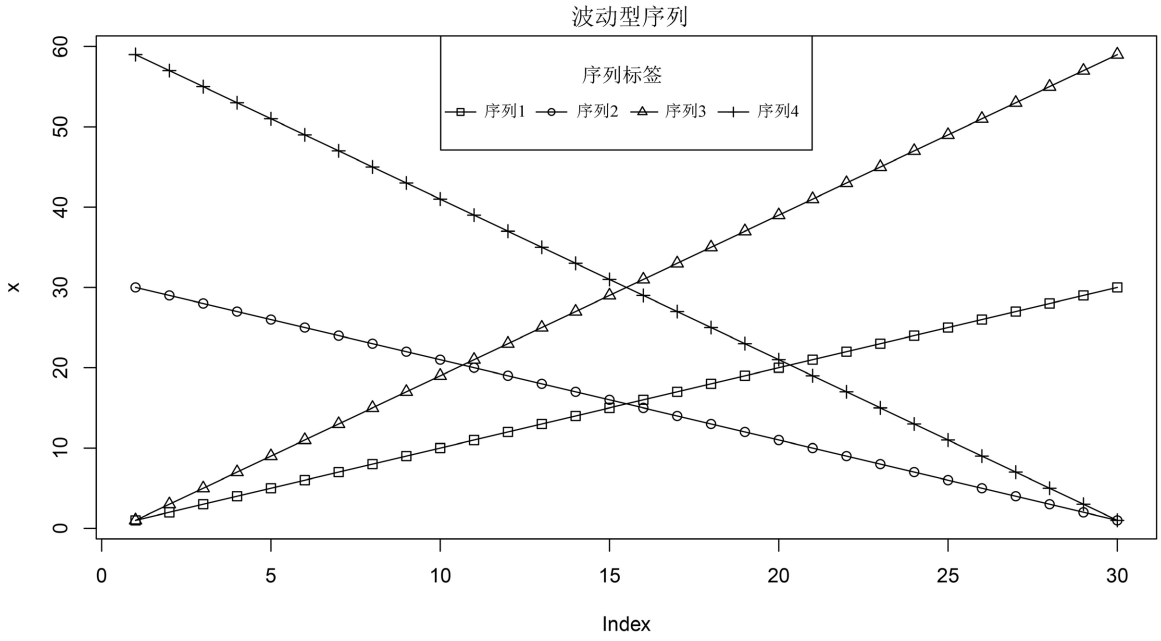

Figure 2. Sequences of wave type
图 2. 波动型序列

趋势性概率关联度与传统统计的相关系数高度相似，见表 2。当相关系数为−1 时，趋势性概率关联度为 0，当相关系数为 1 时，趋势性概率关联度为 1。这是由于传统的相关系数里有负相关的情况，但是在这里两者之间的趋势相似。而灰色相对关联度与其他度量对比，范围较小，系数在 0.5 到 0.75，未能很好的表征状况。特别是与单调序列相比，同样灰色相对关联度为 0.5 左右的序列对，在图 1 与图 2 中呈现截然相反的实际状况，即图 1 中的序列对 x1-x2 和图 2 中的 x1-x2。这时改进灰色相对关联度所表现出的关联更加准确，图 2 中的序列对 x1-x2 关联系数为 0.0006，图 1 中的序列对 x1-x2 关联系数为 0.5167。说明了改进灰色相对关联度在吸收灰色相对关联度的优点时，在波动型情况下体现更加优越，表征了变

化趋势方向上的差异。而且，观察发现当斜率完全相反时，改进灰色相对关联度系数相似，即 x1-x2，x3-x4，他们的关联系数均为 0.0003 左右；当斜率相反，斜率绝对值不相同时，改进灰色相对关联系数相似，即 x2-x3，x3-x4，他们的系数均为 0.0003 左右。而剩余的序列 x1-x3，x2-x4 的关联系数为单调型情况。总的来看，引进灰色关联度来体现序列之间的关联，在波动型序列上的表现优于灰色相对关联度，使此类序列之间的关联度表现更加显著。

Table 2. Correlation of wave sequence
表 2. 波动型序列的关联度

序列对	相关系数	灰色相对关联度	趋势性概率关联度	改进灰色相对关联度
x1-x2	-1	0.5006	0	0.0006
x1-x3	1	0.7501	1	0.7501
x1-x4	-1	0.5006	0	0.0006
x2-x3	-1	0.5003	0	0.0003
x2-x4	1	0.9919	1	0.9919
x3-x4	-1	0.5003	0	0.0003

3.3. 混合型序列

在对单调型与波动型序列讨论过后，继续讨论更加复杂的情况，即波动型与单调型均有的序列，其更加符合一般序列的情况。在混合型序列的模拟中，构建 4 条序列，分别考虑受序列总体斜率，受序列截距项的影响，见图 3。所有序列均存在 6 段趋势，其中序列 1 为的截距项为 0 的序列，与之斜率完全相反的序列 2 的截距项为 10 的序列，序列 3 为截距项为 30，斜率变化与 x1 相同的序列，序列 4 为与序列 3 斜率相反截距项为 40 的序列。

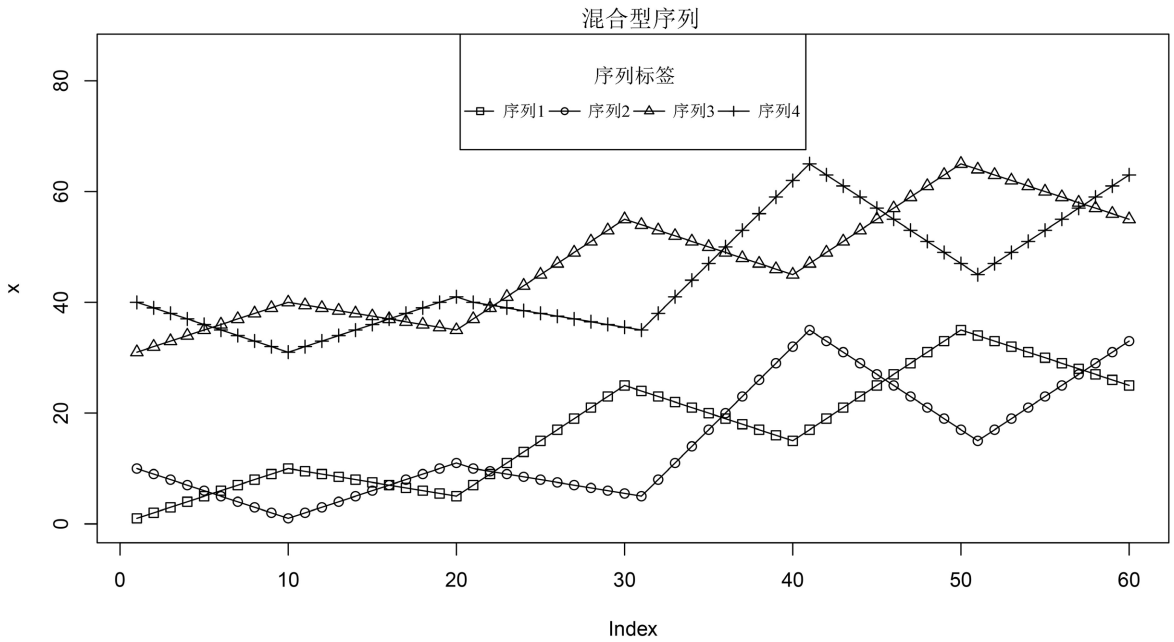

Figure 3. Sequences of mixed type
图 3. 混合型序列

趋势性概率关联度与传统统计的相关系数相似，见表 3。当相关系数为 0.5649 时，趋势性概率关联度为 0.0508，当相关系数为 1 时，趋势性概率关联度为 1。而灰色相对关联度与相关系数和趋势性概率关联度度量对比，发生了差距较大的变化。特别是序列 x2 与 x3，在图 3 中观察发现他们的趋势变化方向是总体递增的，局部变化相反。但是从趋势变化量上对比是高度一致的，因而使灰色相对关联度达到了 0.9361。因此，借助趋势性概率关联度进行调整后，下降到了 0.4615。其他序列之间的关联度都比较正常，序列对 x1-x3、x2-x4 趋势变化方向相同，他们之间的趋势变化量不同，但均超过 0.5。序列对 x1-x2、x1-x4 的趋势变化方向均是相反，他们之间的趋势变化量相似，关联度类似，小于 0.1。序列对 x3-x4 的趋势方向相反，但是趋势变化量较大，因而比序列对 x1-x2 的关联度大。总的来看，引进灰色关联度来体现序列之间的关联，在混合型序列上的表现优于灰色相对关联度，一是可以克服传统相关系数在趋势变化量上特征体现不够的劣势，二是可以通过趋势性概率关联度调整灰色相对关联度在波动型序列上的不足。因而在混合型序列中，改进的方法使混合序列间的关联度更加突出。

Table 3. Correlation of mixed sequence
表 3. 混合型序列的关联度

序列对	相关系数	灰色相对关联度	趋势性概率关联度	改进灰色相对关联度
x1-x2	0.5649	0.5143	0.0508	0.0397
x1-x3	1	0.5164	1	0.5164
x1-x4	0.5649	0.5038	0.0508	0.0292
x2-x3	0.5649	0.9361	0.0508	0.4615
x2-x4	1	0.6318	1	0.6318
x3-x4	0.5649	0.6149	0.0508	0.1404

4. 实证分析

4.1. 数据选取

本文选取了《2017 年中国统计年鉴》中 2016 年分地区交通运输、仓储和邮政业就业人员数作为样本。这一样本包含了 7 个变量，分别表征各类运输行业以及相关的仓储物流装卸行业的从业人数，见表 4。

Table 4. Employment number of transportation, storage and postal industries by region, 2016
表 4. 2016 年分地区交通运输、仓储和邮政业就业人数

地区	铁路运输业	道路运输业	水上运输业	航空运输业	装卸搬运和运输代理业	仓储业	邮政业
北京	279,180	110,033	248	75,996	42,288	10583	61,513
天津	57,471	14,834	18,034	8836	18,072	18107	10,841
河北	148,054	55,748	23,851	4766	10,902	11687	30,525
山西	83,031	114,877	70	5994	2557	6502	21,148
内蒙古	69,790	123,971	26	4486	2343	5580	21,491
辽宁	122,195	109,380	48,621	19,450	18,225	10275	20,334
吉林	52,935	63,592	159	6212	944	17491	18,296
黑龙江	72,898	132,249	3617	8445	3536	19939	30,037

续表

上海	196,395	40,590	53,610	83,947	72,452	28,385	31,065
江苏	259,038	23,204	77,294	15,805	34,174	19,181	57,103
浙江	165,609	27,957	25,711	15,276	21,792	11,455	47,346
安徽	128,272	40,208	12,734	4386	4876	9289	29,540
福建	104,685	39,513	16,310	18,111	21,296	4961	29,113
江西	103,024	60,672	7363	2917	1442	7107	20,346
山东	224,205	83,083	62,645	17,820	35,753	19,972	46,938
河南	254,487	113,594	4219	11,194	12,486	27,587	34,570
湖北	166,210	86,669	16,154	7067	8962	10,639	53,161
湖南	102,405	78,233	2857	8517	5166	4833	37,672
广东	393,904	61,613	50,296	129,470	54,347	28,289	93,163
广西	75,405	64,134	7179	7757	9649	5137	25,013
海南	21,705	6224	6257	22,461	4551	489	7968
重庆	172,431	29,767	11,446	13,173	6744	3158	28,099
四川	191,317	67,132	10,341	36,664	14,955	6807	76,449
贵州	55,171	35,094	626	9635	2622	2388	14,178
云南	79,584	39,099	264	23,817	11,299	2140	16,569
西藏	5255	69	130	844	15	262	2125
陕西	115,492	103,082	137	10,763	6968	9463	35,356
甘肃	48,593	57,970	95	2850	1049	4583	12,738
青海	16,308	20,256	38	2129	433	845	2866
宁夏	11,387	17,333	95	2942	321	669	3971
新疆	79,460	53,951	117	13,571	2205	2073	11,211

4.2. 关联度矩阵比较

Table 5. The correlation matrix of the variables
表 5. 变量的相关系数矩阵

	x1	x2	x3	x4	x5	x6	x7
x1	1.0000	0.2723	0.6143	0.7141	0.7463	0.6857	0.8870
x2	0.2723	1.0000	-0.0894	0.0579	0.0091	0.3040	0.3070
x3	0.6143	−0.0894	1.0000	0.4216	0.7350	0.6069	0.4857
x4	0.7141	0.0579	0.4216	1.0000	0.8323	0.5151	0.6503
x5	0.7463	0.0091	0.7350	0.8323	1.0000	0.6856	0.6061
x6	0.6857	0.3040	0.6069	0.5151	0.6856	1.0000	0.5311
x7	0.8870	0.3070	0.4857	0.6503	0.6061	0.5311	1.0000

根据以上提出的方法，对以上的数据进行计算得到以下三种关联度度量矩阵。主要有以下几点不同：

(1) 观察灰色相对关联度矩阵，所有变量的关联度大于 0.5，其中许多变量之间的关联度达到 0.8 以上，见表 5。这样会导致数据的伪高度相关性，使主成分分析的结果不太客观，与实际情况差距较大。

(2) 观察相关系数矩阵，存在较低的负相关系数，见表 6。整体的相关系数差异与灰色相对关联度差异相比，相关系数差异更大，体现的信息量更多。由以上模拟可以知道，相关系数在趋势性上体现的差异不够。

(3) 观察改进灰色相对关联度矩阵，见表 7。与未改进之前相比，整体系数的范围扩展到了 0~1 之间，能够体现趋势变化方向和变化量上的信息，降低了伪相关的风险。与相关系数矩阵对比，改进灰色相对关联度增加了相关系数所不能够体现的趋势性信息。因而，改进灰色关联度能体现与相关系数所不同的关系，使信息量更加丰富。

Table 6. The grey relative correlation matrix of the variables
表 6. 变量的灰色相对关联度矩阵

	x1	x2	x3	x4	x5	x6	x7
x1	1.0000	0.9100	0.5001	0.8726	0.9166	0.5524	0.9605
x2	0.9100	1.0000	0.5001	0.8056	0.8416	0.5639	0.9453
x3	0.5001	0.5001	1.0000	0.5001	0.5001	0.5001	0.5001
x4	0.8726	0.8056	0.5001	1.0000	0.9472	0.5391	0.8431
x5	0.9166	0.8416	0.5001	0.9472	1.0000	0.5437	0.8837
x6	0.5524	0.5639	0.5001	0.5391	0.5437	1.0000	0.5569
x7	0.9605	0.9453	0.5001	0.8431	0.8837	0.5569	1.0000

Table 7. The improved grey relative correlation matrix of the variables
表 7. 变量的改进灰色相对关联度矩阵

	x1	x2	x3	x4	x5	x6	x7
x1	1.0000	0.7434	0.3835	0.7226	0.8499	0.4524	0.8771
x2	0.7434	1.0000	0.2168	0.5556	0.6416	0.3639	0.7619
x3	0.3835	0.2168	1.0000	0.3335	0.3835	0.3501	0.4001
x4	0.7226	0.5556	0.3335	1.0000	0.8639	0.3557	0.6765
x5	0.8499	0.6416	0.3835	0.8639	1.0000	0.4437	0.7670
x6	0.4524	0.3639	0.3501	0.3557	0.4437	1.0000	0.3736
x7	0.8771	0.7619	0.4001	0.6765	0.7670	0.3736	1.0000

4.3. 主成分聚类结果比较

最后，由以上计算获得的关联度矩阵，进行主成分聚类。为了方便比较聚类结果。将所有的地区聚为 5 类，聚类结果见表 8，具体表现为：

1) 第一类，灰色主成分聚类与改进方法的聚类结果相同，即广东单独成类。广东省是中国改革开放的前沿，经济总量位于全国第一，制造业和新兴产业发达。通过实际数据，可以看到其在交通运输、仓储业和邮政业上明显高于其他地区。而传统主成分聚类将北京与广东聚为一类，观察后发现在一些指标

上还是存在很大的差距，因而基于灰色理论的聚类方法更优越。

2) 第二类，三种方法聚类结果相似，主要区别在于北京和河南，改进的方法所含地区最多，均涵盖的地区为上海、江苏和山东。上海是中国的经济金融中心，江苏是中国的先进制造业基地，并且位于长江三角洲城市群，经济发展综合水平全国第一，山东是北方第一经济强省。而改进方法中多出的河南省，位于中国的中原地区，经济发展强劲，中原城市群初具规模，是最重要的交通物流中转中心，因此有了这样的地位。与以上地区聚为一类，应该说更加稳妥和合理。

3) 第三类，灰色主成分聚类与改进方法的聚类结果相同，即海南、贵州、西藏、甘肃、青海、宁夏和新疆。不难发现，他们的经济实力，人口数量处于倒数的位置，基于他们的地理区位，海南省位于中国的最南部，由于海域的隔绝与其他省份的交通关联不强，且周边存在粤港澳大湾区，承担了大部分交通运输任务。宁夏、甘肃和新疆与西藏和青海地区一样，处于区位的边缘地区，交通环境恶劣，与其他地区的交通关联较弱，且人口相对而言比较稀少。贵州位于西南地区，属于群山之中，同样交通关联较弱，因而这一类基本上属于西部和边疆地区。然而传统主成分聚类中的第五类与之对应，多出了山西、内蒙古、吉林和江西，没有灰色方法得出的结论合理。

4) 第四类，灰色主成分聚类与改进方法的聚类结果相似，即黑龙江、河北、浙江、安徽、湖北和广西聚为一类。观察后发现这些地区与第一、二类地区邻接，接受这些地区的辐射很大。除此之外，这些地区人口稠密，具备一定优势。比较发现，铁路运输业、道路运输业、仓储业有相似之处。但是改进方法比灰色主成分聚类少了河南，将河南的交通枢纽地位更加突出。然而，传统主成分聚类并没有相对应的类别，且未表现出扩散的趋势。

5) 第五类，改进方法与灰色主成分聚类结果相同，即天津、山西、内蒙古、吉林、安徽、福建、江西、湖南、广西、重庆和云南，是剩余的其他东、中、西部地区。与传统主成分聚类相比，改进的方法体现了更多的信息量，且传统主成分聚类结果较难解释。

综上所述，将三种方法得到的结果与实际情况相验证，可知改进灰色主成分聚类方法结果能够体现更多信息量，更加符合实际情况。

Table 8. Five cluster results obtained by the three principal component clustering methods
表 8. 三种主成分聚类方法得到的五类聚类结果

方法	第一类	第二类	第三类	第四类	第五类
传统主成分聚类	北京、广东	上海、江苏、山东	黑龙江、湖南、广西、云南、陕西	天津、河北、辽宁、浙江、安徽、福建、河南、湖北、重庆、四川	山西、内蒙古、吉林、江西、海南、贵州、西藏、甘肃、青海、宁夏、新疆
灰色主成分聚类	广东	北京、上海、江苏、山东	海南、贵州、西藏、甘肃、青海、宁夏、新疆	河北、辽宁、黑龙江、浙江、河南、湖北、四川、陕西	天津、山西、内蒙古、吉林、安徽、福建、江西、湖南、广西、重庆、云南
改进灰色主成分聚类	广东	北京、上海、江苏、山东、河南	海南、贵州、西藏、甘肃、青海、宁夏、新疆	河北、辽宁、黑龙江、浙江、湖北、四川、陕西	天津、山西、内蒙古、吉林、安徽、福建、江西、湖南、广西、重庆、云南

5. 结论

针对灰色理论引入主成分聚类问题，本文从关联度矩阵角度出发，改进了相对关联度矩阵的伪高度关联度和趋势性关联误判问题。从而能够获得更加准确的关联矩阵，可以体现趋势变化量与变化方向上的信息。将关联矩阵代替传统的度量进行主成分聚类，取得了更加优越的表现。

模拟结果表明：当序列均为单调型序列时，改进的灰色关联度保持了灰色理论的优势，突出了趋势变化量上的关联；当序列均为波动型序列时，改进的灰色关联度结合了趋势性概率关联度的优势，突出了趋势变化方向上的关联，解决了灰色相对关联度的关联性误判现象。实证结果表明：改进灰色主成分聚类方法可以体现更多的信息量，得到更加符合实际的结果。

基金项目

本文得到 2024~2025 年度重庆对外经贸学院科研项目：基于不同分组结构的超高维数据特征筛选研究(项目编号：KYZK2024007)的支持。

参考文献

[1] 李雄英, 颜斌. 稳健主成分聚类方法的构建及其比较研究[J]. 数理统计与管理, 2019, 38(5): 849-857.

[2] 刘思峰, 杨英杰, 吴利丰. 灰色系统理论及其应用[M]. 北京: 科学出版社, 2014: 50-95.

[3] 尉雪波, 张辉. 灰色主成分分析及其应用[J]. 山东财政学院学报, 2004(5): 57-59, 63.

[4] Tung, C. and Lee, Y. (2009) A Novel Approach to Construct Grey Principal Component Analysis Evaluation Model. *Expert Systems with Applications*, **36**, 5916-5920. https://doi.org/10.1016/j.eswa.2008.07.007

[5] Zhao, N. and Shao, H.Y. (2014) Robust Principal Component Analysis Algorithm Based on Nonparametric Correlation Coefficient Matrix. *Advanced Materials Research*, **989-994**, 2613-2616. https://doi.org/10.4028/www.scientific.net/amr.989-994.2613

[6] 袁周, 方志耕. 灰色主成分评价模型的构建及其应用[J]. 系统工程理论与实践, 2016, 36(8): 2086-2090.

[7] 傅为忠, 代露露, 潘群群. 基于主成分与灰色聚类相结合的安徽省主导产业选择研究[J]. 华东经济管理, 2013, 27(3): 18-24.

[8] 陈宝平. 基于主成分分析和灰色聚类对我国居民收入差距分析[J]. 数学的实践与认识, 2018, 48(24): 134-143.

[9] 崔立志, 刘思峰, 李致平, 等. 灰色斜率相似关联度研究及其应用[J]. 统计与信息论坛, 2010, 25(3): 56-59.

[10] 黄山松, 曾波. 基于数据增量变化率的邓氏关联度模型的优化[J]. 统计与决策, 2010(22): 4-7.

[11] 刘小妹, 柯林, 于俊杰. 灰色绝对关联度的改进模型[J]. 数学的实践与认识, 2018, 48(10): 16-22.