Statistics and Application 统计学与应用, 2018, 7(2), 271-279
Published Online April 2018 in Hans. http://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2018.72032

Hans汉斯

A Comparative Study of Compositional Data Transformation Methods Based on Spatial Equivalence

Lijuan Guo[1], Rong Guan[2]

[1]School of Economics, Beijing Technology and Business University, Beijing
[2]School of Statistic and Mathematics, Central University of Finance and Economics, Beijing
Email: guolijuan@th.btbu.edu.cn

Received: Apr. 5th, 2018; accepted: Apr. 23rd, 2018; published: Apr. 30th, 2018

Abstract

Traditional statistical analysis method in Euclidean space is not suitable for compositional data, due to its unit-sum constraint in Simplex space. A common solution is to firstly transform compositional data in Simplex space into data in Euclidean space and then perform statistical analysis on the transformed data. This paper proposes to compare three commonly used method, *i.e.*, additive logratiotransformation (alr), centered logratio transformation (clr), and isometric logratio transformation (ilr). Based on Aitchison's algebra, the comparison is carried out to examine whether a transformation method satisfies the properties of linearity and orthogonality. A real dataset, namely the rock data, is used to verify the comparison results. Three transformation methods are used to relax the unit-sum constraint of the rock data, respectively, and a discriminant model is then established on the transformed data. Comparison results from both theory and real-data studies indicate that isometric logratio transformation is superior to the other two transformation methods in two points. First, isometric logratio transformation does not change the geometry concepts, *i.e.*, inner product and distance, which is inevitably caused by additive logratio transformation. Second, isometric logratio transformation successfully relaxes the unit-sum constraint and avoids multicolinearity, which cannot be solved by centered logratio transformation.

Keywords

Compositional Data, Simplex Space, Euclidean Space, Orthogonal Transformation, Fisher Discriminant Analysis

基于空间等价性的成分数据变换方法比较研究

郭丽娟[1]，关　蓉[2]

文章引用: 郭丽娟, 关蓉. 基于空间等价性的成分数据变换方法比较研究[J]. 统计学与应用, 2018, 7(2): 271-279.
DOI: 10.12677/sa.2018.72032

[1]北京工商大学经济学院，北京
[2]中央财经大学统计与数学学院，北京
Email: guolijuan@th.btbu.edu.cn

收稿日期：2018年4月5日；录用日期：2018年4月23日；发布日期：2018年4月30日

摘　要

单形空间的定和约束使得传统统计分析方法对成分数据失效，通常需要采用适当的变换方法将成分数据转化到欧氏空间后再进行统计分析。本文以非对称对数比变换、中心化对数比变化、等距对数比变换等三种常用的变换方法为研究对象，基于成分数据代数体系，从能否实现单形空间到欧氏空间等价转换的角度，比较研究了三种变换方法的合理性，为成分数据变换技术的选择提供理论依据。并选取岩石判别分类问题，分别采用以上方法对原始成分数据进行变换后建立判别模型，比较判别结果的可靠性。实证结果表明，等距对数比变换既克服了非对称对数比变换改变内积及距离等几何概念的缺陷，又避免了中心化对数比变换导致的多重共线性给多元分析方法带来的影响，在保持样本空间形态不发生变化的前提下解除了定和约束，是一种合理的变换方法。

关键词

成分数据，单形空间，欧氏空间，正交变换，Fisher判别分析

1. 引言

成分数据是指任意非负的 D 元向量 $\boldsymbol{x}=[x_1, x_2,\cdots,x_D]$，且满足约束条件

$$\sum_{i=1}^{D} x_i = 1,\ 0 \le x_i \le 1 \tag{1}$$

式(1)被称为定和约束，是成分数据的基本性质。D 元成分数据 $\boldsymbol{x}$ 中的每一个元 $x_i\ (i=1,\cdots,n)$ 代表相对信息，表示该部分在整体中所占比重。与普通数据相比，一方面，从原始绝对数据计算得到的比例结构，能够更进一步揭示绝对数据背后的相对信息；另一方面，成分数据更适合分析整体的各部分比例关系。成分数据作为一种十分重要的数据类型，在经济学、管理学、环境科学、医学等诸多领域都有广泛的应用[1] [2] [3] [4]。

D 元成分数据所张成的向量空间称为单形空间，单形空间中的成分数据需要满足定和约束，给后续的分析工作带来了很大的困难，针对普通数据的传统统计分析方法对于成分数据不再适用，主要表现在以下三个方面：1)单形空间内观察到的直观形态不能按照欧氏空间直角坐标系内的方式来解释；2)由于定和约束，按照传统统计方法计算得到的成分数据协方差矩阵是奇异矩阵，具有明显的负偏性，与普通数据协方差矩阵的内涵截然不同[5]；3)适用于普通数据的统计分析工具基本上都建立在“数据总体服从多元正态分布”的假设之上，而单形空间上的成分数据却缺乏一个适当的参数分布，使得在对数据的变异模式进行分析时存在参数建模的困难。实践证明，不带限制条件的普通数据分析方法对成分数据是失效的。

为了解决成分数据的建模困难，已有文献主要通过某种变换方法将成分数据降维，消除冗余度，转换为欧氏空间上的普通数据，再进行统计建模分析，从而发展了一系列有关成分数据的方法与模型。这些变换方法主要包括非对称对数比变换法、中心化对数比变换法和等距对数比变换法。Aitchison [6]研究了一元成分数据的降维技术，提出基于中心化对数比协方差的对数衬度主成分分析，克服了基于原始成分数据协方差阵的主成分不能准确反映变异最大方向的问题。在成分数据的不同分布假设下，张尧庭[7]采用中心化对数变换方法解约束后，讨论了自变量或因变量为一个成分数据的回归分析方法。王惠文等[8]首次提出用每一个成分数据表示一个主题含义，采用中心化对数变换，利用偏最小二乘路径模型，建立一元成分数据关于多元成分数据的回归模型。李春轩等[9]提出基于等距对数比变换的成分数据空间插值法，与基于另外两种变换方法的插值结果进行对比研究。Wang 等[10]基于成分数据的 Aitchison 代数体系以及对数比变换，研究了多元成分数据的主成分分析方法。Pawlowsky-Glahn *et al.* [11]在单形空间中建立了因变量为一元成分数据，自变量为若干个普通变量的回归模型，创新之处在于回归系数均为成分数据，推导了基于等距对数比变换的等价模型，并给出参数估计的普通最小二乘解及参数的解释含义。将成分数据进行等距对数比变换后，PETRA *et al.* [12]采用向量自回归模型建立了一元成分数据的时间序列分析模型，讨论了参数估计及性质等问题。

以上研究成果都是将成分数据从单形空间变换到欧氏空间后再开展建模分析，主要着眼于如何对变换后的数据创新模型，解决建模问题，而对于所选择的成分数据变换方法却鲜有深入研究。事实上，基于某种变换方法的模型是否准确合理，首先取决于所采用的变换方法是否保持空间的等价性，即是否构成正交变换，使得变换前后数据的代数性质及几何性质保持不变。郭丽娟等[13]提出基于等距对数比变换的成分数据判别分析模型，并在实证研究中发现，基于等距对数比变换的判别结果优于使用其他变换方法所得到的判别模型。然而，目前为止，有关几种常用变换方法的等价性分析尚缺乏理论依据及比较研究。因此，本文将系统地对非对称对数比变换、中心化对数比变换和等距对数比变换等三种方法进行比较分析，引入单形空间代数体系，基于能否构成单形空间到欧氏空间上正交变换的角度，从理论上证明三种变换方法的合理性，以期为成分数据变换方法的选择问题提供理论依据。作为实证分析，选取地质学中岩石分类问题，分别采用以上方法对原始成分数据进行变换后，再建立 Fisher 判别模型，比较研究基于不同变换方法的判别模型可靠性。

2. 成分数据变换方法的理论比较

Aitchison 等在对成分数据统计方法的研究中发现，研究比值的对数统计量可以克服成分数据的定和约束，从而将成分数据转换到普通的欧氏空间。以下对应用比较广泛的三种成分数据变换方法做简要介绍，包括非对称对数比变换、中心化对数比变换和等距对数比变换。

2.1. 非对称对数比变换

记单形空间为

$$S^D=\left\{\boldsymbol{x}=[x_1,\cdots,x_D]\,\middle|\,x_i>0,i=1,\cdots,D;\sum_{i=1}^{D}x_i=1\right\}$$

对任意 D 元成分数据 $\boldsymbol{x}\in S^D$，非对称变换比变换(additive logratio transformation，简称 *alr*)定义为一个将 S^D 映射到 R^{D-1} 的函数

$$alr(\boldsymbol{x})=\boldsymbol{u}=(u_1,\cdots,u_{D-1}),$$

其中

$$u_j = \ln\left(\frac{x_j}{x_D}\right),\ j = 1,\cdots,D-1 \tag{2}$$

选用 $u_j = \ln\left(\frac{x_j}{x_D}\right)$ 作为分析变量有许多便利之处。首先，在式(2)的变换中，成分数据从原来的 D 维空间被降低到 $D-1$ 维空间，原来的 D 个线性相关的变量 $x_i\ (i=1,\cdots,D)$ 被转换成 $D-1$ 个线性无关的变量 $u_j\ (j=1,\cdots,D-1)$，消除了原成分数据中的冗余维度；其次，由于 u_j 在 $(-\infty,+\infty)$ 内取值，在后续的分析中可以更为灵活地选择模型；再次，由于进行了对数变化，有可能把非线性问题线性化；第四，根据 Aitchison 的研究，如果成分数据 $\boldsymbol{x}$ 遵从加法逻辑正态分布，则变换后数据 $\boldsymbol{u}$ 服从正态分布，便可以运用基于正态分布假设的传统统计方法进行分析。

2.2. 中心化对数比变换

由于式(2)中的非对称变换使得变量的物理含义发生较大变化，故而模型的解释意义被削弱了。为了实现对成分数据的对称处理，Aitchison [6]又提出了中心化对数比变换(centered logratio transformation，简称 *clr*)，即

$$clr(\boldsymbol{x}) = \boldsymbol{v} = (v_1,\cdots,v_D),$$

其中

$$v_j = \ln\frac{x_j}{\sqrt[D]{\prod_{i=1}^{D} x_j}},\ j = 1,\cdots,D.$$

采用中心化对数比变换方法，变换后的各分量 $v_j\ (j=1,\cdots,D)$ 仍保持对称性，所建模型的可解释性就更强。但是，很容易验证，变换后各分量之和由“单位和”转化为“零和”，即 $\sum_{j=1}^{D} v_j = 0$。换言之，定和约束没有从本质上被克服，各分量间仍然存在完全相关性，后续的统计建模仍然存在困难。

2.3. 等距对数比变换

为了从本质上认识单形空间及成分数据的性质特征，Aitchison 等[14]经过不断努力，提出了一套完整的单形空间代数体系，被称为 Aitchison 代数体系。

定义 1 对于任意 D 元正实数向量 $\boldsymbol{z} = [z_1,\cdots,z_n] \in R_+^D$，定义关于 $\boldsymbol{z}$ 的闭合运算为

$$C(\boldsymbol{z}) = \left[\frac{z_1}{\sum_{i=1}^{D} z_i}, \frac{z_2}{\sum_{i=1}^{D} z_i}, \cdots, \frac{z_D}{\sum_{i=1}^{D} z_i}\right].$$

任意正实数向量 $\boldsymbol{z} \in R_+^D$ 都可以通过闭合运算转化成 S^D 上的成分数据。Aitchison 代数体系建立在闭合运算的基础之上。

定义 2 对任意成分数据 $\boldsymbol{x}_1 = [x_{11}, x_{12},\cdots,x_{1D}]$, $\boldsymbol{x}_2 = [x_{21}, x_{22},\cdots,x_{2D}] \in S^D$，以及任意 $\alpha \in R$，定义单形空间中的加法、数乘和内积分别为

$$\boldsymbol{x}_1 \oplus \boldsymbol{x}_2 = C(x_{11}x_{21}, x_{12}x_{22},\cdots,x_{1D}x_{2D})$$

$$\alpha \otimes \boldsymbol{x}_1 = C(x_{11}^{\alpha}, x_{12}^{\alpha},\cdots,x_{1D}^{\alpha})$$

$$\langle \boldsymbol{x}_1, \boldsymbol{x}_2 \rangle_{\alpha} = \frac{1}{2D}\sum_{i=1}^{D}\sum_{j=1}^{D} \ln\frac{x_{1i}}{x_{1j}} \ln\frac{x_{2i}}{x_{2j}}.$$

容易证明，基于 Aitchison 代数体系的单形空间是一个希尔伯特空间，有非常好的数学性质。

基于该代数体系，Egozcue 等[15]构建了从单形空间 S^D 到欧氏空间 R^{D-1} 上的形如式(3)所示的映射方式，称为等距对数比变换方法(isometric logratio transformation，简称 *ilr*)

$$ilr(\mathrm{x}) = \left(\langle \boldsymbol{x}, \boldsymbol{e}_1 \rangle_\alpha, \langle \boldsymbol{x}, \boldsymbol{e}_2 \rangle_\alpha, \cdots, \langle \boldsymbol{x}, \boldsymbol{e}_{D-1} \rangle_\alpha\right) \tag{3}$$

其中，$\boldsymbol{e}_i \left(i = 1, \cdots, D-1\right)$ 为单形空间的一组标准正交基，选用不同的标准正交基，可获得不同的变换结果，若选用如式(4)的标准正交基

$$\boldsymbol{e}_i = C\left[\exp\left(\underbrace{\sqrt{\frac{1}{i(i+1)}}, \cdots, \sqrt{\frac{1}{i(i+1)}}}_{i \text{ elements}}, -\sqrt{\frac{i}{i+1}}, 0, \cdots, 0\right)\right] \tag{4}$$

则式(3)可进一步简化为：$ilr(\boldsymbol{x}) = \boldsymbol{y} = (y_1, y_2, \cdots, y_{D-1})$，其中

$$y_j = \sqrt{\frac{j}{j+1}} \ln\left[\frac{g(x_1, \cdots, x_j)}{x_{j+1}}\right], j = 1, \cdots, D-1$$

这里，$g(x_1, \cdots, x_j)$ 表示 $x_1, \cdots, x_j$ 的几何均值。等距对数比变换将单形空间上的成分数据 $\boldsymbol{x} = [x_1, x_2, \cdots, x_D]$ 变换为欧氏空间 R^{D-1} 上的普通向量 $\boldsymbol{y} = (y_1, y_2, \cdots, y_{D-1})$。

Egozcue 等[15]同时还给出了等距对数比变换的逆变，即

$$\boldsymbol{x} = ilr^{-1}(\boldsymbol{y}) = \bigoplus_{i=1}^{D-1} (y_i \otimes \boldsymbol{e}_i) \tag{5}$$

2.4. 变换方法的比较

不论采用何种变换方法对成分数据进行预处理，目的都是为了在变换后的空间中找到合适的统计分析方法对成分数开展建模分析。从这个角度来看，三种对数变换法都将成分数据转换到普通的欧氏空间中，适用于普通数据的统计分析方法都可以被有效地应用。然而，评价一种变换方法时，更为重要的衡量标准是能否构成正交变换，即变换前后是否保持数学性质以及几何特征不发生变化。如果不具有这样的性质，那么在变换后的样本空间中得到的统计分析结论，就不能代表原始变量空间和样本空间的特征。

对于任意成分数据 $\boldsymbol{x}_1, \boldsymbol{x}_2 \in S^D$ 以及任意实数 $\alpha \in R$，以下定理不难证明。

定理 1 等距对数比变换 (*ilr*)是从 S^D 到 R^{D-1} 上的正交变换，不仅保持加法和数乘运算，而且保持内积不发生变化，即

1) $ilr(\boldsymbol{x}_1 \oplus \boldsymbol{x}_2) = ilr(\boldsymbol{x}_1) + ilr(\boldsymbol{x}_2)$ (6)

2) $ilr(\alpha \otimes \boldsymbol{x}_1) = \alpha \cdot ilr(\boldsymbol{x}_1)$ (7)

3) $\langle \boldsymbol{x}_1, \boldsymbol{x}_2 \rangle_a = \langle ilr(\boldsymbol{x}_1), ilr(\boldsymbol{x}_2) \rangle$ (8)

证明：由于基于 Aitchison 代数体系的单形空间是一个希尔伯特空间，因而容易证明

$$\begin{aligned}
&ilr(\boldsymbol{x}_1 \oplus \boldsymbol{x}_2)\\
&= \left(\langle \boldsymbol{x}_1 \oplus \boldsymbol{x}_2, \boldsymbol{e}_1 \rangle_\alpha, \langle \boldsymbol{x}_1 \oplus \boldsymbol{x}_2, \boldsymbol{e}_2 \rangle_\alpha, \cdots, \langle \boldsymbol{x}_1 \oplus \boldsymbol{x}_2, \boldsymbol{e}_{D-1} \rangle_\alpha\right)\\
&= \left(\langle \boldsymbol{x}_1, \boldsymbol{e}_1 \rangle_\alpha + \langle \boldsymbol{x}_2, \boldsymbol{e}_1 \rangle_\alpha, \langle \boldsymbol{x}_1, \boldsymbol{e}_2 \rangle_\alpha + \langle \boldsymbol{x}_2, \boldsymbol{e}_2 \rangle_\alpha, \cdots, \langle \boldsymbol{x}_1, \boldsymbol{e}_{D-1} \rangle_\alpha + \langle \boldsymbol{x}_2, \boldsymbol{e}_{D-1} \rangle_\alpha\right)\\
&= \left(\langle \boldsymbol{x}_1, \boldsymbol{e}_1 \rangle_\alpha, \langle \boldsymbol{x}_1, \boldsymbol{e}_2 \rangle_\alpha, \cdots, \langle \boldsymbol{x}_1, \boldsymbol{e}_{D-1} \rangle_\alpha\right) + \left(\langle \boldsymbol{x}_2, \boldsymbol{e}_1 \rangle_\alpha, \langle \boldsymbol{x}_2, \boldsymbol{e}_2 \rangle_\alpha, \cdots, \langle \boldsymbol{x}_2, \boldsymbol{e}_{D-1} \rangle_\alpha\right)\\
&= ilr(\boldsymbol{x}_1) + ilr(\boldsymbol{x}_2)
\end{aligned}$$

$$
\begin{aligned}
&ilr\left(\alpha\otimes\boldsymbol{x}_1\right)\\
&=\left(\left\langle\alpha\otimes\boldsymbol{x}_1,\boldsymbol{e}_1\right\rangle_\alpha,\left\langle\alpha\otimes x_1,\boldsymbol{e}_2\right\rangle_\alpha,\cdots,\left\langle\alpha\otimes\boldsymbol{x}_1,\boldsymbol{e}_{D-1}\right\rangle_\alpha\right)\\
&=\left(\alpha\left\langle\boldsymbol{x}_1,\boldsymbol{e}_1\right\rangle_\alpha,\alpha\left\langle\boldsymbol{x}_1,\boldsymbol{e}_2\right\rangle_\alpha,\cdots,\alpha\left\langle\boldsymbol{x}_1,\boldsymbol{e}_{D-1}\right\rangle_\alpha\right)\\
&=\alpha\left(\left\langle\boldsymbol{x}_1,\boldsymbol{e}_1\right\rangle_\alpha,\left\langle\boldsymbol{x}_1,\boldsymbol{e}_2\right\rangle_\alpha,\cdots,\left\langle\boldsymbol{x}_1,\boldsymbol{e}_{D-1}\right\rangle_\alpha\right)\\
&=\alpha\cdot ilr\left(\boldsymbol{x}_1\right)
\end{aligned}
$$

以下证明公式(8)。设在单形空间 S^D 上，成分数据 $\boldsymbol{x}_1$ 与 $\boldsymbol{x}_2$ 经过 ilr 变换，分别得到欧氏空间 R^{D-1} 上的向量 $\boldsymbol{y}_1=ilr\left(\boldsymbol{x}_1\right)$ 和 $\boldsymbol{y}_2=ilr\left(\boldsymbol{x}_2\right)$。由公式(5)所示的逆变换公式，可得

$$
\begin{aligned}
\left\langle\boldsymbol{x}_1,\boldsymbol{x}_2\right\rangle_\alpha&=\left\langle\bigoplus_{i=1}^{D-1}\left(y_{1i}\otimes\boldsymbol{e}_i\right),\bigoplus_{j=1}^{D-1}\left(y_{2j}\otimes\boldsymbol{e}_j\right)\right\rangle_\alpha\\
&=\sum_{i=1}^{D}\sum_{j=1}^{D}\left\langle y_{1i}\otimes\boldsymbol{e}_i,y_{2j}\otimes\boldsymbol{e}_j\right\rangle_\alpha\\
&=\sum_{i=1}^{D}\sum_{j=1}^{D}y_{1i}y_{2j}\left\langle\boldsymbol{e}_i,\boldsymbol{e}_j\right\rangle_\alpha\\
&=\sum_{i=1}^{D}y_{1i}y_{2i}=\left\langle ilr\left(\boldsymbol{x}_1\right),ilr\left(\boldsymbol{x}_2\right)\right\rangle
\end{aligned}
$$

定理 2 非对称对数比变换(*alr*)保持加法和数乘运算，但不能保持内积不变，即

1) $alr\left(\boldsymbol{x}_1\oplus\boldsymbol{x}_2\right)=alr\left(\boldsymbol{x}_1\right)+alr\left(\boldsymbol{x}_2\right)$,

2) $alr\left(\alpha\otimes\boldsymbol{x}_1\right)=\alpha\cdot alr\left(\boldsymbol{x}_1\right)$,

3) $\left\langle\boldsymbol{x}_1,\boldsymbol{x}_2\right\rangle_a\neq\left\langle alr\left(\boldsymbol{x}_1\right),alr\left(\boldsymbol{x}_2\right)\right\rangle$.

定理 3 中心化对数比变换(*clr*)是从 S^D 到 R^{D-1} 上的正交变换，保持加法和数乘运算，同时保持内积不发生变化，即

1) $clr\left(\boldsymbol{x}_1\oplus\boldsymbol{x}_2\right)=clr\left(\boldsymbol{x}_1\right)+clr\left(\boldsymbol{x}_2\right)$

2) $clr\left(\alpha\otimes\boldsymbol{x}_1\right)=\alpha\cdot clr\left(\boldsymbol{x}_1\right)$

3) $\left\langle\boldsymbol{x}_1,\boldsymbol{x}_2\right\rangle_a=\left\langle clr\left(\boldsymbol{x}_1\right),clr\left(\boldsymbol{x}_2\right)\right\rangle$

定理 2 和定理 3 的证明与定理 1 相似，本文不再赘述。由以上定理可知：

1) 非对称对数比变换只构成单形空间 S^D 到欧氏空间 R^{D-1} 上的线性变换，而非正交变换。由于不保持内积，变换前后由内积诱导的点点距离、模长以及向量间夹角等几何概念不具有等价性。有可能导致变换后的样本空间形态发生变化。因此，基于该变换的统计分析方法不能从根本上保证结论的合理性和准确性。

2) 中心化对数比变换虽然是一个正交变换，但是该变换不能消除冗余度，经过中心化对数变换后的变量之间仍然存在“和为零”的约束，这种变量之间的完全共线性导致了协方差矩阵不满秩，使得基于协方差结构的统计方法完全失效；此外，完全共线性给多元统计分析方法带来了建模和模型解释上的诸多困难。因此，在实际应用中，应当避免使用中心化对数比变换对成分数据进行预处理。

3) 等距对数比变换不仅消除了成分数据的定和约束，消除了冗余维度，而且构造了单形空间到欧氏空间上的正交变换，不仅能够保持向量的加法运算与数乘运算，还能够保持内积运算，从而保持了向量的模长、夹角以及点点距离不变等良好性质，实现了从单形空间 S^D 到欧氏空间 R^{D-1} 的等价变换，确保在变换后的欧氏空间中应用传统统计分析方法建模的合理性。

3. 实证结果——基于 Fisher 判别模型的比较分析

为验证上述比较研究的实效性，本文选取岩石标本判别分类问题，分别采用不同的变换方法将成分数据转化为普通数据，再利用判别分析模型对标本进行判别归类，并比较不同方法下判别的效率。由于中心化对数比变换仍然无法消除定和约束，变换后变量间的完全共线性可能使判别模型失效，因此本文只采用非对称对数比变换和等距对数比变换两种方法进行数据预处理。

3.1. 数据

选取 20 个岩石标本，其中 10 个属于 A 类岩石，10 个属于 B 类岩石。每个岩石标本所含五种矿物质的百分比与该岩石的类型有着密切的联系，以五种矿物质占比为判别变量对 20 个标本建立 Fisher 判别模型。数据来源于文献[16]。

3.2. 建模与比较

由于含五种矿物质的百分比构成成分数据 $\boldsymbol{x}=\left[x_1,x_2,x_3,x_4,x_5\right]$，变量之间受到“定和约束”的限制，适用于普通数据的 Fisher 判别方法对成分数据不再有效。因此，首先对原始数据进行等距对数比变换，原变量空间转化为 4 维欧氏空间，以变换后的 4 个变量 $\boldsymbol{y}=\left(y_1,y_2,y_3,y_4\right)$ 作为判别变量、20 个岩石标本作为样本，建立 Fisher 判别模型。

根据数据，求得 $\left|\boldsymbol{B}-\lambda\boldsymbol{E}\right|$ 的最大特征值为 $\lambda_1=3.427$，对应的特征向量为 $\boldsymbol{w}_1=\left(-0.739,-0.253,-20.711,16.162\right)$，从而 Fisher 线性判别函数为 $\mu\left(\boldsymbol{y}\right)=-0.739y_1-0.253y_2+20.711y_3+16.162y_4$ 计算两类总体的均值向量并代入判别函数，求出阈值 $\bar{\mu}=23.015$。据此得到判别规则为：对于任意岩石标本 $\boldsymbol{x}=\left[x_1,x_2,x_3,x_4,x_5\right]$，经过等距对数比变换得到向量 $\boldsymbol{y}=\left(y_1,y_2,y_3,y_4\right)$，代入判别函数，求得 $\mu\left(\boldsymbol{y}\right)$。当 $\mu\left(\boldsymbol{y}\right)>\bar{\mu}$ 时，属于 A 类岩石，当 $\mu\left(\boldsymbol{y}\right)<\bar{\mu}$ 时，属于 B 类岩石；$\mu\left(\boldsymbol{y}\right)=\bar{\mu}$ 时，待判。

利用所求的 Fisher 判别模型对训练样本进行样本内的拟合结果见表 1，只有一个岩石样品被错判，准确率为 95%。

接下来，对原始数据进行非对称对数比变换，再利用变换后的数据表建立 Fisher 判别模型，将训练样本回代得到的判别结果如表 2 所示，准确率只有 65%。

Table 1. Results of discriminant analysis based on isometric logratio transformation
表 1. 基于等距对数比变换判别模型的判别结果

实际类别	预测结果		合计
	A 岩石	B 岩石	
A 岩石	9	1	10
B 岩石	0	10	10

Table 2. Results of discriminant analysis based on additive logratio transformation
表 2. 基于非对称对数比变换判别模型的判别结果

实际类别	预测结果		合计
	A 岩石	B 岩石	
A 岩石	6	4	10
B 岩石	3	7	10

从模型的比较结果来看，基于非对称对数变换的判别模型发生误判的可能性较高，其原因在于，非对称对数比变换不构成正交变换，使用该方法对数据进行变换改变了原有样本点距离的空间特征，而Fisher 判别模型是一种基于距离进行判别归类的分类方法，因而在变换后的空间中建立判别模型未能准确地将样本点判别归类。而基于等距对数比变换的模型保证了变换前后样本空间的等价性，在变换后的空间中，样本点特征并未发生任何改变，建立判别模型与在原单形空间中判别归类是等价的，能够客观准确地将未知类对象进行归类。

4. 结论

本文针对三种常见的成分数据变换方法进行比较研究，从理论上分明证明了三种变换方法能否实现单形空间到欧氏空间的等价转换，并选取岩石判别分类问题，比较基于不同变换方法的判别结果。结论表明，基于 Aitchison 代数体系提出的等距对数比变换是一个正交变换，不仅实现了从单形空间到普通欧氏空间的转化，还保持了代数运算、内积以及由内积诱导的几何概念不发生任何变化；等距对数比变换既克服了非对称对数比变换改变内积及距离等几何概念的缺陷，又避免了中心化对数比变换导致的多重共线性给多元分析方法带来的影响。

因此，在经过等距对数比变换后的欧氏空间中讨论成分数据与直接在单形空间中基于 Aitchison 代数体系的研究是完全等价的，该变换方法很好地解决了成分数据统计建模中遇到的困难，具有非常重要的应用价值。

基金项目

国家自然科学基金资助项目(71401192)；首都流通业研究基地资助项目(JD-YB-2018-017)；北京市社会科学基金研究基地项目(15JDJGB076)。

参考文献

[1] 洪冬, 韩晟, 管晓东, 等. 基于成分数据分析法的医院药品费用结构变化预测研究[J]. 中国新药杂志, 2015, 24(9): 965-971.

[2] Buccianti, A. and Pawlowsky-Glahn, V. (2005) New Perspectives on Water Chemistry and Compositional Data Analysis. *Mathematical Geology*, **37**, 703-727. https://doi.org/10.1007/s11004-005-7376-6

[3] Jarautabragulat, E., Hervadasala, C., Egozcue, J.J., *et al*. (2015) Air Quality Index Revisited from a Compositional Point of View. *Mathematical Geosciences*, **48**, 581-593. https://doi.org/10.1007/s11004-015-9599-5

[4] Snyder, R.D., Ord, K., Koehler, A.B., *et al*. (2015) Forecasting Compositional Time Series: A State Space Approach. *Monash Econometrics and Business Statistics Working Papers*, Monash University.

[5] Billheimer, D., Guttorp, P. and Fagan, W.F. (1998) Statistical Analysis and Interpretation of Discrete Compositional Data. National Center for Statistics and the Environment (NRCSE) Technical Report NRCSE-TRS.

[6] Aitchison, J. (1983) Principal Component Analysis of Compositional Data. *Biomertrika*, **70**, 57-65. https://doi.org/10.1093/biomet/70.1.57

[7] 张尧庭. 成分数据统计分析引论[M]. 北京: 科学出版社, 2000.

[8] 王惠文, 张志慧, Tenenhaus, M. 成分数据的多元回归建模方法研究[J]. 管理科学学报, 2006, 9(4): 27-32.

[9] 李春轩, 罗毅, 包安明, 等. 基于对数比转换的成分数据空间插值研究[J]. 中国农业科学, 2012, 45(4): 648-655.

[10] Wang, H., Shangguan, L.Y., Guan, R., *et al*. (2015) Principal Component Analysis for Compositional Data Vectors. *Computational Statistics*, **30**, 1079-1096. https://doi.org/10.1007/s00180-015-0570-1

[11] Pawlowsky-Glahn, V., Egozcue, J.J. and Tolosana-Delgado, R. (2015) Modeling and Analysis of Compositional Data. John Wiley & Sons, Ltd.

[12] Kynclová, P., Filzmoser, P. and Hron, K. (2015) Modeling Compositional Time Series with Vector Autoregressive Models. *Journal of Forecasting*, **34**, 303-314. https://doi.org/10.1002/for.2336

[13] 郭丽娟, 王惠文, 关蓉. 基于等距 logratio 变换的成分数据判别分析方法[J]. 系统工程, 2016, 34(2): 153-158.

[14] Aitchison, J., Barceló-Vidal, C., Egozcue, J.J., *et al*. (2002) A Concise Guide to the Algebraic-Geometric Structure of the Simplex, the Sample Space for Compositional Data Analysis. *Proceedings of IAMG*, **2**, 387-392.

[15] Egozcue, J.J., Pawlowsky-Glahn, V., Mateu-Figueras, G., *et al*. (2003) Isometric Logratio Transformations for Compositional Data Analysis. *Mathematical Geology*, **35**, 279-300. https://doi.org/10.1023/A:1023818214614

[16] Aitchison, J. (1986) The Statistical Analysis of Compositional Data. Chapman and Hall, London. https://doi.org/10.1007/978-94-009-4109-0