基于核磁共振代谢组学的成分数据分析在中药评价中的应用

陈佳佳[1]，李爱平[2]，张晓琴[1*]，秦雪梅[2]，李胜家[1]

1. 山西大学数学科学学院，山西 太原 030006
2. 山西大学 中医药现代研究中心，山西 太原 030006

摘 要：核磁共振代谢组学数据预处理时常常需要对谱峰面积进行归一化，而归一化后数据的相对比例与原始谱图中峰面积的相对比例相同，因此归一化后数据能反映谱峰面积的相对信息。选取每个样本对应的谱峰面积归一化后数据进行分析，由于成分数据是仅含有相对信息的非负向量，故归一化后的数据可以被考虑为是成分数据。基于核磁共振代谢组学的成分数据分析可研究中药评价中各组样本的均一性、寻找影响不同组分类的特征代谢物、对给定的新样本进行判别分析。黄芪质量评价的实例分析结果证明提出的方法是可行的。

关键词：代谢组学；成分数据；Aitchison 距离；均一性；差异代谢物；判别分析

中图分类号：F282 **文献标志码：**A **文章编号：**0253 - 2670(2016)19 - 3522 - 05

DOI: 10.7501/j.issn.0253-2670.2016.19.027

Compositional data analysis method based on NMR metabolomics: An application to evaluate Chinese materia medica

CHEN Jia-jia[1], LI Ai-ping[2], ZHANG Xiao-qin[1], QIN Xue-mei[2], LI Sheng-jia[1]

1. School of Mathematical Sciences, Shanxi University, Taiyuan 030006, China
2. Modern Research Center for Traditional Chinese Medicine, Shanxi University, Taiyuan 030006, China

Abstract: The peak area normalization is often needed as the preprocessing of NMR metabolomic data and the relative ratio of normalized data is the same as that of peak area in original spectrum, so the normalized data can reflect the relative information of peak area. This research selects the normalized data of the peak area for each sample to analyze, compositional data is the nonnegative vector containing only the relative information. Therefore the normalized data can be considered as compositional data. This paper used compositional data analysis based on NMR metabolomics to study the homogeneity for each group's samples, identify the characteristic metabolites which contribute the classification of different groups, and make the discriminate analysis for the given new sample in the evaluation of Chinese materia medica. In the case the analysis of quality evaluation on *Astragali Radix*, it can be seen from the results that the proposed method is feasible.

Key words: metabolomics; compositional data; Aitchison distance; homogeneity; differential metabolites; discriminate analysis

成分数据描述的是整体中的部分，例如岩石的化学成分比例、病人血液的不同细胞类型的浓度、饮料的营养素浓度、投票选举比例等。传统的成分数据定义为含有常数和约束的非负向量[1]。由于成分数据的成分和是无意义的，因此成分数据的定义推广为仅含有相对信息的非负向量[2-3]，一般用行向量 $x=(x_1, x_2, \cdots, x_D)$ 来表示，对应的样本空间是单行空间 S^D。

$$S^D=\left\{x=(x_1,x_2,\cdots,x_D)\,|\,x_i>0,i=1,2,\cdots,D;\sum_{i=1}^{D}x_i=c\right\}$$

c 是任意的常数，依赖于测量的单位，通常取 1 或 100

1986 年，Aitchison[1]系统地研究了成分数据的统计分析，指出成分数据研究的是数据间的相对信息，而非绝对信息，提出用对数比例变换来避免常数和约束。

成分数据最早出现在地理科学中[4-5]，现广泛应用于经济、化学、生物及其他学科中[6]。然而，只有很少的文献把成分数据应用在代谢组学中[7]。代谢组学[8]是对某一生物或细胞在一特定生理时期内所有低相对分子质量代谢产物同时进行定性和定量分析的一门新学科。基于核磁共振（NMR）[9]的代谢组学，谱峰面积数据获得流程：供试样品经测试，

收稿日期：2016-05-10

基金项目：山西省高等学校教学改革项目（J2014006）；山西省自然科学基金资助项目（2015011044）；山西省国际交流合作项目（2015081020）

作者简介：陈佳佳（1991—），女，博士在读，研究方向为成分数据分析。E-mail: chenjiajia0401@163.com

***通信作者** 张晓琴（1975—），女，博士，副教授，研究方向为统计机器学习、成分数据分析。E-mail: zhangxiaoqin@sxu.edu.cn

通过傅里叶变换获得 NMR 指纹图谱，经过定标、相位和基线校正，以合适步长进行积分，导出谱峰面积数据矩阵。通常需要对谱峰面积数据进行归一化处理，常用的归一化方法有线归一化、面归一化、模归一化[10]。无论用哪种归一化方法，归一化后数据的相对比例是不变的，都等于原始谱图中谱峰面积的相对比例，因此归一化后数据是含有相对信息的向量。由于归一化后数据是非负的，所以归一化后数据可以被考虑为是成分数据。

代谢组学领域常用的多元统计分析方法有主成分分析、聚类分析、偏最小二乘法-判别分析。通过主成分分析的得分图，可以直观看出样本的分类情况、比较同类样本的组内间距以及不同类的组间间距。偏最小二乘法-判别分析[11]常被用来寻找差异标志物，基于 VIP 值和 S-plot 图来选择差异代谢物。找出差异代谢物后，通过不同组样本间的 t 检验，进而得到显著的差异代谢物，即特征代谢物。跟欧氏空间上的普通数据相比，成分数据所属的单行空间上有相应的运算和度量，因此传统的统计分析方法不能直接应用于成分数据。考虑到成分数据特有的度量向量空间结构，本文基于成分数据分析研究代谢组学中常常需要涉及的问题：（1）样本初始分布状态，即各组样本的相似性；（2）在确定样本的分类后，需要对分组贡献大的关键差异成分进行表征；（3）建立的分类模型有望对新的未知样本进行预测，即进行判别分析。

本文对成分数据的基本知识进行简要介绍，提出基于核磁共振代谢组学的成分数据分析方法，并以黄芪质量评价为例对该方法进行验证。

1 成分数据的基本知识

成分数据的基本运算、度量及其描述性统计的定义[1-3]如下。对于任意的成分数据 $x\in S^D, y\in S^D$ 和实数 $\alpha\in R$，加法运算和乘法运算定义为 $x\oplus y=C(x_1y_1, x_2y_2, \cdots, x_Dy_D)$，$\alpha\otimes x=C(x_1^\alpha, x_2^\alpha, \cdots, x_D^\alpha)$，其中 C 为闭合运算，即每个成分除以所有成分的总和，再乘以 c。

成分数据 $x\in S^D$ 和 $y\in S^D$ 的 Aitchison 距离定义为 $d_a(x, y)$。

$$d_a(x,y)=\sqrt{\sum_{i=1}^{D}\left(\ln\frac{x_i}{g_m(x)}-\ln\frac{y_i}{g_m(y)}\right)^2}$$

$g_m(x)=\sqrt[D]{x_1x_2\cdots x_D}$ 和 $g_m(y)=\sqrt[D]{y_1y_2\cdots y_D}$ 分别为成分数据 x 和 y 的几何均值

给定成分数据集：

$$X=\begin{pmatrix}x_{(1)}\\x_{(2)}\\\vdots\\x_{(n)}\end{pmatrix}=\begin{pmatrix}x_{11}&x_{12}&\cdots&x_{1D}\\x_{21}&x_{22}&\cdots&x_{2D}\\\vdots&\vdots&\ddots&\vdots\\x_{n1}&x_{n2}&\cdots&x_{nD}\end{pmatrix}=(x_1,x_2,\cdots,x_D)\quad(1)$$

其中 $x_{(i)}=(x_{i1}, x_{i2}, \cdots, x_{iD})\in S^D$ 为第 i（i=1, 2, …, n）个成分数据样本，$x_j=(x_{1j}, x_{2j}, \cdots, x_{nj})^T$ 为 n 个成分数据样本的第 j（j=1, 2, …, D）个成分组成的列向量。

公式（1）中的成分数据集 X 的样本中心定义为：

$$\mathrm{cen}(X)=\frac{1}{n}\otimes\left(\bigoplus_{i=1}^{n}x_{(i)}\right)=C\left(\sqrt[n]{\prod_{i=1}^{n}x_{i1}},\sqrt[n]{\prod_{i=1}^{n}x_{i2}},\cdots,\sqrt[n]{\prod_{i=1}^{n}x_{iD}}\right)$$

公式（1）中的成分数据集 X 的离差可以用方差矩阵来描述，定义为：

$$\mathrm{T}(X)=\begin{pmatrix}t_{11}&t_{12}&\cdots&t_{1D}\\t_{21}&t_{22}&\cdots&t_{2D}\\\vdots&\vdots&\ddots&\vdots\\t_{D1}&t_{D2}&\cdots&t_{DD}\end{pmatrix},\quad t_{ij}=\mathrm{var}\left(\ln\frac{x_i}{x_j}\right)$$

公式（1）中的成分数据集 X 的总离差的测量是总方差，定义为：

$$\mathrm{totvar}(X)=\frac{1}{2D}\sum_{i=1}^{D}\sum_{j=1}^{D}t_{ij}$$

给定成分数据 $x\in S^D$，子成分定义为 $x_S=C(x\mathrm{S})$，其中 S 是 $D\times S$（$D\leqslant S$）的选择矩阵，该矩阵中每列有一个元素为 1，每行最多有一个元素为 1，其余元素为 0。同理，对于公式（1）中的成分数据集 X，子成分数据集 X_S 定义为：

$$X_S=\begin{pmatrix}C(x_{(1)}\mathrm{S})\\C(x_{(2)}\mathrm{S})\\\vdots\\C(x_{(n)}\mathrm{S})\end{pmatrix}$$

2 基于核磁共振代谢组学的成分数据分析方法

代谢组学一般研究多组样本，假定有 G 组样本，则所有样本的谱峰面积归一化后数据可以用 $n\times D$ 的数据集 X 表示。

$$X=\begin{pmatrix}x^1\\x^2\\\vdots\\x^G\end{pmatrix}=(x_1, x_2, \mathrm{K}, x_D),\ X^g=\begin{pmatrix}x_{(1)}^g\\x_{(2)}^g\\\vdots\\x_{(n_g)}^g\end{pmatrix}\quad(g=1, 2, \cdots, G)$$

其中 X 的每行表示每个样本的谱峰面积归一化后数据，每列表示某个代谢物，x_j (j=1, 2, …, D) 为

第 j 个变量，X^g (g=1, 2, …, G) 代表第 g 组样本数据集，n_g 为第 g 组的样本个数，且 $n_1+n_2+\cdots+n_g=n$，$x_{(i)}^g \in S^D$ (i=1, 2, …, n_g) 为第 g 组样本数据集的第 i 个样本。随后研究不同组样本的均一性评价、寻找表征不同组的代谢物谱差异的特征代谢物、根据寻找的特征代谢物对给定的新样本进行判别分析。

2.1 不同组样本的均一性评价

对于同一组的样本数据集 X^g (g=1, 2, …, G)，如果用图形来表示样本点，样本点越密集，即样本点离数据集中心的距离越小，则该样本数据集的均一性越好。因此，所有样本点与中心的 Aitchison 距离的平方和的平均值可以当作是样本数据集的均一性评价指标，即

$$D(X^g)=\frac{1}{n_g}\sum_{i=1}^{n_g} d_a^2\left(x_{(i)}^g, \mathrm{cen}(X^g)\right) \quad (2)$$

如果 $D(X^g)$ 越小，则该组数据集的均一性越好。

2.2 筛选特征代谢物

方差用来衡量数据的波动程度，所以方差大的变量有可能是不同组的差异代谢物。差异代谢物选取方法的基本思想是从方差小的那些变量开始依次剔除，直到上一步和下一步的总方差的相对误差大于给定的临界值时停止剔除变量。不失一般性，假定 var (x_1)≥var (x_2)≥…≥var (x_D)，计算步骤如下：

（1）令 k=0，定义 $X_0=X$ 为原始的样本数据集，计算 totvar (X_0)。

（2）剔除变量 x_{D-k}。记 $X_{k+1}=(x_1, x_2, \cdots, x_{D-k-1})$ 为剔除变量 $x_{D-k}, x_{D-k+1}, \cdots, x_D$ 后的子成分样本数据集，计算 totvar (X_{k+1})。根据子成分一致性原则[1]，有 totvar (X_k)≥totvar (X_{k+1})。

（3）当上一步和下一步总方差的相对误差 $\frac{\mathrm{totvar}(X_k)-\mathrm{totvar}(X_{k+1})}{\mathrm{totvar}(X_k)}<\tau$（$\tau$=10%）时，令 $k=k+1$，重复（2），否则停止剔除变量。

通过以上迭代过程，X_k 为最终选取的子成分数据集，$x_1, x_2, \cdots, x_{D-k}$ 为最终选取的差异变量。接下来用 t 检验对差异变量的任意 2 组样本进行显著性检验，当有 P<0.05 时，该变量即为显著的差异代谢物，即特征代谢物。

2.3 新样本进行判别分析

用基于 Aitchison 距离的 k 近邻法来验证寻找的特征代谢物对新样本的分类能力，基本思路：根据 Aitchison 距离找到离预测实例最近的 k 个样本点，基于多数表决的规则，这 k 个样本点的多数属于某个类，则预测实例就属于这个类。

3 黄芪质量评价的实例分析

研究甘肃黄芪和山西黄芪的化学成分，数据来源于文献报道[12]，分别为 8 批甘肃移栽速生芪和 8 批山西传统野生黄芪。采用传统水煎，冷冻干燥后，用氘代重水溶解进行 ^{1}H-NMR 测试，所得自由衰减信号导入 MestReNova 软件（version 8.0.1，Mestrelab Research，Santiago de Compostella，Spain），以 δ 0.04 积分段对化学位移区间 0.78～9.22 进行分段积分，其中 δ 4.66～5.06 残留水峰不进行积分，导出数据矩阵进行统计分析。样本 X 数据集为：

$$X=\begin{pmatrix}X^1\\X^2\end{pmatrix}=(x_1, x_2, \cdots, x_{201}),\ X^1=\begin{pmatrix}x_{(1)}^1\\x_{(2)}^1\\\vdots\\x_{(8)}^1\end{pmatrix},\ X^2=\begin{pmatrix}x_{(1)}^2\\x_{(2)}^2\\\vdots\\x_{(8)}^2\end{pmatrix}$$

其中 X^1 代表甘肃黄芪，X^2 代表山西黄芪，x_j (j=1, 2, …, 201) 为第 j 个变量。根据公式（2）计算得出 $D(X^1)$=79.360 0，$D(X^2)$=83.806 5，因此甘肃黄芪均一性相比山西的黄芪较好。实际上，甘肃黄芪作为市场上主流商品黄芪，一般生长 2 年，且加工后几乎拥有相同的直径和长度，因而各批所含化学成分的量相对均匀，而山西黄芪作为传统野生黄芪，据传统经验一般生长 5 年以上，但具体年限不详，因而本实验收集的不同批山西黄芪的均一性相对较差。根据“2.2”项方法选取特征代谢物，结果见表 1，该方法除找到与文献报道[12]一样的特征代谢物外，还找到特征代谢物异亮氨酸、缬氨酸、精氨酸、谷氨酰胺、β-木糖、α-葡萄糖、苯丙氨酸。根据找到的特征代谢物对给定的样本进行判别分析，采用留一交叉法验证，每次选取一个样本当作预测样本，其余样本当作训练样本，用基于 Aitchison 距离的 k 近邻法来验证分类的准确性，结果见表 2。从表 2 可以看出，对于每个给定的样本都分类正确。基于文献报道[12]找到的特征代谢物，运用同样的判别分析方法对给定的样本进行判别分析，分类准确度达 100%。

4 结语

迄今为止，对于代谢组学的数据分析有很多方法，但鲜有学者将代谢组学数据考虑为成分数据。本文基于成分数据的知识来研究代谢组学常常关心的问题，实例分析结果表明样本均一性评价与实际情况相符、筛选的特征代谢物与文献报道[12]一致、判别分析的准确性高。在之后的研究中，希望基于

表 1 特征代谢物对应的的变量、化学位移和化合物名称

Table 1 Corresponding variable, chemical shift, and compound of characteristic metabolites

变量	化学位移 (δ)	化合物	变量	化学位移 (δ)	化合物
x_4	0.94	亮氨酸	x_{53}, x_{55}	2.90, 2.98	天冬酰胺
x_5	0.98	异亮氨酸	x_{39}, x_{57}	2.34, 3.06	γ-羟基丁酸
x_7	1.06	缬氨酸	x_{61}	3.22	胆碱
x_{14}	1.34	苏氨酸	x_{63}	3.30	甜菜碱
x_{18}	1.50	丙氨酸	x_{96}	4.62	β-木糖
x_{23}	1.70	精氨酸	x_{102}	5.26	α-葡萄糖
x_{29}	1.94	醋酸	x_{103}	5.30	α-半乳糖
x_{31}, x_{33}	2.02, 2.10	脯氨酸	x_{106}	5.42	蔗糖
x_{35}	2.18	谷氨酰胺	x_{107}	5.46	棉子糖
x_{42}	2.46	琥珀酸	x_{154}	7.34	苯丙氨酸
x_{49}	2.74	天冬氨酸			

表 2 基于 Aitchison 距离的 *k* 近邻法判别分析结果

Table 2 Results of discriminate analysis by *k* nearest neighbor method based on Aitchison distance

预测样本	K (k = 6) 近邻样本	类别	预测样本	K (k = 6) 近邻样本	类别
$x^1_{(1)}$	$x^1_{(2)}$, $x^1_{(3)}$, $x^1_{(4)}$, $x^1_{(5)}$, $x^1_{(6)}$, $x^1_{(8)}$	甘肃	$x^2_{(1)}$	$x^2_{(3)}$, $x^2_{(4)}$, $x^2_{(5)}$, $x^2_{(6)}$, $x^2_{(7)}$, $x^2_{(8)}$	山西
$x^1_{(2)}$	$x^1_{(1)}$, $x^1_{(3)}$, $x^1_{(4)}$, $x^1_{(6)}$, $x^1_{(7)}$, $x^1_{(8)}$	甘肃	$x^2_{(2)}$	$x^2_{(1)}$, $x^2_{(3)}$, $x^2_{(4)}$, $x^2_{(5)}$, $x^2_{(7)}$, $x^2_{(8)}$	山西
$x^1_{(3)}$	$x^1_{(1)}$, $x^1_{(2)}$, $x^1_{(4)}$, $x^1_{(5)}$, $x^1_{(6)}$, $x^1_{(8)}$	甘肃	$x^2_{(3)}$	$x^2_{(1)}$, $x^2_{(4)}$, $x^2_{(5)}$, $x^2_{(6)}$, $x^2_{(7)}$, $x^2_{(8)}$	山西
$x^1_{(4)}$	$x^1_{(1)}$, $x^1_{(2)}$, $x^1_{(3)}$, $x^1_{(6)}$, $x^1_{(7)}$, $x^1_{(8)}$	甘肃	$x^2_{(4)}$	$x^2_{(1)}$, $x^2_{(3)}$, $x^2_{(5)}$, $x^2_{(6)}$, $x^2_{(7)}$, $x^2_{(8)}$	山西
$x^1_{(5)}$	$x^1_{(1)}$, $x^1_{(2)}$, $x^1_{(3)}$, $x^1_{(6)}$, $x^1_{(7)}$, $x^1_{(8)}$	甘肃	$x^2_{(5)}$	$x^2_{(1)}$), $x^2_{(3)}$), $x^2_{(4)}$, $x^2_{(6)}$, $x^2_{(7)}$, $x^2_{(8)}$	山西
$x^1_{(6)}$	$x^1_{(1)}$, $x^1_{(2)}$, $x^1_{(3)}$, $x^1_{(4)}$, $x^1_{(7)}$, $x^1_{(8)}$	甘肃	$x^2_{(6)}$	$x^2_{(1)}$, $x^2_{(3)}$, $x^2_{(4)}$, $x^2_{(5)}$, $x^2_{(7)}$, $x^2_{(8)}$	山西
$x^1_{(7)}$	$x^1_{(1)}$, $x^1_{(2)}$, $x^1_{(3)}$, $x^1_{(4)}$, $x^1_{(6)}$, $x^1_{(8)}$	甘肃	$x^2_{(7)}$	$x^2_{(1)}$, $x^2_{(3)}$, $x^2_{(4)}$, $x^2_{(5)}$, $x^2_{(6)}$, $x^2_{(8)}$	山西
$x^1_{(8)}$	$x^1_{(1)}$, $x^1_{(2)}$, $x^1_{(3)}$, $x^1_{(4)}$, $x^1_{(6)}$, $x^1_{(7)}$	甘肃	$x^2_{(8)}$	$x^2_{(1)}$, $x^2_{(3)}$, $x^2_{(4)}$, $x^2_{(5)}$, $x^2_{(6)}$, $x^2_{(7)}$	山西

成分数据的回归分析或相关分析来研究若干因变量与自变量的相互依赖关系，例如植物代谢组学研究中，环境因素（气温、湿度、日照等）与影响不同产地植物分类的差异代谢物之间的相互依赖关系，或临床生化指标与体液代谢组学的潜在生物标志物之间的相互依赖关系。

参考文献

[1] Aitchison J. *The Statistical Analysis of Compositional Data* [M]. London: Chapman & Hall, 1986.

[2] Pawlowsky-Glahn V, Buccianti A. *Compositional Data Analysis: Theory and Applications* [M]. Chichester: John Wiley & Sons Ltd., 2011.

[3] Pawlowsky-Glahn V, Egozcue J J, Tolosana-Delgado R. *Modeling and Analysis of Compositional Data* [M]. Chichester: John Wiley & Sons Ltd., 2015.

[4] Sun X L, Wu Y J, Wang H L, *et al*. Mapping soil particle size fractions using compositional kriging, cokriging and additive log-ratio cokriging in two case studies [J]. *Math Geosci*, 2014, 46(4): 429-443.

[5] Tolosana-Delgado R, Von Eynatten H. Simplifying compositional multiple regression: Application to grain size controls on sediment geochemistry [J]. *Comput Geosci*, 2010, 36(5): 577-589.

[6] Lin W, Shi P, Feng R, *et al*. Variable selection in regression with compositional covariates [J]. *Biometrika*, 2014, 101(4): 785-797.

[7] Kalivodová A, Hron K, Filzmoser P, *et al*. PLS-DA for compositional data with application to metabolomics [J]. *J Chemometr*, 2015, 29(1): 21-28.

[8] Nicholson J K, Connelly J, Lindon J C, *et al*.

Metabonomics: a platform for studying drug toxicity and gene function [J]. *Nat Rev Drug Discov*, 2002, 1(2): 153-161.

[9] Kim H K, Choi Y H, Verpoorte R. NMR-based metabolomic analysis of plants [J]. *Nat Prot*, 2010, 5(3): 536-549.

[10] 温锦波, 杨叔禹, 肖 娴, 等. 基于核磁共振的代谢组学数据预处理 [J]. 厦门大学学报: 自然科学版, 2007, 46(6): 783-787.

[11] Xiong A Z, Yang L, Ji L L, *et al*. UPLC-MS based metabolomics study on *Senecio scandens* and *S. vulgaris*: an approach for the differentiation of two Senecio herbs with similar morphology but different toxicity [J]. *Metabolomics*, 2011, 8(4): 614-623.

[12] Li A P, Li Z Y, Sun H F, *et al*. Comparison of two different *Astragali Radix* by a ^{1}H-NMR-based metabolomic approach [J]. *J Proteome Res*, 2015, 14(5): 2005-2016.