成分数据的空间自回归模型

Tingting Huang, Huiwen Wang, Gilbert Saporta

▶ **To cite this version:**

Tingting Huang, Huiwen Wang, Gilbert Saporta. 成分数据的空间自回归模型. Journal of Beijing University of Aeronautics and Astronautics, 2019, 45 (1), pp.93-98. ⟨10.13700/j.bh.1001-5965.2018.0253⟩. ⟨hal-02471589⟩

HAL Id: hal-02471589

https://cnam.hal.science/hal-02471589v1

Submitted on 8 Feb 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2018.0253

成分数据的空间自回归模型

黄婷婷[1,2], 王惠文[1,3,*], SAPORTA Gilbert[4]

(1. 北京航空航天大学 经济与管理学院, 北京 100083; 2. 城市运行应急保障模拟技术北京市重点实验室, 北京 100083;
3. 北京航空航天大学 大数据科学与脑机智能高精尖创新中心, 北京 100083;
4. 法国国立工艺学院 计算机和通信研究中心, 巴黎 75003)

摘 要: 针对已有成分数据线性回归模型对研究对象相互独立的严格要求,提出了含有成分数据和普通数据的空间自回归模型,在此基础上提出了成分数据空间自回归模型的估计方法。新模型结合了空间自回归模型处理因变量之间相互依赖的优势,可同时处理成分数据和普通数据。通过利用等距对数比(ilr)变换将成分数据解约束,得到了新模型的参数估计量。蒙特卡罗模拟实验验证了所提估计方法的有效性。

关 键 词: 成分数据; 等距对数比(ilr)变换; 极大似然估计; 空间依赖; 空间自回归模型
中图分类号: F222
文献标识码: A **文章编号**: 1001-5965(2019)01-0093-06

数据搜集技术的快速发展不仅带来了海量的数据,也带来了类型越来越复杂的数据,如函数数据[1-3]、成分数据[4]和符号数据[5-6]等。在这些类型复杂的数据中,成分数据由于关注部分在总体中的占比信息,受到愈来愈广泛的关注。如 Fry 等[7]利用住户开支统计调查结果研究预算分配模型,Pawlowsky-Glahn 和 Egozcue[8]利用成分数据比较东欧和西欧国家在食物消费结构上的习惯差异,Pawlowsky-Glahn[9]等利用成分数据回归模型分析了巴西宗教信仰构成的变化。

成分数据分析主要研究活动对象结构变化产生的规律及其对其他对象产生的影响。关于成分数据的理论研究,标志性的成果是 1986 年 Aitchison 撰写的《成分数据统计分析》[10],该书详细阐述了成分数据统计分析方法建立的数学基础。在成分数据分析中,线性回归模型是一种常用的分析技术。现有的成分数据线性回归模型可以分为两大类:第 1 类因变量是普通数据[11-12],第 2 类因变量是成分数据[13-15]。Hron 等[12]利用第 1 类成分数据线性回归模型研究了 GDP 组成与预期寿命的关系;而 Wang 等[14]利用第 2 类模型研究了地区总产值与就业和投资的关系。本文在因变量是普通数据的成分数据回归模型基础上进行研究。在成分数据回归模型中,通常以样本之间独立同分布作为前提。而在实际应用中,独立同分布的假设往往是不成立的。如何对现有的成分数据线性回归模型进行改进,使之适应实际应用的需求,是一个值得深入研究的问题。

在空间计量经济学[16]中,空间自回归模型通过引入空间依赖项,打破了因变量相互独立的假设,使得许多与空间地理位置或社交网络有关的现象得到解释。利用空间自回归模型,可以对区域经济发展的问题[17-18]、溢出性问题[19-20]等进行分析。现有的空间自回归模型在普通数据的基础上已经发展得相对完善,已有的对空间自回归模型进行估计的方法包括 Ord[21]和 Lee[22]提出的极

收稿日期: 2018-05-03; **录用日期**: 2018-07-28; **网络出版时间**: 2018-08-23 10:38
网络出版地址: kns.cnki.net/kcms/detail/11.2625.V.20180821.1326.003.html
基金项目: 国家自然科学基金 (71420107025)
* **通信作者**. E-mail: wanghw@vip.sina.com

引用格式: 黄婷婷, 王惠文, SAPORTA Gilbert. 成分数据的空间自回归模型[J]. 北京航空航天大学学报, 2019, 45(1): 93-98.
HUANG T T, WANG H W, SAPORTA Gilbert. Spatial autoregressive model for compositional data[J]. Journal of Beijing University of Aeronautics and Astronautics, 2019, 45(1): 93-98 (in Chinese).

http://bhxb.buaa.edu.cn

大似然估计法、Kelejian、Prucha[23] 和 Lee[24] 提出的广义矩估计法、Lesage 和 Pace[25] 从贝叶斯的角度提出的马尔可夫链蒙特卡罗方法(Markov chain Monte Carlo method)。

因此,针对经典成分数据线性回归模型假设样本间相互独立的严格要求,研究因变量之间具有空间依赖的成分数据回归模型,通过在普通数据的空间自回归模型中,引入成分数据的协变量,提出了同时含有成分数据和普通数据的空间自回归模型。并依据成分数据的特点,给出了混合 2 种数据的空间自回归模型的估计方法。提出的新模型比已有的成分数据线性回归模型具有更强的灵活性,可以处理更加复杂的空间依赖问题。

1 基础理论

本节主要介绍成分数据的代数空间——单形空间(simplex)中的基本运算,以及与成分数据联系紧密的几种变换,利用这些变换可以将具有约束的成分数据转化成易于处理的普通数据。

1.1 单形空间

对于含有 d 个成分的成分数据,对应的单形空间 S^d(上标 d 表示成分数据有 d 个成分,因此实际是 $d-1$ 维的)定义为

$$S^d = \{\boldsymbol{x} = (x_1, x_2, \cdots, x_d)^{\mathrm{T}}, \quad x_i > 0, i = 1, 2, \cdots, d; \sum_{i=1}^{d} x_i = k\} \tag{1}$$

式中:$\boldsymbol{x}$ 为一个 d 维的成分数据;$x_i > 0$ 表示成分数据的每一个成分都是非负的;$\sum_{i=1}^{d} x_i = k$ 为成分数据必须满足的约束条件,即各成分累加和是个定值。不失一般性,在本文中令 $k = 1$。在 S^d 中,基本的运算包括加法运算、数乘运算、内积运算。

现有单形空间 S^d 中的任意 2 个成分数据 $\boldsymbol{x}$、$\boldsymbol{y}$ 以及实数 α,记 $\boldsymbol{x} = (x_1, x_2, \cdots, x_d)^{\mathrm{T}} \in S^d$,$\boldsymbol{y} = (y_1, y_2, \cdots, y_d)^{\mathrm{T}} \in S^d$,$\alpha \in \mathbf{R}$,则 $\boldsymbol{x}$ 和 $\boldsymbol{y}$ 的加法 $\oplus$ 及 α 和 $\boldsymbol{x}$ 数乘运算 $\odot$ 可分别定义为

$$\boldsymbol{x} \oplus \boldsymbol{y} = C(x_1 y_1, x_2 y_2, \cdots, x_d y_d) \tag{2}$$

$$\alpha \odot \boldsymbol{x} = C(x_1^{\alpha}, x_2^{\alpha}, \cdots, x_d^{\alpha}) \tag{3}$$

式中:$C(\cdot)$ 表示闭合运算,定义为

$$C(x_1, x_2, \cdots, x_d) = \left(\frac{x_1}{\sum_{i=1}^{d} x_i}, \frac{x_2}{\sum_{i=1}^{d} x_i}, \cdots, \frac{x_d}{\sum_{i=1}^{d} x_i}\right)^{\mathrm{T}} \tag{4}$$

不难看出,闭合运算保证了运算结果仍在 S^d 中。基于运算 $\oplus$ 和 $\odot$,可以导出 $\boldsymbol{x}$ 和 $\boldsymbol{y}$ 的减法运算,

$$\boldsymbol{x} - \boldsymbol{y} = \boldsymbol{x} \oplus (-1 \odot \boldsymbol{y}) = C\left(\frac{x_1}{y_1}, \frac{x_2}{y_2}, \cdots, \frac{x_d}{y_d}\right)^{\mathrm{T}} \tag{5}$$

$\boldsymbol{x}$ 和 $\boldsymbol{y}$ 的内积运算 $\langle \boldsymbol{x}, \boldsymbol{y} \rangle_{\mathrm{a}}$ 定义为

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle_{\mathrm{a}} = \sum_{i=1}^{d} \ln \frac{x_i}{g_m(\boldsymbol{x})} \ln \frac{y_i}{g_m(\boldsymbol{y})} \tag{6}$$

式中:$g_m(\boldsymbol{x}) = \left(\prod_{i=1}^{d} x_i\right)^{1/d}$ 为 $\boldsymbol{x}$ 各个成分的几何平均值;同理可定义 $g_m(\boldsymbol{y}) = \left(\prod_{i=1}^{d} y_i\right)^{1/d}$;内积符号 $\langle \boldsymbol{x}, \boldsymbol{y} \rangle_{\mathrm{a}}$ 的下标 a 表示该运算在单形空间 S^d 中。内积运算还可以导出单形空间中任意一个成分数据 $\boldsymbol{x}$ 的范数 $\|\boldsymbol{x}\|_{\mathrm{a}}$ 及任意 2 个成分数据 $\boldsymbol{x}$ 和 $\boldsymbol{y}$ 之间的距离 $d_{\mathrm{a}}(\boldsymbol{x}, \boldsymbol{y})$,其定义分别为

$$\|\boldsymbol{x}\|_{\mathrm{a}} = \sqrt{\langle \boldsymbol{x}, \boldsymbol{y} \rangle_{\mathrm{a}}} = \sqrt{\sum_{i=1}^{d} \left(\ln \frac{x_i}{g_m(\boldsymbol{x})}\right)^2} \tag{7}$$

$$d_{\mathrm{a}}(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|_{\mathrm{a}} = \sqrt{\sum_{i=1}^{d} \left(\ln \frac{x_i}{g_m(\boldsymbol{x})} - \ln \frac{y_i}{g_m(\boldsymbol{y})}\right)^2} \tag{8}$$

可以证明,含有内积运算的单形空间是一个希尔伯特空间。

1.2 等距对数比变换

需注意,因约束 $\sum_{i=1}^{d} x_i = 1$ 的存在,成分数据 $\boldsymbol{x} = (x_1, x_2, \cdots, x_d)^{\mathrm{T}}$ 的 d 个成分之间不相互独立,直接将多元统计方法应用到成分数据上会产生矛盾的结果。已有的对成分数据进行变换的方法主要为对数比变换,具体类型包括加法对数比(alr)变换,中心对数比(clr)变换和等距对数比(ilr)变换。由于 alr 变换不是等距变换,而 clr 变换得到的变量是线性相关的,不便于直接用于回归建模,因此此处仅介绍 ilr 变换。

ilr 变换是 Egozcue 等[26] 提出的。该变换将 d 维的单形空间 S^d 映射到 $d-1$ 维的欧几里得空间 $\mathbf{R}^{d-1}$ 上,得到的实数向量消除了原成分数据中不同成分之间的共线性,可以直接用于建模。该变换利用标准正交基的正交性和单位长度性质,将成分数据变换成易于处理的标准正交基的系数。设标准正交基为 $\{\boldsymbol{e}_k\}_{k=1}^{d-1}$,$\boldsymbol{e}_k = (e_{k1}, e_{k2}, \cdots, e_{kd})^{\mathrm{T}}$,则任意一个成分数据 $\boldsymbol{x}$ 都可以表示为 $\boldsymbol{x} = \langle \boldsymbol{x}, \boldsymbol{e}_1 \rangle_{\mathrm{a}} \odot \boldsymbol{e}_1 \oplus \langle \boldsymbol{x}, \boldsymbol{e}_2 \rangle_{\mathrm{a}} \odot \boldsymbol{e}_2 \oplus \cdots \oplus \langle \boldsymbol{x}, \boldsymbol{e}_{d-1} \rangle_{\mathrm{a}} \odot \boldsymbol{e}_{d-1}$,相应地,$\boldsymbol{x}$ 的 ilr 变换坐标 $\mathrm{ilr}(\boldsymbol{x})$ 为

$$\mathrm{ilr}(\boldsymbol{x}) = (\langle \boldsymbol{x}, \boldsymbol{e}_1 \rangle_{\mathrm{a}}, \langle \boldsymbol{x}, \boldsymbol{e}_2 \rangle_{\mathrm{a}}, \cdots, \langle \boldsymbol{x}, \boldsymbol{e}_{d-1} \rangle_{\mathrm{a}})^{\mathrm{T}} \tag{9}$$

Egozcue 等[26] 证明,ilr 变换是保内积的变换,即对于含有 d 个成分的成分数据 $\boldsymbol{x}$ 和 $\boldsymbol{y}$,有

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle_{\mathrm{a}} = \langle \mathrm{ilr}(\boldsymbol{x}), \mathrm{ilr}(\boldsymbol{y}) \rangle \tag{10}$$

下面给出具体的 ilr 变换过程。

http://bhxb.buaa.edu.cn

已知观测到样本量 n 的 d 维成分数据 $\{\boldsymbol{C}_i\}_{i=1}^n$,其中 $\boldsymbol{C}_i=(x_{i1},x_{i2},\cdots,x_{id})^{\mathrm{T}}$,则 $\boldsymbol{C}_i$ 进行 ilr 变换后的坐标为

$$\boldsymbol{\xi}_i=\mathrm{ilr}(\boldsymbol{C}_i)=\mathrm{clr}(\boldsymbol{C}_i)\boldsymbol{\Psi}^{\mathrm{T}}=\ln(\boldsymbol{C}_i)\boldsymbol{\Psi}^{\mathrm{T}} \tag{11}$$

式中:

$$\mathrm{clr}(\boldsymbol{C}_i)=\left(\ln\frac{C_{i1}}{g_m(\boldsymbol{C}_i)},\ln\frac{C_{i2}}{g_m(\boldsymbol{C}_i)},\cdots,\ln\frac{C_{id}}{g_m(\boldsymbol{C}_i)}\right)$$

$\boldsymbol{\Psi}$ 为 $(d-1)\times d$ 维的矩阵,具体表达式为

$$\boldsymbol{\Psi}=\begin{bmatrix}\varphi_{11}&\cdots&\varphi_{1d}\\ \vdots&&\vdots\\ \varphi_{(d-1)1}&\cdots&\varphi_{(d-1)d}\end{bmatrix}$$

$$\varphi_{ij}=\begin{cases}\sqrt{\dfrac{1}{(d-i)(d-i+1)}} & j\leqslant d-1\\ -\sqrt{\dfrac{d-i}{d-i+1}} & j=d-i+1\\ 0 & \text{其他}\end{cases}$$

由于 ilr 变换是保内积的变换,因此在第 3 节的估计方法中,将使用变换后的坐标 $\{\boldsymbol{\xi}_i\}_{i=1}^n$ 代替原来的成分数据 $\{\boldsymbol{C}_i\}_{i=1}^n$ 进行参数估计。

2　模型的提出

借鉴 Qu 和 Lee[27] 对空间自回归模型的背景假设,考虑空间关系发生在一个非均匀分布的格子 L,$L\subset\mathbf{R}^p$,$p\geqslant 1$ 上,格子上的点相互可分,即任意 2 点的距离大于 0。从格子 L 上观测到了 n 个对象,每个对象的观测数据为 $\{y_i,x_{i1},\cdots,x_{id},x_{id+1},\cdots,x_{ip}\}_{i=1}^n$,其中 $x_{ij}(j=1,2,\cdots,d)$ 共同组成 d 个成分的成分数据 $\boldsymbol{C}_i=(x_{i1},x_{i2},\cdots,x_{id})^{\mathrm{T}}$,且每个 $\boldsymbol{C}_i$ 是随机成分数据 $\boldsymbol{C}$ 的独立同分布观测;$x_{ij}(j=d+1,d+2,\cdots,p)$ 为普通数据,它们是随机变量 $X_j(j=d+1,d+2,\cdots,p)$ 的独立同分布观测,标记 $\boldsymbol{X}_i=(x_{id+1},x_{id+2},\cdots,x_{ip})^{\mathrm{T}}$。记 $\boldsymbol{Y}=(y_1,y_2,\cdots,y_n)^{\mathrm{T}}$,$\boldsymbol{C}=(\boldsymbol{C}_1,\boldsymbol{C}_2,\cdots,\boldsymbol{C}_n)^{\mathrm{T}}$,$\boldsymbol{X}=(\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_n)^{\mathrm{T}}$,则因变量 $\boldsymbol{Y}$ 符合以下回归模型:

$$\boldsymbol{Y}=\alpha\boldsymbol{\tau}_n+\rho\boldsymbol{W}\boldsymbol{Y}+\langle\boldsymbol{C},\boldsymbol{B}\rangle_{\mathrm{a}}+\boldsymbol{X}\boldsymbol{\Gamma}+\boldsymbol{E} \tag{12}$$

式中:$\alpha\boldsymbol{\tau}_n$ 为截距项,$\boldsymbol{\tau}_n$ 为所有元素均为的 1 的维度为 n 的向量;ρ 为未知的空间自相关参数,取值在区间 $(-1,1)$ 内;$\boldsymbol{W}=\{w_{ij}\}_{n\times n}$ 为外生的空间矩阵,w_{ij} 为对象 i 与 j 之间的权重;$\boldsymbol{B}$ 为待估的成分数据系数,具有 p 个成分;$\boldsymbol{\Gamma}$ 为普通数据的待估系数;$\boldsymbol{E}$ 为独立于 $\boldsymbol{X}$ 的误差项,服从均值为 0,方差为 $\sigma^2\boldsymbol{I}_n$ 多元正态分布,$\boldsymbol{I}_n$ 为 $n\times n$ 的单位矩阵。

需强调的是,式(12)中 $\boldsymbol{C}$ 和回归系数 $\boldsymbol{B}$ 都为成分数据,$\langle\boldsymbol{C},\boldsymbol{B}\rangle_{\mathrm{a}}$ 为一个实数。在 Aitchison 内积空间中,$\langle\boldsymbol{C},\boldsymbol{B}\rangle_{\mathrm{a}}$ 代表 $\boldsymbol{X}$ 对 $\boldsymbol{Y}$ 解释性最强的投影方向。

当 $\rho=0$ 时,式(12)退化为普通的成分数据线性模型。在这个意义上,式(12)比经典的成分数据线性模型具有更强的灵活性,可以处理更加复杂的数据关系。

3　估计方法

为估计模型式(12)中的参数 $\alpha,\rho,\boldsymbol{B},\boldsymbol{\Gamma}$,首先需将相互不独立的成分数据转化为相互独立的普通数据,1.2 节中已作详细介绍;其次,要解决因变量 y_i 之间不相互独立的问题,此处采用极大似然估计法 ilr 变换后的模型进行估计。

同样利用 1.2 节中的 ilr 变换,可得到成分数据系数 $\boldsymbol{B}$ 的变换坐标 $\boldsymbol{b}=\mathrm{ilr}(\boldsymbol{B})$。

由于 $\boldsymbol{B}$ 是需估计的参数,因此变换后的坐标 $\boldsymbol{b}$ 是未知的。记 $\boldsymbol{\xi}=(\boldsymbol{\xi}_1,\boldsymbol{\xi}_2,\cdots,\boldsymbol{\xi}_n)^{\mathrm{T}}$,则模型式(12)可写为

$$\boldsymbol{Y}=\alpha\boldsymbol{\tau}_n+\rho\boldsymbol{W}\boldsymbol{Y}+\boldsymbol{\xi}\boldsymbol{b}+\boldsymbol{X}\boldsymbol{\Gamma}+\boldsymbol{E} \tag{13}$$

为描述简便,记:$\boldsymbol{\delta}=(\boldsymbol{b},\boldsymbol{\Gamma})^{\mathrm{T}}$,$\boldsymbol{Z}=(\boldsymbol{\xi},\boldsymbol{X})$,则式(13)可表示为

$$\boldsymbol{Y}=\alpha\boldsymbol{\tau}_n+\rho\boldsymbol{W}\boldsymbol{Y}+\boldsymbol{Z}\boldsymbol{\delta}+\boldsymbol{E} \tag{14}$$

由于模型式(12)中误差项服从多元正态分布,因变量 $\boldsymbol{Y}$ 的似然函数为

$$l(\rho,\boldsymbol{\delta},\sigma^2)=-\frac{n}{2}\ln(\pi\sigma^2)+\ln|\boldsymbol{I}_n-\rho\boldsymbol{W}|-\frac{\boldsymbol{e}^{\mathrm{T}}\boldsymbol{e}}{2\sigma^2} \tag{15}$$

式中:$\boldsymbol{e}=\boldsymbol{Y}-\alpha\boldsymbol{\tau}_n-\rho\boldsymbol{W}\boldsymbol{Y}-\boldsymbol{Z}\boldsymbol{\delta}$。因式(15)有 3 个未知参数 ρ、$\boldsymbol{\delta}$ 和 σ^2,直接对这 3 个变量求导存在一定的计算困难。现假若已得到 ρ 的估计值 $\hat{\rho}$,那么利用极大似然估计法,可以相应得到 $\boldsymbol{\delta}$ 和 σ^2 的估计量,它们分别为

$$\hat{\boldsymbol{\delta}}=(\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z})^{-1}\boldsymbol{Z}^{\mathrm{T}}(\boldsymbol{I}_n-\boldsymbol{W})\boldsymbol{Y} \tag{16}$$

$$\hat{\sigma}^2=\frac{1}{n}(\boldsymbol{Y}-\alpha\boldsymbol{\tau}_n-\rho\boldsymbol{W}\boldsymbol{Y}-\boldsymbol{Z}\boldsymbol{\delta})^{\mathrm{T}}\cdot(\boldsymbol{Y}-\alpha\boldsymbol{\tau}_n-\rho\boldsymbol{W}\boldsymbol{Y}-\boldsymbol{Z}\boldsymbol{\delta}) \tag{17}$$

考虑将似然函数式(15)中的变量 $\boldsymbol{\delta}$ 和 σ^2 分别用估计量 $\hat{\boldsymbol{\delta}}$ 和 $\hat{\sigma}^2$ 代替,那么似然函数中的 3 个变量就变成一个变量。式(15)替换后的表达式为

$$l(\rho)=c-\frac{n}{2}\ln(\hat{\sigma}^2)+\ln|\boldsymbol{I}_n-\rho\boldsymbol{W}| \tag{18}$$

式中:c 为一个常数。利用牛顿法等数值解法,可以得到的 ρ 的估计值 $\hat{\rho}$。相应地,通过式(16)和式(17)可以分别得到 $\boldsymbol{\delta}$ 和 σ^2 的估计量。

由于得到 $\hat{\boldsymbol{\delta}}$ 以后,可以得到 $\boldsymbol{b}$ 的估计量 $\hat{\boldsymbol{b}}$;再通过 ilr 变换的逆变换 ilr^{-1},就可得到 $\boldsymbol{B}$ 的估计量 $\hat{\boldsymbol{B}}$ 为

http://bhxb.buaa.edu.cn

$$\hat{\boldsymbol{B}} = \mathrm{ilr}^{-1}(\hat{\boldsymbol{b}}) = C(\exp(\hat{\boldsymbol{b}}\boldsymbol{\Psi})) \tag{19}$$

至此,所有参数都可以估计出来。

4 数值模拟

为评估所提出估计方法的统计性质,下面设计了几组数值模拟实验检验估计量的表现。所有的计算过程都是在 R 软件中实现,用到的包有“spdep”和“compositions”。

关于空间自回归模型的空间网络结构,采取最常见的“车”相邻(rook matrix)。假设 n 个样本点随机地散落在一个 R 行 T 列的格子棋盘上,每个样本点占据棋盘上的一个方格,那么在棋盘上共享一条边的 2 个样本点就是相邻的。在这样的情况下,处在棋盘中间的任意样本点都有 4 个邻居,处在棋盘边上的样本点有 3 个邻居,而处在棋盘角上的样本点只有 1 个邻居。分别设置 $R = 10,20,30$, $T = 30,25,30$,相应地样本量 $n = R \times T = 300,500,900$。为了查看空间依赖的强弱是否对估计量有影响,同样设计了 3 组不同的 ρ 值,$\rho = 0,0.5,0.8$。

关于混合数据的空间自回归模型,由于截距项不是主要关注的参数,此处设 $\alpha = 0$,其他参数设置如下:$\boldsymbol{Y} = \rho\boldsymbol{W}\boldsymbol{Y} + \langle \boldsymbol{B}, \boldsymbol{C} \rangle_a + \Gamma X + 0.8\boldsymbol{E}$;$\boldsymbol{C} \sim N_s(\boldsymbol{\mu}, \boldsymbol{\Sigma})$;$X \sim N(1, 0.8)$;$\boldsymbol{E} \sim N(\boldsymbol{0}, \boldsymbol{I}_n)$;$\boldsymbol{\mu} = (0.49, 0.61)^{\mathrm{T}}$;$\boldsymbol{\Sigma} = \begin{bmatrix} 2 & -1.5 \\ -1.5 & 2 \end{bmatrix}$;$\Gamma = 1$;$\boldsymbol{B} = (b_1, b_2, b_3)^{\mathrm{T}} = (2,1,1.5)^{\mathrm{T}}$。其中:$\boldsymbol{C} \sim N_s(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ 表示 $\boldsymbol{C}$ 进行 ilr 变换后的坐标服从均值为 $\boldsymbol{\mu}$,协方差为 $\boldsymbol{\Sigma}$ 的多元正态分布。此处 $\mathrm{ilr}^{-1}(\boldsymbol{\mu}) = (1,2,3)^{\mathrm{T}}$。

在每一种情形下,重复实验次数 $k = 100$。对于参数 ρ 和 Γ,用样本均值偏离真值的大小和样本标准差衡量估计量的表现。对于成分数据系数,用成分数据均值 $\bar{\boldsymbol{B}}$ 与真值的偏差以及成分数据的总方差 totvar($\boldsymbol{X}$)衡量估计结果的优劣。其中,样本均值的计算公式为

$$\bar{\boldsymbol{B}} = C(\bar{b}_1, \bar{b}_2, \cdots, \bar{b}_d), \quad \bar{b}_j = \frac{1}{k}\sum_{i=1}^{k} \ln(b_{ij}) \tag{20}$$

样本的总方差的计算公式为

$$\mathrm{totvar}(\boldsymbol{X}) = \frac{1}{2d}\sum_{j,k=1}^{d} \hat{v}_{jk} \tag{21}$$

其中:$\hat{v}_{jk} = \frac{1}{n}\sum_{i=1}^{n}\left(\ln\frac{b_{ij}}{b_{ik}} - \ln\frac{\bar{b}_j}{\bar{b}_k}\right)^2$。

估计结果如图 1 ~ 图 3 所示。可以得到如下结论:

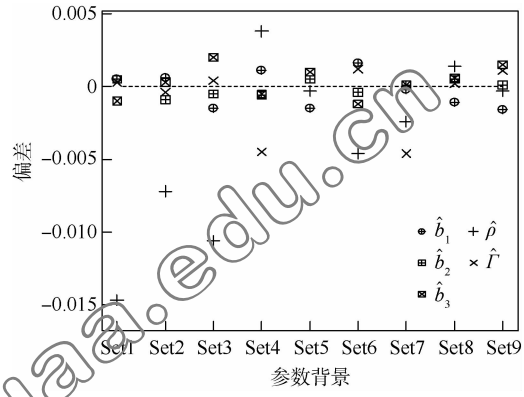
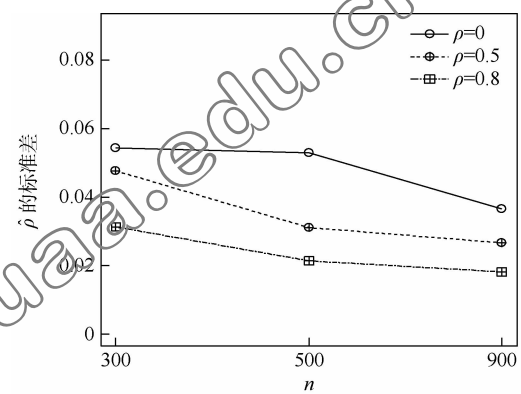
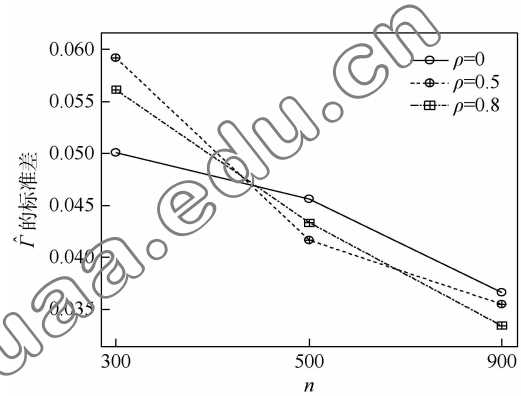


图 1 $\hat{\rho}$、$\hat{\Gamma}$、$\hat{b}_1$、$\hat{b}_2$ 和 $\hat{b}_3$ 的样本偏差

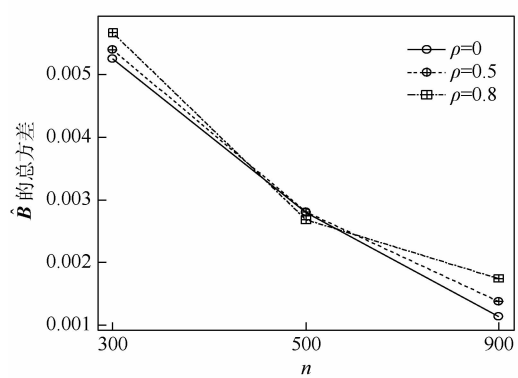
Fig. 1 Sample deviation of $\hat{\rho}$, $\hat{\Gamma}$, $\hat{b}_1$, $\hat{b}_2$ and $\hat{b}_3$



(a) ρ 取不同值时,$\hat{\rho}$ 的标准差随样本量的变化情况



(b) ρ 取不同值时,$\hat{\Gamma}$ 的标准差随样本量的变化情况



(c) ρ 取不同值时,$\hat{\boldsymbol{B}}$ 的总方差随样本量的变化情况

图 2 $\hat{\rho}$、$\hat{\Gamma}$的标准差及$\hat{\boldsymbol{B}}$的总方差

Fig. 2 Standard deviation of $\hat{\rho}$, $\hat{\Gamma}$ and total variance of $\hat{\boldsymbol{B}}$

http://bhxb.buaa.edu.cn

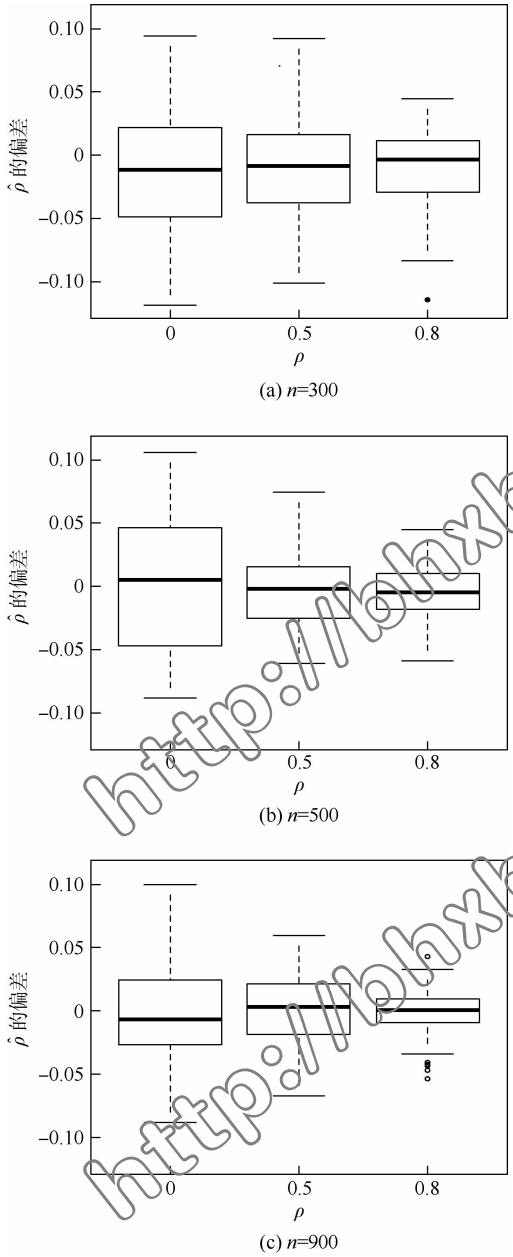


图 3　n 和 ρ 取不同值时,$\hat{\rho}$偏差箱线图

Fig. 3　Boxplots of deviation of $\hat{\rho}$ when n and ρ change

1) $\hat{\rho}$、$\hat{\Gamma}$、$\hat{b}_1$、$\hat{b}_2$、$\hat{b}_3$ 的样本均值在所有的参数设置背景下偏离实际值均较小。图 1 中给出了不同参数设置背景(Set1 ~ Set9 对应的(ρ,n)的取值分别为(300,0)、(300,0.5)、(300,0.8)、(500,0)、(500,0.5)、(500,0.8)、(900,0)、(900,0.5)和(900,0.8)共 9 种情况)下不同参数估计值偏差的散点图,可以发现,偏差的绝对值不超过 0.015,成分数据系数 3 个成分$\hat{b}_1$、$\hat{b}_2$、$\hat{b}_3$ 的偏差相对于$\hat{\rho}$和$\hat{\Gamma}$均较小。

2) $\hat{\rho}$和$\hat{\Gamma}$样本标准差及$\hat{\boldsymbol{B}}$的总方差随着样本量的增大而减小。从图 2 中可以看出,不论 ρ 取何值,随着 n 的增加,估计量的标准差或总方差折线都是减小的趋势。

3) 当样本量大小相同时,$\hat{\rho}$的样本标准差随着 ρ 值的增大而减小。从图 3 中可以看出,当 n 值固定时,随着 ρ 从 0 增加到 0.8,箱子越来越窄。

5　结　论

针对普通成分数据线性回归模型要求样本间相互独立的局限性,在空间自回归模型的基础上,提出了混合成分数据与普通数据的空间自回归模型,所提出的模型及估计方法具有如下优点:

1) 新提出的模型不仅能够同时处理成分数据和普通数据,还能表达数据中因变量之间相互依赖的问题。特别地,新模型可以处理地理空间中的依赖性。

2) 新模型所提出的估计量具有相合性。随着样本量的增大,可以发现估计值的标准差在逐渐减小。除此之外,新提出的估计方法操作简单,可以在 R 软件上直接实现。

在实际应用中,新模型可处理社交网络、地理空间等含有网络结构的依赖问题。而针对其他情况造成成分数据线性模型样本之间不相互独立的问题,则需要分情况进行深入分析。

参考文献(References)

[1] RAMSAY J O, SILVERMAN B W. Functional data analysis [M]. Berlin: Springer, 1997.

[2] RAMSAY J O, SILVERMAN B W. Applied functional data analysis: Methods and case studies [M]. Berlin: Springer, 2002.

[3] VIEU P, FERRATY F. Nonparametric functional data analysis [M]. Berlin: Springer, 2006.

[4] PAWLOWSKY-GLAHN V, BUCCIANTI A. Compositional data analysis: Theory and applications [M]. Chichester: Wiley-Blackwell, 2011.

[5] BILLARD L, DIDAY E. Symbolic regression analysis [M] // JAJUGA K, SOKOLOWSKI A, BOCK H. Classification, clustering, and data analysis. Berlin: Springer, 2002: 281-288.

[6] BILLARD L, DIDAY E. Regression analysis for interval-valued data [M]. Berlin: Springer, 2000: 369-374.

[7] FRY J M, FRY T R L, MCLAREN K R. Compositional data analysis and zeros in micro data [J]. Applied Economics, 2000, 32(8): 953-959.

[8] PAWLOWSKY-GLAHN V, EGOZCUE J J. Exploring compositional data with the CoDa-dendrogram [J]. Austrian Journal of Statistics, 2011, 40(1&2): 103-113.

[9] PAWLOWSKY-GLAHN V, EGOZCUE J J, TOLOSANA-DELGADO R. Modelling and analysis of compositional data [J]. Hoboken: John Wiley & Sons, Ltd., 2015: 152-154.

[10] AITCHISON J. The statistical analysis of compositional data

http://bhxb.buaa.edu.cn

[M]. Berlin:Springer,1986.
[11] AITCHISON J. The statistical analysis of compositional data [J]. Journal of the Royal Statistical Society Series B,1982,44(2):139-177.
[12] HRON K, FILZMOSER P, THOMPSON K. Linear regression with compositional explanatory variables[J]. Journal of Applied Statistics,2012,39(5):1115-1128.
[13] ATCHISON J, SHEN S M. Logistic-normal distributions: Some properties and uses[J]. Biometrika,1980,67(2):261-272.
[14] WANG H,SHANGGUAN L, WU J, et al. Multiple linear regression modeling for compositional data [J]. Neurocomputing, 2013,122:490-500.
[15] TOLOSANA-DELGADO R, EYNATTEN H V. Simplifying compositional multiple regression: Application to grain size controls on sediment geochemistry [J]. Computers & Geosciences, 2010,36(5):577-589.
[16] ANSELIN L. Spatial econometrics: Methods and models[M]. Berlin:Springer,1988.
[17] 林光平,龙志和,吴梅. 中国地区经济 σ 收敛的空间计量实证分析[J]. 数量经济技术经济研究,2006,23(4):14-21.
LIN G P, LONG Z H, WU M. A spatial investigation of σ-convergence in China[J]. The Journal of Quantitative & Technical Economics,2006,23(4):14-21(in Chinese).
[18] 郭金龙,王宏伟. 中国区域间资本流动与区域经济差距研究[J]. 管理世界,2003(7):45-58.
GUO J L, WANG H W. Study on the regional capital flows and regional economic differences in China [J]. Management World,2003(7):45-58(in Chinese).
[19] TOPA G. Social interactions, local spillovers and unemployment [J]. Review of Economic Studies,2010,68(2):261-295.
[20] BAICKER K. The spillover effects of state spending[J]. Journal of Public Economics,2005,89(2-3):529-544.
[21] ORD H. Estimation methods for models of spatial interaction [J]. Publications of the American Statistical Association,1975, 70(349):120-126.
[22] LEE L F. Asymptotic distributions of quasi-maximum likelihood estimators for spatial autoregressive models[J]. Econometrica, 2004,72(6):1899-1925.
[23] KELEJIAN H, PRUCHA I R. A generalized moments estimator for the autoregressive parameter in a spatial model[J]. International Economic Review,1999,40(2):509-533.
[24] LEE L F. GMM and 2SLS estimation of mixed regressive, spatial autoregressive models[J]. Journal of Econometrics,2007,137(2):489-514.
[25] LESAGE J P, PACE R K. Introduction to spatial econometrics [M]. New York:CRC Press,2009:513-514.
[26] EGOZCUE J J, PAWLOWSKY-GLAHN V, MATEUFIGUERAS G, et al. Isometric logratio transformations for compositional data analysis[J]. Mathematical Geology,2003,35(3):279-300.
[27] QU X, LEE L F. Estimating a spatial autoregressive model with an endogenous spatial weight matrix[J]. Journal of Econometrics,2015,184(2):209-232.

作者简介:

黄婷婷 女,博士研究生。主要研究方向:复杂数据的回归模型建模方法。

王惠文 女,博士,教授,博士生导师。主要研究方向:经济管理中复杂数据统计分析的理论、方法与应用。

Spatial autoregressive model for compositional data

HUANG Tingting[1,2], WANG Huiwen[1,3,*], SAPORTA Gilbert[4]

(1. School of Economics and Management, Beihang University, Beijing 100083, China;
2. Beijing Key Laboratory of Emergency Support Simulation Technologies for City Operations, Beijing 100083, China; 3. Beijing Advanced Innovation Center for Big Data and Brain Computing, Beihang University, Beijing 100083, China; 4. Centre d'études et de Recherche en Informatique et Communications, Conservatoire National des Arts et Métiers, Paris 75003, France)

Abstract: The existing compositional linear models assume that samples are independent, which is often violated in practice. To solve this problem, we put forward a spatial autoregressive model for compositional data, which contains both compositional covariates and scalar predictors. Furthermore, a new estimation method is proposed. The new model has advantages of coping with mixed compositional and numerical data and expressing dependence between the responses. And the parameter estimators are obtained through isometric logratio (ilr) transformation, which transforms dependent compositional data into independent real vector. A Monte-Carlo simulation experiment verifies the effectiveness of the proposed estimation method.

Keywords: compositional data; isometric logratio (ilr) transformation; maximum likelihood estimation; spatial dependence; spatial autoregressive model

Received: 2018-05-03; **Accepted**: 2018-07-28; **Published online**: 2018-08-23 10:38
URL: kns.cnki.net/kcms/detail/11.2625.V.20180821.1326.003.html
Foundation item: National Natural Science Foundation of China (71420107025)
* **Corresponding author**. E-mail: wanghw@vip.sina.com

http://bhxb.buaa.edu.cn