

一种基于Epanechnikov二次核的成分数据缺失值填补法 *

张晓琴^{*} 康菊 荆文君

(山西大学数学科学学院, 太原, 030006)

摘要

核函数方法已经被成功的用于各种函数的估计. 本文利用核函数的思想, 针对缺失数据造成现有的成分数据统计方法失效和k近邻填补法(KNNI)在利用缺失数据的k个近邻估计缺失数据时没有考虑到它们各自不同的贡献, 提出了一种基于Epanechnikov二次核的成分数据缺失值填补法(EKI)和对其进行修正后的Epanechnikov核成分数据缺失值填补法(MEKI). 实验结果表明, 基于修正的Epanechnikov二次核的成分数据缺失值填补法比k近邻填补法能够得到更为准确的估计.

关键词: 成分数据, 缺失值填补, k近邻填补法, Epanechnikov二次核, Aitchison距离.

学科分类号: O212.1.

§1. 引言

成分数据的概念最早来自于Ferrers (1866), 满足

$$S^D = \left\{ \boldsymbol{x} = (x_1, x_2, \dots, x_D); x_i > 0, i = 1, 2, \dots, D, \sum_{i=1}^D x_i = c \right\} \quad (1.1)$$

的空间称之为D维成分数据空间, 这里c是任意常数, S^D 中的元素是D维行向量. 在之后的讨论中, 本文取c = 1. 跟普通数据比较, 成分数据满足定和限制, 因此, 将传统的统计方法直接应用于成分数据将会导致不合理的结论. 对此, Aitchison (1986)一书中的核心思想就是对成分数据做对数比变换, 将成分数据映射到欧几里得空间中, 从而使得经典的统计方法可以适用于变换后的数据分析. 之后Egozcue等(2003)提出了等距对数比变换进行相应的研究. 本文采用等距对数比变换, 因为等距变换保证了变换前两组成分数据的Aitchison距离等于变换后两组数据的欧式距离, 这可以给计算上带来很大的方便.

下面给出本文所涉及到的一些相关概念和性质.

定义 1.1 (Egozcue等, 2003, 等距对数比变换) 设 $\boldsymbol{x} = (x_1, x_2, \dots, x_D) \in S^D$ 是一个D维成分数据, 令

$$z_i = \sqrt{\frac{D-i}{D-i+1}} \ln \frac{\sqrt[D-i]{\prod_{l=i+1}^D x_l}}{x_i}, \quad i = 1, \dots, D-1. \quad (1.2)$$

*国家自然科学基金重点项目(71031006)、国家青年基金项目(41101440)、山西省教育厅专项项目(20120301)和国家自然科学基金项目(81173366)资助.

^{*}通讯作者, E-mail: zhangxiaoqin@sxu.edu.cn.

本文2013年11月18日收到, 2014年7月1日收到修改稿.

doi: 10.3969/j.issn.1001-4268.2014.06.004

公式(1.2)把 D 维成分数据 \mathbf{x} 变换为 $D - 1$ 维向量 $\mathbf{z} = (z_1, z_2, \dots, z_{D-1})$, 这个变换称为“等距对数比变换”, 记作 $\mathbf{z} = ilr(\mathbf{x})$.

定义 1.2 (Egozcue等, 2003, Aitchison距离) 设 $\mathbf{x} = (x_1, x_2, \dots, x_D)$, $\mathbf{y} = (y_1, y_2, \dots, y_D) \in S^D$, 其Aitchison距离定义为

$$d_A(\mathbf{x}, \mathbf{y}) = \sqrt{\frac{1}{D} \sum_{i=1}^{D-1} \sum_{j=i+1}^D \left(\ln \frac{x_i}{x_j} - \ln \frac{y_i}{y_j} \right)^2}. \quad (1.3)$$

可以证明:

- (1) 对 $\forall \mathbf{x}, \mathbf{y} \in S^D$, 有 $d_A(\mathbf{x}, \mathbf{y}) = d_E(ilr(\mathbf{x}), ilr(\mathbf{y}))$, 其中 $d_E(\cdot, \cdot)$ 是指欧式距离.
- (2) 对 $\forall a, b \in \mathbf{R}$, $\mathbf{x}, \mathbf{y} \in S^D$, $a\mathbf{x} = (ax_1, ax_2, \dots, ax_D)$, $b\mathbf{y} = (by_1, by_2, \dots, by_D)$, 有 $d_A(\mathbf{x}, \mathbf{y}) = d_A(a\mathbf{x}, b\mathbf{y})$.

从定义可以看出, 若成分数据中包含缺失值, 将不能做对数比变换, 进而影响进一步的数据分析. 对于普通的数据, 常用的缺失值填补法可分为单一填补法和多重填补法, 其中单一填补法包括均值填补法, 回归填补法, EM算法等, 多重填补法包括回归预测方法(regression predict method), 倾向得分法(propensity score method), 蒙特卡洛的马氏链方法(MCMC). 刘鹏等(2004), 金勇进和邵军(2009)详细介绍了一般常用的缺失数据处理方法及其优缺点. 但是由于成分数据满足非负定和的性质, 所以成分数据的缺失数据为非负且相关, 这就导致了一般的填补方法不再适用. 对于成分数据的缺失值填补, Hron等(2010)提出一种针对成分数据的 k 近邻缺失值填补法和基于 k 近邻的迭代回归的缺失值填补法, 该方法首先用 k 近邻填补法得到缺失数据的一个初始填补值, 再用一种迭代回归得到最终的填补值; 石丽(2012)将多重插补法应用到了成分数据中, 该方法首先对成分数据做对数比变换得到正态分布, 然后用多重插补法对缺失数据进行填补, 最后用对数比变换的逆变换得到成分数据的缺失估计值; 孙志猛等(2011), Qin等(2007)提出了利用半参数方法对缺失数据进行估计. 受到范明等(2004), 王星(2009), 谢中华(2010)中提到的非参数密度估计方法和Hron等(2010), 何亮等(2009)中的 k 近邻预测方法的启发, 并且考虑到 k 近邻填补法在利用缺失数据的 k 个近邻估计缺失数据时没有区别它们各自不同的贡献, 本文提出一种基于Epanechnikov二次核的成分数据缺失值填补法. 实验证明, 该方法能够得到比KNN填补法更为准确的估计, 且方法上要比多重插补法简单.

本文的结构安排如下: 第2节分别介绍了三种成分数据的缺失值填补法: k 近邻填补法, 基于Epanechnikov二次核的缺失值填补法和修正的基于Epanechnikov二次核的成分数据缺失值填补法; 第3节和第4节分别通过模拟实验和实证分析对三种方法做了比较; 第5节对本文的工作做了一个总结.

§2. 基于Epanechnikov二次核的缺失值填补法

设 $\mathbf{X} = (x_{i,j})_{n \times D}$, $i = 1, 2, \dots, n$; $j = 1, 2, \dots, D$ 是成分数据的样本观测矩阵, \mathbf{X} 的每

一行对应于一个观测, 是一个成分数据, 所以每一行的行和都为1, 每一列对应于一个成分的观测. $\boldsymbol{\delta} = (\delta_{i,j})_{n \times D}$ 是与 \mathbf{X} 对应的指示矩阵, 其中 $\delta_{i,j} = 1$ 表示 $x_{i,j}$ 具有观测, $\delta_{i,j} = 0$ 表示 $x_{i,j}$ 是缺失的. 为讨论方便, 假设我们已经对观测矩阵的列进行了调整, 使得含缺失值的列在前 $s(1, 2, \dots, s)$ 列, 不含缺失值的列在后 $D - s$ (即 $s + 1, s + 2, \dots, D$) 列. 记 $\mathbf{X} = (\mathbf{X}_m, \mathbf{X}_o)$, 其中 \mathbf{X}_o 是 $n \times (D - s)$ 的矩阵且不含缺失值, 是由 \mathbf{X} 的所有不含缺失值的成分构成的矩阵, \mathbf{X}_m 是 $n \times s$ 的矩阵且每列至少有一个缺失值, 是由 \mathbf{X} 的所有含缺失值的成分构成的矩阵. 相应的记 $\mathbf{x}_i^o = (x_{i,s+1}, x_{i,s+2}, \dots, x_{i,D})$, $i = 1, 2, \dots, n$ 为 \mathbf{x}_i 在后 $D - s$ 个位置上的观测值. 由于数据含有缺失值无法精确的计算各个样本到缺失数据的距离, 在计算样本到缺失数据的距离时只用样本的后 $D - s$ 列数据, 因为所有的样本在后 $D - s$ 列都含有观测值, 涉及到距离的计算, 本文都采用 Aitchison 距离(见公式(1.3)). 当 $\delta_{i,j} = 0$, $j = 1, 2, \dots, s$ 时, 对应的 $x_{i,j}$ 没有观测, 本节之后的工作就是对此 $x_{i,j}$ 进行填补.

2.1 k 近邻填补法(KNNI)

k 近邻填补法的思想是确定缺失数据的 k 个最近邻, 用这 k 个最近邻在缺失数据相应变量的取值的均值作为缺失数据的估计. 本文借助 k 近邻填补法的思想对成分数据观测矩阵 \mathbf{X} 中的缺失数据进行填补, 具体步骤如下:

(1) 对 \mathbf{X}_o 做变换使之成为成分数据 \mathbf{X}_o^* : $\mathbf{X}_o \rightarrow \mathbf{X}_o^* = (x_{i,j}^*)_{n \times (D-s)}$, 其中

$$x_{i,j}^* = \frac{x_{i,j}}{\sum_{j=s+1}^D x_{i,j}}, \quad j = s+1, s+2, \dots, D, \quad i = 1, 2, \dots, n.$$

这样 \mathbf{X}_o^* 每行和都为1, 每个 $\mathbf{x}_i^* = (x_{i,s+1}^*, x_{i,s+2}^*, \dots, x_{i,D}^*)$, $i = 1, 2, \dots, n$ 是一个成分数据.

(2) 当 $\delta_{i,j} = 0$ ($j = 1, 2, \dots, s$) 时, 计算 \mathbf{X}_o^* 中第 m ($m = 1, 2, \dots, i-1, i+1, \dots, n$) 行与第 i 行的 Aitchison 距离 $d_A(\mathbf{x}_i^*, \mathbf{x}_m^*)$ (见公式(1.3)), 找到 \mathbf{x}_i^* 的 k 个最近邻, 记作 $\mathbf{x}_{i[1]}^*, \mathbf{x}_{i[2]}^*, \dots, \mathbf{x}_{i[k]}^*$. 将 \mathbf{x}_i^* 的 k 个最近邻记作 $N_k(\mathbf{x}_i^*)$, 即 $N_k(\mathbf{x}_i^*) = \{\mathbf{x}_{i[1]}^*, \mathbf{x}_{i[2]}^*, \dots, \mathbf{x}_{i[k]}^*\}$. 然后用 \mathbf{x}_i^* 的 k 个最近邻所在的行对应的第 j 列数据(指对应于原来数据阵 \mathbf{X} 中的第 j 列)的均值作为 $x_{i,j}$ 的填补值 $x'_{i,j}$, 即

$$x'_{i,j} = \text{Ave}(x_{l,j} | \mathbf{x}_l^* \in N_k(\mathbf{x}_i^*)) = \frac{1}{k} \sum_{l=1}^k x_{i[l],j}.$$

(3) 当所有的缺失数据都已经进行填补后, 对每个填补值 $x'_{i,j}$ 进行调整, 使每行都满足定和限制. 调整公式为

$$\hat{x}_{i,j} = \frac{x'_{i,j}}{\sum_{j \in \{j | \delta_{i,j}=0\}} x'_{i,j}} \left(1 - \sum_{j=1}^D \delta_{i,j} x_{i,j} \right). \quad (2.1)$$

2.2 基于Epanechnikov二次核的成分数据缺失值填补法(EKI)

k 近邻填补法能够很好的对缺失数据进行填补, 但它其实是将相同的权值赋给了缺失数据的 k 个近邻, 并没有考虑不同的样本对缺失数据的贡献, 利用核函数方法可以弥补这种不足. 核函数方法可以根据样本离缺失数据的距离将不同的权值赋给样本. 核函数方法已被成功的用于函数估计等问题, 这里我们利用核函数的思想对成分数据的缺失数据进行填补, 本文选择Epanechnikov二次核, 因为它是一种常用的局部核函数, 操作容易, 并且在函数估计方面得到了很好的应用. 下面将详细介绍Epanechnikov二次核估计.

2.2.1 基于Epanechnikov二次核的估计

给定数据 $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_n, y_n)$, 其中 $\mathbf{x}_i \in \mathbf{R}^n$ 是自变量, $y_i \in \mathbf{R}$ 是响应变量. 当给出新的数据 \mathbf{x}_0 , Epanechnikov二次核估计方法(见范明等, 2004)提供了一种如何根据给定数据预测 y_0 的方法, y_0 的预测值记作 $\hat{f}(\mathbf{x}_0)$.

$\hat{f}(\mathbf{x}_0)$ 的Nadaraya-Watson核加权平均估计是

$$\hat{f}(\mathbf{x}_0) = \frac{\sum_{i=1}^n K_\lambda(\mathbf{x}_0, \mathbf{x}_i) y_i}{\sum_{i=1}^n K_\lambda(\mathbf{x}_0, \mathbf{x}_i)} \doteq \sum_{i=1}^n \beta_i y_i, \quad (2.2)$$

其中

$$\begin{aligned} \beta_i &= \frac{K_\lambda(\mathbf{x}_0, \mathbf{x}_i)}{\sum_{i=1}^n K_\lambda(\mathbf{x}_0, \mathbf{x}_i)} \quad (i = 1, 2, \dots, n), \\ K_\lambda(\mathbf{x}_0, \mathbf{x}) &= D\left(\frac{\|\mathbf{x} - \mathbf{x}_0\|}{h_\lambda(\mathbf{x}_0)}\right), \end{aligned} \quad (2.3)$$

$$D(t) = \begin{cases} \frac{3}{4}(1 - t^2), & |t| \leq 1; \\ 0, & \text{其他.} \end{cases} \quad (2.4)$$

(2.3)和(2.4)式被称为Epanechnikov二次核. (2.3)式中的 λ 是光滑参数, $h_\lambda(\mathbf{x}_0)$ 是一个窗宽函数, 它确定 \mathbf{x}_0 的邻域宽度. 将 \mathbf{x}_i ($i = 1, 2, \dots, n$)到 \mathbf{x}_0 的距离从小到大排序, 将排在第 l 位的标记为 $\mathbf{x}_{[l]}$, 则对于 k 最近邻域, $\lambda = k$, $h_\lambda(\mathbf{x}_0) = \|\mathbf{x}_0 - \mathbf{x}_{[k]}\|$. 不同于 k 近邻填补法, 这样可以把随其到目标点 \mathbf{x}_0 的距离平滑衰减的权赋给邻域中的点, 而不是将相同的权赋给邻域中的所有的点. β_i 可以看做是 \mathbf{x}_i 在估计 $\hat{f}(\mathbf{x}_0)$ 时的权值.

2.2.2 基于Epanechnikov二次核的成分数据缺失值填补法

本节将利用基于Epanechnikov二次核估计的思想对成分数据缺失值进行填补. 对成分数据的标记同2.1节一致. 具体步骤如下:

- (1) 同成分数据的 k 近邻填补方法中的第(1)步;
 (2) 当 $\delta_{i,j} = 0$ ($j = 1, 2, \dots, s$)时, 对 $x_{i,j}$ 进行填补. 计算初始填补值 $x'_{i,j}$, 且

$$x'_{i,j} = \frac{\sum_{l=1}^n \delta_{l,j} K_\lambda(\mathbf{x}_i^*, \mathbf{x}_l^*) x_{l,j}}{\sum_{l=1}^n \delta_{l,j} K_\lambda(\mathbf{x}_i^*, \mathbf{x}_l^*)} \hat{=} \sum_{l=1}^n \beta_l x_{l,j}. \quad (2.5)$$

在利用(2.3)式时, $\|\mathbf{x}_i^*, \mathbf{x}_l^*\| = d_A(\mathbf{x}_i^*, \mathbf{x}_l^*)$. $\beta_l = \delta_{l,j} K_\lambda(\mathbf{x}_i^*, \mathbf{x}_l^*) / \sum_{l=1}^n \delta_{l,j} K_\lambda(\mathbf{x}_i^*, \mathbf{x}_l^*)$ 可以看做是 \mathbf{x}_l 的权值.

(3) 当对每个缺失数据都进行了初始填补后, 利用公式(2.1)对填补值 $x'_{i,j}$ 进行调整, 得到 $\hat{x}_{i,j}$, 使每行都满足“定和限制”.

2.3 修正的Epanechnikov核成分数据缺失值填补法(MEKI)

由于实验中发现基于Epanechnikov二次核的缺失值填补法没有得到理想的结果, 甚至填补结果有点令人失望. 这可能是由于成分数据中各成分数据间的距离都很小, k 个近邻间的距离近乎相等, 所以它们对应的核权值也相差不大, 起不到理想的作用. 为了增大各权值之间的差距, 所以尝试对Epanechnikov二次核取对数后再进行加权估计, 即在利用(2.5)式计算初始填补值时令

$$\tilde{K}_\lambda(\mathbf{x}_0, \mathbf{x}) = -\log \left\{ \frac{3}{4} \left[1 - \left(\frac{\|\mathbf{x} - \mathbf{x}_0\|}{h_\lambda(\mathbf{x}_0)} \right)^2 \right] \right\}, \quad (2.6)$$

同样, 令 $h_\lambda(\mathbf{x}_0) = \|\mathbf{x}_0 - \mathbf{x}_{[k]}\|$. 利用修正的Epanechnikov二次核对成分数据进行缺失值填补时得到了意想不到的结果.

值得注意的是, 与基于Epanechnikov二次核的缺失值填补法不同, 修正的Epanechnikov二次核的缺失值填补法中目标点 \mathbf{x}_0 的 k 个近邻的权值随其到 \mathbf{x}_0 的距离的靠近反而减小. 这种出乎意料的结果可能与成分数据的特征, 各成分非负且满足定和限制有关. 由于成分数据满足非负且各成分之和等于1, 这就使得各缺失数据填补值之间满足一定的关系, 并非独立的进行填补.

基于修正的Epanechnikov核的成分数据缺失值填补法(MEKI)分为以下两个步骤:

- (1) 同成分数据的 k 近邻填补方法中的第(1)步;
 (2) 当 $\delta_{i,j} = 0$ ($j = 1, 2, \dots, s$)时, 对 $x_{i,j}$ 进行填补, 计算初始填补值 $x'_{i,j}$

$$x'_{i,j} = \frac{\sum_{l=1}^n \delta_{l,j} \tilde{K}_\lambda(\mathbf{x}_i^*, \mathbf{x}_l^*) x_{l,j}}{\sum_{l=1}^n \delta_{l,j} \tilde{K}_\lambda(\mathbf{x}_i^*, \mathbf{x}_l^*)} \hat{=} \sum_{l=1}^n \tilde{\beta}_l x_{l,j}, \quad (2.7)$$

其中 $\tilde{\beta}_l = \delta_{l,j} \tilde{K}_\lambda(\mathbf{x}_i^*, \mathbf{x}_l^*) / \sum_{l=1}^n \delta_{l,j} \tilde{K}_\lambda(\mathbf{x}_i^*, \mathbf{x}_l^*)$ 可以看做是样本 \mathbf{x}_l ($l = 1, 2, \dots, n$) 在估计 $x_{i,j}$ 时的权值。可以看出，只有当 $\delta_{l,j} \neq 0$ 时， \mathbf{x}_l 的权值 $\beta_l \neq 0$ ，也就是用第 j 个分量含有观测的值的加权平均来对 $x_{i,j}$ 进行估计。

(3) 当所有的缺失数据都已经进行填补后，对每个填补值 $x'_{i,j}$ 进行调整，使每行都满足定和限制。即，

$$\hat{x}_{i,j} = \frac{x'_{i,j}}{\sum_{j \in \{j | \delta_{i,j} \neq 0\}} x'_{i,j}} \left(1 - \sum_{j=1}^D \delta_{i,j} x_{i,j} \right). \quad (2.8)$$

可以看出，修正的Epanechnikov二次核成分数据缺失值填补法同基于Epanechnikov核的成分数据缺失值填补法一样，都是加权 k 近邻的填补法，将目标点的 k 个最近邻加权后来进行估计。

在后面的模拟实验和实例分析中，本文将用成分数据中缺失数据的填补值与真实值之间平均的Aitchison距离来对三种方法进行比较。即误差 e 定义为缺失数据的填补值与真实值之间平均的Aitchison距离：

$$e = \frac{1}{n_m} \sum_{i \in I} d_A(\mathbf{x}_i, \hat{\mathbf{x}}_i),$$

其中， I 是由观测矩阵 \mathbf{X} 中含缺失数据的行的下标组成集合， n_m 是观测矩阵 \mathbf{X} 含缺失数据的行的个数，也就是 I 中元素的个数。

§3. 模拟实验

在模拟实验中，模拟5维的成分数据 \mathbf{x} 。令 $\mathbf{z} = ilr(\mathbf{x})$ 服从均值为 μ ，方差为 Σ 的正态分布。这里取

$$\mu = (0, 0, 0, 0), \quad \Sigma = \begin{bmatrix} 1 & 0.5 & 0.5 & 0.5 \\ 0.5 & 1 & 0.5 & 0.5 \\ 0.5 & 0.5 & 1 & 0.5 \\ 0.5 & 0.5 & 0.5 & 1 \end{bmatrix}.$$

实验过程为：(1) 在均值为 μ ，方差为 Σ 的正态分布中随机抽取100个样本，记为样本矩阵 \mathbf{Z} ；(2) 根据等距对数比变换的逆变换算出原始的成分数据 \mathbf{X} ；(3) 去除 \mathbf{X} 的 $x_{1,1}, x_{1,2}$ ，即令第一个样本的第一个元素和第二个元素为缺失数据。分别用三种缺失值填补法来对 $x_{1,1}, x_{1,2}$ 进行估计。(4) 计算估计的成分数据与原始成分数据之间的误差。误差用Aitchison距离度量。(5) 对上述过程循环100次，对三种方法分别计算平均误差 e 。

$$e = \frac{1}{100} \sum_{i=1}^{100} d_A(\mathbf{x}_1^i, \hat{\mathbf{x}}_1^i),$$

$\mathbf{x}_1^i, \hat{\mathbf{x}}_1^i$ 分别是第 i 次循环的第一个成分数据的真实值及其估计值。

完成上述5个过程算一次循环,总共循环10000次,结果共9210次修正的核填补法比k近邻填补法的误差小,而只有1842次基于Epanechnikov二次核的缺失值填补法比k近邻填补法的误差小.图1是10000次中随机抽取100次的实验结果误差图.

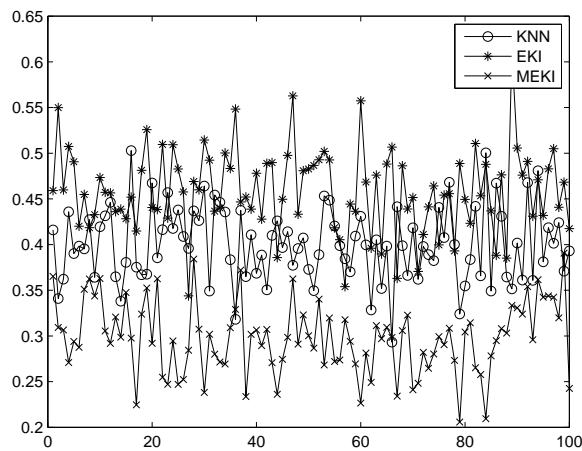


图1 随机抽取100次的实验结果误差图

从图1中也可以明显地看出,基于修正的Epanechnikov二次核的缺失值填补法比基于Epanechnikov二次核的缺失值填补法和k近邻填补法的结果都要准确,基于Epanechnikov二次核的缺失值填补法的填补效果误差最大.

再计算三种填补法10000次循环误差的平均, KNN, EKI, MEKI的平均误差分别记为:
 e_1, e_2, e_3 .

$$e_1 = 0.3881, \quad e_2 = 0.4537, \quad e_3 = 0.3013.$$

通过计算可以得出,修正的Epanechnikov核的缺失值填补法比k近邻填补法准确率提高了

$$\frac{|0.3013 - 0.3881|}{0.3881} \times 100\% = 22.37\%,$$

比基于Epanechnikov二次核的缺失值填补法的准确率提高了

$$\frac{|0.3013 - 0.4537|}{0.4537} \times 100\% = 33.59\%.$$

§4. 实例分析

Aitchison (2003)分别给出了hongite和kongite的25个样本,每个样本包含5个成分,分别标记为A, B, C, D, E,各个成分的和为100.对每个样本每行除以100,使每行和为1,每行是一个成分数据.本文首先对Aitchison (2003)中的表1.1.1a和表1.1.1b的成分数据的第一,二个成分进行填补,再把两张表合起来对其进行填补.采用留一法,即每次利用其余样本对其中一个样本进行填补,最后用误差的平均值作为最后的误差.表1是三种情况下误差的结果:

表1 三个表格在三种填补方法下的误差

方法	数据		
	表a	表b	表a, b
MEKI	0.0216	0.0352	0.0243
EKI	0.0570	0.0903	0.0576
KNNI	0.0518	0.0832	0.0530

对表a, MEKI比KNNI准确率提高了58.3%, 比EKI准确率提高了62.11%. 对表b, MEKI比KNNI准确率提高了57.6%, 比EKI准确率提高了61.02%. 将表a, b合并后, MEKI比KNNI准确率提高了54.11%, 比EKI准确率提高了57.81%. 可以看出, 在实例分析中得到了与模拟实验一致的结论.

§5. 总 结

本文针对缺失数据造成现有的成分数据统计方法失效和 k 近邻填补法在利用缺失数据的 k 个近邻估计缺失数据时没有考虑到它们各自不同的贡献, 提出了一种基于Epanechnikov二次核的成分数据缺失值填补法和修正的Epanechnikov核成分数据缺失值填补法. 其中, 基于Epanechnikov二次核的成分数据缺失值填补法把目标点的 k 个近邻随其到目标点 x_0 的距离平滑衰减的权赋给邻域中的点, 而不是将相同的权赋给领域中的所有的点, 基于修正的Epanechnikov二次核的成分数据缺失值填补法的 k 个近邻的权值随其到 x_0 的距离的靠近反而减小. 实验结果表明修正的Epanechnikov核的缺失值填补法比 k 近邻填补法准确率提高了22.37%, 比Epanechnikov二次核的缺失值填补法的准确率提高了33.59%. 修正后的Epanechnikov核成分数据缺失值填补法(MEKI)得到了较好的结果.

本文的创新之处在于将密度估计的核思想应用到成分数据的缺失值处理中, 它利用缺失数据的局部信息来对缺失值进行估计, 不受样本分布的影响. 此外, 利用了针对成分数据的Aitchison距离, 因此不用先将成分数据做对数比变换转化为正态分布后进行填补, 最后再经过逆变换得到成分数据的缺失数据估计值, 相对比较简单.

此外, 如何选取窗宽使得更合理的权值赋给目标点的近邻也是一个值得研究的问题. 在今后的工作中将会重视是否有更好的核函数对缺失数据进行估计.

参 考 文 献

- [1] Ferrers, N.M., *An Elementary Treatise on Trilinear Coordinates*, London: Macmillan, 1866.
- [2] Aitchison, J., *The Statistical Analysis of Compositional Data*, London: Chapman and Hall, 1986.
- [3] Egozcue, J.J., Pawlowsky-Glahn, V., Mateu-Figueras, G. and Barceló-Vidal, C., Isometric logratio transformations for compositional data analysis, *Mathematical Geology*, **35**(3)(2003), 279–300.

- [4] 刘鹏, 雷蕾, 张雪凤, 缺失数据处理方法的比较研究, *计算机科学*, **31(10)**(2004), 155–156.
- [5] 金勇进, 邵军, 缺失数据的统计处理, 北京: 中国统计出版社, 2009.
- [6] Hron, K., Templ, M. and Filzmoser, P., Imputation of missing values for compositional data using classical and robust methods, *Computational Statistics and Data Analysis*, **54(12)**(2010), 3095–3107.
- [7] 石丽, 多重插补在成分数据缺失值补全中的应用, 山西大学硕士学位论文, 2012.
- [8] 孙志猛, 张忠占, 杜江, 缺失数据下半参数单调回归模型的估计, *数理统计与管理*, **30(6)**(2011), 979–988.
- [9] Qin, Y.S., Zhang, S.C., Zhu, X.F., Zhang, J.L. and Zhang, C.Q., Semi-parametric optimization for missing data imputation, *Applied Intelligence*, **27(1)**(2007), 79–88.
- [10] 范明, 柴玉梅, 翁红英等译, 统计学习基础—数据挖掘, 推理与预测, 北京: 电子工业出版社, 2004.
- [11] 王星, 非参数统计, 北京: 清华大学出版社, 2009.
- [12] 谢中华, MATLAB统计分析与应用: 40个案例分析, 北京: 北京航空航天大学出版社, 2010.
- [13] 何亮, 宋擒豹, 沈钧毅, 海震, 一种新的组合 k -邻近预测方法, *西安交通大学学报*, **43(4)**(2009), 5–9.
- [14] Aitchison, J., A concise guide to compositional data analysis, in *Compositional Data Analysis Workshop*, Girona, 2003.
- [15] Stewart, C. and Field, C., Managing the essential zeros in quantitative fatty acid signature analysis, *Journal of Agricultural, Biological, and Environmental Statistics*, **16(1)**(2011), 45–69.

An Imputation Method for Missing Data in Compositional Based on Epanechnikov Kernel

ZHANG XIAOQIN KANG JU JING WENJUN

(School of Mathematics Science, Shanxi University, Taiyuan, 030006)

Kernel function method has been successfully used for the estimation of a variety of function. By using the kernel function theory, an imputation method based on Epanechnikov kernel and its modification were proposed to solve the problem that missing data in compositional caused the failures of existing statistical methods and the k -nearest imputation didn't consider the different contributions of the k nearest samples when it used them to estimated the missing data. The experimental results illustrate that the modified imputation method based on Epanechnikov kernel get a more accurate estimation than k -nearest imputation for compositional data.

Keywords: Compositional data, imputation for missing data, k -nearest imputation, Epanechnikov kernel, Aitchison distance.

AMS Subject Classification: 62G05, 62P05.