Modeling and Simulation 建模与仿真, 2023, 12(6), 5682-5692
Published Online November 2023 in Hans. https://www.hanspub.org/journal/mos
https://doi.org/10.12677/mos.2023.126516

Hans汉斯

基于不完全数据缺失值的非参数插补改进

汪子同

南京信息工程大学数学与统计学院，江苏 南京

收稿日期：2023年10月18日；录用日期：2023年11月20日；发布日期：2023年11月27日

摘 要

在数据分析研究中，数据的质量越高，数据集整体越完整，那么得到的研究结果往往越有价值。可是现实中常常面临含有大量不完全数据的数据集，如果直接删除不完全数据进行分析研究就会直接损失大量的样本信息。针对不完全数据的缺失值估计问题，基于非参数插补的思想，本文提出了两种回归函数估计量，给出了两种估计量的推导过程，在模拟研究中验证了在不同数据分布以及数据缺失率下，两个改进的非参数插补法对比其他经典的非参数插补法以及加权估计法在总体均值估计方面具有优势。

关键词

不完全数据，非参数插补法，加权估计，最小二乘

Nonparametric Imputation Improvements under Missing Values for Incomplete Data

Zitong Wang

School of Mathematics and Statistics, Nanjing University of Information Science & Technology, Nanjing Jiangsu

Received: Oct. 18th, 2023; accepted: Nov. 20th, 2023; published: Nov. 27th 2023

Abstract

In data analysis research, the higher the quality of the data and the more complete the overall data set, the more valuable the results are often obtained. However, in reality, we often face datasets containing a large amount of incomplete data, and if the incomplete data is directly deleted for analysis and research, a large amount of sample information will be directly lost. Aiming at the problem of missing value estimation of incomplete data, based on the idea of nonparametric im-

文章引用: 汪子同. 基于不完全数据缺失值的非参数插补改进[J]. 建模与仿真, 2023, 12(6): 5682-5692.
DOI: 10.12677/mos.2023.126516

putation, this paper proposes two regression function estimators, gives the derivation process of two estimators, and verifies in simulation studies that the two improved nonparametric imputation methods have advantages over other classical nonparametric imputation methods and weighted estimation methods in the estimation of the overall mean under different data distribution and data loss rate.

Keywords

Incomplete Data, Nonparametric Imputation, Weighted Estimation, Least Squares

1. 引言

在当前大数据时代背景下，搜集数据的技术水平相较以前大幅提高，各领域数据的获取也变得越来越容易。但是现实中真正完整的数据集并不常见，更多时候我们获取得到的数据集是不完全的，含有或多或少的缺失数据值。不完全数据的处理问题一直都是统计学领域的研究热点。

处理不完全数据的方法有很多，大致可分为删除法和插补法。删除法是将含有缺失值的数据删除掉，该方法操作简单、易于理解，但是仅仅适用于样本数量大并且数据缺失率较小时的场景，并且或多或少会损失掉样本的完整信息。插补法是将缺失值进行估计并插补到数据集中。插补法主要分为单一插补和多重插补。单一插补就是对每个缺失值进行一次估计；多重插补在单一插补的基础上，对缺失值进行多次估计，然后将多个估计值插补进数据集中形成多个“完整”数据集，最后利用评分函数确定最终的估计值。相较于单一插补，多重插补的插补方式是随机抽取的，所以估计效率更高，但多重插补的操作要求较高，需要更多的精力。

关于缺失值插补的研究最早可以追溯到 Yates [1]提出的一种缺失值的估计方法，该方法在方差分析中表现出很好的效果。Cheng 和 Wei [2]提出了一种叫做核加权回归的非参数插补方法，他们还证明了该插补在估计总体均值时的渐近性质。Cheng [3]提出了一种与核加权回归方法类似的基于最近邻回归加权的插补方法。Horvitz 和 Thompson [4]针对抽样调查的缺失数据，认为可以赋予完全观测值适当的倾向函数，提出了一种基于倾向函数的逆概率加权估计法，目的是重现完整的数据集。后期提出的新加权方法基本上是承袭早期的这些思想改进而来。Robins 等[5]将逆概率加权估计用于数据缺失条件下的半参数回归函数估计，发现该估计方法当参数回归模型或者倾向函数任意一种被正确指定时，估计结果都是渐近有效的，这种性质被称为双稳健性质。Ning 等[6]依据 HT 估计的原理，改进了核密度插补估计，构造了逆概率加权插补估计量。Ning 等[7]又结合了核密度估计和最近邻估计，提出了一种新的非参数的双稳健插补方法，并比较了各类非参数回归插补方法在正则条件下的渐近性质。祝恒坤[8]提出了一种基于逆概率加权插补和完全插补的 Mallows 模型平均方法用于非随机缺失情形，并证明了相关估计量在实现最小平方误差的意义下能渐近地达到最优。丁先文等[9]研究了响应变量随机缺失下，基于分位数回归半参数模型的稳健估计问题，提出了一种新的插补方法对缺失的响应变量进行多重插补。刘沙等[10]提出了一个基于统计度量的缺失值填补算法，利用数据点的类中心和标准差来填补缺失值。

本文在现有理论基础和前人相关工作的基础上，针对缺失值估计问题，将非参数插补法里的两种经典方法核密度插补法(KR)和最近邻插补法(KNN)得到的回归函数估计量进行加权组合，得到新的混合回归函数估计量，记为 WM (Weighted Mixture)，权重系数由基于完全数据的最小二乘法确定。再在 WM 估计量基础上加入一个基于完全数据的纠偏项，构造一个基于偏差逆概率加权(Deviance Inverse Probability weighting)的混合函数估计量，记为 DIPW。

2. 方法

2.1. 数据缺失机制

在进行不完全数据处理之前，了解数据发生缺失的原因是很有必要的。统计学者在早期研究不完全数据的时候，并未在意部分数据发生缺失的原因，直到 1976 年 Rubin [11]首次提出缺失机制的概念，用缺失机制代表数据发生缺失的原因，广大学者才逐渐围绕缺失机制展开研究。我们用三元组 $(X_i, Y_i, \delta_i), i=1,\cdots,n$ 表示不完全数据集，其中 X 表示完全观测的特征变量；Y 表示存在缺失值的响应变量；δ 为指示变量，$\delta_i=1$ 时 Y_i 的值可以观测到，$\delta_i=0$ 时 Y_i 的值缺失。

随机缺失(MAR)指的是响应变量 Y 是否发生缺失依赖于特征变量 X，而不依赖于 Y 本身。Rosenbaum and Rubin [12]将利用基线协变量进行治疗分配的概率定义为倾向函数。倾向函数与缺失机制在本质上描述的都是变量被某一固定值观测到的条件概率。那么可以用倾向函数表示 MAR：

$$P(\delta=1|X,Y)=P(\delta=1|X)=p(X), \tag{1}$$

其中，$P(\delta=1|X)$ 为响应变量 Y 被固定观测到的概率，$p(X)$ 为已知的倾向函数。

2.2. 经典非参数插补法

在 MAR 的假设下，我们重点研究不同的非参数插补法对响应变量 Y 的均值 $\mu(\mu=EY)$ 的估计效果。以一元协变量为例，以下插补方法共同假设回归函数 $m(x)=E(Y|X=x)$。

第一种非参数插补法为核密度插补法(KR)，它的原理是通过某个已知的核函数对数据点进行加权求和，选取的数据点需要满足跟含有缺失值的点的距离小于给定的带宽 h。利用 KR 得到的 μ 的估计量为：

$$\hat{\mu}_{KR}=\frac{1}{n}\sum_{i=1}^{n}\left[\delta_i Y_i+(1-\delta_i)\hat{m}_{KR}(X_i)\right], \tag{2}$$

其中，

$$\hat{m}_{KR}(X_i)=\frac{\sum_{j=1}^{n}K_h(X_j,X_i)\delta_j Y_j}{\sum_{j=1}^{n}K_h(X_j,X_i)\delta_j}, i=1,\cdots,n. \tag{3}$$

$K_h(u,v)=h^{-1}K((u-v)/h)$。由公式(4)可以看出只有与 X_i 距离小于带宽 h，并且对应响应变量未发生缺失的 X_j 才会进行局部加权步骤，这种局部加权通过核函数 K_h 实现。核函数 K_h 提前给定，那么该插补方法唯一的未知参数是带宽 h。$\hat{m}_{KR}(X_i)$ 是光滑的，h 控制的是 $\hat{m}_{KR}(X_i)$ 的复杂程度，并且 h 越小 $\hat{m}_{KR}(X_i)$ 越不光滑。综上可知 h 的取值会直接影响 KR 效果的好坏。

当样本量较少或者数据较稀疏时，如果选择 h 的值比较小，在对 $\hat{m}_{KR}(X_i)$ 插补的过程中可能会找不到观测值未缺失的 X_j 来进行局部加权，那么 $\hat{m}_{KR}(X_i)$ 的值就没办法估计出来。同时当面临高维数据，KR 会遇到“维数祸根”的问题。

最近邻插补法(KNN)沿用了 KR 的局部加权思想，它的原理是通过距离筛选出用于估计缺失值的数据点，优先选取距离最近的点。利用 KNN 得到 μ 的估计量为：

$$\hat{\mu}_{KNN}=\frac{1}{n}\sum_{i=1}^{n}\left[\delta_i Y_i+\left(1-\delta_i\right)\hat{m}_{KNN}\left(X_i\right)\right], \tag{4}$$

其中，

$$\hat{m}_{KNN}\left(X_i\right)=\frac{1}{k}\sum_{j=1}^{k}Y_{i(j)}, i=1,\cdots,n. \tag{5}$$

给定的近邻数 k 为 KNN 中的数据个数。通过距离远近筛选出合适的近邻，再对这些近邻观测值求均值来得到 $\hat{m}_{KNN}\left(X_i\right)$。$Y_{i(j)}$ 表示观测值缺失的响应变量 Y_i 的第 j 个近邻。$\hat{m}_{KNN}\left(X_i\right)$ 不连续，并且 k 越大，KNN 模型的复杂程度越高。显然 k 直接决定了 KNN 效果的好坏。

由于 KNN 中对于近邻的选取是依据距离，所以 KNN 可以很好地解决数据稀疏以及维度较高的问题。但也有可能会存在某个近邻 $Y_{i(j)}$ 存在缺失值的问题，那么 KNN 的估计效果就会弱于 KR。

最早的一种逆概率加权估计法 HT 估计是一种对总体均值的估计方法，在分层抽样中有着广泛应用。某种意义上来说，含有缺失值的不完全数据类似于分层抽样中的总体，其中未含有缺失值的数据为一层，含有缺失值的数据为另外一层。针对不完全数据，HT 估计原理为响应变量 Y_i 以一固定概率 ω_i 被观测到，对 ω_i 取倒数视为 Y_i 的权重，这样就可以通过未含有缺失值的数据来对总体均值进行估计。利用 HT 估计得到的 μ 的估计量为：

$$\hat{\mu}_{HT}=\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i Y_i}{\omega_i}, \tag{6}$$

其中，

$$\omega_i=\frac{\sum_{j=1}^{n}K_h\left(X_j,X_i\right)\delta_j}{\sum_{j=1}^{n}K_h\left(X_j,X_i\right)}, i=1,\cdots,n. \tag{7}$$

ω_i 为基于核平滑(KS)得到的倾向函数 $p\left(X_i\right)$ 的估计值。将(7)中的样本量 n 用有效样本量替代，得到估计效果更好的 HTR 估计：

$$\hat{\mu}_{HTR}=\frac{\sum_{i=1}^{n}\delta_i Y_i/\omega_i}{\sum_{i=1}^{n}\delta_i/\omega_i}. \tag{8}$$

Ning 基于 KR 和 HT 估计，构造了逆概率加权插补法(IPW)，它的原理是对核密度估计出的回归函数进行一个纠偏操作，具体方法是将回归函数加上一个基于完全数据的纠偏项。利用 IPW 得到的 μ 的估计量为：

$$\hat{\mu}_{IPW}=\frac{1}{n}\sum_{i=1}^{n}\hat{m}_{IPW}\left(X_i\right)=\frac{1}{n}\sum_{i=1}^{n}\left\{\hat{m}_{KR}\left(X_i\right)+\frac{\delta_i\left[Y_i-\hat{m}_{KR}\left(X_i\right)\right]}{\omega_i}\right\}, \tag{9}$$

其中，

$$\hat{m}_{IPW}\left(X_i\right)=\hat{m}_{KR}\left(X_i\right)+\frac{\delta_i\left[Y_i-\hat{m}_{KR}\left(X_i\right)\right]}{\omega_i}. \tag{10}$$

$\hat{m}_{KR}\left(X_i\right)$ 详见公式(4)，ω_i 详见公式(8)。同样的，将(10)中的 n 用有效样本量替代，可以得到一个新的关于总体均值 μ 的双稳健估计量：

$$\hat{\mu}_{DR}=\frac{1}{n}\sum_{i=1}^{n}\hat{m}_{KR}\left(X_i\right)+\frac{1}{\sum_{i=1}^{n}\delta_i/\omega_i}\sum_{i=1}^{n}\frac{\delta_i\left[Y_i-\hat{m}_{KR}\left(X_i\right)\right]}{\omega_i}. \tag{11}$$

2.3. 改进的非参数插补法

结合以上 KR 以及 KNN 的各自特点，对这两种经典的非参数插补方法加权组合有望构造出一个包含两者优点的新回归函数估计量 $\hat{m}_{WM}\left(X_i\right)$。在后续的模拟中证实了当数据全部来自同一分布的情况下，新估计量的估计效果优于任何一个单一的插补估计量。

新回归函数定义为：

$$\hat{m}_{WM}\left(X_i\right)=\alpha\hat{m}_{KR}\left(X_i\right)+\left(1-\alpha\right)\hat{m}_{KNN}\left(X_i\right),i=1,\cdots,n. \tag{12}$$

这样变量的缺失值可以估计为 $Y_i=\hat{m}_{WM}\left(X_i\right)$。

对于非参数插补新回归函数的未知参数 α 的求解可以借助最小二乘法思想，令误差项的平方最小，用所有完全数据进行估计，假设完全数据个数为 m：

$$\begin{aligned}\alpha&=\arg\min\sum_{i=1}^{m}\left[Y_i-\hat{m}_{\alpha}\left(X_i\right)\right]^2\\&=\arg\min\sum_{i=1}^{m}\left[Y_i-\alpha\hat{m}_{KR}\left(X_i\right)-\left(1-\alpha\right)\hat{m}_{KNN}\left(X_i\right)\right]^2\\&=\arg\min\sum_{i=1}^{m}\left[Y_i-\hat{m}_{KNN}\left(X_i\right)+\left(\hat{m}_{KNN}\left(X_i\right)-\hat{m}_{KR}\left(X_i\right)\right)\alpha\right]^2.\end{aligned} \tag{13}$$

进而可以得到估计方程为：

$$\sum_{i=1}^{m}\left[Y_i-\hat{m}_{KNN}\left(X_i\right)+\left(\hat{m}_{KNN}\left(X_i\right)-\hat{m}_{KR}\left(X_i\right)\right)\alpha\right]\left(\hat{m}_{KNN}\left(X_i\right)-\hat{m}_{KR}\left(X_i\right)\right)=0. \tag{14}$$

数据缺失率、样本量 n、带宽 h 以及近邻数 k 都可以直接影响未知参数 α 的求解。

参考公式(11)的逆概率加权思想，我们同样为新回归函数加入一个基于完全数据的纠偏项，得到另一个新的回归函数估计量。该估计量定义为：

$$\hat{m}_{DIPW}\left(X_i\right)=\hat{m}_{WM}\left(X_i\right)+\frac{\delta_i\left[Y_i-\hat{m}_{WM}\left(X_i\right)\right]}{\omega_i}. \tag{15}$$

$\hat{m}_{WM}\left(X_i\right)$ 详见公式(13)，ω_i 详见公式(8)。

3. 模拟研究

3.1. 模拟设定

为了比较本文提出的两种基于新回归函数的非参数插补与其他经典的非参数插补的插补效果，我们假定三个例子进行数值模拟验证。考虑模型 $Y=g\left(X\right)+\varepsilon$，其中偏差项 $\varepsilon\sim N\left(0,\sigma^2\left(x\right)\right)$。样本量 $n=50,200,500$，模拟次数 $N=1000$。

为了比较本文所提出方法的有限样本性质，选择核密度插补法(KR)、最近邻插补法(KNN)、逆概率加权插补法(IPW)、双稳健插补法(DR)、HT 估计和 HTR 估计进行对比。并通过 MAD、MSE、CCI 和 ZS 四个定性指标进行结果评估，定性指标见(16)~(19)所示。MAD 为插补值均值与观测值均值的绝对差值。MSE 为插补值均值与总体均值离差的平方，由于实际 MSE 数值非常小，在这里我们将其乘以 n 作为评价结果。CCI 为总体均值的收敛比例，若总体均值落入插补值均值置信度为 95%的置信区间内，则 CCI 为 1；否则为 0。ZS 为平均偏差与偏差标准误的比值。

$$\mathrm{MAD}=\frac{1}{N}\sum_{j=1}^{N}\left(\left|\hat{\mu}_j-\bar{\mu}_j\right|\right), \tag{16}$$

$$n\times \mathrm{MSE}=n\cdot\frac{1}{N}\sum_{j=1}^{N}\left(\hat{\mu}_j-\mu\right)^2, \tag{17}$$

$$\mathrm{CCI}=\frac{1}{N}\sum_{j=1}^{N}I\left(\mu\in\left[\hat{\mu}_j\pm t_{(0.975,N)}\right]\right), \tag{18}$$

$$\mathrm{ZS}=\frac{\overline{Bias}}{\sqrt{\sum_{j=1}^{N}\left(Bias_j-\overline{Bias}\right)^2/N}}, \tag{19}$$

其中，$\hat{\mu}_j$ 为第 j 次模拟的插补值均值，$\bar{\mu}_j$ 为第 j 次模拟的观测值均值，μ 为总体的理论均值，$\overline{Bias}=\sum_{j=1}^{N}\left(\hat{\mu}_j-\bar{\mu}_j\right)/N$。

模拟过程注意以下几点要求：

1) 核函数选用 Epanechnikov 多项式核函数

$$K(u)=\begin{cases}\frac{3}{4}\left(1-u^2\right), & \text{for } |u|\le 1,\\ 0, & \text{otherwise.}\end{cases}$$

2) 当 h 取值较小时，缺失值附近可能没有观测值未缺失的点，那么该缺失值不能用 KR 进行插补估计，于是我们直接舍弃该点，这也就意味着 KR 的有效样本量可能小于 n；相应地，该缺失值的 WM 插补值直接用 KNN 插补值替换。

3) 由于 HT 估计和 HTR 估计未对缺失值进行插补，于是这两种方法下的 CCI 是用观测值均值代替插补值均值进行计算的，所以这两个方法下的 CCI 会普遍高于其他方法，数值非常接近 1。

三种模拟假设如下：

$$\text{Case 1}=\begin{cases}g(x)=4x,\\ p(x)=\dfrac{\mathrm{e}^{2.5x}}{1+\mathrm{e}^{2.5x}},\\ X\sim U(0,1),\\ \sigma^2(x)=1,\\ \mu=2,\\ P(\delta=0)=0.245,\end{cases}\quad \text{Case 2}=\begin{cases}g(x)=4-9(x-0.4)^2\\ p(x)=\begin{cases}0.6, & 0\le x<0.3\\ 0.3, & 0.3\le x<0.7\\ 0.6, & 0.7\le x\le 1\end{cases}\\ X\sim U(0,1),\\ \sigma^2(x)=1,\\ \mu=3.16,\\ P(\delta=0)=0.52,\end{cases}\quad \text{Case 3}=\begin{cases}g(x)=2x+1\\ p(x)=\dfrac{\mathrm{e}^{2.5x}}{1+\mathrm{e}^{2.5x}},\\ X\sim 0.4U(-3,0)+0.6U(0,4),\\ \sigma^2(x)=\begin{cases}0.6 & x<0\\ 1 & x\ge 0\end{cases}\\ \mu=2.2,\\ P(\delta=0)=0.291\ .\end{cases}$$

模拟 1 假设 $g(x)$ 为线性的简单函数，倾向函数光滑，数据缺失率较小。模拟 2 将 $g(x)$ 替换为多项式函数，该函数有一条垂直于 x 轴的对称轴，且对称轴位于 X 服从的均匀分布范围内，这样数据分布整体远远比模拟 1 的数据复杂得多，且距离对于数据关系的解释能力会变弱；同时倾向函数为分段函数，数据缺失率较大，分布较稀疏。模拟 3 在模拟 1 的基础上假设 X 服从混合分布，并且将误差项的方差替换为分段函数，这样数据整体更为复杂，数据缺失率也比模拟 1 略大。

3.2. 数值模拟

数值模拟结果如表 1~3 所示。

表 1 的模拟结果显示：当 n 较小时，HT 估计的 MAD 略大于其他插补法，ZS 远高于其他插补法，该方法效果最差；当 n 较大时，选取恰当的 k 或者 h，六种经典方法的插补效果接近；当 n 固定时，在相同的 h 下，HTR 估计相对于 HT 估计，MAD 略微减小，ZS 显著减小，这种加权估计法对总体均值的估计效果与其他经典非参数插补法相当；在相同的 h 下，IPW 和 DR 相对于 KR，MAD 整体略微减小，CCI 显著增大，

插补效果更好；当 n 较小时，如果 k 取值较大(k = 16, 32) KNN 的插补效果堪称“灾难”，如果 h 取值较小(h = 0.1)HT 估计的插补效果也远远差于其他取值下的插补效果。而不论 n 的大小，相对于单一插补的 KNN 以及 KR，WM 的 MAD 都要更小，尤其是当 n 较小时 MAD 显著小于所有单一插补方法；DIPW 在 WM 的基础上加入了一个纠偏项，由于数据缺失率较小，数据分布较简单，WM 的 CCI 在 0.95 附近，DIPW 的 CCI 在 0.96 附近，所以整体来说插补均值置信区间收敛概率比较稳定，DIPW 的插补效果略好于 WM。

Table 1. Comparison of the results of all imputation methods in Simulation 1 at different sample sizes
表 1. 模拟 1 中所有插补方法在不同样本量下的结果对比

		n = 50				n = 200				n = 500			
Method	k/h	MAD	n × MSE	CCI	ZS	MAD	n × MSE	CCI	ZS	MAD	n × MSE	CCI	ZS
KNN	4	0.071	2.687	0.953	−0.002	0.035	2.663	0.955	−0.007	0.023	2.719	0.950	−0.005
	8	0.067	2.682	0.952	−0.058	0.032	2.675	0.957	0.024	0.022	2.786	0.944	−0.032
	16	0.073	3.216	0.922	0.012	0.033	2.552	0.953	0.027	0.021	2.600	0.948	0.025
	32	0.088	3.197	0.905	−0.001	0.032	2.501	0.954	0.033	0.021	2.682	0.948	−0.005
KR	0.1	0.068	2.699	0.943	0.013	0.032	2.653	0.955	0.024	0.020	2.566	0.959	0.022
	0.15	0.070	2.729	0.948	−0.062	0.033	2.859	0.948	0.034	0.021	2.905	0.927	−0.041
	0.2	0.068	2.438	0.958	−0.024	0.033	2.736	0.937	0.015	0.021	2.678	0.946	0.029
	0.25	0.067	2.656	0.947	0.013	0.033	2.618	0.948	−0.009	0.021	2.793	0.947	−0.014
HT	0.1	0.078	2.784	0.996	−0.327	0.033	2.783	0.992	−0.136	0.021	2.477	0.998	−0.119
	0.15	0.073	2.529	0.996	−0.203	0.034	2.723	0.994	−0.059	0.022	2.807	0.996	−0.092
	0.2	0.070	2.678	0.994	−0.121	0.034	2.627	0.992	−0.070	0.022	2.670	0.994	−0.018
	0.25	0.070	2.620	0.994	−0.150	0.032	2.689	0.998	−0.048	0.021	2.783	0.994	−0.042
HTR	0.1	0.071	2.679	0.997	−0.054	0.032	2.739	0.993	0.020	0.021	2.476	0.998	−0.043
	0.15	0.070	2.459	0.996	−0.094	0.033	2.713	0.993	0.021	0.021	2.816	0.996	−0.045
	0.2	0.068	2.661	0.994	0.004	0.033	2.623	0.993	−0.012	0.021	2.667	0.995	0.021
	0.25	0.067	2.600	0.996	−0.057	0.031	2.691	0.998	0.008	0.021	2.778	0.994	−0.011
IPW	0.1	0.068	2.706	0.972	−0.069	0.033	2.702	0.971	0.071	0.021	2.519	0.982	−0.266
	0.15	0.067	2.873	0.966	−0.034	0.032	2.839	0.973	0.026	0.020	2.673	0.983	−0.017
	0.2	0.069	2.536	0.977	−0.028	0.032	2.855	0.967	0.008	0.020	2.619	0.973	0.014
	0.25	0.068	2.766	0.975	−0.020	0.033	2.518	0.983	0.013	0.020	2.550	0.976	0.011
DR	0.1	0.068	2.707	0.972	−0.069	0.033	2.702	0.971	0.070	0.021	2.519	0.982	−0.266
	0.15	0.067	2.874	0.966	−0.034	0.032	2.839	0.973	0.026	0.020	2.673	0.983	−0.017
	0.2	0.069	2.535	0.977	−0.028	0.031	2.854	0.967	0.008	0.020	2.619	0.973	0.014
	0.25	0.068	2.766	0.975	−0.020	0.033	2.518	0.983	0.012	0.020	2.550	0.976	0.011
WM	4/0.15	0.066	2.664	0.946	−0.006	0.032	2.647	0.947	0.014	0.019	2.616	0.951	0.015
	4/0.2	0.066	2.497	0.952	−0.036	0.033	2.505	0.957	−0.048	0.021	2.712	0.951	−0.049
	8/0.15	0.067	2.658	0.961	0.038	0.031	2.798	0.949	−0.013	0.021	2.669	0.952	−0.043
	8/0.2	0.066	2.657	0.951	−0.067	0.032	2.803	0.945	−0.019	0.021	2.700	0.953	−0.035
DIPW	4/0.15	0.069	2.516	0.952	0.005	0.033	2.578	0.950	0.004	0.020	2.705	0.961	0.255
	4/0.2	0.068	2.583	0.956	0.057	0.034	2.882	0.959	0.093	0.021	2.966	0.962	−0.040
	8/0.15	0.067	2.986	0.974	−0.001	0.032	2.769	0.963	0.048	0.019	2.619	0.957	0.011
	8/0.2	0.068	2.687	0.963	−0.043	0.033	2.820	0.962	0.036	0.021	2.966	0.962	−0.040

Table 2. Comparison of the results of all imputation methods in Simulation 2 at different sample sizes
表 2. 模拟 2 中所有插补方法在不同样本量下的结果对比

		n = 50				n = 200				n = 500			
Method	k/h	MAD	$n \times$ MSE	CCI	ZS	MAD	$n \times$ MSE	CCI	ZS	MAD	$n \times$ MSE	CCI	ZS
KNN	4	0.139	3.358	0.81	0.235	0.063	3.004	0.854	−0.003	0.038	3.056	0.867	−0.019
	8	0.149	4.037	0.734	0.341	0.063	3.053	0.847	0.091	0.039	2.869	0.857	−0.005
	16	0.172	4.186	0.695	0.452	0.064	3.264	0.813	0.262	0.039	2.918	0.844	0.126
	32	0.442	13.365	0.435	−1.402	0.079	3.809	0.736	0.729	0.040	3.171	0.809	0.303
KR	0.1	0.137	3.231	0.817	−0.077	0.060	3.109	0.829	0.101	0.039	2.990	0.823	0.116
	0.15	0.133	3.501	0.814	0.125	0.060	2.873	0.828	0.148	0.039	3.238	0.811	0.168
	0.2	0.127	3.463	0.803	0.119	0.060	3.074	0.823	0.189	0.039	3.076	0.826	0.259
	0.25	0.130	3.237	0.792	0.129	0.064	2.943	0.819	0.227	0.041	3.008	0.803	0.392
HT	0.1	0.202	4.980	0.977	−0.716	0.066	2.807	1	−0.376	0.039	3.049	1	−0.192
	0.15	0.162	3.882	1	−0.484	0.062	2.933	1	−0.202	0.039	3.230	1	−0.144
	0.2	0.140	3.049	1	−0.396	0.061	3.013	1	−0.166	0.038	2.904	1	−0.157
	0.25	0.131	3.046	0.999	−0.352	0.061	2.834	1	−0.161	0.037	2.754	1	−0.116
HTR	0.1	0.134	3.405	1	0.031	0.060	2.731	1	−0.028	0.039	2.994	1	0.041
	0.15	0.134	3.269	1	0.008	0.060	2.823	1	0.046	0.039	3.204	1	0.003
	0.2	0.127	2.800	1	−0.018	0.060	3.039	1	−0.015	0.038	2.944	1	−0.038
	0.25	0.122	2.960	0.999	−0.017	0.061	2.841	1	−0.012	0.038	2.848	1	−0.011
IPW	0.1	0.137	3.152	0.943	0.028	0.062	2.850	0.968	0.040	0.038	2.985	0.977	−0.016
	0.15	0.131	2.998	0.948	0.050	0.061	2.648	0.965	0.049	0.038	2.937	0.967	0.045
	0.2	0.129	3.142	0.956	0.036	0.061	2.946	0.967	−0.020	0.036	2.695	0.981	−0.033
	0.25	0.130	3.178	0.966	0.080	0.062	2.685	0.977	0.049	0.040	3.026	0.971	0.024
DR	0.1	0.138	3.162	0.943	0.027	0.062	2.850	0.968	0.039	0.038	2.985	0.977	−0.016
	0.15	0.131	3.005	0.948	0.049	0.061	2.648	0.965	0.049	0.038	2.937	0.967	0.045
	0.2	0.129	3.139	0.957	0.033	0.061	2.946	0.967	−0.021	0.036	2.695	0.981	−0.033
	0.25	0.130	3.178	0.967	0.077	0.061	2.683	0.977	0.047	0.040	3.026	0.972	0.024
WM	4/0.15	0.130	3.183	0.820	0.079	0.060	2.976	0.822	0.138	0.037	2.967	0.827	0.166
	4/0.2	0.129	3.038	0.816	0.071	0.061	2.909	0.832	0.132	0.038	2.875	0.823	0.271
	8/0.15	0.131	3.108	0.824	0.197	0.060	3.061	0.823	0.066	0.039	2.954	0.821	0.277
	8/0.2	0.129	3.045	0.824	0.158	0.059	2.930	0.826	0.174	0.040	3.210	0.803	0.206
DIPW	4/0.15	0.137	3.334	0.830	0.013	0.060	2.983	0.832	0.017	0.037	2.918	0.831	−0.064
	4/0.2	0.135	3.371	0.814	0.043	0.061	2.910	0.846	0.065	0.037	2.789	0.843	−0.008
	8/0.15	0.135	3.396	0.789	0.048	0.059	3.013	0.831	0.036	0.039	2.899	0.836	0.044
	8/0.2	0.136	3.205	0.813	0.028	0.058	2.892	0.832	0.022	0.039	3.113	0.826	−0.007

表 2 的模拟结果显示：当 n 较小时，选取不同的 k 或者 h 对六种经典方法的插补效果都有显著影响，这是因为数据缺失率较大，数据分布较复杂，距离并不能很好地体现数据间的关系，更多更远的其他观测值与缺失值的实际值误差较大，而由于 KNN 是根据绝对距离选择观测值，KR 是根据距离给予观测值适当的权重，所以综合来看 KNN 插补效果比 KR 差，并且 KNN 的 MAD 是随着 k 的增大而增大的，KR 的 MAD 在特定情况($h \le 0.2$)下随着 h 的增大而减小；当 n 较大时，选取恰当的 k 或者 h，六种经典方法的插补效果接近；不论 n 的大小，HT 估计的 MAD 随着 h 的增大而减小，这是由于 h 的变大使得 ω_i 的估

计更加准确，所以 HT 估计效果变好；当 n 固定时，由于数据缺失率较大，HT 估计效果相比 HT 估计有非常显著的提升；在相同的 h 下，相对于 KR，IPW 和 DR 的 CCI 更大，插补效果更好。不论 n 的大小，相对于单一插补的 KNN 以及 KR，WM 的 MAD 都要更小，并且当 n 较小时 $n \times$ MSE 更小，CCI 更大；由于数据缺失率较大，数据分布较复杂，如果插补值与实际值有较大偏差，那么纠偏项就会使插补效果降低；当 n 较小时，相对于 WM，DIPW 的 MAD 和 $n \times$ MSE 都要更大，ZS 更小；当 n 较大时，相对于 WM，DIPW 的 MAD 更小，CCI 更大，并且 ZS 也会更小。

Table 3. Comparison of the results of all imputation methods in Simulation 3 at different sample sizes
表 3. 模拟 3 中所有插补方法在不同样本量下的结果对比

		n = 50				n = 200				n = 500			
Method	k/h	MAD	$n \times$ MSE	CCI	ZS	MAD	$n \times$ MSE	CCI	ZS	MAD	$n \times$ MSE	CCI	ZS
KNN	4	0.071	3.769	0.952	0.038	0.035	3.724	0.954	0.058	0.023	3.795	0.959	−0.001
	8	0.073	3.825	0.946	−0.014	0.034	3.769	0.953	−0.006	0.022	3.397	0.960	0.027
	16	0.086	4.217	0.916	0.013	0.032	3.513	0.959	0.042	0.022	3.485	0.959	−0.051
	32	0.124	4.946	0.868	−0.052	0.036	3.764	0.950	0.017	0.021	3.618	0.961	−0.035
KR	0.1	0.072	3.819	0.950	−0.017	0.032	3.651	0.953	0.032	0.021	3.703	0.953	0.010
	0.15	0.070	3.749	0.954	−0.009	0.033	3.747	0.953	−0.007	0.021	3.366	0.959	0.026
	0.2	0.071	3.784	0.944	0.030	0.032	3.492	0.958	0.024	0.022	3.539	0.951	−0.054
	0.25	0.074	3.767	0.932	−0.029	0.036	3.716	0.947	0.007	0.022	3.676	0.952	−0.047
HT	0.1	0.087	4.122	0.993	−0.388	0.035	3.765	0.999	−0.174	0.021	3.704	0.998	−0.103
	0.15	0.082	3.979	0.993	−0.222	0.035	3.823	0.999	−0.106	0.022	3.420	0.993	−0.050
	0.2	0.078	3.964	0.986	−0.105	0.035	3.542	0.997	−0.044	0.023	3.541	0.999	−0.096
	0.25	0.084	3.949	0.993	−0.141	0.038	3.702	0.997	−0.041	0.024	3.688	0.999	−0.052
HTR	0.1	0.078	3.893	0.997	0.042	0.034	3.693	0.999	0.028	0.021	3.700	0.997	0.024
	0.15	0.077	3.861	0.993	0.006	0.034	3.773	0.999	−0.002	0.022	3.404	0.993	0.015
	0.2	0.076	3.894	0.989	0.032	0.034	3.536	0.997	0.037	0.023	3.524	0.999	−0.048
	0.25	0.081	3.855	0.995	−0.035	0.037	3.664	0.997	0.005	0.023	3.672	0.999	−0.036
IPW	0.1	0.072	3.780	0.964	0.057	0.033	3.665	0.97	0.037	0.021	3.709	0.978	0.011
	0.15	0.071	3.727	0.972	0.008	0.033	3.730	0.972	−0.009	0.021	3.349	0.979	0.028
	0.2	0.071	3.781	0.962	0.028	0.032	3.457	0.984	0.030	0.021	3.471	0.981	−0.050
	0.25	0.072	3.681	0.965	−0.037	0.033	3.623	0.976	0.009	0.021	3.595	0.974	−0.037
DR	0.1	0.072	3.781	0.964	0.058	0.033	3.665	0.97	0.037	0.021	3.709	0.978	0.011
	0.15	0.071	3.727	0.972	0.008	0.033	3.730	0.972	−0.009	0.021	3.349	0.979	0.028
	0.2	0.071	3.781	0.962	0.027	0.032	3.457	0.984	0.030	0.021	3.471	0.981	−0.050
	0.25	0.071	3.680	0.965	−0.036	0.033	3.623	0.976	0.008	0.021	3.595	0.974	−0.037
WM	4/0.15	0.071	3.746	0.957	0.019	0.032	3.391	0.965	−0.023	0.021	3.817	0.951	−0.035
	4/0.2	0.067	3.735	0.953	−0.004	0.034	3.656	0.951	−0.041	0.022	3.611	0.948	−0.024
	8/0.15	0.068	3.769	0.943	0.008	0.032	3.450	0.956	−0.063	0.021	3.714	0.950	0.041
	8/0.2	0.071	3.737	0.948	−0.064	0.034	3.829	0.945	0.018	0.021	3.349	0.949	0.028
DIPW	4/0.15	0.072	3.760	0.951	0.027	0.036	3.392	0.968	−0.030	0.022	3.845	0.954	−0.044
	4/0.2	0.072	3.866	0.947	−0.011	0.038	3.712	0.952	−0.040	0.023	3.626	0.953	−0.020
	8/0.15	0.069	3.785	0.941	0.006	0.033	3.485	0.955	−0.072	0.022	3.759	0.955	0.042
	8/0.2	0.073	3.749	0.942	−0.066	0.038	3.812	0.948	0.019	0.023	3.380	0.956	0.033

表 3 的模拟结果显示：当 n 较小时，k 的取值对 KNN 的插补效果影响较大，选取较大的 k 会使得 KNN 的 MAD 显著增大；当 n 较大时，选取恰当的 k 或者 h，六种经典方法的插补效果接近；不论 n 的大小，HT 估计效果都是最差的；当 n 固定时，HT 估计效果相比 HT 估计有提升，但其 MAD 依旧比其他非参数插补法大，这是由于数据分布较分散，数据缺失率较小，所以加权估计法的对于总体均值的估计效果不如非参数插补法；在相同的 h 下，相对于 KR，IPW 和 DR 的 CCI 显著变大，插补效果更好。不论 n 的大小，相对于单一插补的 KNN 以及 KR，WM 的插补效果都要更好一点；由于数据分布较分散，插补值与缺失值的真实值可能存在较大的偏差，当 n 较小时，DIPW 整体插补效果都要略差于 WM；当 n 较大时，DIPW 的 CCI 在 0.95 左右，WM 的 CCI 在 0.945 左右，表明 DIPW 的插补均值置信区间收敛概率更大，但是在其余指标上 DIPW 都要略差于 WM。

对所有方法的 MAD 数值进行可视化，如图 1 所示。

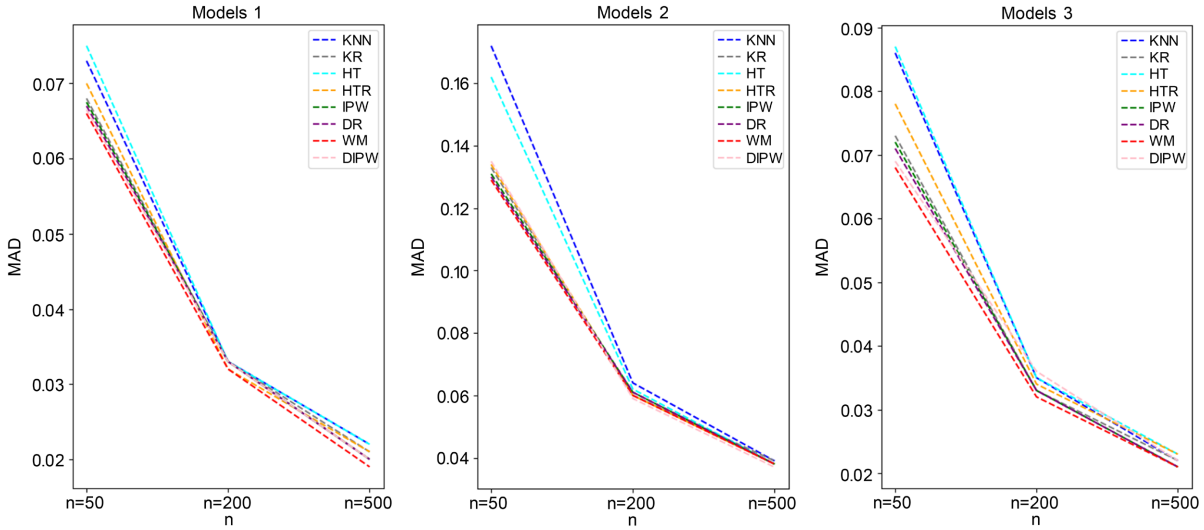

Figure 1. Comparison of MAD for all methods under different sample sizes under three simulation settings
图 1. 三种模拟设置下所有方法在不同的样本量下 MAD 的对比

4. 结论

当不完全数据的响应变量随机缺失时，我们考虑用非参数插补法对缺失值进行估计插补。两种经典非参数插补法核密度插补法(KR)和最近邻插补法(KNN)都有各自的优点以及局限性，针对不同数据分布以及数据缺失率情况，我们想到将两种经典方法进行加权组合，这样构造的新的回归函数估计量就具有更好的插补效果。针对不同假设下的数据，我们对缺失值进行多种插补估计，得出以下结论：

1) 不论数据分布情况以及数据缺失率大小，在相同的 n 下 WM 的 MAD 都要比单一插补的 KNN 以及 KR 的更小，整体插补效果更好；

2) 当数据分布情况复杂或者样本缺失率较大时，DIPW 的纠偏项可能具有较大偏差，此时的 DIPW 的 MAD 虽然要大于 WM 的，但是整体的 CCI 更大一点，也就是插补均值置信区间收敛概率更高；

3) 如果 n 较小，那么选取不恰当的 k 或者 h 会导致 KNN 或 KR 的插补效果很差；但是 k 或者 h 的取值对 WM 以及 DIPW 的插补效果并没有很显著的影响，所以 WM 以及 DIPW 的插补稳定性更好。

参考文献

[1] Yates, F. (1933) The Analysis of Replicated Experiments When the Field Results Are Incomplete. *Empire Journal of Experimental Agriculture*, **1**, 129-142.

[2] Cheng, P.E. and Wei, L.J. (1986) Nonparametric Inference under Ignorable Missing Data Process and Treatment Assignment. *International Statistical Symposium*, **1**, 97-112.

[3] Cheng, P.E. (1984) Strong Consistency of Nearest Neighbor Regression Function Estimators. *Journal of Multivariate Analysis*, **15**, 63-72. https://doi.org/10.1016/0047-259X(84)90067-8

[4] Horvitz, D.G. and Thompson, D.J. (1952) A Generalization of Sampling without Replacement from a Finite Population. *Journal of the American Statistical Association*, **47**, 663-685. https://doi.org/10.1080/01621459.1952.10483446

[5] Robins, J.M., Rotnitzky, A.G. and Lue, P.Z. (1994) Estimation of Regression Coefficients When Some Regressors Are Not Always Observed. *Journal of the American Statistical Association*, **89**, 846-886. https://doi.org/10.1080/01621459.1994.10476818

[6] Ning, J.H. and Cheng, P.E. (2012) A Comparison Study of Nonparametric Imputation Methods. *Statistics and Computing*, **22**, 273-285. https://doi.org/10.1007/s11222-010-9223-y

[7] Ning, J., Liou, M. and Cheng, P.E. (2019) Convex Mixtures Imputation and Applications. *StatisticaSinica*, **29**, 329-351. https://doi.org/10.5705/ss.202015.0204

[8] 祝恒坤, 张海丽. 基于逆概率加权和插补的 Mallows 模型平均方法[J]. 系统科学与数学, 2022, 42(4): 1032-1059.

[9] 丁先文, 张文, 袁红. 含缺失数据的半参数模型的稳健估计[J]. 统计与决策, 2022, 38(1): 25-28.

[10] 刘莎, 杨有龙. 基于灰色关联分析的类中心缺失值填补方法[J]. 四川大学学报(自然科学版), 2020, 57(5): 871-878.

[11] Rubin, D.B. (1976) Inference and Missing Data. *Biometrika*, **63**, 581-592. https://doi.org/10.1093/biomet/63.3.581

[12] Rosenbaum, P.R. and Rubin, D.B. (1983) The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika*, **70**, 41-55. https://doi.org/10.1093/biomet/70.1.41