分 类 号 ： C8
论文编号：2019071400002

单位代码： 10672
密　　级： 公开

2022 届硕士研究生学位论文

不同缺失机制下数据填补算法的比较研究

学位申请人姓名 郑智泉

培　养　学　院 数据科学与信息工程学院

一导姓名及职称 黄介武　教授

二导姓名及职称

学科代码及名称 071400　统计学

研　究　方　向 统计模型与统计计算

论 文 提 交 日 期 2022 年 06 月 13 日

中国·贵州·贵阳

贵州民族大学学位论文原创性声明

本人郑重声明：所呈交的学位论文是本人在导师的指导下独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不含任何其他个人或集体已经发表或撰写过的作品成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。因本学位论文引起的法律后果完全由本人承担。

论文题目：不同缺失机制下数据填补算法的比较研究

学位论文作者签名：

签字日期：2022年6月13日

贵州民族大学学位论文版权使用授权书

本学位论文作者完全了解贵州民族大学有关保留、使用学位论文的规定，有权保留并向国家有关部门或机构送交论文的纸质版和电子版，允许论文被查阅和借阅。本人授权贵州民族大学可以将学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文，可以公开学位论文的全部或部分内容。

学位论文作者签名：

签字日期：2022年6月13日

导师签名：

签字日期：2022年6月13日

摘　要

数据缺失在统计调查与研究中普遍存在，数据缺失往往会导致统计推断结果不可靠。对缺失数据处理的常用方法有删除含缺失值的样本点或观测变量、不处理、对缺失值进行填补等。本文主要研究不同缺失机制、不同缺失率下，常见填补算法的适用场景和参数优化问题，并通过实例对改进算法进行比较研究，以期为实际应用提供参考。

首先通过数学归纳法给出了完全随机缺失、随机缺失、非随机缺失三种缺失机制的数学描述，并给出了三类缺失在特定情形下的模拟代码。然后基于交叉验证法和高斯函数加权法对K近邻算法进行优化，提出了加权K近邻填补算法。同时，针对加权K近邻填补算法中为不同缺失值样本点计算得到的近邻距离差异而导致权重分配不合理问题，通过数据集中的具体观测值对高斯函数的参数进行动态调节，提出了基于动态调参的加权K近邻填补算法，理论研究和实证分析显示，该方法在提升填补效果的同时具备一定的通用性。最后，针对加权K近邻填补算法中过分依赖最邻近样本点而导致的算法稳定性下降问题，本文创造性采用缺失森林算法对其填补结果进行校准，针对缺失率增大而导致的填补算法效果逐步下降的问题，本文使用迭代法将填补过程中非完整数据集的缺失率进行逐步降低，提出加权K近邻与缺失森林混合迭代填补算法，实证分析显示，基于不同缺失机制和缺失率前提，该方法在填补准确性方面继承了加权K近邻算法，而在稳定性方面继承了缺失森林算法。

关键词：缺失机制　K近邻　迭代法　缺失森林　动态调参

Comparative Research on Data Filling Algorithms Under Different Missing Mechanisms

Zheng Zhiquan (Statistics)

Directed by Professor Huang Jiewu

Abstract： Missing Data is a common phenomenon in statistical investigation and research, which often leads to unreliable statistical inference results. Common methods of missing data processing include deleting sample points or observation variables containing missing values, not processing, and filling missing values, etc. This paper mainly studies the application scenarios and parameter optimization problems of common filling algorithms under different missing mechanisms and different missing rates, and compares the improved algorithms through examples, in order to provide reference for practical application.

Firstly, the mathematical description of three kinds of missing mechanisms, namely Missing Completely at Random, Missing at Random and Not Missing at Random, is given by mathematical induction, and the simulation codes of three kinds of missing mechanisms are given. Then the K-Nearest Neighbor algorithm is optimized by cross validation method and Gaussian function weighting method, and a weighted K-Nearest Neighbor filling algorithm is proposed. At the same time, in view of the K-Nearest Neighbor of Weighted fill algorithm for different sample points missing value calculated neighbor distance difference caused by unreasonable weights allocation problem, through specific observation data set to dynamic adjustment, the parameters of the Gaussian function was proposed based on dynamic K-Nearest Neighbor of Weighted fill algorithm, theoretical research and empirical analysis shows that, This method can improve the filling effect and has certain universality. Finally, the K-Nearest Neighbor of Weighted fill algorithm to rely too much on the adjacent sample points and led to the decrease of the algorithm stability problem, this paper creatively using the algorithm of Missing Forest for its fill the calibration results, fill algorithm against loss rate increases and lead to decline gradually, In this paper, iterative method is used to gradually reduce the missing rate of incomplete data sets in the filling process, The K-Nearest Neighbor of Weighted and Missing Forest Hybrid Iterative filling algorithm is proposed. Empirical analysis shows that, based on the premise of different missing mechanisms and missing rates, this method not only inherits K-Nearest Neighbor of Weighted algorithm in filling accuracy, but also inherits Missing Forest algorithm in stability.

Keywords: Missing Mechanism　K-Nearest Neighbor　Iteration Method　Missing Forest　Dynamical Parameters Regulating

目 录

1 绪 论

1.1 研究背景和意义

当下，数据对各行业的重要性愈发显著，利用数据进行分析，挖掘背后的关联逻辑，并以此进行决策判断的领域越来越多。然而，在真实的数据处理过程中，数据预处理往往伴随着数据缺失、噪声、异常值、数据来源不一致等因素，这给后续的数据处理工作带来了极大困扰，其分析结果的质量也因此受损[1]。数据缺失即是数据预处理过程中最为常见的问题之一，数据缺失是指数据集的某一个或多个观测变量存在缺失值[2-3]。

数据缺失普遍存在于社会调查、生物研究、经济分析[4]、临床数据[5]等各个领域，比如技术上的无法获取[6]，最直观的案例便是在医疗领域，由于医学技术当下的局限性导致一些医疗仪器无法检测到某些疾病相关的属性信息而导致的数据的缺失数据[7-8]；也有因获取代价过高、数据采集设备故障或由于某种原因，比如说隐私问题而造成的单元无回答等情况也会影响样本集的完整性[9]。此外，样本数据收集完成后，因工作人员失误而导致的数据丢失[10]，或是收集上来的部分数据有误、数据不可用也会间接造成样本集的不完整；又如社会调查领域也普遍存在无回答现象[11]，这不仅降低了样本本身所具有的代表性，有些甚至会出现线性回归估计量出现严重偏差[12]。

针对无回答的处理方式，有事前预防和事后补救两种措施。事前预防主要体现在收集问卷的手段以及收集人本身的素质是否可靠，然而这只能在一定程度上避免缺失值的出现，并不能完全杜绝缺失值情况的发生；事后补救最直接的方法是有针对性的进行再次调查，以便获得更为完整的数据集[13]，而这将支出额外的人力、物力、财力。在社会调查领域，无论采用哪种方式，缺失值问题仍不可避免。

针对具体的数据缺失要选用合适的处理方法，比如填补缺失数据，而具体填补方法的选取依据是缺失变量的丢失类型[14]。面对繁冗复杂的缺失情况，1987 年，Little 和 Rubin[15]两位学者提出了三种缺失机制，定义了数据缺失的类型，即完全随机缺失（MCAR，Missing Completely at Random）、随机缺失（MAR，Missing at Random）、非随机缺失（NMAR，Not Missing at Random），并在后续的研究中[16]进行了更为详实的阐述说明。值得一提的是，数据缺失是客观存在的，往往

会给统计分析工作带来阻碍。比如使小样本数据集可用信息变得更少[17]，进而导致参数估计误差增大，置信区间出错，且这些问题都会影响统计推断的有效性，从而引发错误的判断[18-19]。此外，很多统计分析手段都是基于完整数据集进行的，数据集的不完整性会导致统计分析方法适用性降低。如何对非完整数据集进行有效分析备受人们的关注，其中最为重要的便是对数据集中缺失值的处理方式。

1.2 国内外研究现状

对缺失值的处理方法一般有不处理、删除、缺失值填补三种方式。其中删除法包括样本删除、变量删除、成对删除三种，比如在流行病学的研究中，处理缺失值的常用方法之一便是删除法，这被称为“完整病例分析”[20]。样本删除是指将含有缺失值的样本直接删除，这在样本量很大且缺失率较小时可以考虑使用[6]，而当样本量小或缺失率较大时，该方法会丢掉过多信息，从而导致误差增大[21]；变量删除是指在数据收集过程中，某一个观测变量的观测值缺失率过高，其原因可能是问题设计的不合理等原因，则可以将该变量整体删除；成对删除又叫完全变量分析，即只分析与特定研究问题相关的变量，无需使用含有缺失值的变量[6]。由于不处理会影响统计推断结果，而删除法会造成信息的浪费，因此在更多情况下，人们将关注点聚焦在缺失值填补上。

缺失值填补按填补值个数可分为单值填补和多值填补两类[22]，单值填补表示为缺失值填补一个可能的估计值，而多值填补表示为缺失值填补多个可能的估计值。单值填补方法有很多，且每种填补方法均有特定的使用条件和范围，如均值填补、中位数填补、回归填补等。均值填补是利用观测变量中已观测部分的值对该变量中的缺失值进行填补处理，具体又可细分为无条件均值填补、分层均值填补两种，然而均值填补在一定程度上掩盖了数据的分布特征，且根据缺失机制的定义，该方法只在缺失率较小且满足MCAR机制前提下适用[23]，不适合处理更为复杂的数据情形。中位数填补与均值填补类似，且中位数填补和均值填补更加适用于离散型数据，针对连续型数据的缺失值处理，1960年，学者Buck首次回归填补思想。回归填补的核心是利用变量之间的关系拟合一个回归模型，并利用该模型以及含缺失值样本点的完全观测部分求取填补值。回归填补最大的问题是构造的回归方程没有残差项，对于自变量完全一致的样本点，它们的估计值也会相同，这会对原始样本造成不同程度的扭曲[24]，从而影响最终的判断。针对该问题，

部分学者结合随机抽样和回归法进行填补，即给普通的回归填补值增加一个随机项，提出随机回归填补法。此外，采用同期调查中的完整数据集对现有缺失值进行填补处理，或使用同类调查的历史数据进行填补也是常用的填补手段。Nordbotten和Chapman分别在1963年和1976年探讨了基于分析历史数据对缺失值进行填补处理的冷卡插补法，该方法又可细分为冷卡替代和比率替代两种[25]。Ernst和Ford于1980年和1983年利用同期调查数据对缺失值填补问题进行了讨论和改进[26]，并提出热卡插补概念，该方法又可分为随机热卡插补、分层热卡插补、最近距离热卡插补和序贯热卡插补[27]。

为每一个含缺失值样本点寻找一个或多个相似的临近样本点也是处理缺失数据的方法之一，理论上，如果相似程度越高，则基于这些样本点分析得到的填补结果就越准确。1951年，Fix和Hodges首次提出最近邻法对缺失值进行填补，该方法适合处理离散型数据,其原理是通过选取没有缺失数据的变量作为辅助变量，利用距离函数求取含缺失值样本点与其他完全观测样本点的距离，进而选择最近邻样本点对应的观测值作为插补值即可。该思想可以说是K近邻填补的前身，相较于K近邻填补,最近邻法填补效果稳定性差。基于这种局部相似性的理念,2001年O Troyanskaya等学者[28]提出了K近邻填补，该方法的核心就是为每一个含缺失值的样本点在完整数据集中寻找k个近邻，进而分析并生成最终的填补值[29-30]。在K近邻填补算法中，距离公式的选取和k值的大小是该算法的核心，以k值选取为例，过大会导致填补结果受无关样本点影响从而影响算法精度，过小则可能受异常值影响[31]。针对k值的选取问题，可以采用交叉验证等方法优化，并进一步采用加权的方式来降低k值选取不合理所带来的影响[32]。

2015 年，Tutz和Ramzan[33]提出了一种基于Lp距离的加权最近邻插补方法，针对高维情况，它能够自动选择有助于距离测量的相关变量，且不依赖于参数k，这种方法适用于相关性较高的数据缺失问题。2016 年，Hiroshi 和 Shehan[34]提出一种结合遗传算法和K近邻算法的数据插补方法，称为进化K近邻填补法。相比于普通的K近邻填补，这种方法引入了遗传算法对k的选择进行了优化，它能够根据给定数据集产生优化的k，并为数据集中的每个特征分配权重。由于K近邻填补在使用过程中，实质上是将原始数据集分为了完整数据集和非完整数据集两部分，针对于非完整数据集中的每一个样本点，均需要从完整数据集中取寻找

k个近邻，这就会存在当缺失率较大时，从原始数据集分离出的完整数据集规模变得很小，从而导致难以为非完整数据集中的样本点寻找出真正的近邻。针对该问题，杨日东等[35]于2019年提出局部K近邻填补，该方法的核心是通过切片的方式使选取近邻样本时可参考的完整数据集规模增大，从而解决了在缺失率较大时无法找到有效的样本近邻点问题。

基于决策树所演化而来的数据填补算法是目前研究的热门领域之一。决策树是机器学习领域中经典的分类算法之一,根据不同的属性对样本点进行逐层分类，直到不再可分或满足分类条件为止[36]。决策树算法具有模式简单、效率高等特点，但由于只会产生一个分类结果，当训练集样本较小或类别较大时，算法效果并不稳定。为此，可以通过构造多颗决策树的思想进行解决。Bootstrap方法是构造生成多颗决策树的前提，采用有放回的抽样方式对原始数据集进行抽样，并生成若干个与原始数据集规模相同的数据集，进而采用这些数据集生成决策树。其中，Bagging法是早期生成树的方式之一[37]，由于使用所有观测变量的缘故，采用该方法生成的决策树会出现大量相似或一致的现象。在此基础上，Breiman等[38]学者于2001年提出了随机森林，这解决了Bagging法中出现大量相似的问题。随机森林依然需要通过Bootstrap方法对原始数据集进行有放回的抽样，从而得到若干个与原始数据集规模一致的用于生成多颗决策树Bootstrap样本集，然后综合分析所有决策树的结果并得出相应结论。大量的研究表明，随机森林拥有良好的预测效果，且对数据集本身的要求很少，算法稳定性优良[39-40]。相较于Bagging法，随机森林最大的不同在于当前分裂节点变量的选择并不是从原始数据集的所有变量中进行选择的，而是通过原始数据集中随机抽取的变量库中随机选择的[41]，且该变量库中的变量个数小于原始数据集中的变量个数。基于该思想，学者Stekhoven在2012年提出缺失森林法来应对复杂多变的数据缺失问题[42]。由于缺失森林在填补开始时要先对缺失值部分进行初始估算,然后在此基础上进行迭代，直至填补过程结束，初始估算常用的方法是均值填补、众数填补等，陈婉娇等[43]尝试使用K近邻算法作为缺失森林填补过程的首轮估算方法,提出KNN-MF填补方法，并验证了其可行性。

仅考虑单值填补方法在填补性能方面所带来的提升是不够的，单值插补方法的不足也很明显，有些方法会明显改变数据的原有分布，比如均值填补等[44]降低

数据的方差，且缺失数据本身所具有的不确定性也无法得到体现[45]。为此，Rubin[46-47]等学者提出了多重插补的思想，这解决了单值插补方法的部分缺陷，是在单值填补的基础上衍生而来的填补方法[27]，结合不同单值填补方法所衍生出的多重插补法目前已经被广泛应用于各个领域，常见的多重插补方法有马尔科夫链蒙特卡洛方法MCMC[48-49]、倾向得分法PS[50]、预测均值匹配PMM[51-52]、EMB多重插补法[53-54]等，而MCMC多重插补法具体又包含DA多重插补[55]和EM多重插补[56]等，在之后的研究中，学者针对不同缺失机制下的常用多重插补算法插补重数选择问题及应用场景问题进行了详细探讨[57]。具体而言，多重插补法过程大致分为三步，一是对数据集中的每一个缺失值填补m个可能的近似值，从而生成m个完整数据集；二是采用处理完整数据集的统计方法对上述m个完整数据集进行分析，处理这些数据集的方法是一致的；三是综合分析上述结果，整合得到最终的插补结果，如采用相同的参数估计方法对m个完成数据集进行回归分析，进而得到m组参数估计，最后对m组估计值采用求平均的方式得到最终的参数值即可。

1.3 论文结构

本文围绕数据缺失展开论述，对填补算法进行逐步改进，探索不同算法的填补性质，各章的组织结构如下：

第一章 绪论：主要介绍研究背景及意义，阐述国内外研究内容以及相关进展情况。

第二章 数据缺失的原因及模拟方法：先详细阐述三种缺失机制，并对其进行数学描述，基于此展开讨论，分单变量和多变量两种缺失情形，为每一种缺失机制进行数学归纳；进一步的，对本文所用到的完全随机缺失、随机缺失、非随机缺失的计算机模拟方法进行了详细描述，通过实证分析验证该模拟方法的有效性和可靠性，给出构造三种缺失机制的核心代码。

第三章 不同缺失机制下K近邻填补及其改进算法的模拟与实证研究：通过实证分析的手段，采用均值填补、中位数填补、K近邻、加权K近邻对不同缺失机制、不同缺失率下的非完整数据集进行填补，在填补过程中分析K近邻、加权K近邻算法核心参数的选取依据，并探讨加权K近邻中最优的权重分配方案。实证分析环节，先考虑对样本集中的某单一变量进行缺失处理，进而在不同评价准则下对填补效果的优良性进行比较，然后将其推广至多变量缺失情形，并使用相同

的实验方法对其进行对比分析，得出相关实验结论。

第四章 不同缺失机制下缺失森林及其改进算法的模拟与实证研究：本章是对第三章内容的延伸，考虑到改进后的加权 K 近邻所面临的不足，以及缺失率增大而出现的填补算法效果下降问题，本章采用缺失森林对加权 K 近邻的每一个填补结果进行校准，采用迭代法解决缺失率变大所带来的填补算法效果下降问题，从而提出加权 K 近邻与缺失森林混合迭代填补法。在实证分析环节，通过与第三章一致的实验方法和步骤对本章的三种填补算法进行对比分析。

第五章 总结与展望：对全文进行总结，并给出未来工作展望及研究构想。

2 数据缺失机制及模拟方法

本章先对三种数据缺失机制进行数学描述，然后分单变量缺失和多变量缺失两种情形对每一种缺失机制进行数学归纳，进而对本文使用的缺失机制模拟方法给出了核心代码，最后通过实证分析，验证本章提供模拟方法的有效性和可靠性。

2.1 数据缺失的机制

记$M=(M_{obs,}M_{mis})$为$n\times m$的观测样本，其中M_{obs}为已观测部分，M_{mis}为未观测部分。记R为$n\times m$的指示变量矩阵，对于M中能够被观测到的部分，R中对应的元素值记为1，否则记为0。在缺失机制的研究中考虑概率$P(R|M,\phi)$，ϕ是与缺失机制有关的未知参数，基于此，若$P(R|M_{obs},M_{mis},\phi)=P(R|M_{obs},\phi)$，则缺失机制为MAR，即数据的缺失只与能够观测到的部分有关，而与未观测部分无关；若$P(R|M_{obs},M_{mis},\phi)=P(R|\phi)$，则缺失机制为MCAR，即数据是否缺失与完全变量和不完全变量均没有关系；若$P(R|M_{obs},M_{mis},\phi)=P(R|M_{mis},\phi)$或$P(R|M_{obs},M_{mis},\phi)=P(R|M_{obs},M_{mis},\phi)$，则缺失机制为NMAR，即数据是否缺失与未观测到的数据有关，或许还与已观测到的数据有关。

2.2 完全随机缺失

根据定义，MCAR机制下数据的缺失与已经观测到的变量和未被观测到的变量均是无关的，因此，样本点是否含有缺失值是完全随机产生的，可以采用随机数生成法对其进行模拟实现。在实际问题中，由于每一个样本点均有若干个观测指标，即缺失值可能存在于某一个或某几个指标当中，所以，在完全随机缺失这一前提下，本文考虑单变量完全随机缺失和多变量完全随机缺失两种情形。

2.2.1 单变量完全随机缺失

有且仅有X_j中含有缺失值的情形为单变量数据缺失，为了模拟MCAR，先求解样本集M中含有缺失值的样本数量，如下：

$$n_{mis}=g(n\times p) \tag{2.1}$$

式(2.1)中，n_{mis}代表M含缺失值的样本数量，n代表M样本总数，p代表缺

失率，$g(\cdot)$为取整函数。进一步，通过随机数生成法计算出含有缺失值的行，公式如下：

$$i_{mis} = f(1:n, n_{mis}) \tag{2.2}$$

式(2.2)中，$f(\cdot)$代表等概率抽样函数，即从整数序列1到n中等概率随机抽取n_{mis}个数，i_{mis}代表M中含有缺失值的样本点的i值集合。然后将$M_{i_{mis}j}$重新赋值为NA即可。

2.2.2 多变量完全随机缺失

有且仅有$X_J, J \subseteq [1,m]$中含有缺失值的情形为多变量数据缺失，为了模拟MCAR，同样需要求解式(2.1)和式(2.2)中的n_{mis}和i_{mis}两个结果。由于实际问题中待观测样本点的具体缺失指标是随机发生的，针对每一个$i_{mis}^{\zeta}, \zeta \in [1, n_{mis}]$，$i_{mis}^{\zeta}$代表$i_{mis}$中的第$\zeta$个值，通过如下公式计算得出第$i_{mis}^{\zeta}$个样本点具体的缺失列集合$\tau$：

$$\tau = f\left(J, f\left(1:h(J),1\right)\right) \tag{2.3}$$

式(2.3)中，J代表M中所有含有缺失值的列的集合，$h(\cdot)$代表计算向量长度函数。然后将$M_{i_{mis}^{\alpha}\tau}$重新赋值为NA即可，依此类推，通过公式(2.3)将所有i_{mis}中对应的缺失列进行数据缺失处理。

特别的，当式(2.3)中的集合J只有一个元素时，此时多变量完全随机缺失将退化为单变量完全随机缺失。若将式(2.3)作如下变换

$$\tau = f\left(J, f\left(1:1,1\right)\right) \tag{2.4}$$

此时M中的缺失情况仍然满足多变量完全随机缺失情形，但并不完全符合现实生活中的实际问题，本文将这种情况称为伪多变量完全随机缺失。

2.3 随机缺失

根据定义，MAR机制下数据的缺失与已经观测到的变量是有关的，因此，M中样本点是否含有缺失值受完全观测变量的影响，即X_J的观测值是否缺失与X_{φ}存在某种关联，$X_{\varphi}, \varphi \in [[1,m]-J]$为不包含缺失值变量中的第$\varphi$个变量。同MCAR类似，为了更好的模拟现实问题，在MAR下，本文仍考虑单变量随机缺失和多变量

随机缺失两种情形。

2.3.1 单变量随机缺失

有且仅有X_j中含有缺失值的情形为单变量数据缺失，为了模拟MAR，先通过公式(2.1)先求解样本集M中含有缺失值的样本数量n_{mis}，进而假设X_j的数据缺失与X_φ有关。从观测值的分布特征来看，X_φ中的观测值y_φ可能近似服从均匀分布或其他分布；从观测值的数值大小来看，排序后的y_φ为一组递增或递减序列。

（1）依据观测值的分布特征

如果采用y_φ的分布特征作为X_j中观测值缺失的依据，应分为两种情况看待：

第一，y_φ近似服从均匀分布时，此时采用随机抽样法对y_φ随机抽取n_{mis}个值，记录其在M样本集中X_φ变量对应的i值作为i_{mis}，然后将$M_{i_{mis}j}$重新赋值为NA即可，由于y_φ近似服从均匀分布，因此，在该种情况下，MAR将退化为MCAR；

第二，y_φ近似服从其他分布时，此时仍采用随机抽样法对y_φ随机抽取n_{mis}个值，记录其在M样本集中X_φ变量对应的i值作为i_{mis}，最后将$M_{i_{mis}j}$重新赋值为NA即可，由于y_φ不是等概率分布，因此通过该方法得到的i_{mis}具有与y_φ类似的分布特征，该种情况满足MAR机制的定义。然而，不论y_φ的值近似服从什么分布，当缺失率p增大时，该种情形下的MAR将会逐步向MCAR情形靠拢，最终退化为MCAR。

（2）依据观测值的数值大小

如果采用排序后y_φ所生成的一组递增或递减趋势的序列作为X_j中观测值缺失的依据，需要考虑分层抽样缺失和分组抽样缺失两种情形，为了便于讨论，先对y_φ进行排序处理，然后按序逐一取出该观测值排序前在M中所对应的i值，形成新的集合i_o，进而将i_o划分为z个非空子集$i_o^\beta,\beta\in[1,z]$，i_o^β与i_o满足如下公式：

$$\begin{cases} i_o=\bigcup_{\beta=1}^{z} i_o^\beta \\ \sum_{a=1}^{z-1}\sum_{b=a+1}^{z} i_a\cap i_b=\varnothing \end{cases},\beta\in[1,n] \tag{2.5}$$

具体模拟过程如下：

第一，在分层抽样缺失情形中，先按照预定规则从不同i_o^β中随机抽取$n_{mis}^\beta, n_{mis}^\beta \in [0, n_{mis}]$个元素作为$i_{mis}$集合的子集$i_{mis}^\beta$，$n^\beta$、$n_{mis}^\beta$、$n_{obs}^\beta$、$n_{mis}$、$i_{mis}^\beta$和$i_{mis}$存在如下关系：

$$\begin{cases} n_{mis} = \sum n_{mis}^\beta \\ i_{mis} = \bigcup i_{mis}^\beta \quad , \beta \in [1, z] \\ n_{mis}^\beta + n_{obs}^\beta = n^\beta \end{cases} \tag{2.6}$$

式(2.6)中，n^β表示第β个子集i_o^β中样本总数，n_{obs}^β表示第β个子集i_o^β中未被抽取的样本总数且$n_{obs}^\beta \in [0, n^\beta]$，$i_{mis}^\beta$表示从第$\beta$个子集$i_o^\beta$中抽取的样本点的集合，$i_{mis}$表示从所有$i_o^\beta$抽取的样本点的集合，最后将$M_{i_{mis}j}$重新赋值为NA即可。特别的，当$z=1$时，此时采用该方法模拟的MAR将退化为MCAR。

第二，分组抽样缺失是分层抽样缺失的一种特殊形式，即$n_{mis}^\beta = n_{mis}$时，此时n^β、n_{mis}^β、n_{obs}^β、n_{mis}、i_{mis}^β和i_{mis}存在如下关系：

$$\begin{cases} n_{mis} = n_{mis}^\beta \\ i_{mis} = i_{mis}^\beta \quad , \beta \in [1, z], n_{obs}^\beta \in [0, n^\beta - n_{mis}] \\ n_{mis}^\beta + n_{obs}^\beta = n^\beta \end{cases} \tag{2.7}$$

式(2.7)表明，对i_o的第β个子集i_o^β中的全部元素或部分元素进行抽样便可满足设定缺失率，为了完全契合分组抽样缺失过程，本文对i_o采用等分的手段进行分组，生成若干子样本集$i_o^1, i_o^2, \cdots, i_o^{z-1}, i_o^z$，使$n^1 = n^2 = \cdots = n^{z-1} = n_{mis}$，$z$的值通过式(2.8)求解得出：

$$\begin{cases} z = g\left(\dfrac{n}{n_{mis}}\right) + 1, h\left(i_o^z\right) > 0 \\ z = g\left(\dfrac{n}{n_{mis}}\right), h\left(i_o^z\right) = 0 \end{cases} \tag{2.8}$$

根据式(2.8)结果，此时，公式(2.7)中$n_{obs}^\beta = 0$，可化简为如下形式：

$$\begin{cases} n_{mis} = n_{mis}^{\beta} \\ i_{mis} = i_{mis}^{\beta} \\ n_{mis}^{\beta} = n^{\beta} \end{cases}, \beta \in [1, z] \tag{2.9}$$

由于M中样本点数量的不确定性以及模拟缺失率的不同，最后一组子样本集M_z中样本点数量存在如下关系式：

$$n^z \le n_{mis} \tag{2.10}$$

式(2.10)中，当$n^z < n_{mis}$时，本文对i_o的最后一个分组做截尾处理，当$n^z = n_{mis}$时为分组时的特殊情况，i_{mis}通过式(2.9)直接进行确定，最后在M样本集中将$M_{i_{mis}j}$重新赋值为NA即可。

分组抽样缺失是本文重点讨论的MAR模式，其详细的计算机模拟过程如下：

第一，先通过式(2.1)求解n_{mis}；

第二，通过排序函数对y_φ进行升序排列，然后从小到大依次取出X_φ变量的观测值在排序前所对应的i值，并组成新的集合i_o；

第三，对i_o进行采用等分的手段进行分组，生成若干子样本集$i_o^1, i_o^2, \cdots, i_o^{z-1}, i_o^z$，使$n^1 = n^2 = \cdots = n^{z-1} = n_{mis}$，并依据公式(2.10)对$i_o$的最后一个分组是否做截尾处理进行判定；

第四，随机选择一个i_o^β，β的具体值由公式(2.11)计算得出：

$$\beta = f(1:(z-1),1) \tag{2.11}$$

令$i_{mis} = i_o^\beta$，最后在M样本集中将$M_{i_{mis}j}$重新赋值为NA即可。

2.3.2 多变量随机缺失

有且仅有x_J中含有缺失值的情形为多变量数据缺失，同多变量完全随机缺失的模拟方法类似，仍需要通过式(2.1)求解n_{mis}，然后根据单变量随机缺失中描述的方法得到i_{mis}。由于x_J中各具体变量观测值的缺失与否可能都与某一个X_φ变量相关，也可能与某几个X_φ变量相关，针对每一个$i_{mis}^\zeta, \zeta \in [1, n_{mis}]$，通过公式(2.3)

计算得出第i_{mis}^{ζ}个样本点具体的缺失列集合τ，然后将$M_{i_{mis}^{\zeta}\tau}$重新赋值为NA即可，依此类推，将所有i_{mis}中每个元素值对应的缺失列进行数据缺失处理即可。

特别的，当式(2.3)中的集合J只有一个元素时，此时多变量随机缺失将退化为单变量随机缺失；若采用式(2.4)对每个含缺失值样本点的τ进行求解，此时M中的缺失情况仍然满足多变量随机缺失情形，但并不完全符合现实生活中的实际问题，本文将这种情况称为伪多变量随机缺失。

2.4 非随机缺失

根据定义，NMAR机制下数据的缺失与自身或其他未观测到的变量是有关的，即X_J中观测值是否缺失与X_J自身相关或与其他未知变量Ω相关，Ω代表除M中观测变量以外的其他还未观测到的变量集合。同前文类似，在NMAR下，本章仍考虑单变量非随机缺失和多变量非随机缺失两种情形。

2.4.1 单变量非随机缺失

有且仅有X_j中含有缺失值的情形为单变量数据缺失，为了模拟NMAR，仍需先求解公式(2.1)得到样本集M中含有缺失值的样本数量n_{mis}，进而假设X_j中观测值缺失与X_j自身相关或与其他未知变量Ω相关。

（1）观测值的缺失与自身相关

设X_j中的所有观测值为y_j，假设观测变量X_j中的观测值是否缺失与其自身存在某种关联，本文仍从y_j的近似分布特征和数值大小两个方面探讨计算机模拟NMAR的过程。从观测值的分布特征来看，y_j同样会近似服从均匀分布或其他分布；从观测值的数值大小来看，排序后的y_j为一组递增或递减序列。

当y_j近似服从均匀分布时，此时采用随机抽样法对y_j随机抽取n_{mis}个值，记录其在M样本集中X_j变量对应的i值作为i_{mis}，然后将$M_{i_{mis}j}$重新赋值为NA即可，由于y_j近似服从均匀分布，因此，在该种情况下，NMAR将退化为MCAR；

当y_j近似服从其他分布时，此时仍采用随机抽样法对y_j随机抽取n_{mis}个值，

记录其在M样本集中X_j变量对应的i值作为i_{mis}，最后将$M_{i_{mis}j}$重新赋值为NA即可，由于y_j不是等概率分布，因此通过该方法得到的i_{mis}具有与y_j类似的分布特征，该种情况满足NMAR的定义。然而，不论y_j的值近似服从什么分布，当缺失率p增大时，该种情形下的NMAR将会逐步向MCAR情形靠拢，最终退化为MCAR。

当y_j没有显著的分布特征时，同MAR一样，先将y_j进行排序，以此取出排序后的值在M对应的i值组成i_o，进而通过分层抽样缺失或分组抽样缺失两种手段确定i_{mis}，最后在$M_{i_{mis}j}$中将对应的观测值进行挖空处理即可。在该部分，本文仍采用分组抽样缺失的方式进行NMAR模拟，详细过程如下：

第一，仍需先通过式(2.1)求解M样本集中含有缺失值的样本点总数n_{mis}；

第二，通过排序函数对y_j进行升序排列，然后从小到大依次排序后的y_j在M中所对应的i值，并组成新的集合i_o；

第三，依据式(2.7)(2.8)(2.9)对i_o进行分组处理并得到i_o^{β}，再依据式(2.10)对i_o^{β}是否需要进行结尾处理进行判定；

第四，采用式(2.11)确定β值，令$i_{mis}=i_o^{\beta}$，最后在M样本集中将$M_{i_{mis}j}$重新赋值为NA即可。

（2）观测值的缺失与其他未知变量相关

引入协变量X_t，X_t为Ω的其中一个因素，假设X_j中样本点是否含有缺失值依赖于X_t，本文利用X_t生成M样本集中含有缺失值的样本点，具体步骤如下：

第一，生成一个X_t的观测值$y_t^{\zeta},\zeta\in[1,n_{mis}]$，令$y_t^{\zeta}\sim N(\mu,\sigma^2)$，其中$\mu=\frac{1}{n}\sum_{i=1}^{n}i$，$\sigma^2=\frac{1}{n}\sum_{i=1}^{n}(i-\mu)^2$；

第二，通过取整函数将y_t^{ζ}整数部分保留，当$1\le y_t^{\zeta}\le n$，且y_t^{ζ}不等于y_t中的任何一个元素时，将y_t^{ζ}的值存放至集合y_t中，y_t代表协变量X_t所有观测值的一

个子集，且该集合中元素个数为n_{mis}；

第三，重复上述两个步骤，直至集合y_t中有n_{mis}个元素为止；

第四，令$i_{mis}=y_t$，在M样本集中将$M_{i_{mis}j}$重新赋值为NA即可。

值得说明的是，该方法下模拟的NMAR会随着缺失率p的增大逐步向MCAR靠拢。

上述NMAR模拟方法中，观测值的缺失是服从正太分布，而且该方法可以推广至其他分布。然而在观测值是否缺失与其他未观测变量有关的情形中，还有一类常见的数据类型是时间序列数据，如病人在某个时间节点住院，或中途离开等原因都会导致某一项或几项观测指标中的观测值出现缺失，这种缺失情形称之为左删失或右删失数据，本章不做重点考虑。

2.4.2 多变量非随机缺失

有且仅有中X_J含有缺失值的情形为多变量数据缺失。结合NMAR机制的定义，将单变量非随机缺失推广到多变量非随机缺失时，即使采用分层抽样缺失或分组抽样缺失的方法对M数据集进行缺失处理，多变量非随机缺失的计算机模拟过程与多变量完全随机缺失、多变量随机缺失的计算机模拟方法依然有所不同，详细过程如下：

第一，仍需先通过式(2.1)求解M样本集中含有缺失值的样本点总数n_{mis}，此时的n_{mis}为含有缺失值样本点总数的初始值；

第二，令$X_J^{\omega},\omega\in(1,h(J))$为$X_J$中的第$\omega$个含缺失值的变量，在使用分组缺失的方法对NMAR进行模拟时，X_J^{ω}所对应的观测值y_J^{ω}是进行排序的依据，X_J^{ω}所含缺失值的具体个数n_{mis}^{ω}是进行分组时n^{β}大小的依据，即$n^{\beta}=n_{mis}^{\omega}$。在含有缺失值的样本点总数$n_{mis}$的基础上，本文通过式(2.12)对$X_J^{\omega}$所含的缺失值样本点占$n_{mis}$的比重进行指定，具体做法如下：

当缺失变量指定后，任意一个X_J^{ω}中所含缺失值的个数n_{mis}^{ω}应满足如下关系：

$$1\le n_{mis}^{\omega}\le n_{mis}-h(J)+1$$

进而，采用如下公式，先对每一个X_J^{ω}变量中的初始缺失数量p^{ω}进行随机指定，然后再求取W^{ω}：

$$\begin{cases} p^{\omega}=f\left(\frac{1}{n}:\frac{n_{mis}-h(J)+1}{n},h(J)\right) \\ W^{\omega}=\frac{p^{\omega}}{\Sigma(p^{\omega})} \end{cases},\omega\in(1,h(J)) \tag{2.12}$$

式(2.12)中，$h(\cdot)$是统计集合中元素个数的函数，W^{ω}代表变量x_j^{ω}中所含缺失值的数量占n_{mis}的比重。在多变量非随机缺失的计算机模拟中，本文使用如下公式求解具体的n_{mis}^{ω}：

$$n_{mis}^{\omega}=g(n_{mis}\times W^{\omega}),\omega\in(1,h(J)) \tag{2.13}$$

式(2.13)中，n_{mis}^{ω}代表所含缺失值的第ω个变量中含有缺失值个数。由于式(2.13)中采用$g(\cdot)$对阶段结果进行取整，此时的n_{mis}^{ω}与n_{mis}会有如下关系：

$$n_{mis}\geq\sum n_{mis}^{\omega},\omega\in(1,h(J))$$

为了保证M样本集缺失值的总数为n_{mis}，现对n_{mis}^{ω}中值最小的元素n_{mis}^{σ}采用如下公式进行修正，假设该元素在集合J中所对应的下标为σ，则：

$$n_{mis}^{\sigma}=n_{mis}-\sum n_{mis}^{\omega}+\sigma(n_{mis}^{\omega}),\omega\in(1,h(J)) \tag{2-14}$$

式(2.13)中，$\sigma(\cdot)$代表取最小值函数。经过修正后的使n_{mis}^{ω}和n_{mis}满足如下关系：

$$n_{mis}=\sum n_{mis}^{\omega},\omega\in(1,h(J))$$

针对每一个X_J^{ω}，依据式(2.7)(2.8)(2.9)将对应的i_o进行分组处理并得到i_o^{β}，再依据公式(2.10)对i^{β}是否需要进行结尾处理进行判定，然后采用公式(2.11)确定β值，令$i_{mis}=i_o^{\beta}$，进而在M样本集中将$M_{i_{mis}j}$重新赋值为NA即可。

特别的，由此方法模拟的多变量非随机缺失，会出现以下两种结果：

第一，依据不同的X_J^{ω}观测值进行排序、分组，进而求得的集合i_o^{β}会出现不同

程度的重叠部分，因此在 M 样本集中，会出现含缺失值的样本点总数量会小于初始值 n_{mis} 的现象，且随着缺失率 p 的增大，该现象会更加突出。在 R 平台上使用 Boston 数据集对该结果进行验证，将该方法用计算机随机模拟一次，所得结果如图 2-1 所示：

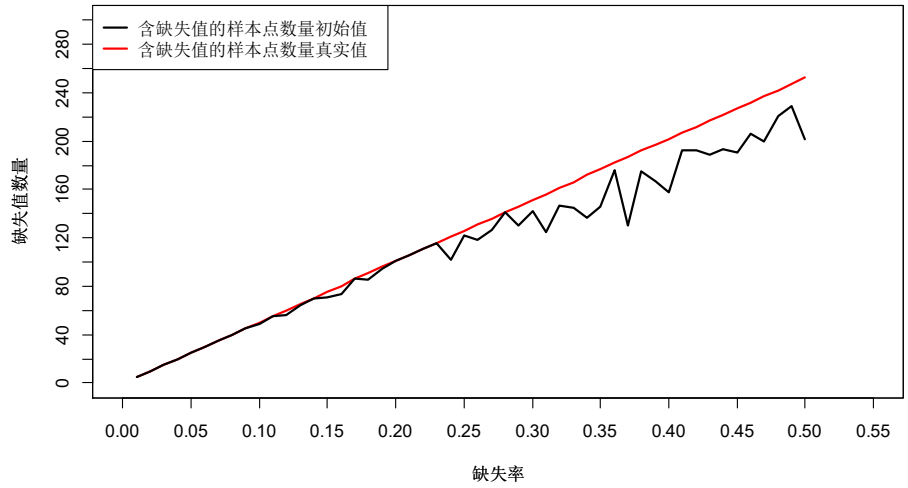

图 2-1：不同缺失率下随机执行 1 次程序的缺失值数量结果变化趋势图

图 2-1 显示，在使用分组缺失方法进行多变量非随机缺失的模拟过程中，随着缺失率 p 由 1%逐步增加，模拟后含缺失值的样本点总数量会小于初始值 n_{mis} 的现象会愈加明显，但与结果一所描述的并不完全吻合。考虑到模拟过程具有随机性，为不失一般性，现将每一种缺失率情况重复执行 100 次，然后求其均值，缺失率 p 仍由 1%逐步递增至 50%，模拟结果如图 2-2 所示：

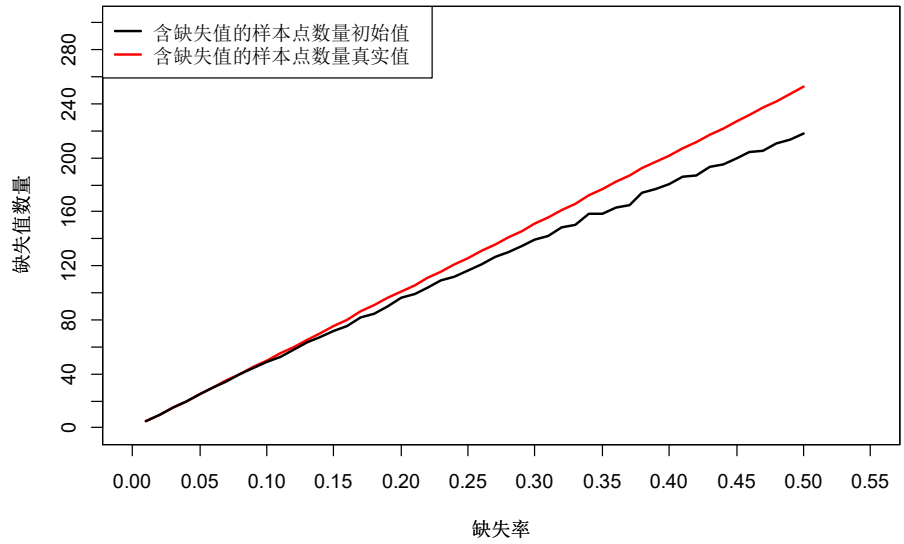

图 2-2：不同缺失率下随机执行 100 次程序的结果均值变化趋势图

图 2-2 显示，第一，实验结果验证了本文所描述的分组缺失方法进行多变量非随机缺失的计算机模拟过程中，结果一的客观真实性。第二，缺失率 p 较小时，含缺失值的样本点中有且仅有一个观测值缺失的情况占绝大多数，随着缺失率的增大，含缺失值的样本点中有超过两个及以上的观测值缺失情况占比会上升。

2.5 三种缺失方法的计算机模拟

本章使用 R 语言 MASS 包中的公开数据集 Boston 对三种缺失机制下的单变量、多变量两种情形进行计算机模拟，实验运行环境为 win10 系统，R 平台版本为 R version 4.0.5，程序执行版本为 RStudio Version 1.4.1106，计算机模拟程序核心代码段及模拟结果如下：

2.5.1 完全随机缺失

程序关键代码行如下：

```
n <- floor(p*length(Boston[,Column[sample(1:length(Column),1)]])) #求解 n_mis ;
row <- sort(sample(1:length(Boston[,Column[sample(1:length(Column),1)]]),n)) #求解 i_mis
for (i in 1:length(row )) {
    Column1 <- sort(Column[sample(1:length(Column),sample(1:length(Column),1))])
    Boston[row [i],Column1 ] <- NA
}#循环进行缺失处理。
```

运行结果：

在单变量完全随机缺失中，令 $p=0.3, j=11$，实验结果如图 2-3 所示：

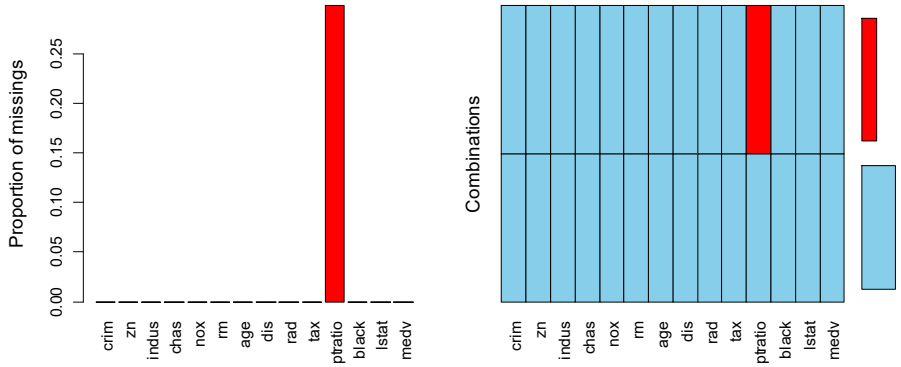

图 2-3：p=0.3,j=11 下单变量完全随机缺失分布图

在多变量完全随机缺失中，令 $p=0.2, J=(3,9,11)$，实验结果如图 2-4 所示：

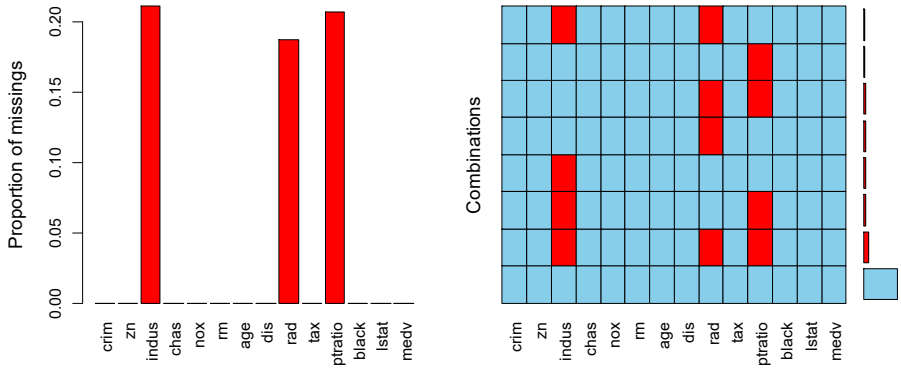

图 2-4：p=0.2,J=(3,9,11)下多变量完全随机缺失分布图

2.5.2 随机缺失

程序关键代码行如下：

```
n <- floor(p*length(Boston[,Column[sample(1:length(Column),1)]]))#求解 n_mis ;
for (i in 1:length(unique(sort(Boston[,14])))) {
    row_sort <- c(row_sort,which(Boston[,14] == unique(sort(Boston[,14]))[i]))
}#求解 i* ;
z <- sample(1:floor(1/p),1)#求解 z ;
row <- row_sort[(z*n-n+1):(z*n)]#求解 i_mis ;
for (i in 1:length(row )) {
    Column1 <- sort(Column[sample(1:length(Column),sample(1:length(Column),1))])
    Boston[row [i],Column1 ] <- NA
}#循环进行缺失处理。
```

运行结果：

在单变量随机缺失中，令 $p=0.3, j=11, \varphi=14$，实验结果如图 2-5 所示：

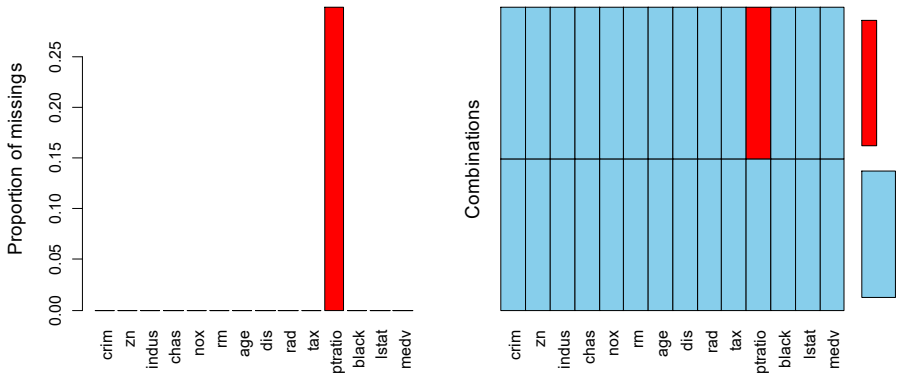

图 2-5：p=0.3,j=11, $\varphi=14$ 下单变量完全随机缺失分布图

在多变量完全随机缺失中，令 $p=0.3, J=(3,9,11), \varphi=14$，实验结果如图 2-6 所示：

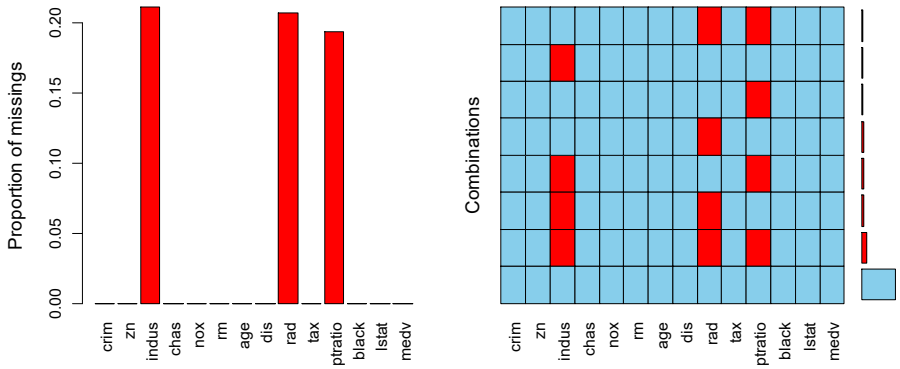

图 2-6：p=0.3,J=(3,9,11), $\varphi=14$ 下多变量完全随机缺失分布图

2.5.3 非随机缺失

程序关键代码行如下：

```
n <- floor(p*length(Boston[,Column[sample(1:length(Column),1)]]))#求解 n_mis ；
p_Column<- sample((1/length(Boston[,1])):(n-2)/length(Boston[,1]),length(Column))#求解 p^ω ；
n_Column<- c(n_Column,floor(n*(p_Column/sum(p_Column))))#求解 n_mis^ω ；
n_Column[which(n_Column==min(n_Column))[1]] <- n-sum(w_Column[-which(n_Column==
min(n_Column))[1]])#修正 n_mis^σ 的值；
for (i in 1:length(n_Column)) {
    row_every_column <-c()
    row_sort <- c()
    for (j in 1:length(unique(sort(Boston[,Column[i]])))) {
    row_sort <- c(row_sort,which(Boston[,Column[i]] == unique(sort(Boston[,Column[i]]))[j]))
    }
    z <- sample(1:floor(length(Boston[,Column[i]])/n_Column[i]),1)#求解 z ；
    row <- row_sort[(z*n_Column[i]-n_Column[i]+1):(z*n_Column[i])]#求解 i_mis ；
    for (j in 1:length(row )) {
        Boston[row [j],row [i]] <- NA
    }#循环进行缺失处理；
```

}

运行结果：

在单变量非随机缺失中，令 $p=0.3, j=11$，实验结果如图 2-7 所示：

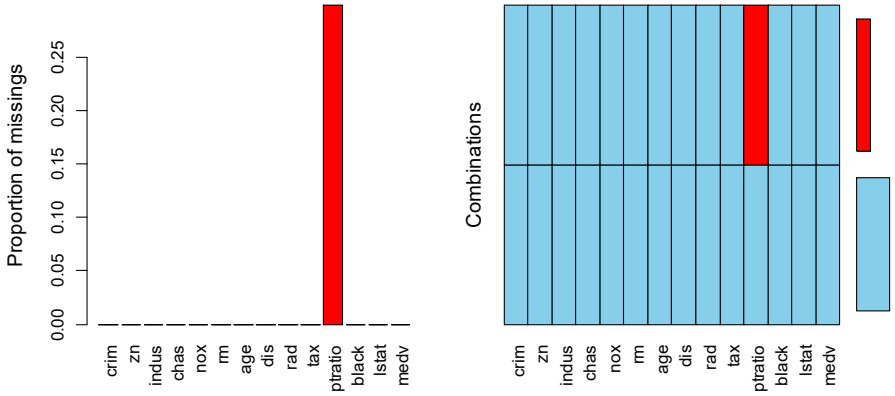

图 2-7：p=0.3,j=11 下单变量完全随机缺失分布图

在多变量非随机缺失中，令 $p=(0.1,0.3,0.5), J=(3,9,11)$，实验结果如图 2-8、图 2-9、图 2-10 所示：

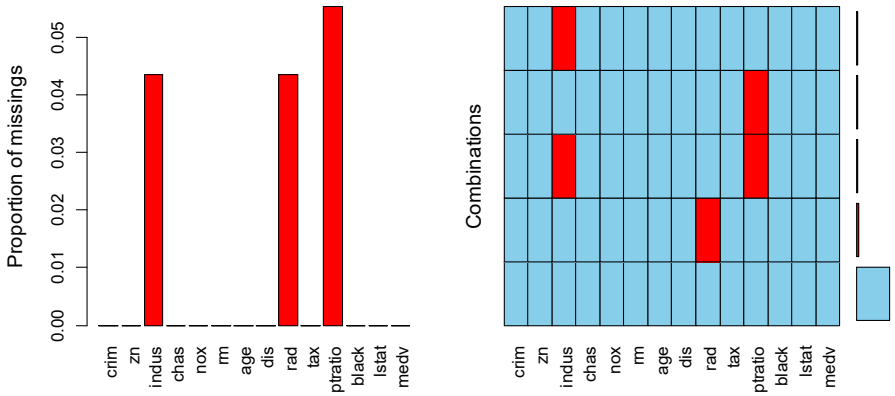

图 2-8：p=0.1,J=(3,9,11)下多变量完全随机缺失分布图

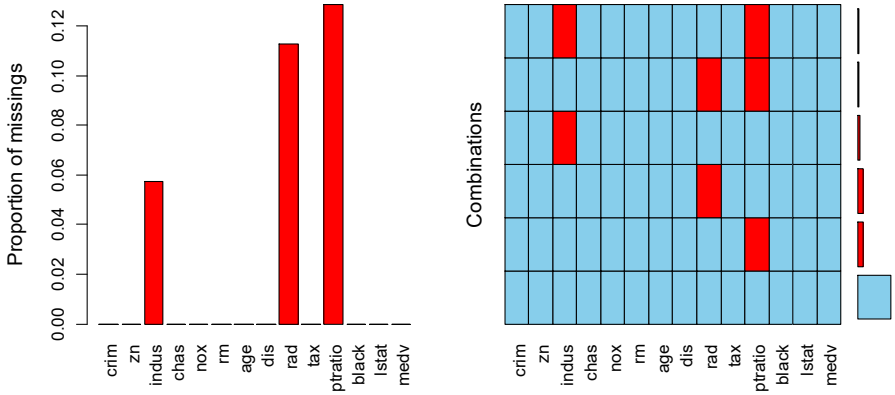

图 2-9：p=0.3,J=(3,9,11)下多变量完全随机缺失分布图

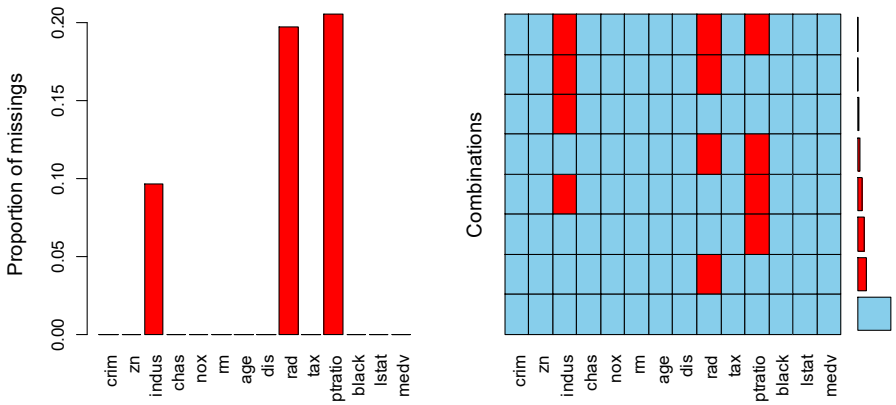

图 2-10：p=0.3,J=(3,9,11)下多变量完全随机缺失分布图

2.6 本章小结

本章针对三种缺失机制下的单变量和多变量两种缺失情形做了深入探讨，分析了随机缺失、非随机缺失机制下计算机模拟可能出现的结果和退化情况。

3 不同缺失机制下KNN及其改进算法的实证研究

K近邻中k值选取过大导致的分类模糊和k值选取过小导致的分类错误问题普遍存在，加权K近邻在一定程度上缓解了k值选取问题，但加权K近邻中权重分配是否合理将直接影响算法的填补效果。本章使用交叉验证法对k值进行优化，利用高斯函数进行权重函数的构造，针对权重分配问题，提出动态调参法为每一个待分类样本点进行动态的权重分配，以达到相对最优结果。

3.1 算法介绍

K近邻（K- Nearest Neighbor，简称KNN）核心思想是根据测试集中已有的样本信息为待分类样本点在训练集中寻找$k,k\geq1$个近邻，进而依据被选出的k个近邻来判断待分类样本点的类别。根据KNN算法的核心思想，如何选取k个近邻和选取几个近邻成了该算法的核心问题。

针对如何选取近邻点的问题，可以通过度量样本点之间的距离来解决。度量的距离公式有很多，常用的有欧式距离、马氏距离、曼哈顿距离等，本章选取曼哈顿距离公式进行度量，公式如下：

$$D_{\left(M_{\zeta}^{mis},M_{\gamma}^{obs}\right)}=\sum\left|M_{\zeta,\varphi}^{mis}-M_{\gamma,\varphi}^{obs}\right|,\zeta\in[1,n_{mis}],\gamma\in[1,n-n_{mis}],\varphi\in[[1,m]-J] \quad (3.1)$$

式（3.1）中，M_{ζ}^{mis}代表测试集样本点，M_{γ}^{obs}代表训练集样本点，$D_{\left(M_{\zeta}^{mis},M_{\gamma}^{obs}\right)}$代表$M_{\zeta}^{mis}$和$M_{\gamma}^{obs}$两个样本点之间的距离大小，值越小代表两个样本点越相似，φ代表不包含缺失值的变量对应的j值。

KNN具有简单直观的优点，然而，算法需要为每个测试集中的样本点寻找k个近邻，为了计算当前待分类样本与训练集中每个的样本点的距离，往往需要遍历整个训练集，时间开销巨大，尤其是在训练集样本点或测试集样本点数量庞大时。此外，KNN中k值过大，可能会导致分类结果模糊，而k值过小，可能会导致分类结果出错，而不同的数据样本，KNN中k值的最优解往往也不相同。综上，考虑具体问题中k值优化问题是算法优化的核心。

3.2 评价准则

本文假设$\hat{y}=\{\hat{y}_1,\hat{y}_2,\hat{y}_3,\ldots,\hat{y}_{\Phi}\}$，$y=\{y_1,y_2,y_3,\ldots,y_{\Phi}\}$。其中，$\hat{y}$代表填补值，

y代表真实值，Φ代表样本集中缺失值的总数量。为了评价填补算法的优良性，本文在平均绝对误差（Mean Absolute Error，MAE）、均方根误差（Root Mean Square Error，$RMSE$）、平均绝对误差百分比（Mean Absolute Percentage Error，$MAPE$）3个评价准则下比较填补算法的优良性。

其中，MAE、$RMSE$分别代表了填补值$\hat{y}$和真实值y的一种绝对误差，定义如下：

$$MAE=\frac{1}{\Phi}\sum_{\zeta=1}^{\Phi}|\hat{y}_{\zeta}-y_{\zeta}|$$

$$RMSE=\sqrt{\frac{1}{\Phi}\sum_{\zeta=1}^{\Phi}(\hat{y}_{\zeta}-y_{\zeta})^{2}}$$

由上式可以看出，MAE、$RMSE$取值范围为$[0,+\infty)$，且值越趋近于0，说明填补效果越好。相较于MAE而言，$RMSE$还反应了系统稳定性，对于异常点比较敏感，该值的越小不仅可以说明填补效果越好，也说明对各缺失值而言，填补效果更加稳定。

在实际生活中，由于被填补缺失变量真实值的大小在具体案例中相差会很大，如果只考虑MAE、$RMSE$，往往不能够反映出填补值与真实值之间相对于填补值本身的误差程度，因此，我们需要采用能够代表填补值$\hat{y}$和真实值y的一种相对误差，即$MAPE$，定义如下：

$$MAPE=\frac{100\%}{\Phi}\sum_{\zeta=1}^{\Phi}\left|\frac{\hat{y}_{\zeta}-y_{\zeta}}{y_{\zeta}}\right|$$

由上式可以看出，$MAPE$取值范围为$[0,+\infty)$，且值越趋近于0，说明填补效果越好。进行多次填补之后会得到不同的多组不同的评价结果，即

$$MAE=\{MAE_1,MAE_2,MAE_3,\cdots,MAE_N\}$$

$$RMSE=\{RMSE_1,RMSE_2,RMSE_3,\cdots,RMSE_N\}$$

$$MAPE=\{MAPE_1,MAPE_2,MAPE_3,\cdots,MAPE_N\}$$

进而对评价结果求其平均值，公式定义如下：

$$\overline{MAE}=\frac{1}{N}\sum_{\lambda=1}^{N}MAE_{\lambda}$$

$$\overline{RMSE}=\frac{1}{N}\sum_{\lambda=1}^{N}RMSE_{\lambda}$$

$$\overline{MAPE}=\frac{1}{N}\sum_{\lambda=1}^{N}MAPE_{\lambda}$$

其中，N 代表实验总次数，λ 代表第 λ 次实验，MAE_{λ}、$RMSE_{\lambda}$、$MAPE_{\lambda}$ 分别代表第 λ 实验得到的 MAE、$RMSE$、$MAPE$ 评价结果的值。

3.3 算法改进

3.3.1 交叉验证法

采用交叉验证法对 KNN 中 k 值选取进行优化，其核心思想是先对训练集数据进行随机排序，这可以消除原始数据结构化顺序带来的影响；然后对随机排序后的数据进行等额分组，每组样本量可以依据测试集样本数量与训练集样本数量的比值来确定；进一步的，在计算过程中设定 k 的取值范围，使 k 按既定规则逐步增大，然后基于具体的 k 值，随机选取一组子样本作为新的测试集数据，将其余组子样本作为新的训练集数据进行计算，并对所得评价结果求平均值；最后通过最优结果来确定 k 的取值。具体实现步骤如下：

第一，将训练集样本进行随机排序得到 M_o，然后将 M_o 等分为 z 个子样本集 $M_o^1,M_o^2,M_o^3,\cdots,M_o^z$，其中 $z=g\left(\frac{1}{p}\right)$，$p$ 为缺失率，$g(\cdot)$ 为取整函数；

第二，令 $M_o^{\beta},\beta\in(1,z)$ 作为新的测试集样本，并将 M_o^{β} 中对应的缺失变量 X_J 的真实值 y_J 进行删除处理，M_o 中的其余子样本集作为新的训练集样本，使用 KNN 算法对 X_J 的值进行填补，得到填补值 $\hat{y}_J$，其中 β 的值由式（2.11）随机产生。

第三，此时，M_o^{β} 中样本点数量为 n_{mis}，z 为子样本集数量，令 N 为每种 k 值下程序执行的次数；

第四，令 $MAE_{CV}=\{MAE_{CV}^1,MAE_{CV}^2,\cdots,MAE_{CV}^{\lambda},\cdots MAE_{CV}^N\}$，其中 MAE_{CV}^{λ} 代表第

λ 次运算得到的均方误差，$MAE_{CV}^{\lambda}=\frac{1}{\Phi}\sum|\hat{y}_J-y_J|,\lambda\in(1,N)$，$\Phi$ 代表 M_o^{β} 中观测值缺失总数；

第五，令 $\overline{MAE_{CV}}=mean(MAE_{CV})$，$MAE\_SD_{CV}=sd(MAE_{CV})$，其中 $mean(\cdot)$ 为求均值函数，$sd(\cdot)$ 为求标准差函数；

第六，本文将 k 的取值设为 1 到 20，步长为 1，重复第二步至第五步，寻找最小的 $\overline{MAE_{CV}}$，MAE_SD_{CV} 即可。

采用 Boston 数据集，基于上述步骤进行实验，本文所有实验过程均采用原始数据进行，不对其进行任何标准化操作。由于实验结果与数据集中的具体观测值相关，在交叉验证法中，我们预设所含缺失值变量列的集合 $J=3,9,11$，缺失率 $p=0.05,0.1,0.2$，并在单变量缺失与多变量缺失两种情况下来考虑 k 值的优化问题，且实验过程均在 MCAR 下进行。考虑到计算机模拟具有随机性，对每一个 k 值进行 100 次程序模拟，具体实验结果如图 3-2-1 所示：

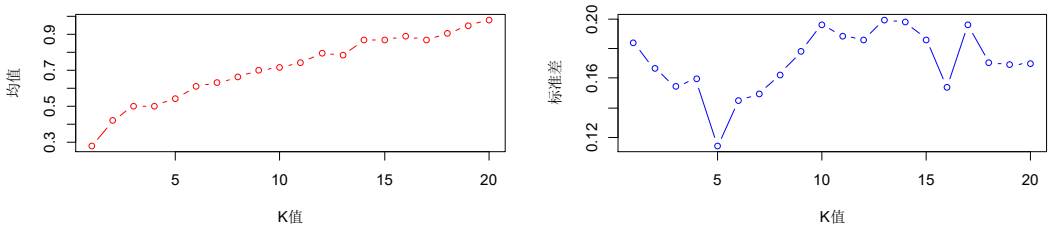

图 3-3-1：p=0.05, j=11 时不同 k 值下 100 次 MAE 结果的均值与标准差

根据图 3-3-1 实验结果可得，随着 k 的不断增大，$\overline{MAE_{CV}}$ 结果也逐步变大；当 $1\le k\le 5$ 时，MAE_SD_{CV} 结果逐步减小，这也符合了 k 选取过小导致选错临近点的情况时有发生，从而导致结果的不稳定性；当 $k\ge 10$ 时，MAE_SD_{CV} 结果没有明显规律，且此时的 $\overline{MAE_{CV}}$ 结果远大于 $k\in[1,5]$ 时的 $\overline{MAE_{CV}}$ 结果；当 $k\ge 18$ 时，MAE_SD_{CV} 结果趋于稳定，此时的 $\overline{MAE_{CV}}$ 结果也达到最大，这印证了在 KNN 中，k 选取过大导致的分类模糊的情况出现，即算法失效。综上，当 $k=5$ 时，$\overline{MAE_{CV}}$ 结果相对较小，但 MAE_SD_{CV} 结果显著优于其他情况，在完全随机缺失下，当 $J=11$，$p=0.05$ 时，优化后的 $k=5$。

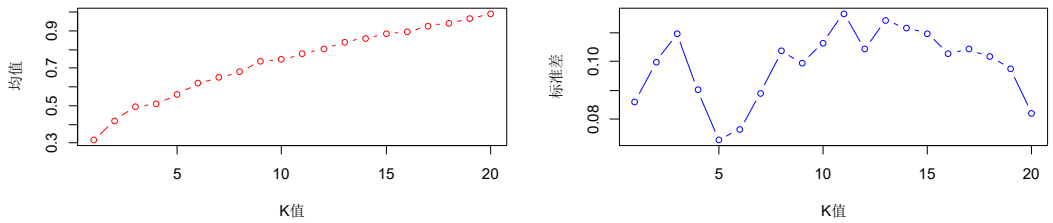

图 3-3-2：p=0.1, j=11 时不同 k 值下 100 次 MAE 结果的均值与标准差

根据图 3-3-2 实验结果可得，随着 k 的不断增大，$\overline{MAE_{CV}}$ 结果仍逐步变大；当 $1 \le k \le 5$ 时，MAE_SD_{CV} 结果先逐步增大，随后逐步减小，$k=5$ 时达到最小，这基本符合图 3-3-1 的结论；当 $k>5$ 时，MAE_SD_{CV} 结果有逐步变大的趋势，并在 $k=11$ 时达到最大值，随后 MAE_SD_{CV} 的值又逐步减小，且此时的 $\overline{MAE_{CV}}$ 结果远大于 $k \in [1,5]$ 时的 $\overline{MAE_{CV}}$ 结果，同样印证了在 KNN 中，k 选取过大导致算法失效的问题；由于此时的缺失率大于图 3-3-1 中所使用的缺失率，因此会出现图 3-3-2 中当 $k \ge 19$ 时，MAE_SD_{CV} 结果会出现显著性降低的情形，且随着缺失率的增大，MAE_SD_{CV} 结果出现显著性降低的 k 值会越来越小，而下降的趋势则会减缓。综上，当 $k=5$ 时，$\overline{MAE_{CV}}$ 结果相对较小，但 MAE_SD_{CV} 结果显著优于其他情况，在 MCAR 下，当 $J=11$，$p=0.1$ 时，优化后的 $k=5$。

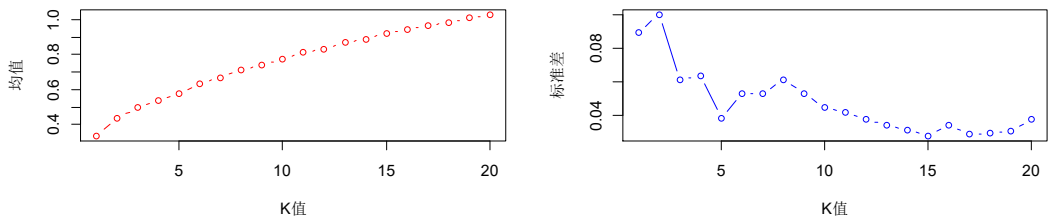

图 3-3-3：p=0.2, j=11 时不同 k 值下 100 次 MAE 结果的均值与标准差

根据图 3-3-3 实验结果，仍可以得到与图 3-3-2 中类似的实验结论；此外，当 $k \ge 8$ 时，MAE_SD_{CV} 结果已经出现明显降低的情形，并且下降的趋势有所放缓，这一点加固了图 3-3-2 得到的实验结果。综上，当 $k=5$ 时，$\overline{MAE_{CV}}$ 结果相对较小，而 MAE_SD_{CV} 结果也相对优于其他情况，在完全随机缺失下，当 $J=11$，$p=0.2$ 时，优化后的 $k=5$。

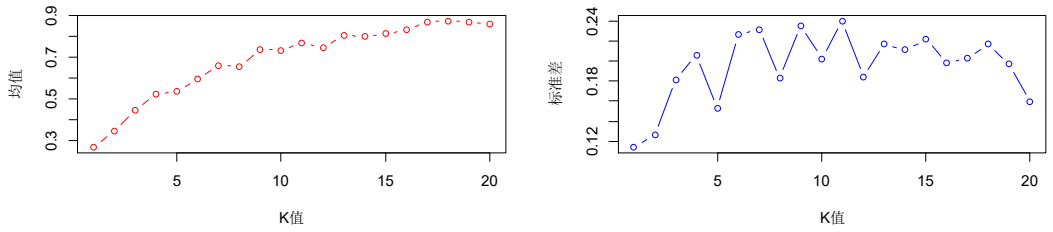

图 3-3-4：p=0.1, j=9 时不同 k 值下 100 次 MAE 结果的均值与标准差

根据图 3-3-4 实验结果可得，随着 k 的不断增大，$\overline{MAE_{CV}}$ 结果仍逐步变大；当 $1 \le k \le 4$ 时，MAE_SD_{CV} 结果逐步增大；当 k 分别取 4 和 5 时，所得到的 $\overline{MAE_{CV}}$ 结果几乎相等，但 MAE_SD_{CV} 结果相差很大，这印证了 $k=5$ 时，算法仍然有效，且稳定性相对优良；从 $k=4$ 开始，MAE_SD_{CV} 结果呈现出震荡状态，$k=18$ 时出现 MAE_SD_{CV} 结果逐步减小的趋势。上述结果基本符合 KNN 的特征。综上，当 $J=9$，$p=0.1$ 时，若选择 $k=1$ 会得到最小的 $\overline{MAE_{CV}}$ 和 MAE_SD_{CV} 结果，然而这与 Boston 数据集中第 9 列数据的特殊性是分不开的；若选择 $k=5$，则会得到相对较小的 MAE_SD_{CV} 结果，但此时的 k 会令算法的填补效果变差。

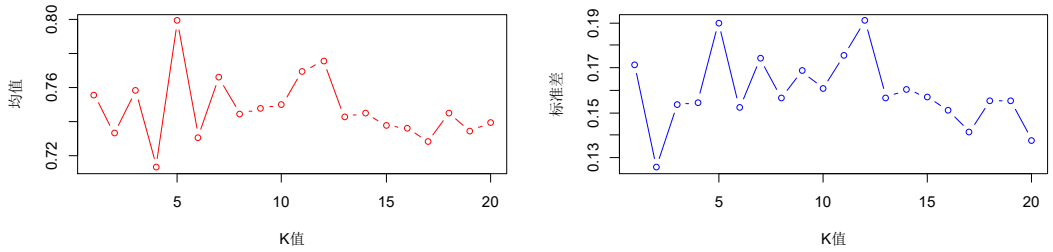

图 3-3-5：p=0.1, J=(9, 11)时不同 k 值下 100 次 MAE 结果的均值与标准差

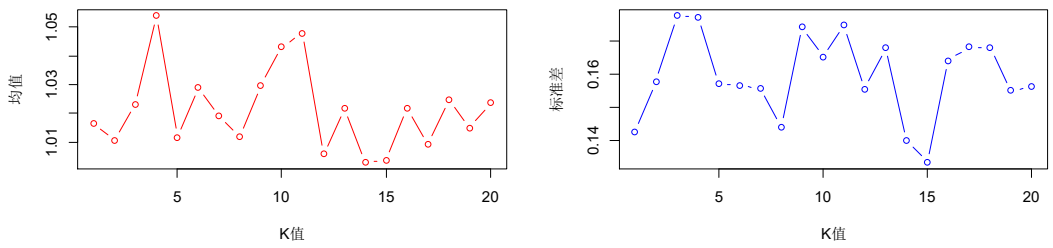

图 3-3-6：p=0.1, J=(3, 9, 11)时不同 k 值下 100 次 MAE 结果的均值与标准差

基于多变量缺失情形，当 p=0.1, J=(9, 11)时，根据图 3-3-5 实验结果可得，随着 k 值的不断增大，$\overline{MAE_{CV}}$ 结果也逐步变大；当 $2 \le k \le 5$ 时，MAE_SD_{CV} 结果

呈现逐步上升趋势，算法稳定性随 k 值的增大而降低；当$k>7$时，MAE_SD_{CV}结果开始呈现震荡状态，算法稳定性与k值的变化趋势呈现出不一致性；当$k=4$时，模型结果相对最好，且稳定性相对较佳。当 p=0.1, J=(3,9,11)时，此时会得到图 3-3-6 所示的实验结果。结果显示，该种情况下$k=15$会得到最小的MAE_SD_{CV}结果以及近乎最小的$\overline{MAE_{CV}}$结果；整体来看，两个趋势图并无明显的规律特征。综上，在缺失列 J=(3,9,11)的前提下，KNN 中$k=15$时效果达到最佳。

综上，使用交叉验证法对k进行优化，并不能得到一个固定的最优解，优化后的k会因为具体的样本集、缺失变量、缺失率而发生改变。综合上述实验结果结果，在后续章节的实证分析中，本文选择$k=5$。

3.3.2 加权 k 近邻

KNNW（K-Nearest Neighbor of Weighted，简称 KNNW）是对 KNN 的一种改进，核心思想是为最邻近待分类样本点的近邻赋予更高权重，依此类推。

KNNW 首先需要构造权重函数，无论权重函数的形式如何，依据权重函数为每个临近点求得的对应的权重值$w_l, l\in[1,k]$均应满足$\sum_{l=1}^{k} w_l=1$的前提。权重函数的构造是否符合实际问题的需要，将会直接影响算法的填补效果。

本章使用高斯函数为基础进行加权处理，高斯函数形式如下：

$$f(x)=ae^{-(x-b)^2/2\alpha^2} \tag{3.2}$$

式(3.2)中，a,b,α为实数常数，且$a>0$。基于此进行权重函数的构造，如下：

$$w_l=\frac{ae^{-(D_l-b)^2/2\alpha^2}}{\sum_{l=1}^{k} ae^{-(D_l-b)^2/2\alpha^2}}, l\in[1,k] \tag{3.3}$$

式(3.3)中，a,b,α为实数常数，且$a>0$，D_l为第l个近邻与待分类样本点之间的距离，因此$D_l\geq 0$，k为选取的近邻数量。上式中，需要为权重函数指定合理的参数值，即a,b,α的具体值。式(3.3)经过简单化简后可得：

$$w_l=\frac{e^{-(D_l-b)^2/2\alpha^2}}{\sum_{l=1}^{k} e^{-(D_l-b)^2/2\alpha^2}}, l\in[1,k] \tag{3.4}$$

式(3.4)中，参数 b 代表权重函数在水平方向上的偏移量，参数 α 代表函数的峰值以及 w_l 随 D_l 值的增大而衰减快慢程度。在相同的数据前提下，α 值越大，函数的峰值越小，且随 D_l 值的增大，w_l 衰减速度较缓；α 值越小，函数的峰值越大，且随 D_l 值的增大，w_l 衰减速度较快。下面采用计算机模拟的方式对该权重函数进行图形绘制，如图 3-3-7：

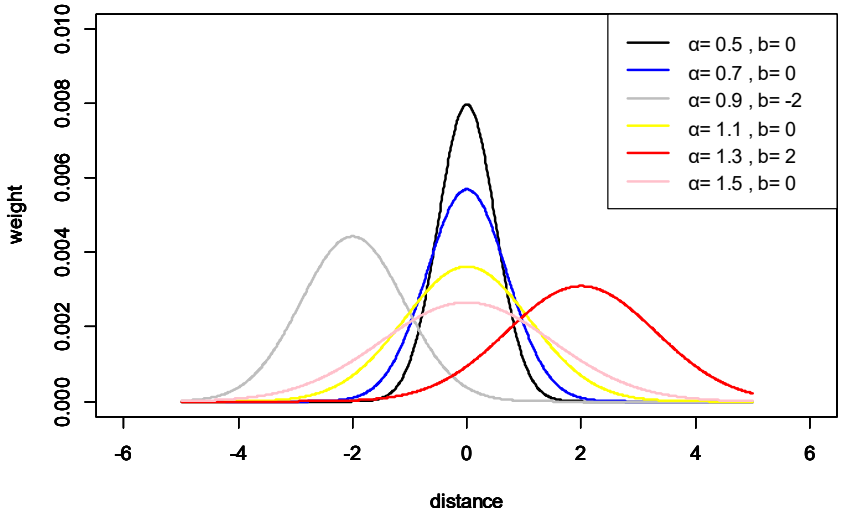

图 3-3-7：基于高斯函数的权重函数图像

图 3-3-7 印证了前文所述的结论。然而在 KNNW 中，依据不同的距离 D_l，使用公式(3.4)计算的出相应的 w_l 才是算法要解决的核心问题，而偏移量 b 的大小不会影响最终的权重结果，因此本文将 b 置为 0，公式(3.4)进一步化简为：

$$w_l = \frac{e^{-(D_l)^2/2\alpha^2}}{\sum_{l=1}^{k} e^{-(D_l)^2/2\alpha^2}} \tag{3.5}$$

式(3.5)中，在参数 α 确定的情况下，不同的 D_l 会对函数所求的 w_l 产生很大的影响。在 KNNW 中，加权函数需要为每一个近邻依据其对应的 D_l 分配合适的权重，且需要避免为最近邻的点分配无限接近于 1 的权重值，或为相对较远的一个或几个临近点分配无限接近于 0 的权重值。

在实际的填补问题中，基于不同的数据集，通过式（3.1）求解得到的 D_l 值差别通常会很大，即便是相同的数据集，为不同样本点求解得到的 D_l 值差别依然不

容忽视。若采用固定参数α的方式对实际问题进行求解，必然会出现在 KNN 中k选取不合理所引发的类似问题。因此，本文提出不固定参数α，依据具体问题进行动态调节参数α的方式进行实际问题求解。

3.3.3 动态调参

对于每一个待填补的样本点，使用 K 近邻算法为其找到k个近邻$M_l^{obs}, l \in [1,k]$，且每个近邻对应的距离为$D_l, D_1 \le D_2 \le \cdots \le D_k$。通过公式(3.5)计算对应的权重值$w_l, l \in [1,k]$，权重值$w_l$呈现出$w_1 \ge w_2 \ge \cdots \ge w_k$的规律。详细的动态调节参数$\delta$如下：

对同一待分类样本的所选出的第 1 个近邻点的权重w_1和第k个近邻点的权重w_k做如下限定，令其满足如下不等式：

$$\frac{w_1}{w_k} \ge \xi \tag{3.6}$$

将公式(3.5)待入不等式(3.6)可得：

$$\frac{e^{-(D_1)^2/2\alpha^2}}{\sum_{l=1}^{k} e^{-(D_l)^2/2\alpha^2}} \Big/ \frac{e^{-(D_k)^2/2\alpha^2}}{\sum_{l=1}^{k} e^{-(D_l)^2/2\alpha^2}} \ge \xi, l \in [1,k] \tag{3.7}$$

对式(3.7)进行简单化简如下：

$$\frac{e^{-(D_1)^2/2\alpha^2}}{e^{-(D_k)^2/2\alpha^2}} \ge \xi, l \in [1,k] \tag{3.8}$$

对式(3.8)中的参数α进行求解并整理可得如下不等式：

$$\alpha \le \sqrt{\frac{(D_k)^2 - (D_1)^2}{2\ln \xi}} \tag{3.9}$$

利用不等式(3.9)，通过控制ξ值的大小来对加权函数的参数α进行动态调整，以便适应具体问题中因观测值不同所引起的权重配置不合理的情况。不等式(3.7)中，ξ值越大，表明第 1 个近邻点在对缺失值进行填补时所占比重越大；反之ξ值越小表明第 1 个近邻点在对缺失值进行填补时所占比重相对越小，且w_1和w_k的值满足如下不等式：

$$\frac{1}{k}<w_1<1,0<w_k<\frac{1}{k}$$

在真实的问题求解过程中，当w_1值越接近于 1，此时的 KNNW 填补效果越接近于$k=1$时的 KNN 填补法；当w_k值越接近于$\frac{1}{k}$，此时的 KNNW 填补效果越接近于 KNN。因此，为选定的k个临近点分配合适的权重组合是整个动态调参的核心问题，即w_1值不要过大，w_k值不要过小。

本章基于 Boston 数据集，令缺失率$p=0.1$，缺失列$J=11$，采用式(3.1)为含缺失值的任意样本点寻找$k=5$个近邻，采用动态调参法不同α值下的权重分布情况进行比较分析，结果如图 3-3-8：

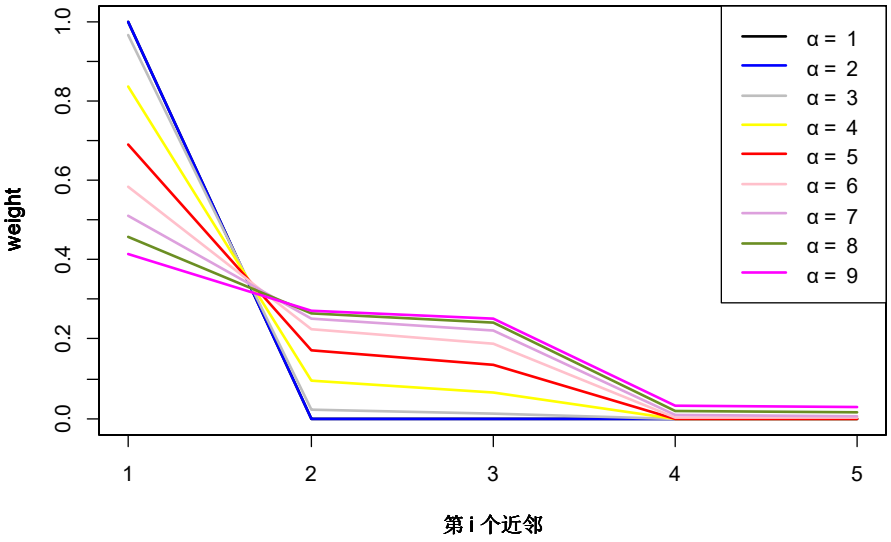

图 3-3-8：不同α值下的k个近邻点权重分布图

图 3-3-8 实验结果显示，不同α值对同一组数据的权重分配情况大不相同。令$\xi=1000$，对α值进行赋初值，通过迭代的方式寻找令$w_k\geq 0.1,0.01,0.001,0.0001$、$\xi=1.1$时的$\alpha$值点，并将上述$\alpha$值作为权重函数参数，进而采用 KNN、KNNW 进行数据填补，考虑到计算机模拟的随机性，将不同参数下的程序随机执行 100 次，并在 MAE、RMSE、MAPE 下对实验结果进行对比分析，结果如表 3-3-1 所示：

表 3-3-1 的实验结果显示，第一，当$\xi=1.1$时，KNNW 和 KNN 在 MAE、RMSE、MAPE 评价准则下填补效果近乎一致，且 KNNW 的$\overline{MAE}$、$\overline{RMSE}$、$\overline{MAPE}$值略微小于 KNN，这表明，在w_1和w_k的比值逐步接近于 1 时，KNNW 将逐步退化为 KNN；第

二，基于当前的实验环境即数据，随着 w_k 值的减小，KNNW 在三种评价准则下的值也逐步减小，并在 $w_k = 0.001$ 时达到最佳。

表 3-2-1：不同的权重配比下 2 种填补算法在三种评价准则下填补 100 次误差结果的均值

评价准则	填补方法	ξ	w_k			
		1.1	0.1	0.01	0.001	0.0001
$\overline{MAE}$	KNN	0.561	0.569	0.557	0.551	0.555
	KNNW	0.554	0.468	0.346	0.285	0.295
$\overline{RMSE}$	KNN	1.031	1.040	1.055	1.035	1.035
	KNNW	1.021	0.905	0.852	0.796	0.821
$\overline{MAPE}$	KNN	0.033	0.033	0.032	0.032	0.032
	KNNW	0.032	0.027	0.020	0.016	0.017

针对 w_k 取值不同所带来的算法稳定性问题从表 3-3-2 的实验结果可以看出：

表 3-3-2：不同的权重配比下 2 种填补算法在三种评价准则下填补 100 次误差结果的方差

评价准则	填补方法	ξ	w_k			
		1.1	0.1	0.01	0.001	0.0001
MAE_SD	KNN	0.133	0.123	0.127	0.117	0.115
	KNNW	0.131	0.112	0.113	0.118	0.103
RMSE_SD	KNN	0.225	0.231	0.228	0.224	0.196
	KNNW	0.225	0.237	0.261	0.290	0.238
MAPE_SD	KNN	0.008	0.007	0.008	0.007	0.007
	KNNW	0.008	0.007	0.007	0.007	0.006

表 3-3-2 的实验结果显示，第一，在 w_1 和 w_k 的比值逐步接近于 1 时，KNNW 与 KNN 结果接近一致，这在此印证了表 3-3-1 的实验结果；第二，在 $w_k = 0.01$ 时，基于 MAE、MAPE 两个评价准则下的 KNNW 稳定性明显高于 KNN；第三，在 $w_k = 0.0001$ 时，在 MAE、MAPE 两个评价准则下均获得了最小的标准差值，但结合表 1 的实验结果，在 $w_k = 0.0001$ 时，三种评价准则的结果并非最小；第四，依据在 RMSE 准

则下获取的最小值来看，当$\xi=1.1$时，由于 KNNW 最接近于 KNN，因此也获得了最小的 RMSE_SD 值，这在一定程度上反映出加权 KNN 算法的因权重配置是否合理所带来的问题。综上，在后续实证分析环节，本文 KNN、KNNW 的参数设定为$k=5$，每轮程序运行过程中通过$\xi=1000$来对α赋初值，通过$w_k=0.01$寻求α的最优解。

3.4 基于单变量缺失的实证分析

在 MCAR 机制下，本文采用的具体方法步骤是将 Boston 数据集中的 506 个样本进行编号，依据事先设定好的缺失率在 R 平台上生成对应数量的随机数n_{mis}，$n_{mis}\in[1,506]$，然后将该编号对应样本点中“城镇师生比例”指标列 ptratio 中对应的观测值置为“NA”即可。然而在 MAR 和 NMAR 机制下，计算机模拟数据缺失的过程略有不同。以 MAR 为例，随机缺失代表观测值是否缺失与该样本中已观测到的数据有关，因此本文挑选“业主自住住宅的价值中位数”指标列 medv 作为 ptratio 中的对应的观测值是否缺失的关联变量，而缺失列仍为“城镇师生比例”。

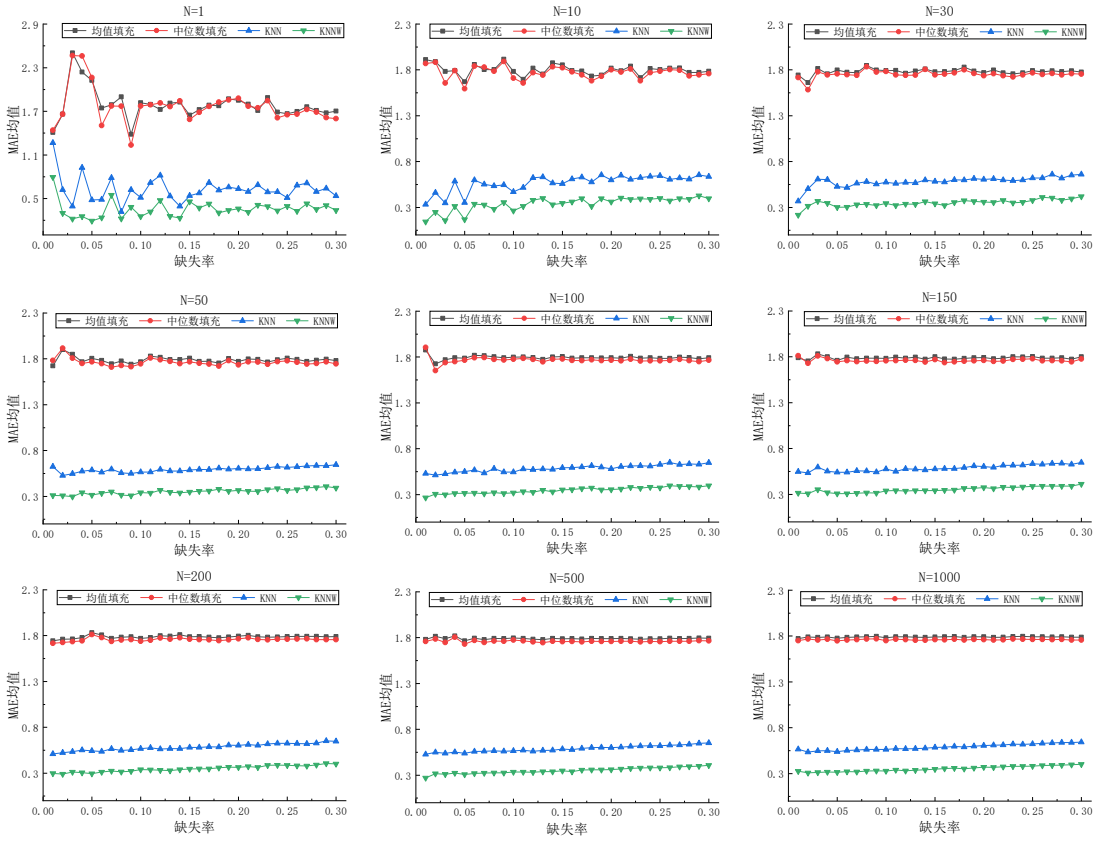

图 3-4-1：四种算法在不同缺失率前提下执行不同次数的 MAE 均值结果折线图（单变量）

本文将指标列 medv 中的观测值进行升序排列，依照缺失比例的不同对其标注分位

点，进而将数据划分为若干段，再通过随机数生成法进行数据段的随机选择，遍历当前数据段中的所有值，记录该值在原始数据集中的行号，最后将该行号对应样本点“城镇师生比例”指标列中的数值置为“NA”即可。模拟NMAR机制的情况与MAR类似，只需将随机挑选的不含缺失值本身的指标列变更为含缺失值指标列自身即可，其余步骤不变。为了验证算法的有效性，基于三种缺失机制，在MAE、RMSE、MAPE评价准则下对KNNW、MF、KNNWMF算法进行评估，考虑到计算机模拟具有随机性，将实验重复执行若干次，然后使用MAE、RMSE、MAPE结果的均值、方差、置信区间等数字特征进行对比分析。首先需要对不同缺失率前提下，程序执行不同次数所得结果的一致性进行评价。设定缺失率p由1%逐步递增至30%，步长设定为1%，程序执行次数N分别取1、10、30、50、100、200、500、1000次，并在完全随机缺失机制下求取对应的$\overline{MAE}$值，实验结果如图3-4-1。

根据图3-4-1的实验结果，在MCAR前提下，KNN、KNNW在MAE准则下均显著优于均值填补和中位数填补；此外，由于计算机模拟非完整数据集的随机性特点，当$N=1$时，四种填补方法在三种评价准则下的结果均呈现出震荡状态，而随着N的增大，基于三种评价结果所得的均值曲线逐步变得平滑，当$N\geq 100$时基本处于稳定状态；当缺失率$p<0.05$时，此时由于缺失率很小，导致每次模拟得到的非完整数据集中的缺失值部分几乎不会产生重叠现象，因此会出现图中当缺失率很小时，N取值不同而导致的结果不完全一致的现象发生，然而随着N取值的不断增大，该现象会逐渐消失。特别的，在RMSE、MAPE准则下具有相同的实验结果，且在MAR、NMAR中，基于不同的评价准则仍有类似图3-4-1的实验结论，此处不再赘述。综上，在本章的实证分析环节，取$p=0.05,0.1,0.15,0.2,0.25,0.3$，$N=100$。

3.4.1 完全随机缺失

首先考虑完全随机缺失下单变量缺失的情形，令缺失列$J=11$，然后在每一个缺失率前提下分别采用均值填补、中位数填补、KNN、KNNW对随机构造的含有缺失值的数据集进行填补，然后将上述实验重复执行100次，进而求其对应的$\overline{MAE}$、$\overline{RMSE}$、$\overline{MAPE}$值，详情如表3-4-1所示：

根据表3-4-1的实验结果，第一，基于不同的缺失率情况，由$\overline{MAE}$、$\overline{RMSE}$、

$\overline{MAPE}$ 结果来看，KNN、KNNW 明显小于均值填补、中位数填补，且均值填补、中

表 3-3-1：不同缺失率下、4 种填补算法在 MCAR 前提下填补 100 次误差结果的均值(单变量)

评价准则	填补算法	缺失率					
		5%	10%	15%	20%	25%	30%
MAE	均值	1.784	1.800	1.803	1.773	1.783	1.789
	中位数	1.755	1.773	1.772	1.739	1.752	1.761
	KNN	0.531	0.572	0.597	0.605	0.619	0.656
	KNNW	0.304	0.343	0.354	0.362	0.384	0.406
RMSE	均值	2.149	2.171	2.174	2.149	2.148	2.163
	中位数	2.230	2.247	2.249	2.218	2.216	2.233
	KNN	0.994	1.075	1.095	1.104	1.130	1.168
	KNNW	0.715	0.833	0.872	0.871	0.932	0.967
MAPE	均值	0.104	0.104	0.105	0.103	0.103	0.103
	中位数	0.105	0.106	0.106	0.104	0.104	0.105
	KNN	0.031	0.033	0.035	0.035	0.036	0.038
	KNNW	0.017	0.020	0.021	0.021	0.022	0.023

位数填补的结果随缺失率的增大没有呈现出规律特征，这表明传统的填补方法并不能高度拟合数据填补问题中缺失值的真实情况，对于缺失值填补问题并不能有效解决，而 KNN、KNNW 随着缺失率的增大呈现出的结果递增缺失恰恰反映了基于机器学习算法而来的填补方法对缺失值填补问题的适用性；第二，随着缺失率的递增，KNN、KNNW 的 $\overline{MAE}$ 、 $\overline{RMSE}$ 、 $\overline{MAPE}$ 结果均值出现了递增趋势，这表明随着缺失率的增大，算法将面临失效，从而变得不适用；第三，不同缺失率下，KNNW 的 $\overline{MAE}$ 、 $\overline{RMSE}$ 、 $\overline{MAPE}$ 结果明显低于 KNN，特别的，针对于同一缺失数据集，使用 KNNW 填补缺失率 $p=0.3$ 时的 $\overline{MAE}$ 、 $\overline{RMSE}$ 、 $\overline{MAPE}$ 结果依然低于使用 KNN 在缺失率 $p=0.05$ 时所填补后的 $\overline{MAE}$ 、 $\overline{RMSE}$ 、 $\overline{MAPE}$ ，甚至在 MAE、MAPE 两种评价准则下，出现了显著的填补优势。

结合上述结论发现经过 k 值优化、动态调参后的 KNNW 在填补效果和稳定性方

面均有显著提升，此处，只针对 KNN、KNNW 在不同缺失率下填补的每一次 MAE、RMSE、MAPE 结果进行统计，以便对不同算法填补结果的整体分布情况有所了解，现对 KNN、KNNW 绘制 100 次填补结果对应的 MAE、RMSE、MAPE 箱线图，结果如图 3-4-2：

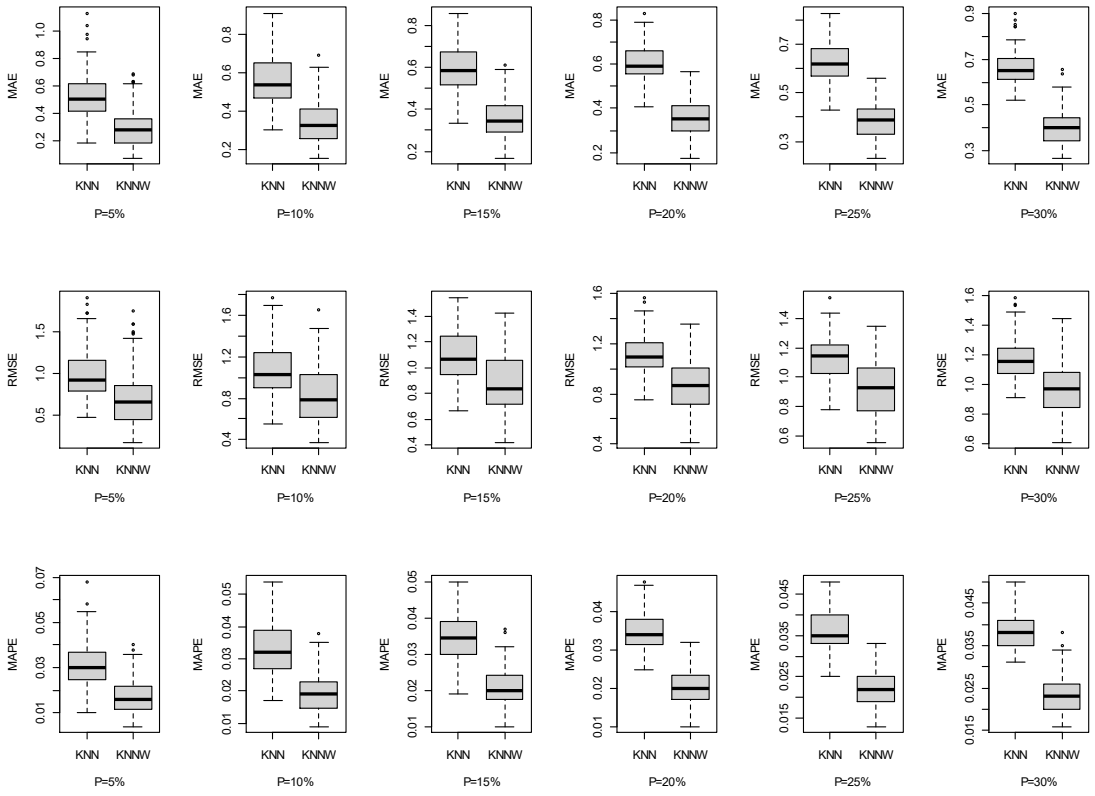

图 3-4-2：不同缺失率下、KNN 与 KNNW 在 MCAR 前提下填补 100 次的误差结果箱线图（单变量）

图 3-4-2 的实验结果表明，第一，基于不同的评价准则和缺失率，KNN、KNNW 对应的 100 次填补评价结果大多出现异常点，这点再次表明了使用单次评价结果对算法的优良性进行评估是不合理的；第二，在 MCAR 下，基于不同的缺失率和评价准则，KNNW 的评价结果最大值、最小值、四分位数、中位数均小于 KNN，这表明 KNNW 的有效性；第三，MAE、MAPE 评价准则下，基于不同缺失率前提，KNNW 填补结果的四分间距接近或小于 KNN，这进一步加强了第二点实验结果，然而在 RMSE 评价准则下，基于不同的缺失率，KNNW 的四分间距均大于 KNN，这代表经改进后的 KNNW 在提高精度的同时，算法填补的稳定性出现下降。

此外，为了验证算法填补的总体效果，本文构造了四种填补算法在不同缺失率前提下，基于不同评价准则的 95%置信区间，实验结果如表 3-4-2 所示：

根据表 3-4-2 的实验结果显示，第一，不同缺失率下，四种算法基于多次 MAE、

表 3-4-2：不同缺失率下、4 种填补算法在 MCAR 前提下填补 100 次误差结果的置信区间及其长度(单变量)

评价准则	填补算法	结果类型	缺失率					
			5%	10%	15%	20%	25%	30%
MAE	均值	区间	[1.740,1.827]	[1.767,1.833]	[1.779,1.828]	[1.752,1.794]	[1.765,1.801]	[1.774,1.805]
		长度	0.086	0.066	0.049	0.042	0.037	0.031
	中位数	区间	[1.704,1.807]	[1.736,1.810]	[1.744,1.799]	[1.714,1.763]	[1.731,1.773]	[1.743,1.779]
		长度	0.103	0.074	0.055	0.049	0.042	0.036
	KNN	区间	[0.498,0.565]	[0.544,0.600]	[0.576,0.618]	[0.589,0.621]	[0.603,0.635]	[0.641,0.671]
		长度	0.067	0.056	0.043	0.032	0.032	0.030
	KNNW	区间	[0.275,0.333]	[0.319,0.366]	[0.336,0.372]	[0.346,0.378]	[0.369,0.399]	[0.391,0.421]
		长度	0.058	0.047	0.036	0.032	0.030	0.030
RMSE	均值	区间	[2.098,2.200]	[2.132,2.210]	[2.145,2.203]	[2.125,2.172]	[2.127,2.170]	[2.144,2.181]
		长度	0.102	0.077	0.058	0.047	0.043	0.037
	中位数	区间	[2.165,2.294]	[2.199,2.295]	[2.212,2.287]	[2.187,2.248]	[2.188,2.244]	[2.209,2.258]
		长度	0.129	0.096	0.075	0.062	0.056	0.049
	KNN	区间	[0.932,1.056]	[1.027,1.123]	[1.059,1.132]	[1.075,1.134]	[1.102,1.158]	[1.141,1.195]
		长度	0.124	0.097	0.074	0.059	0.056	0.054
	KNNW	区间	[0.644,0.785]	[0.779,0.886]	[0.828,0.916]	[0.833,0.909]	[0.897,0.967]	[0.933,1.001]
		长度	0.141	0.107	0.089	0.076	0.071	0.068
MAPE	均值	区间	[0.100,0.107]	[0.102,0.107]	[0.103,0.106]	[0.101,0.104]	[0.102,0.104]	[0.102,0.105]
		长度	0.006	0.005	0.004	0.003	0.003	0.002
	中位数	区间	[0.101,0.109]	[0.103,0.109]	[0.104,0.108]	[0.102,0.106]	[0.103,0.106]	[0.104,0.106]
		长度	0.008	0.006	0.004	0.004	0.003	0.003
	KNN	区间	[0.029,0.033]	[0.032,0.035]	[0.034,0.036]	[0.034,0.036]	[0.035,0.037]	[0.037,0.039]
		长度	0.004	0.003	0.003	0.002	0.002	0.002
	KNNW	区间	[0.016,0.019]	[0.019,0.021]	[0.020,0.022]	[0.020,0.022]	[0.021,0.023]	[0.023,0.024]
		长度	0.003	0.003	0.002	0.002	0.002	0.002

RMSE、MAPE 评价结果所构造的置信区间长度会随着缺失率的增大而变小，单独分析其对应的置信区间起点或终点来看，KNN、KNNW 会随着缺失率的增大呈现出规

律的递增趋势，而均值填补和中位数填补也呈现出一定的增大趋势，这表明随着缺失率的增大，四种填补算法的填补效果均在逐步下降，且由于缺失率增大的缘故，填补过程中所参照的完整数据集规模会逐步减小，这导致填补的不确定性降低，填补精度反而出现一定的提升；第二，在三种评价准则下，基于相同的缺失率前提，KNN、KNNW 的置信区间长度、对应的区间起点和终点均显著小于均值填补和中位数填补，而 KNNW 在这三个方面更是显著优于 KNN，这表明，经过改进后 KNNW 在填补效果、精度上均有了显著提升。

3.4.2 随机缺失

仅仅考虑完全随机缺失的情形并不能验证算法在不同缺失机制下的稳健程度，在该部分考虑单变量随机缺失情形，仍将缺失率 $p=0.05,0.1,0.15,0.2,0.25,0.3$，缺失列 $J=11$，采用与单变量完全随机缺失情形中相同的实验过程的评价标准继续对四种算法进行评估，结果如表 3-4-3 所示：

表 3-4-3：不同缺失率下、4 种填补算法在 MAR 前提下填补 100 次误差结果的均值(单变量)

评价准则	填补算法	缺失率					
		5%	10%	15%	20%	25%	30%
MAE	均值	1.843	1.883	1.765	1.980	1.955	1.857
	中位数	1.908	1.914	1.729	2.047	2.035	2.008
	KNN	0.675	0.665	0.665	0.847	0.944	0.858
	KNNW	0.372	0.386	0.377	0.526	0.629	0.545
RMSE	均值	2.209	2.209	2.069	2.304	2.294	2.165
	中位数	2.305	2.276	2.077	2.406	2.411	2.355
	KNN	1.095	1.097	1.120	1.326	1.483	1.381
	KNNW	0.788	0.828	0.816	1.014	1.248	1.125
MAPE	均值	0.109	0.109	0.098	0.114	0.112	0.103
	中位数	0.116	0.113	0.099	0.121	0.119	0.115
	KNN	0.040	0.039	0.038	0.052	0.056	0.048
	KNNW	0.022	0.023	0.021	0.032	0.037	0.030

根据表 3-4-3 的实验结果可以看出，第一，相较于表 3-4-1 的实验结果，在 MAE、RMSE、MAPE 评价准则下，基于单变量随机缺失这一情形，四种算法在不同

缺失率前提下得到的 $\overline{MAE}$、$\overline{RMSE}$、$\overline{MAPE}$ 结果均略大于与之对应的单变量完全随机缺失所得到的结果，这也表明当缺失机制为随机缺失时，算法受到了一定的影响，但观察其对应的填补结果发现，四种填补法仍适用于该类缺失情形；第二，在单变量随机缺失情形下，KNN、KNNW 基于不同的缺失率前提所得到的 $\overline{MAE}$、$\overline{RMSE}$、$\overline{MAPE}$ 结果依然远小于均值和中位数填补，结论与表 3-4-1 结论一致；第三，针对相同的缺失率前提，KNNW 的评价结果总是显著优于 KNN，且这种优势比单变量完全随机缺失情形下的填补效果更加明显；第四，随着缺失率的增大，KNN、KNNW 的 $\overline{MAE}$、$\overline{RMSE}$、$\overline{MAPE}$ 结果呈现出整体变大的趋势，这也表明算法的填补效果会随着缺失率的增大而面临失效的问题依然没有改变。

同样的，需要对单变量随机缺失情形下填补算法的整体效果进行评估，仍然采用多次程序运行结果来计算出与之对应的 95%置信区间，并绘制相应的箱线图，以此分析四种算法的填补精度和稳健性，实验结果如表 3-4-4、图 3-4-3 所示。

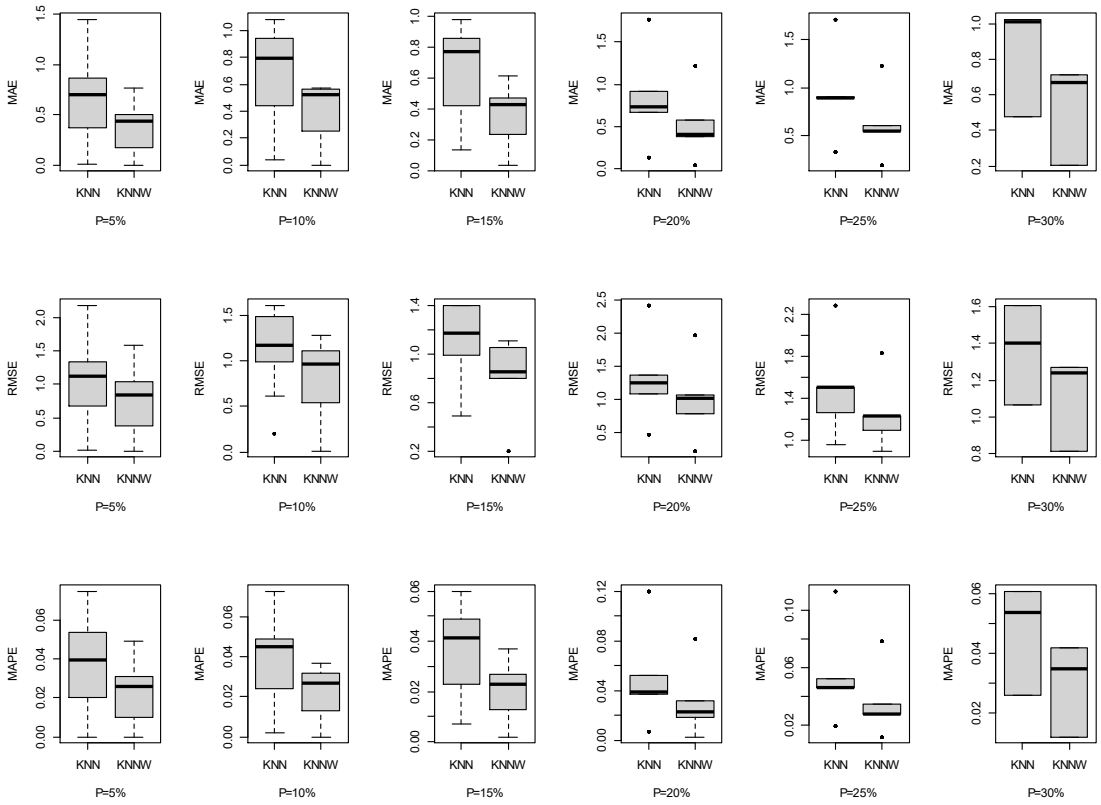

图 3-4-3：不同缺失率下、KNN 与 KNNW 在 MAR 前提下填补 100 次的误差结果箱线图（单变量）

根据图 3-4-3 的实验结果显示，第一，在 MAR 前提下，基于不同的缺失率前提，KNNW 在不同评价准则下的结果最大值、最小值、四分位数、中位数均小于 KNN，

表 3-4-4：不同缺失率下、4 种填补算法在 MAR 前提下填补 100 次误差结果的置信区间及其长度(单变量)

评价准则	填补算法	结果类型	缺失率					
			5%	10%	15%	20%	25%	30%
MAE	均值	区间	[1.748,1.938]	[1.806,1.960]	[1.709,1.821]	[1.887,2.073]	[1.881,2.029]	[1.792,1.922]
		长度	0.190	0.154	0.112	0.186	0.148	0.130
	中位数	区间	[1.781,2.034]	[1.814,2.015]	[1.676,1.781]	[1.911,2.184]	[1.916,2.153]	[1.908,2.107]
		长度	0.253	0.201	0.105	0.273	0.237	0.199
	KNN	区间	[0.607,0.744]	[0.597,0.733]	[0.608,0.723]	[0.737,0.956]	[0.851,1.036]	[0.808,0.907]
		长度	0.137	0.136	0.115	0.219	0.185	0.099
	KNNW	区间	[0.329,0.416]	[0.344,0.428]	[0.340,0.414]	[0.447,0.606]	[0.560,0.698]	[0.501,0.589]
		长度	0.087	0.084	0.074	0.160	0.139	0.088
RMSE	均值	区间	[2.101,2.317]	[2.117,2.301]	[2.022,2.115]	[2.197,2.411]	[2.211,2.378]	[2.116,2.215]
		长度	0.216	0.184	0.093	0.214	0.167	0.099
	中位数	区间	[2.163,2.448]	[2.153,2.399]	[2.011,2.144]	[2.254,2.559]	[2.272,2.550]	[2.233,2.478]
		长度	0.285	0.245	0.133	0.305	0.278	0.245
	KNN	区间	[1.000,1.191]	[1.009,1.185]	[1.058,1.183]	[1.196,1.456]	[1.391,1.576]	[1.337,1.425]
		长度	0.191	0.176	0.125	0.260	0.185	0.089
	KNNW	区间	[0.703,0.874]	[0.746,0.911]	[0.757,0.876]	[0.898,1.130]	[1.183,1.312]	[1.085,1.165]
		长度	0.171	0.165	0.119	0.233	0.130	0.080
MAPE	均值	区间	[0.102,0.117]	[0.103,0.115]	[0.095,0.101]	[0.107,0.122]	[0.106,0.117]	[0.100,0.106]
		长度	0.015	0.012	0.006	0.014	0.011	0.006
	中位数	区间	[0.106,0.126]	[0.105,0.122]	[0.095,0.103]	[0.111,0.132]	[0.110,0.128]	[0.108,0.122]
		长度	0.020	0.016	0.008	0.021	0.018	0.014
	KNN	区间	[0.036,0.044]	[0.035,0.044]	[0.034,0.041]	[0.044,0.059]	[0.049,0.062]	[0.045,0.051]
		长度	0.008	0.009	0.007	0.015	0.013	0.006
	KNNW	区间	[0.019,0.025]	[0.020,0.025]	[0.019,0.023]	[0.027,0.038]	[0.032,0.042]	[0.028,0.033]
		长度	0.005	0.005	0.004	0.011	0.009	0.005

这表明 KNNW 在该缺失机制下仍然具有显著优势；第二，当缺失率 $p \le 0.15$ 时，在

MAE、MAPE准则下，KNNW算法的四分间距小于或接近KNN，而在RMSE准则下正好出现相反现象；第三，当$p \geq 0.2$时，两种算法的四分间距没有明显规律特征，此时，基于不同评价准则所得到的多次实验结果出现最大值与上四分位点重合、最小值与下四分位点重合、中位数与上四分位点与下四分位点重合的现象时有发生，且每次实验结果的值相较于缺失率较低的情况均出现增大的现象，此外当$p = 0.3$时，两种算法基于评价准则结果所得四分间距明显高于其他缺失率前提下的对应结果，这表明，使用本文提供的模拟方法营造随机缺失机制，在缺失率较大的情况下会多次出现非完整数据集一致的情形，这会导致KNNW、KNN多次填补所得的评价结果中存在大量相等现象。

根据表3-4-4的实验结果可以看出，第一，基于不同的评价准则和缺失率，使用KNN、KNNW所得到的置信区间长度始终小于均值和中位数填补，且对应的置信区间起点和终点远小于均值填补和中位数填补所得结果；第二，不同缺失率前提下，KNNW基于三种评价准则的置信区间长度和对应的起点、终点值均小于KNN，这表明经过k值优化和动态调参后的KNN在单变量随机缺失情形下依然能够保证算法的填补精度和准确性；第三，随着缺失率的增大，四种算法所得置信区间的起点和终点具有整体右移的趋势，而区间长度没有明显的递增或递减现象，这依然表明，随着缺失的增大，填补算法面临失效的问题，与单变量完全随机缺失结果近似，不再赘述；第四，对比分析表3-4-2的实验结果可以看出，在单变量随机缺失前提下，对任意缺失率而言，四种算法依据MAE、RMSE、MAPE评价结果所得的95%置信区间，其对应的起点和终点均出现右移现象，这表明，在本文所模拟的随机缺失机制下，数据填补算法的有效性和精度均受到了一定的影响。

3.4.3 非随机缺失

基于相同的评价准则和实验方法对NMAR下的单变量数据缺失情形进行模拟，实验结果如表3-4-5所示：

根据表3-4-5的实验结果，第一，相较于单变量完全随机缺失和单变量随机缺失两种情形，基于MAE、RMSE、MAPE评价准则进行评估时，四种填补算法效果在任何缺失率前提下均出现不同程度的下降趋势，其中，KNN、KNNW填补结果偏差尤为明显，这是由于在非随机缺失情形下算法不再适用所导致的结果，然而，对比表3-4-1、表3-4-3可以发现，处于NMAR下的KNN、KNNW填补后的$\overline{MAE}$、$\overline{RMSE}$、

表 3-4-5：不同缺失率下、4 种填补算法在 NMAR 前提下填补 100 次误差结果的均值(单变量)

评价准则	填补算法	缺失率					
		5%	10%	15%	20%	25%	30%
MAE	均值	1.914	2.031	1.881	2.296	2.332	2.147
	中位数	2.013	2.164	2.114	2.402	2.693	2.658
	KNN	1.468	1.838	1.534	1.977	1.862	2.530
	KNNW	1.435	1.854	1.624	1.994	1.875	2.733
RMSE	均值	1.925	2.061	1.930	2.370	2.423	2.251
	中位数	2.023	2.188	2.152	2.463	2.764	2.754
	KNN	1.681	2.035	1.749	2.376	2.365	2.877
	KNNW	1.783	2.133	1.901	2.421	2.405	3.134
MAPE	均值	0.114	0.120	0.110	0.132	0.133	0.126
	中位数	0.122	0.130	0.125	0.140	0.155	0.156
	KNN	0.090	0.112	0.094	0.117	0.110	0.146
	KNNW	0.088	0.113	0.099	0.118	0.111	0.157

$\overline{MAPE}$ 结果依然优于 MCAR、MAR 下均值填补和中位数填补的评价结果均值，也更优于 NMAR 下这两种传统填补法的评价结果；第二，KNN、KNNW 在不同缺失率前提下的 $\overline{MAE}$ 、 $\overline{RMSE}$ 、 $\overline{MAPE}$ 结果不再具备明显的优劣关系，这表明在 NMAR 下，不仅 KNN 失效，对其采用 k 值优化和动态调参改进后的 KNNW 也面临失效。为了进一步验证 KNNW、KNN 在 NMAR 下的区别，仍然采用前文所述方法构造置信区间，并绘制相应的箱线图，结果如图 3-4-4、表 3-4-6。

根据图 3-4-4 的实验结果显示，第一，相较于前两种缺失机制，在 NMAR 下，基于不同的缺失率前提，KNNW 在不同评价准则下的结果最大值、最小值、四分位数、中位数均大于于 KNN，且这些值相较于 MCAR、MAR 下对应的值，几乎全部出现增大现象,而对应的四分间距之间也无明显规律,这表明在 NMAR 下,KNN 和 KNNW 不仅完全失效，且 KNNW 所具备的填补优势也完全消失；第二，当 $p \geq 0.2$ 时，实验结果与图 3-4-3 第三点结果类似,这表明,使用本文提供的模拟方法营造 NMAR，在缺失率较大的情况下会多次出现非完整数据集一致的情形,这会导致 KNNW、KNN

多次填补所得的评价结果中存在大量相等现象。

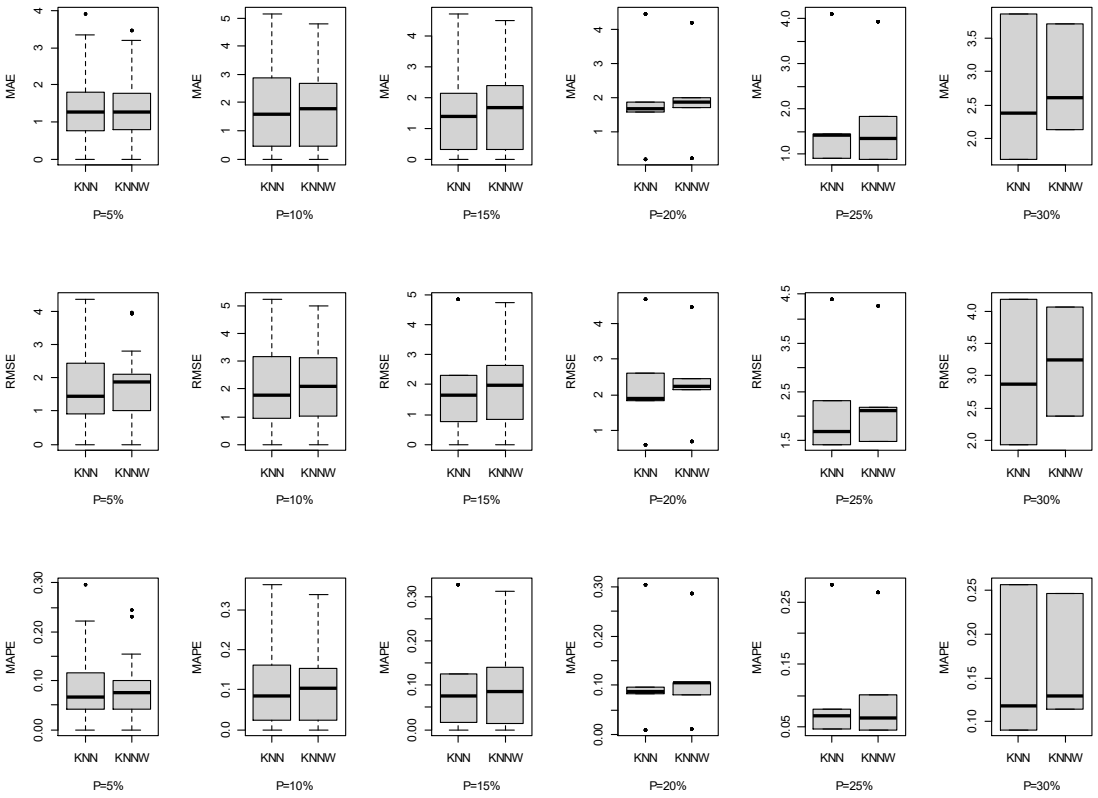

图 3-4-4：不同缺失率下、KNN 与 KNNW 在 NMAR 前提下填补 100 次的误差结果箱线图（单变量）

从表 3-4-6 的实验结果来看，第一，基于不同的评价准则和缺失率，在 NMAR 前提下，KNNW 在置信区间长度上始终小于 KNN，而对应的置信区间起点和终点并没有呈现出明显的规律；第二，随着缺失率的增大，四种算法基于不同评价准则求得的置信区间长度没有呈现出严格的递增或递减缺失，但整体却存在变小的趋势，此外，其对应的置信区间起点和终点呈现出整体变大的趋势，这表明算法在 NMAR 下，随着缺失率的增大同样面临失效的问题。综合上述两点结论发现，在单变量非随机缺失机制下，KNN、KNNW 将变得不再适用，而 KNNW 的填补精度依然优于 KNN。

3.5 基于多变量缺失的实证分析

仅考虑单变量缺失情形是不够的，现实生活中，样本点的观测值缺失可能是一个或多个，因此，本节仍采用与单变量缺失相同的实验方法和评价准则对填补算法的有效性进行验证。同单变量缺失情形一样，首先对多变量缺失情形下不同缺失率，不同程序执行次数所得到的实验结果均值做一致性分析，实验结果如图 3-5-1：

根据图 3-5-1 的实验结果可以看出，在多变量缺失情形下依然存在与单变量

表 3-4-6：不同缺失率下、4 种填补算法在 NMAR 前提下填补 100 次误差结果的置信区间及其长度(单变量)

评价准则	填补算法	结果类型	缺失率					
			5%	10%	15%	20%	25%	30%
MAE	均值	区间	[1.649,2.178]	[1.764,2.299]	[1.631,2.131]	[2.037,2.554]	[2.081,2.584]	[1.864,2.431]
		长度	0.530	0.534	0.500	0.517	0.503	0.566
	中位数	区间	[1.726,2.300]	[1.899,2.430]	[1.848,2.380]	[2.145,2.659]	[2.453,2.933]	[2.416,2.899]
		长度	0.574	0.530	0.532	0.515	0.480	0.484
	KNN	区间	[1.239,1.698]	[1.535,2.142]	[1.251,1.816]	[1.708,2.247]	[1.623,2.102]	[2.357,2.702]
		长度	0.458	0.607	0.565	0.539	0.479	0.345
	KNNW	区间	[1.219,1.652]	[1.575,2.132]	[1.350,1.899]	[1.747,2.242]	[1.649,2.100]	[2.607,2.858]
		长度	0.433	0.557	0.550	0.495	0.450	0.252
RMSE	均值	区间	[1.659,2.191]	[1.792,2.330]	[1.681,2.179]	[2.111,2.628]	[2.169,2.677]	[1.961,2.542]
		长度	0.532	0.538	0.498	0.517	0.508	0.581
	中位数	区间	[1.734,2.311]	[1.919,2.456]	[1.885,2.419]	[2.202,2.724]	[2.517,3.012]	[2.501,3.006]
		长度	0.577	0.537	0.534	0.521	0.494	0.505
	KNN	区间	[1.430,1.932]	[1.727,2.344]	[1.469,2.029]	[2.115,2.638]	[2.141,2.588]	[2.698,3.056]
		长度	0.502	0.617	0.561	0.522	0.448	0.357
	KNNW	区间	[1.533,2.034]	[1.839,2.426]	[1.623,2.179]	[2.187,2.655]	[2.203,2.608]	[2.998,3.269]
		长度	0.501	0.587	0.556	0.468	0.405	0.270
MAPE	均值	区间	[0.095,0.134]	[0.101,0.139]	[0.092,0.128]	[0.114,0.150]	[0.115,0.150]	[0.107,0.145]
		长度	0.039	0.038	0.036	0.037	0.035	0.038
	中位数	区间	[0.101,0.144]	[0.110,0.149]	[0.106,0.145]	[0.121,0.160]	[0.136,0.174]	[0.138,0.174]
		长度	0.043	0.040	0.039	0.039	0.037	0.036
	KNN	区间	[0.074,0.106]	[0.091,0.133]	[0.075,0.114]	[0.098,0.137]	[0.092,0.128]	[0.133,0.160]
		长度	0.032	0.042	0.040	0.039	0.035	0.028
	KNNW	区间	[0.073,0.103]	[0.093,0.132]	[0.080,0.118]	[0.100,0.136]	[0.094,0.127]	[0.145,0.168]
		长度	0.030	0.039	0.038	0.036	0.033	0.022

缺失情形中相似的实验结论，此处不再具体阐述，且在多变量缺失前提下的后续

实证分析中，仍然令缺失率 $p=0.05,0.1,0.15,0.2,0.25,0.3$，$N=100$。

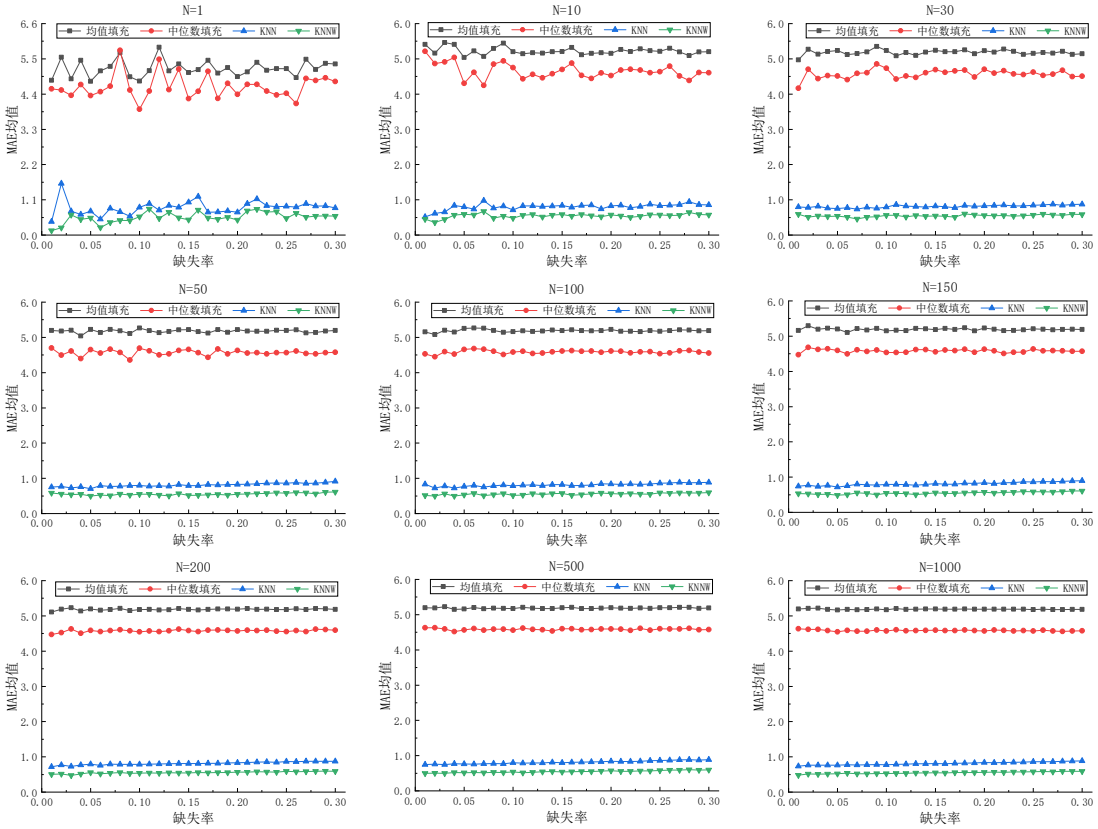

图 3-5-1：四种算法在不同缺失率前提下执行不同次数的 MAE 均值结果折线图（多变量）

3.5.1 完全随机缺失

仍先考虑 MCAR 下多变量缺失的情形，并预设缺失列 $J=3,9,11$，同样的，在每一个缺失率前提下分别采用均值填补、中位数填补、KNN、KNNW 对随机构造的非完整数据集进行填补，并将上述实验重复执行 100 次，进而求其对应的均值作为最终的评价结果，详情如表 3-5-1 所示：

根据表 3-5-1 的实验结果显示，第一，在多变量完全随机缺失情形下，基于不同缺失率，KNN、KNNW 的 $\overline{MAE}$、$\overline{RMSE}$、$\overline{MAPE}$ 结果均显著优于传统的均值填补和中位数填补；第二，随着缺失率的增大，KNN、KNNW 的 $\overline{MAE}$、$\overline{RMSE}$、$\overline{MAPE}$ 结果并没有出现严格的递增趋势，但整体呈现出值变大的情形，这表明在多变量缺失情形下，算法的填补效果依然会随着缺失率的增大而变差；第三，基于不同的评价准则，在任何缺失率前提下，KNNW 均大幅优于 KNN，这证明了 k 值优化和动态调参对算法改进始终起到正向作用。

表 3-5-1：不同缺失率下、4 种填补算法在 MCAR 前提下填补 100 次误差结果的均值(多变量)

评价准则	填补算法	缺失率					
		5%	10%	15%	20%	25%	30%
MAE	均值	5.227	5.183	5.152	5.185	5.213	5.147
	中位数	4.601	4.584	4.499	4.555	4.615	4.524
	KNN	0.739	0.805	0.822	0.836	0.854	0.880
	KNNW	0.496	0.571	0.568	0.566	0.573	0.605
RMSE	均值	6.525	6.517	6.461	6.507	6.554	6.479
	中位数	7.048	7.091	6.983	7.056	7.149	7.048
	KNN	1.514	1.695	1.688	1.724	1.775	1.824
	KNNW	1.316	1.573	1.504	1.508	1.528	1.594
MAPE	均值	0.897	0.856	0.901	0.892	0.888	0.878
	中位数	0.561	0.533	0.557	0.548	0.548	0.539
	KNN	0.167	0.169	0.183	0.185	0.185	0.186
	KNNW	0.131	0.131	0.139	0.136	0.136	0.138

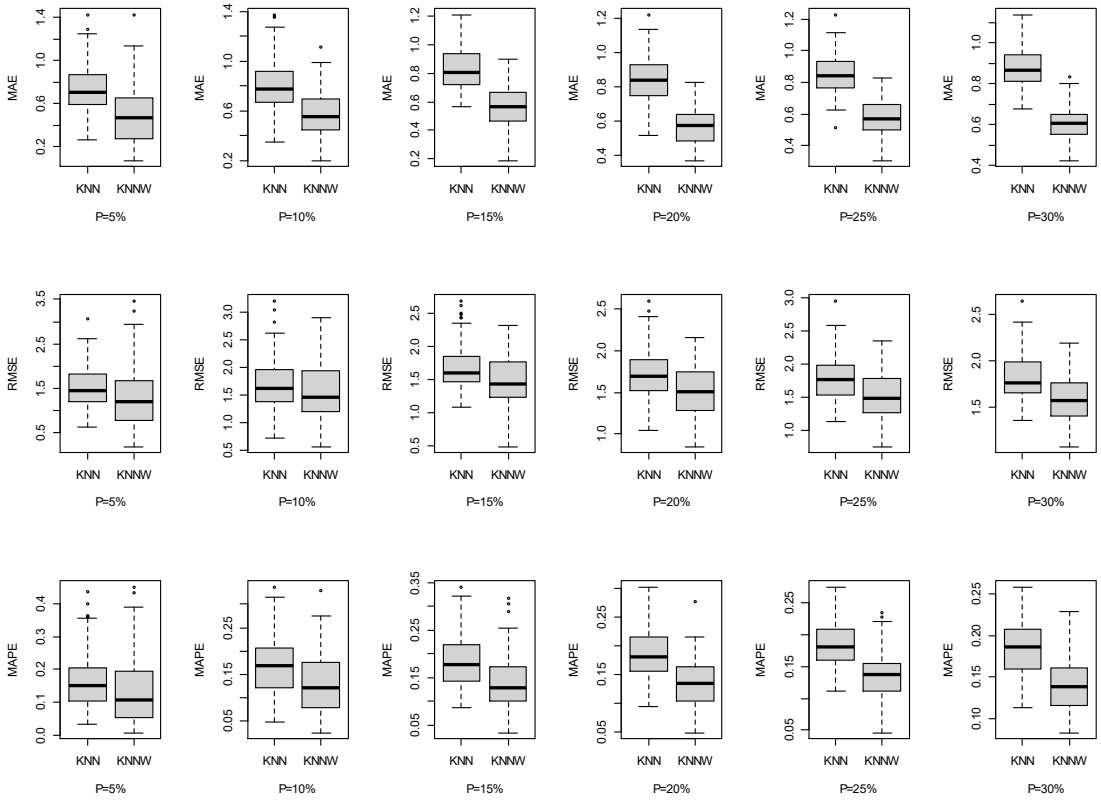

图 3-5-2：不同缺失率下、KNN 与 KNNW 在 MCAR 前提下填补 100 次的误差结果箱线图（多变量）

表 3-5-2：不同缺失率下、4 种填补算法在 MCAR 前提下填补 100 次误差结果的置信区间及其长度(多变量)

评价准则	填补算法	结果类型	缺失率					
			5%	10%	15%	20%	25%	30%
MAE	均值	区间	[5.125,5.329]	[5.123,5.243]	[5.102,5.202]	[5.137,5.232]	[5.173,5.254]	[5.116,5.179]
		长度	0.204	0.119	0.100	0.095	0.081	0.063
	中位数	区间	[4.442,4.761]	[4.480,4.688]	[4.420,4.579]	[4.474,4.637]	[4.544,4.686]	[4.471,4.577]
		长度	0.319	0.208	0.159	0.162	0.142	0.106
	KNN	区间	[0.693,0.786]	[0.767,0.844]	[0.794,0.850]	[0.811,0.861]	[0.830,0.877]	[0.861,0.899]
		长度	0.093	0.077	0.055	0.050	0.048	0.038
	KNNW	区间	[0.445,0.547]	[0.536,0.606]	[0.539,0.596]	[0.543,0.588]	[0.551,0.595]	[0.590,0.620]
		长度	0.101	0.070	0.057	0.044	0.044	0.031
RMSE	均值	区间	[6.404,6.645]	[6.441,6.592]	[6.399,6.523]	[6.446,6.569]	[6.501,6.607]	[6.437,6.520]
		长度	0.241	0.151	0.124	0.123	0.105	0.083
	中位数	区间	[6.831,7.265]	[6.951,7.231]	[6.874,7.091]	[6.945,7.166]	[7.055,7.244]	[6.971,7.125]
		长度	0.433	0.280	0.216	0.220	0.189	0.154
	KNN	区间	[1.418,1.611]	[1.605,1.786]	[1.620,1.755]	[1.667,1.781]	[1.711,1.839]	[1.775,1.874]
		长度	0.193	0.181	0.135	0.115	0.129	0.099
	KNNW	区间	[1.176,1.456]	[1.470,1.676]	[1.428,1.580]	[1.447,1.569]	[1.460,1.595]	[1.546,1.642]
		长度	0.279	0.207	0.153	0.121	0.135	0.096
MAPE	均值	区间	[0.846,0.947]	[0.824,0.889]	[0.874,0.929]	[0.872,0.912]	[0.869,0.907]	[0.861,0.895]
		长度	0.101	0.065	0.055	0.040	0.038	0.034
	中位数	区间	[0.527,0.595]	[0.512,0.553]	[0.538,0.575]	[0.535,0.561]	[0.535,0.560]	[0.528,0.550]
		长度	0.068	0.041	0.037	0.026	0.025	0.022
	KNN	区间	[0.150,0.185]	[0.157,0.181]	[0.172,0.194]	[0.177,0.194]	[0.178,0.192]	[0.180,0.193]
		长度	0.034	0.024	0.021	0.016	0.014	0.013
	KNNW	区间	[0.112,0.150]	[0.118,0.144]	[0.128,0.150]	[0.128,0.143]	[0.129,0.143]	[0.132,0.143]
		长度	0.038	0.026	0.022	0.016	0.014	0.011

同样需要对不同算法的填补精度进行对比分析，基于不同的缺失率，利用多次填补所得到的 MAE、RMSE、MAPE 评价结果来构造对应的 95%置信区间，并绘制

相应的箱线图，结果如图 3-5-2，表 3-5-2。

图 3-5-2 的实验结果表明，第一，基于不同的缺失率和评价准则，KNN、KNNW 对应的 100 次填补评价结果仍有异常点出现，这再次表明了使用单次评价结果对算法的优良性进行评估是不合理的；第二，在 MCAR 下，基于不同的缺失率和评价准则，KNNW 的评价结果最大值、最小值、四分位数、中位数均小于 KNN，这表明 KNNW 的有效性；第三，基于不同缺失率，在 MAE、MAPE 评价准则下，KNNW 填补结果绘制出的四分间距接近或小于 KNN，这进一步加强了第二点实验结果，然而在 RMSE 评价准则下，基于不同的缺失率，KNNW 的四分间距均大于 KNN，这代表经改进后的 KNNW 在提高精度的同时，算法填补的稳定性出现下降。

根据表 3-5-2 的实验结果，第一，整体来看，四种算法对应的 MAE、RMSE、MAPE 置信区间长度会随着缺失率的递增而出现整体减小的趋势，而对应的置信区间起点和终点会随着缺失率的增大出现整体变大的趋势，这表明，随着缺失率的增大，算法的填补效果变差，而填补精度反而出现提升，这也印证了前文所述的结论；第二，基于不同的缺失率，KNN、KNNW 在不同评价准则下的填补精度也显著优于传统的均值填补和中位数填补；第三，基于不同的评价准则，在不同的缺失率前提下分析 KNN、KNNW 对应的置信区间及其长度发现，KNNW 所有的置信区间起点和终点远小于对应的 KNN，甚至出现 KNNW 在缺失率 $p=0.3$ 时得到的 MAE、RMSE、MAPE 置信区间，其对应的起点和终点值依然小于或接近 KNN 在缺失率 $p=0.05$ 时得到的 MAE、RMSE、MAPE 置信区间的对应值，这表明 KNNW 在填补效果上相较于 KNN 获得了显著改善；第四，进一步观察不同评价准则下，基于不同缺失率前提的置信区间长度发现，KNNW 的区间长度始终大于 KNN，这表明，KNNW 在加权的过程中，受最邻近样本点的影响过大从而导致的不稳定性现象加剧。

3.5.2 随机缺失

在多变量缺失前提下，采用计算机随机模拟的方式进行随机缺失机制下的算法评估，采用与多变量完全随机缺失相同的实验方法进行分析，结果如表 3-5-3：

根据表 3-5-3 的实验结果，第一，对比表 3-5-1 的结果来看，在 MAR 下，基于不同的缺失率，四种算法对应的 $\overline{MAE}$、$\overline{RMSE}$、$\overline{MAPE}$ 结果均有不同程度的增大现象，这与单变量随机缺失情形中的结果基本一致；第二，在多变量随机缺失

表 3-5-3：不同缺失率下、4 种填补算法在 MAR 前提下填补 100 次误差结果的均值(多变量)

评价准则	填补算法	缺失率					
		5%	10%	15%	20%	25%	30%
MAE	均值	5.279	5.228	5.356	5.751	5.632	5.574
	中位数	4.654	4.367	4.68	5.219	4.876	4.916
	KNN	0.849	1.002	0.998	1.043	1.287	1.249
	KNNW	0.566	0.673	0.675	0.741	1.035	0.937
RMSE	均值	6.477	6.404	6.659	7.156	6.949	6.943
	中位数	6.711	6.362	6.904	7.561	7.008	7.103
	KNN	1.666	1.883	1.990	2.253	2.473	2.410
	KNNW	1.363	1.583	1.713	1.938	2.35	2.224
MAPE	均值	0.892	0.988	0.866	0.832	0.984	0.948
	中位数	0.556	0.601	0.518	0.513	0.620	0.618
	KNN	0.184	0.219	0.199	0.209	0.269	0.226
	KNNW	0.128	0.158	0.150	0.167	0.226	0.190

情形下，KNN、KNNW 基于不同的缺失率所得到的 $\overline{MAE}$ 、$\overline{RMSE}$ 、$\overline{MAPE}$ 结果依然远小于均值和中位数填补，结论与表 3-5-1 结论一致；第三，针对相同的缺失率前提，KNNW 的评价结果总是显著优于 KNN，且这种优势比单变量完全随机缺失情形下的填补效果更加明显；第四，随着缺失率的增大，KNN、KNNW 的 $\overline{MAE}$ 、$\overline{RMSE}$ 、$\overline{MAPE}$ 结果呈现出整体变大的趋势，这也表明算法的填补效果会随着缺失率的增大而面临失效的问题依然没有改变。

同样的，需要对单变量随机缺失情形下填补算法的整体效果进行评估，仍然采用多次程序运行结果来计算出与之对应的 95%置信区间，并绘制相应的箱线图，以此分析四种算法的填补精度和稳健性，实验结果如表 3-5-4、图 3-5-3：

根据图 3-5-3 的实验结果显示，第一，在 MAR 下，基于不同的缺失率，KNNW 在不同评价准则下的结果最大值、最小值、四分位数、中位数绝大部分仍小于 KNN，只有当 $p=0.1$ 时 KNNW 所对应的 RMSE 结果的上四分位数不满足该项规律，这表明 KNNW 在该缺失机制下仍然具有显著优势；第二，基于不同的缺失率和评价准则，

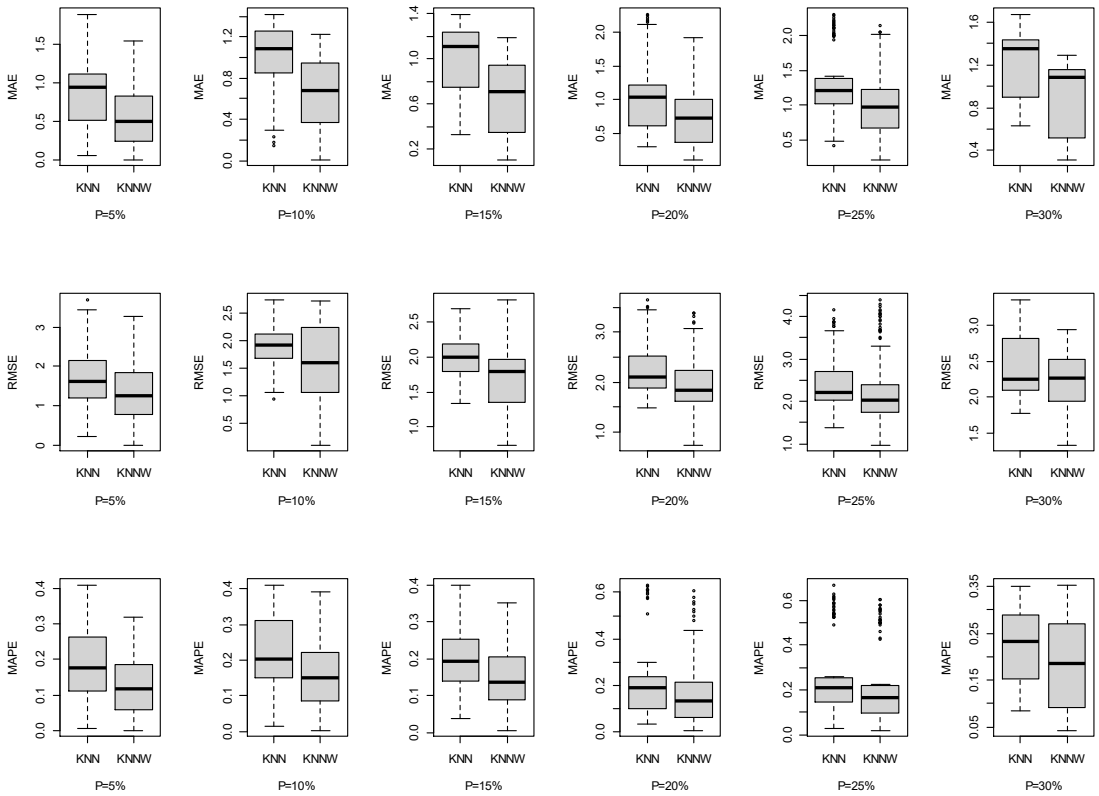

图 3-5-3：不同缺失率下、KNN 与 KNNW 在 MAR 前提下填补 100 次的误差结果箱线图（多变量）

KNNW 的四分间距始终大于或接近 KNN；第三，受多变量缺失情形的影响，计算机模拟非完整数据集的可能性增加，此时，基于不同评价准则所得到的多次实验结果不会出现最大值与上四分位点重合、最小值与下四分位点重合、中位数与上四分位点与下四分位点重合的现象，但当 $p=0.2,0.25$ 时，异常值出现明显增多现象，且对应的评价结果正常值出现逐步增大现象，当 $p=0.3$ 时，四分间距明显增大，且对应的中位数明显高于 $p=0.2,0.25$ 时所得的实验结果，这表明，使用本文提供的模拟方法构造 MAR 在缺失率较大的情况下几乎不会出现非完整数据集一致的情形，但在面对 NMAR 中缺失率较大的情形时，KNNW、KNN 将会收到较大影响，不推荐在该种情形下使用该算法进行填补处理。

根据表 3-5-4 的实验结果可以看出，第一，基于不同的评价准则和缺失率，使用 KNN、KNNW 所得到的置信区间长度始终小于均值和中位数填补，且对应的置信区间起点和终点远小于均值填补和中位数填补所得结果；第二，不同缺失率前提下，KNNW 基于三种评价准则的置信区间长度和对应的起点、终点值均小于 KNN，这表明经过 k 值优化和动态调参后的 KNN 在单变量随机缺失情形下依然能够保证

表 3-5-4：不同缺失率下、4 种填补算法在 MAR 前提下填补 100 次误差结果的置信区间及其长度(多变量)

评价准则	填补算法	结果类型	缺失率					
			5%	10%	15%	20%	25%	30%
MAE	均值	区间	[5.048,5.511]	[5.022,5.434]	[5.129,5.582]	[5.469,6.032]	[5.396,5.868]	[5.363,5.786]
		长度	0.462	0.412	0.454	0.562	0.473	0.422
	中位数	区间	[4.238,5.070]	[4.015,4.719]	[4.290,5.070]	[4.769,5.670]	[4.478,5.274]	[4.548,5.283]
		长度	0.832	0.704	0.780	0.901	0.797	0.735
	KNN	区间	[0.770,0.928]	[0.937,1.067]	[0.941,1.055]	[0.952,1.135]	[1.183,1.392]	[1.192,1.306]
		长度	0.159	0.131	0.114	0.183	0.209	0.113
	KNNW	区间	[0.491,0.641]	[0.607,0.738]	[0.615,0.735]	[0.653,0.829]	[0.928,1.143]	[0.874,1.000]
		长度	0.151	0.131	0.121	0.175	0.215	0.125
RMSE	均值	区间	[6.204,6.751]	[6.156,6.652]	[6.379,6.939]	[6.806,7.507]	[6.636,7.262]	[6.642,7.244]
		长度	0.547	0.496	0.560	0.701	0.626	0.602
	中位数	区间	[6.208,7.214]	[5.924,6.799]	[6.446,7.361]	[7.066,8.057]	[6.552,7.464]	[6.672,7.534]
		长度	1.006	0.875	0.915	0.991	0.912	0.862
	KNN	区间	[1.519,1.813]	[1.812,1.954]	[1.932,2.049]	[2.157,2.350]	[2.339,2.607]	[2.328,2.491]
		长度	0.294	0.142	0.117	0.193	0.268	0.163
	KNNW	区间	[1.198,1.529]	[1.453,1.714]	[1.619,1.808]	[1.836,2.040]	[2.176,2.524]	[2.148,2.301]
		长度	0.331	0.261	0.189	0.203	0.349	0.153
MAPE	均值	区间	[0.808,0.976]	[0.907,1.070]	[0.808,0.925]	[0.758,0.907]	[0.901,1.066]	[0.880,1.016]
		长度	0.168	0.162	0.118	0.149	0.165	0.136
	中位数	区间	[0.502,0.610]	[0.550,0.652]	[0.489,0.548]	[0.470,0.556]	[0.565,0.676]	[0.574,0.662]
		长度	0.108	0.102	0.058	0.085	0.110	0.088
	KNN	区间	[0.164,0.204]	[0.197,0.240]	[0.182,0.215]	[0.181,0.238]	[0.234,0.303]	[0.212,0.241]
		长度	0.040	0.043	0.033	0.057	0.069	0.029
	KNNW	区间	[0.110,0.145]	[0.138,0.178]	[0.133,0.167]	[0.140,0.195]	[0.191,0.260]	[0.173,0.207]
		长度	0.035	0.039	0.033	0.055	0.069	0.034

算法的填补精度和准确性；第三，随着缺失率的增大，四种算法所得置信区间的起点和终点具有整体右移的趋势，而区间长度没有明显的递增或递减现象，这依

然表明，随着缺失的增大，填补算法面临失效的问题，与单变量完全随机缺失结果近似，不再赘述；第四，对比分析表 3-5-2 的实验结果可以看出，在多变量随机缺失前提下，基于不同的缺失率，四种算法依据 MAE、RMSE、MAPE 评价结果所得的 95%置信区间，其对应的起点和终点均出现右移现象，这表明，在本文所模拟的 MAR 下，数据填补算法的有效性和精度均受到了一定的影响。

3.5.3 非随机缺失

基于相同的评价准则和实验方法对 NMAR 下的单变量数据缺失情形进行模拟，实验结果如表 3-5-5 所示：

表 3-5-5：不同缺失率下、4 种填补算法在 NMAR 前提下填补 100 次误差结果的均值(多变量)

评价准则	填补算法	缺失率					
		5%	10%	15%	20%	25%	30%
MAE	均值	5.520	5.072	5.100	5.058	5.565	5.859
	中位数	4.750	4.510	4.377	4.195	4.751	5.396
	KNN	1.155	1.616	1.602	1.865	2.036	2.243
	KNNW	1.010	1.531	1.608	1.903	2.098	2.249
RMSE	均值	6.333	5.863	5.995	5.993	6.521	6.900
	中位数	6.039	5.665	5.610	5.448	6.192	6.830
	KNN	1.821	2.381	2.404	2.789	3.189	3.383
	KNNW	1.706	2.436	2.522	2.915	3.372	3.465
MAPE	均值	0.983	0.799	0.960	1.008	0.989	1.142
	中位数	0.566	0.485	0.587	0.616	0.638	0.803
	KNN	0.251	0.264	0.338	0.368	0.399	0.451
	KNNW	0.233	0.251	0.343	0.376	0.415	0.459

根据表 3-5-5 的实验结果，第一，相较于多变量完全随机缺失和多变量随机缺失两种情形，四种填补算法效果在不同缺失率下均出现不同程度的下降趋势，其中，KNN、KNNW 填补结果偏差尤为明显，这是由于在 NMAR 下算法不再适用所导致的结果，然而，对比表 3-5-1、表 3-5-3 可以发现，处于 NMAR 下的 KNN、KNNW 的 $\overline{MAE}$ 、$\overline{RMSE}$ 、$\overline{MAPE}$ 结果依然优于 MCAR、MAR 下均值填补和中位数填补的

$\overline{MAE}$、$\overline{RMSE}$、$\overline{MAPE}$，也更优于 NMAR 下这两种传统填补法的评价结果；第二，KNN、KNNW 在不同缺失率下的$\overline{MAE}$、$\overline{RMSE}$、$\overline{MAPE}$结果不再具备明显的优劣关系。为了进一步验证 KNNW、KNN 在 NMAR 下的区别，仍然采用前文所述方法构造置信区间，并绘制相应的箱线图，结果如图 3-5-4、表 3-5-6：

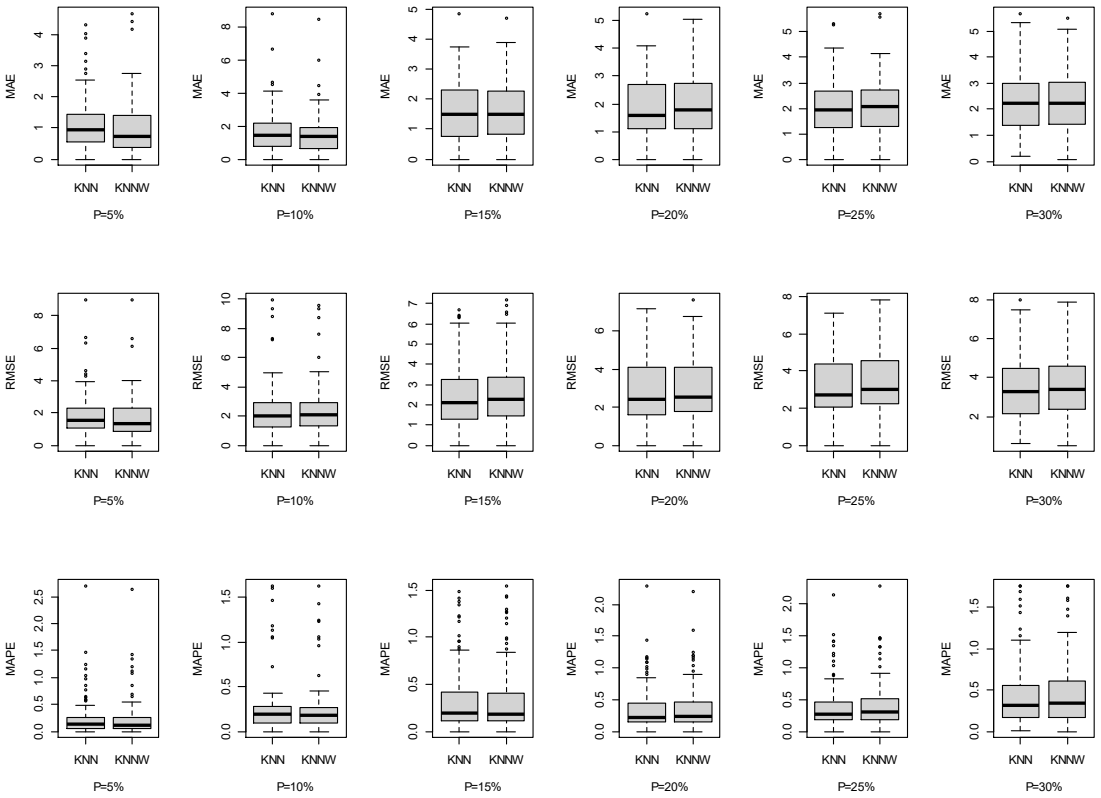

图 3-5-4：不同缺失率下、KNN 与 KNNW 在 NMAR 前提下填补 100 次的误差结果箱线图（多变量）

根据图 3-5-4 的实验结果显示，第一，在 NMAR 下，基于不同的缺失率，KNNW 在不同评价准则下的结果最大值、最小值、四分位数、中位数相较于 KNN 并无明显优势，这表明在 NMAR 情形下使用动态调参的方法失效；第二，对比表 3-5-5 的结果可以发现，NMAR 下基于三种评价准则所得$\overline{MAE}$、$\overline{RMSE}$、$\overline{MAPE}$结果急速膨胀的原因之一在于异常值太多而导致的，从上图可以看出，不同缺失率、不同评价准则下的中位数、四分位数相较于 MCAR、MAR 出现显著增大现象。综上，在多变量非随机缺失情形下，KNN、KNNW 将变得不再使用，且受 NMAR 的影响，填补效果的不稳定性剧增。

从表 3-5-6 的实验结果来看，第一，基于不同的评价准则和缺失率，KNNW 在置信区间长度上相较于 KNN 并无优势，而对应的置信区间起点和终点并没有呈现

表 3-5-6：不同缺失率下、4 种填补算法在 NMAR 前提下填补 100 次误差结果的置信区间及其长度(多变量)

评价准则	填补算法	结果类型	缺失率					
			5%	10%	15%	20%	25%	30%
MAE	均值	区间	[5.130,5.909]	[4.667,5.478]	[4.698,5.501]	[4.692,5.424]	[5.145,5.984]	[5.484,6.233]
		长度	0.779	0.811	0.803	0.733	0.839	0.749
	中位数	区间	[4.120,5.381]	[3.923,5.097]	[3.798,4.956]	[3.690,4.701]	[4.177,5.324]	[4.882,5.909]
		长度	1.261	1.174	1.158	1.011	1.146	1.026
	KNN	区间	[0.980,1.331]	[1.351,1.881]	[1.402,1.803]	[1.652,2.078]	[1.822,2.250]	[2.003,2.483]
		长度	0.351	0.530	0.401	0.426	0.427	0.481
	KNNW	区间	[0.829,1.191]	[1.284,1.777]	[1.405,1.811]	[1.690,2.116]	[1.879,2.317]	[2.015,2.484]
		长度	0.362	0.493	0.406	0.426	0.438	0.469
RMSE	均值	区间	[5.888,6.778]	[5.389,6.337]	[5.518,6.472]	[5.571,6.415]	[6.048,6.994]	[6.472,7.329]
		长度	0.891	0.947	0.954	0.844	0.946	0.857
	中位数	区间	[5.283,6.794]	[4.927,6.404]	[4.889,6.331]	[4.812,6.084]	[5.501,6.882]	[6.207,7.453]
		长度	1.511	1.477	1.442	1.272	1.381	1.246
	KNN	区间	[1.546,2.096]	[2.016,2.746]	[2.098,2.709]	[2.487,3.091]	[2.868,3.510]	[3.056,3.711]
		长度	0.549	0.731	0.612	0.605	0.642	0.655
	KNNW	区间	[1.429,1.984]	[2.085,2.787]	[2.206,2.838]	[2.605,3.224]	[3.037,3.706]	[3.138,3.792]
		长度	0.555	0.702	0.632	0.620	0.670	0.654
MAPE	均值	区间	[0.783,1.183]	[0.658,0.940]	[0.769,1.152]	[0.828,1.188]	[0.819,1.160]	[0.939,1.345]
		长度	0.400	0.282	0.383	0.360	0.341	0.407
	中位数	区间	[0.456,0.675]	[0.398,0.572]	[0.464,0.711]	[0.489,0.743]	[0.514,0.761]	[0.651,0.956]
		长度	0.218	0.174	0.247	0.254	0.247	0.304
	KNN	区间	[0.177,0.324]	[0.202,0.326]	[0.266,0.409]	[0.295,0.441]	[0.325,0.473]	[0.369,0.533]
		长度	0.147	0.124	0.143	0.145	0.148	0.164
	KNNW	区间	[0.160,0.307]	[0.193,0.309]	[0.270,0.416]	[0.303,0.450]	[0.338,0.492]	[0.377,0.541]
		长度	0.147	0.117	0.147	0.147	0.153	0.163

出明显的规律；第二，随着缺失率的增大，四种算法基于不同评价准则求得的置信区间长度没有呈现出严格的递增或递减缺失，而对应的置信区间起点和终点呈

现出整体变大的趋势，这表明算法在NMAR下，随着缺失率的增大同样面临失效的问题。综合上述两点结论发现，在多变量非随机缺失机制下，KNN、KNNW将变得不再适用，且KNNW也不再优于KNN。

3.6 本章小结

本章的实证分析结果显示：第一，在不同的缺失率、缺失列前提下，k值的最优解是会有所不同，且单变量缺失和多变量缺失前提下k值的最优解也会有所不同，为了后续实验的对比分析，本文选择一个相对最优解$k=5$；第二，在完全随机缺失和随机缺失机制下，基于不同的缺失率前提，KNN、KNNW的填补结果始终优于均值填补和中位数填补，且经过k值优化和动态调参后的KNNW填补效果始终优于KNN，但稳定性和填补精度相较于KNN会有所降低；第三，在不同的缺失机制下，随着缺失率的增大，KNN、KNNW的填补结果均出现不同程度的下降；第四，在NMAR下，基于不同的缺失率，KNN、KNNW填补方法虽然面临失效，但仍显著优于均值填补和中位数填补。本文所使用的交叉验证法对k值进行优化选取的过程具有一定的参考价值，而针对权重分配问题提出的动态调参法是一种通用型办法，对不同的样本集数据均能有效的寻找最优的权重分配方案。

4 不同缺失机制下KNNW及其改进算法的实证研究

由于KNNW的结构性问题，导致该算法在填补过程中过分依赖最邻近样本点，从而表现出算法的稳定性有所下降，此外，随着缺失率增大而导致的填补算法逐步失效问题也是本章考虑的重点。针对上述问题，本章采用缺失森林对KNNW填补结果进行校准，并结合迭代法使缺失率逐步降低，从而提出加权K近邻和缺失森林混合迭代填补算法。

4.1 算法介绍

4.1.1 加权KNN

相较于KNN，KNNW极大改善了因k值选取不合理所带来的影响，然而，无论是KNN和KNNW的核心仍然是为每一个含有缺失值的样本点选择k个近邻，进而对这k个样本点中缺失变量所对应的观测值进行组合分析，如求其均值等，最终确定该样本点缺失值部分的预估值。KNN最大的缺点在于k值选取不合理导致的分类模糊或误分类的问题时有发生。KNNW虽然对该问题进行了改进，但算法所计算出的最终结果在很大程度上仍然依赖于最临近样本点所对应的观测值大小，这仍然会出现KNN中所出现的填补效果不稳定的现象。此外，基于不同的缺失率和缺失机制，KNN、KNNW的填补效果会有不同程度的的改变，尤其是在缺失率较大时，算法的填补效果会有不同程度的下降。

针对KNNW中过度依赖最临近样本点所对应观测值大小的问题，可以通过对权重函数参数进行设置来降低对首个临近点的依赖程度，这在一定程度上可以缓解过度依赖所造成的部分样本点填补不合理的情形，但这会使KNNW逐步退化至KNN算法。针对该问题，本文采用缺失森林对KNNW的每一个填补值进行校准，然后选择本轮校准中的最优值进行缺失值填补，其余缺失值保持不变，进而基于填补后的数据集重复该步骤，直至填补完最后一个缺失值即可。

针对缺失率变大所带来的填补效果下降的问题，本文提出迭代法对原有的KNNW进行改进，基本思想是：第一，原有缺失率下，将缺失森林校准后的最优KNNW填补值填入含缺失值样本集的对应位置；第二，以新的样本集为基础，重新使用KNNW对其进行填补，使用缺失森林对本轮填补值进行校准，选出本轮填补值的最优解，进而继续填补到当前样本集的对应位置；第三，重复执行前两个步骤，直至最后一个缺失值填补完成即可。

4.1.2 缺失森林

缺失森林（Missing Forest filling method，简称MF）是近年来国内外学者依托随机森林提出的一种数据填补方法，作为一种非参数方法，该算法对数据本身的要求极少，且具有良好的稳健性和准确性。该算法是包含多棵决策数的算法模型，每颗决策树都是一个分类器，对待分类样本点进行判断，最后综合每棵树的判断结果进行分析，最终确定待分类样本点的类别。MF需要在测试样本集中进行有放回的抽样来生成新的样本集，并随机选择样本集的特征维数，以此来生成每一颗树，特征维数大小会影响每棵树的分类能力与错误率。

MF开始之初，首先使用均值填补等方法对含有缺失值的样本集做一个初始估算，并生成第一个填补后的完整数据集；然后将原始数据集中含有缺失值的变量按缺失率由小到大进行排序，进而使用MF对每一个含有缺失值的变量进行填补，填补次序按变量缺失率的大小由小到大进行处理，生成第二个填补后的完整数据集；通过迭代的方式反复进行，直到最新的填补结果与上次的填补结果没有差别或差别很小时结束；按照类似的方法对含有缺失值的其他变量进行填补，直至数据集中没有缺失值为止。

4.2 算法改进

4.2.1 迭代法

迭代法是指将上一次计算的结果作为下一次计算的初始值，如此往复，直到结果收敛为止。在数据填补问题中，考虑到缺失率较小时，几乎所有的填补算法均能获得相对较优的结果，而随着缺失率的增大，填补算法均面临失效。针对缺失率较大时填补算法面临失效的问题，本文尝试采用迭代法对前文所述算法进行改进，在缺失数据填补过程中，通过预设填补条件来逐步降低缺失率，直至对非完整数据集填补完毕即可。迭代法思路具体如下：

第一，对完整数据集M进行缺失率为p，缺失列为J的随机挖空处理，得到最初的非完整数据集M^{new}；

第二，采用至少两种数据填补算法对M^{new}进行首次缺失值填补；

第三，逐个对比两种填补算法预估的缺失值，并制定相应的填补条件；

第四，将满足填补条件的两个对应预估值进行选取或组合，生成最终的估计

值，然后将其存放至 M^{new} 中对应的缺失值位值，并生成新的 M^{new}；

第五，重复第二至四步，直至填充完所有缺失值即可。

4.2.2 加权 KNN 与缺失森林混合迭代填补法

考虑到 KNNW 所具备的填补效果好，但稳定性有所欠缺的情形，本章采用 MF 对其每一个填补值进行校准，使其具备缺失森林的稳定性优势，并通过迭代法进一步提升填补效果,最终提出加权 KNN 与缺失森林混合迭代填补法（K-Nearest Neighbor of Weighted and Missing Forest Hybrid Iterative filling algorithm, 简称 KNNWMF），其具体实验步骤和方法如下：

第一，根据预先设定的数据缺失机制，对完整数据集 M 进行缺失率为 p，缺失列为 J 的随机挖空处理，得到非完整数据集 M^{new}；

第二，对 M^{new} 中的空缺位置在 M 对应的真实值集合 y 进行按序保存，排序方式以样本点为主要依据，并采用行列交叉的方式进行，具体如下：

当缺失情形为单变量缺失，按照预先设定的缺失率，此时含缺失值的样本点个数为 n_{mis}，缺失列为 $j, j \in [1, m]$，此时含缺失值的样本点数量为 n_{true} 个，其对应的 y 中元素个数为 n_{mis}，此时 $n_{true} = n_{mis}$，具体的排序形式如表 4-2-1：

表 4-2-1：单变量缺失下观测值排序保存形式表

行列号	观测值					
	y_1	y_2	y_3	……	$y_{n_{mis}-1}$	$y_{n_{mis}}$
行标	i_{mis}^{1}	i_{mis}^{2}	i_{mis}^{3}	……	$i_{mis}^{n_{mis}-1}$	$i_{mis}^{n_{mis}}$
列标	j	j	j	……	j	j

当缺失情形为多变量缺失，按照预先设定的缺失率，此时含缺失值的样本点个数 $n_{true} \le n_{mis}$，缺失列为 J，而每个样本点中真正的缺失列集合 $\tau, \tau \subseteq J$ 是通过函数随机生成，其中 $\tau_{\omega}, \omega \in [1, h(\tau)]$ 代表每一个样本点中的第 ω 个缺失列标，假设此时含缺失值的样本点数量仍为 n_{true} 个，其对应的 y_{true} 中元素个数为 n_{mis}，此时 $n_{true} \le n_{mis}$，具体的排序形式如表 4-2-2：

表 4-2-2：单变量缺失下观测值排序保存形式表

行列号	观测值								
	y_1	……	y_α	$y_{\alpha+1}$	$y_{\alpha+2}$	$y_{\alpha+3}$	……	$y_{n_{mis}-1}$	$y_{n_{mis}}$
行标	i_{mis}^1	……	i_{mis}^1	i_{mis}^2	i_{mis}^3	i_{mis}^3	……	$i_{mis}^{n_{true}-1}$	$i_{mis}^{n_{true}}$
列标	τ_1	……	$\tau_{h(\tau)}$	τ_1	τ_1	τ_2	……	$\tau_{h(\tau)-1}$	$\tau_{h(\tau)}$

第三，分别使用 KNNW 和 MF 对非完整数据集 M^{new} 进行缺失值填补，进而得到对应的填补值集合 $\hat{y}_{KNNW}$ 和 $\hat{y}_{MF}$；

第四，取 $|\hat{y}_{KNNW}-\hat{y}_{MF}|$ 中最小值所对应的估计值 $\hat{y}_{KNNW}^{\zeta},\zeta\in[1,n_{mis}]$ 作为本轮填补的估计值，记录该值所对应的行标和列标，并将该值填补至 M^{new} 中的对应位置，进而生成新的 M^{new} 数据集；

第五，重复第二至四步的实验步骤，直至 M^{new} 中不包含缺失值为止。

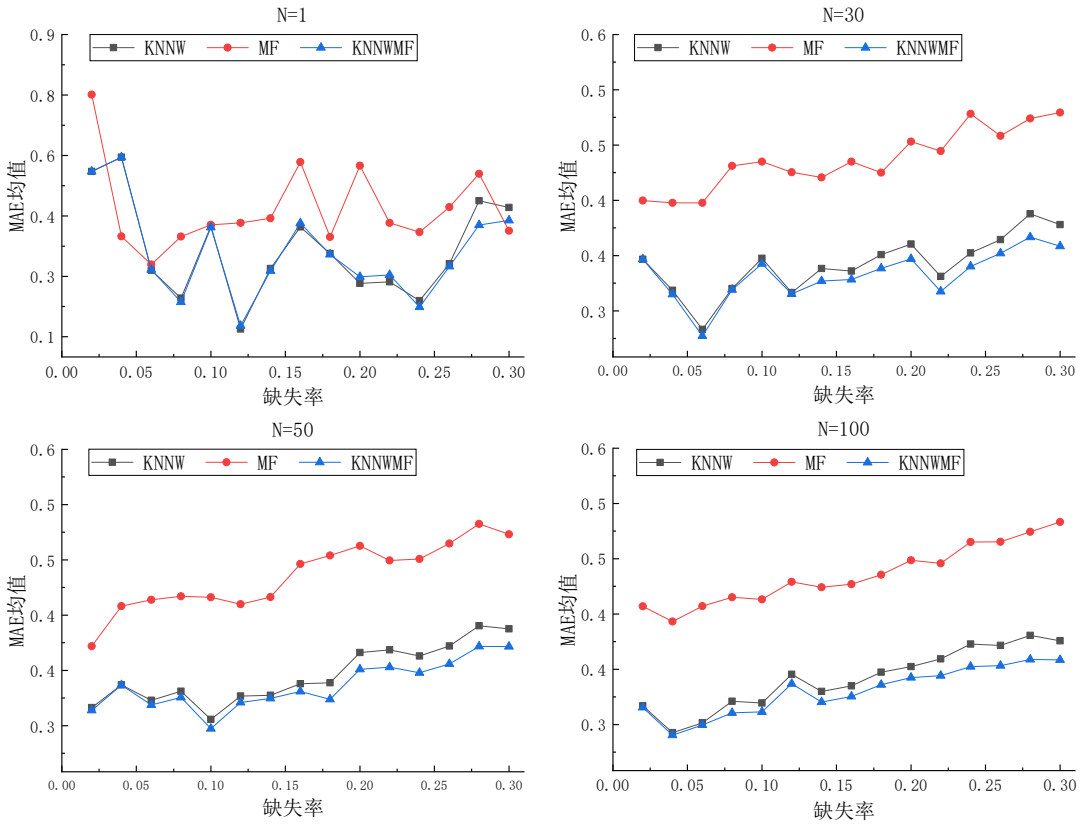

图 4-3-1 四种算法在不同缺失率前提下执行不同次数的 MAE 均值结果折线图（单变量）

为了验证算法的有效性，并同时考虑数据缺失过程中的多变因素，本文基于

MCAR、MAR、NMAR 机制，采用 MAE、RMSE、MAPE 评价准则对 KNNW、MF、KNNWMF 进行评估，具体实验方法同第三章。首先对程序执行不同次数所得结果的一致性进行评价，结合第三章实验结果，在该部分，设定缺失率 p 由 1%逐步递增至 30%，步长设定为 2%，程序执行次数 N 分别取 1、30、50、100 次，并在 MCAR 下对不同缺失率前提、不同 N 求取 $\overline{MAE}$ 、$\overline{RMSE}$ 、$\overline{MAPE}$ 值，实验结果如图 4-3-1、4-3-2 所示：

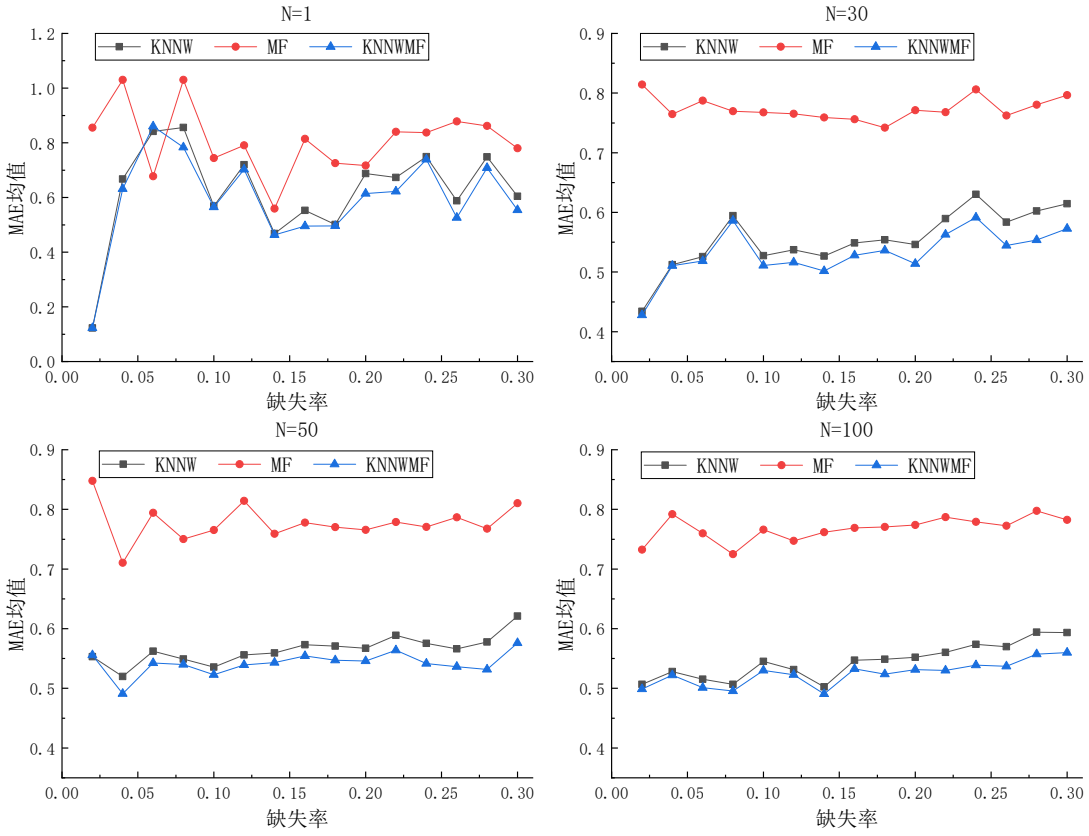

图 4-3-2 四种算法在不同缺失率前提下执行不同次数的 MAE 均值结果折线图（多变量）

根据图 4-3-1、4-3-2 的实验结果，在 MCAR 前提下，KNNW、KNNWMF 基于 MAE、MAPE 评价准则所得实验结果显著优于 MF，而在 RMSE 评价准则下，MF 最优、KNNWMF 次之；此外，由于计算机模拟非完整数据集的随机性特点，当 $N=1$ 时，三种填补方法在三种评价准则下的结果均呈现出震荡状态，当缺失率较小、以及随着缺失率的增大，同样得到了与第三章相同的实验结论，此处不再赘述。综上，在本章节的实验分析环节，同样取 $p=0.05,0.1,0.15,0.2,0.25,0.3$ ， $N=100$ 。

4.3 基于单变量缺失的实证分析

4.3.1 完全随机缺失

本部分先考虑完全随机缺失机制下单变量数据缺失的情况，实验结果如表 4-3-1 所示：

表 4-3-1：不同缺失率下、3 种填补算法在 MCAR 前提下填补 100 次误差结果的均值(单变量)

评价准则	填补算法	缺失率					
		5%	10%	15%	20%	25%	30%
MAE	KNNW	0.321	0.338	0.353	0.360	0.384	0.399
	MF	0.445	0.445	0.451	0.476	0.499	0.514
	KNNWMF	0.318	0.330	0.341	0.343	0.362	0.381
RMSE	KNNW	0.748	0.822	0.862	0.883	0.930	0.950
	MF	0.709	0.720	0.736	0.771	0.792	0.811
	KNNWMF	0.739	0.803	0.834	0.839	0.877	0.905
MAPE	KNNW	0.018	0.019	0.020	0.021	0.022	0.023
	MF	0.026	0.026	0.026	0.028	0.029	0.030
	KNNWMF	0.018	0.019	0.020	0.020	0.021	0.022

根据表 4-3-1 的结果显示：第一，随着缺失率 p 的增加，三种填补算法下的 $\overline{MAE}$、$\overline{RMSE}$、$\overline{MAPE}$ 结果会随缺失率的增加呈现递增趋势，这代表三种填补算法的填补效果会随着缺失率的增加而变差；第二，在 MAE、MAPE 评价准则下，基于不同的缺失率，KNNW 的结果始终优于 MF，而 KNNWMF 的结果始终保持最优，这表明 KNNWMF 具备最佳的填补精度；第三，在 RMSE 评价准则下，MF 的结果始终最小，KNNWMF 的 $\overline{RMSE}$ 结果均低于 KNNW，这表明，MF 在数据填补过程中始终具备优良的稳定性，但填补准确度不及 KNNW。综上，本文提出的 KNNWMF 不仅继承了 KNNW 的准确性，也继承了 MF 的稳健性，在略微提升 KNNW 填补效果的同时提升了算法的稳定性。

为了进一步探讨算法在多次填补中的具体情况，本部分仍采用不同缺失率下使用 KNNW、MF、KNNWMF 填补的每一次 MAE、RMSE、MAPE 结果进行箱线图绘制，结果如图 4-3-3。

图 4-3-3 的实验结果表明，第一，在不同的评价准则和缺失率前提下，MF 的

四分间距均显著小于 KNNW、KNNWMF，这印证了 MF 具备优良的稳定性；第二，在不同的评价准则和缺失率前提下，三种算法对应的 100 次填补评价结果大多出现异常点，这表明了使用单次评价结果对算法的优良性进行评估是不合理的；第三，基于不同的缺失率和评价准则，采用 KNNWMF 所得的评价结果的最大值、最小值、四分位数、中位数均小于 KNNW，且随着缺失率的增大，该现象愈发显著，这表明 KNNWMF 在继承 MF 稳定性的同时也继承了 KNNW 的准确性，也说明本文所提出的迭代法对缺失率较大的情况有着良好的填补效果。

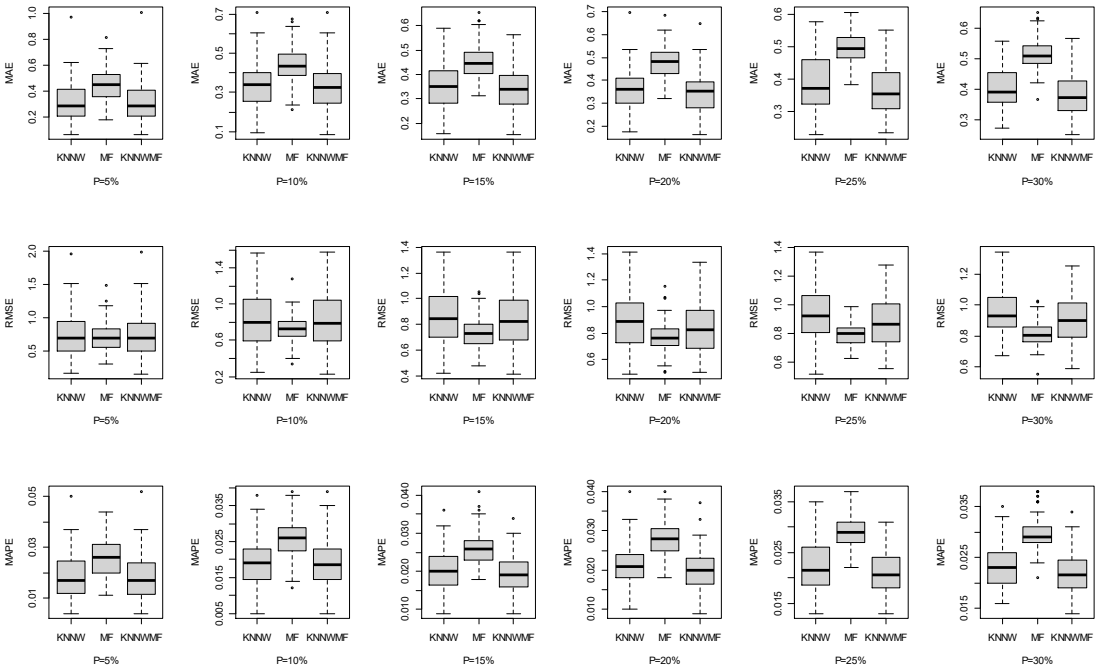

图 4-3-3 不同缺失率下、三种算法在 MCAR 前提下填补 100 次的误差结果箱线图（单变量）

为了进一步探究三种算法在 MAE、RMSE、MAPE 评价准则下的整体优良性，在表 4-3-1 的基础上，本文以 100 次实验结果对应的 MAE、RMSE、MAPE 值为基础，分别对 MAE、RMSE、MAPE 求取不同缺失率下的 95%置信区间，以此来分析算法在填补精度方面的优良性，实验结果如表 4-3-2 所示：

根据表 4-3-2 的结果显示：第一，在不同的缺失率前提下，MF 的 MAE、RMSE、MAPE 结果的 95%置信区间最窄，这进一步证明了 MF 良好的稳定性，此外，KNNWMF 在不同缺失率、评价准则下对应的 95%置信区间长度均小于 KNNW，这代表经过 MF 校准后的 KNNWMF 在填补精度方面有了明显的提升；第二，从置信区间的起止位置来看，在 MAE、MAPE 评价准则下，基于不同的缺失率，KNNWMF 的区间起点始终小于 KNNW 和 MF，而 KNNWMF 的区间终点也始终小于其他两种算法，这表明 KNNWMF

表 4-3-2：不同缺失率下、3 种填补算法在 MCAR 前提下填补 100 次误差结果的置信区间及其长度(单变量)

评价准则	填补算法	结果类型	缺失率 5%	10%	15%	20%	25%	30%
MAE	KNNW	区间	[0.290,0.352]	[0.315,0.361]	[0.335,0.371]	[0.344,0.377]	[0.368,0.400]	[0.387,0.411]
		长度	0.062	0.047	0.036	0.033	0.032	0.025
	MF	区间	[0.420,0.471]	[0.427,0.463]	[0.438,0.465]	[0.464,0.489]	[0.489,0.508]	[0.504,0.524]
		长度	0.052	0.036	0.027	0.025	0.019	0.020
	KNNWMF	区间	[0.287,0.348]	[0.307,0.353]	[0.325,0.358]	[0.327,0.359]	[0.348,0.377]	[0.368,0.393]
		长度	0.061	0.046	0.033	0.032	0.030	0.025
RMSE	KNNW	区间	[0.678,0.817]	[0.766,0.879]	[0.818,0.906]	[0.845,0.921]	[0.892,0.967]	[0.922,0.978]
		长度	0.138	0.113	0.087	0.077	0.075	0.056
	MF	区间	[0.668,0.750]	[0.691,0.749]	[0.713,0.759]	[0.749,0.792]	[0.776,0.808]	[0.795,0.827]
		长度	0.082	0.058	0.046	0.044	0.032	0.032
	KNNWMF	区间	[0.670,0.807]	[0.747,0.859]	[0.793,0.874]	[0.801,0.876]	[0.842,0.912]	[0.877,0.933]
		长度	0.137	0.112	0.081	0.076	0.070	0.056
MAPE	KNNW	区间	[0.017,0.020]	[0.018,0.021]	[0.019,0.021]	[0.020,0.022]	[0.021,0.023]	[0.022,0.024]
		长度	0.003	0.003	0.002	0.002	0.002	0.001
	MF	区间	[0.024,0.028]	[0.025,0.027]	[0.025,0.027]	[0.027,0.029]	[0.029,0.030]	[0.029,0.030]
		长度	0.003	0.002	0.002	0.001	0.001	0.001
	KNNWMF	区间	[0.016,0.020]	[0.018,0.020]	[0.019,0.021]	[0.019,0.021]	[0.020,0.022]	[0.021,0.023]
		长度	0.003	0.003	0.002	0.002	0.002	0.001

在保证填补精度的同时也具备了最佳的填补效果；第三，随着缺失率的增大，三种算法对应置信区间的起点和终点均有不同程度的增大现象，但与之对应的区间长度却出现逐步变窄的趋势，由此可以看出，随着缺失率的增大，三种填补算法的填补效果确有降低，而当缺失率增大时，基于完全随机缺失机制对变量进行随机的挖空挖空处理将会呈现出相似、重叠的特征，而依据程序最终的填补结果所计算的 MAE、RMSE、MAPE 结果会随着缺失率的增大而呈现出聚拢现象。

4.3.2 随机缺失

为了验证算法在不同缺失情形下的稳健性和有效性，依据 MAR 机制的定义，

考虑缺失率 $p=0.05,0.1,0.15,0.2,0.25,0.3$，缺失列 $J=11$，而缺失变量的观测值是否缺失与 Boston 数据集的第 14 列完全观测变量有关，实验仍重复运行 100 次，且结果仍基于 MAE、RMSE、MAPE 三种评价准则进行分析，详情如下：

表 4-3-3：不同缺失率下、3 种填补算法在 MAR 前提下填补 100 次误差结果的均值(单变量)

评价准则	填补算法	缺失率					
		5%	10%	15%	20%	25%	30%
MAE	KNNW	0.368	0.382	0.360	0.589	0.637	0.523
	MF	0.507	0.516	0.491	0.772	0.749	0.684
	KNNWMF	0.358	0.365	0.359	0.601	0.625	0.492
RMSE	KNNW	0.762	0.845	0.767	1.129	1.251	1.102
	MF	0.733	0.769	0.720	1.084	1.106	0.972
	KNNWMF	0.744	0.812	0.751	1.098	1.147	0.998
MAPE	KNNW	0.022	0.022	0.020	0.036	0.038	0.029
	MF	0.030	0.030	0.028	0.047	0.044	0.038
	KNNWMF	0.021	0.021	0.021	0.037	0.037	0.027

根据表 4-3-3 的结果显示：第一，随着缺失率 p 的增加，三种填补算法下的 $\overline{MAE}$、$\overline{RMSE}$、$\overline{MAPE}$ 结果随缺失率的增加没有呈现出逐步递增或递减趋势，然而，随着缺失率的增大，$\overline{MAE}$、$\overline{RMSE}$、$\overline{MAPE}$ 结果仍然表现出变大的情况，这代表在 MAR 下，三种填补算法的填补效果仍会随着缺失率的增加而变差；第二，在 MAE、MAPE 评价准则下，基于不同的缺失率，KNNWMF 的结果依然能够保持整体最优，仅在缺失率 $p=0.15,0.2,0.25$ 时，相较于 KNNW，KNNWMF 填补效果呈现出微弱的劣势，该情况可能与计算机模拟的随机性有关；第三，在 RMSE 评价准则下，MF 的结果始终最小，KNNWMF 的 $\overline{RMSE}$ 结果均低于 KNNW，这表明，在 MAR 前提下，本文提出的 KNNWMF 仍继承了 KNNW 的准确性和 MF 的稳健性，兼具了两种填补算法的优势。

图 4-3-4 是 MAR 前提下，基于不同的评价准则和缺失率，采用 KNNW、MF、KNNWM 填补的每一次 MAE、RMSE、MAPE 结果绘制的箱线图，图 4-3-4 的实验结果表明，

第一，相较于单变量完全随机缺失情形，在不同的缺失率前提和评价准则下，三种算法所得结果的异常值显著增多，尤其是在缺失率较大时该现象更为显著；第二，在 MAE、MAPE 评价准则下，基于不同缺失率，KNNW、KNNWMF 对应的评价结果最大值、中位数、四分位数、最小值均要小于 MF，而在 RMSE 准则下的中位数几乎表现出相反的现象；第三，在缺失率 $p=0.3$ 时，KNNWMF 的 MAE、RMSE、MAPE 的中位数均小于 KNNW 和 KNNWMF，而在其他缺失率前提下，只是呈现出小于或接近的情况；第四，当 $p\geq 0.15$ 时，三种算法基于不同评价结果绘制的箱线图均出现最大值、中位数、四分位数、最小值部分重合现象，且缺失率越大，重合现象明显，这是本文使用的缺失机制模拟方法所导致的，在缺失率较大时，基于单变量缺失得到的非完整数据集会出现完全一致的现象，从而导致最终的评价结果出现多次相等的现象，这进一步印证了前文所述模拟方法的结论。综上，在 MAR 下，KNNW、KNNWMF 的准确性优于 MF，而 MF 的稳定性表现相对较优，且 KNNWMF 继承了 KNNW 的准确性和 MF 的稳定性，尤其在缺失率较大时，填补性能表现优异。

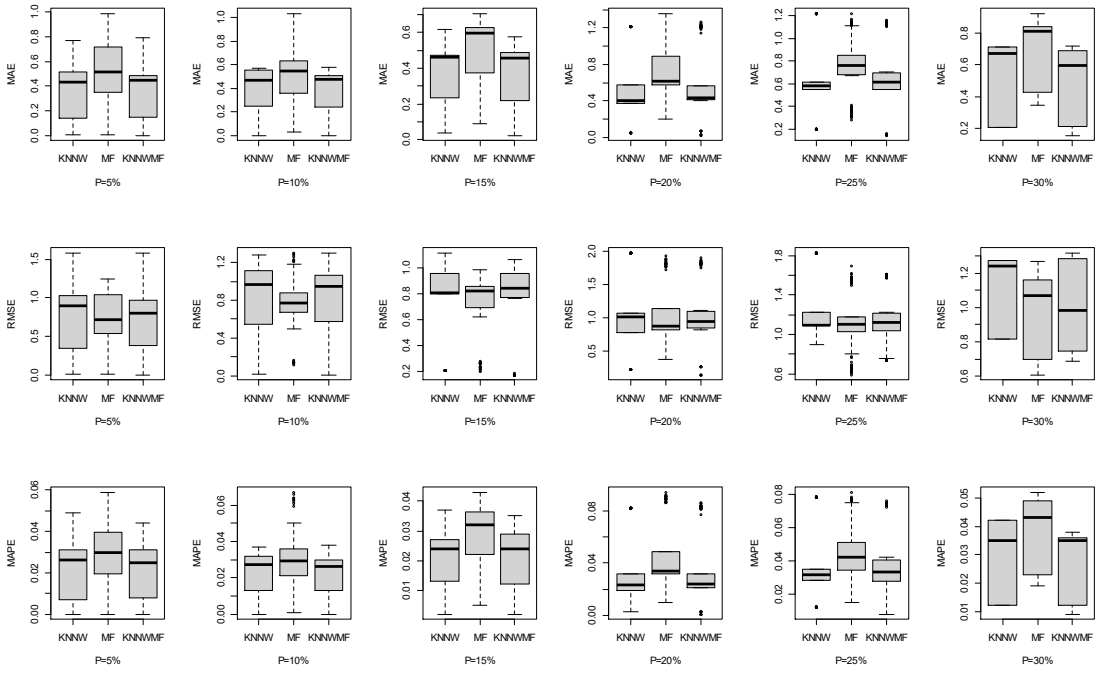

图 4-3-4 不同缺失率下、三种算法在 MAR 前提下填补 100 次的误差结果箱线图（单变量）

同样的，在随机缺失前提下，基于表 4-3-3 的实验结果，本部分仍考虑不同缺失率下运行 100 次实验所得 MAE、RMSE、MAPE 结果对应的 95%置信区间，以此来分析算法在填补精度方面的优良性，实验结果如表 4-3-4 所示：

表 4-3-4：不同缺失率下、3 种填补算法在 MAR 前提下填补 100 次误差结果的置信区间及其长度(单变量)

评价准则	填补算法	结果类型	缺失率					
			5%	10%	15%	20%	25%	30%
MAE	KNNW	区间	[0.323,0.414]	[0.343,0.420]	[0.323,0.397]	[0.515,0.664]	[0.567,0.707]	[0.477,0.569]
		长度	0.091	0.077	0.074	0.148	0.140	0.092
	MF	区间	[0.456,0.557]	[0.472,0.561]	[0.451,0.531]	[0.708,0.837]	[0.695,0.802]	[0.645,0.724]
		长度	0.101	0.090	0.080	0.129	0.107	0.079
	KNNWMF	区间	[0.315,0.400]	[0.328,0.402]	[0.322,0.396]	[0.528,0.673]	[0.558,0.692]	[0.450,0.535]
		长度	0.085	0.074	0.074	0.145	0.133	0.084
RMSE	KNNW	区间	[0.673,0.852]	[0.770,0.919]	[0.709,0.825]	[1.026,1.233]	[1.185,1.317]	[1.060,1.143]
		长度	0.180	0.150	0.116	0.207	0.132	0.083
	MF	区间	[0.671,0.795]	[0.720,0.819]	[0.672,0.767]	[0.998,1.170]	[1.046,1.167]	[0.930,1.015]
		长度	0.124	0.100	0.095	0.173	0.121	0.085
	KNNWMF	区间	[0.660,0.828]	[0.739,0.885]	[0.694,0.808]	[1.006,1.190]	[1.089,1.206]	[0.953,1.042]
		长度	0.168	0.146	0.115	0.184	0.117	0.089
MAPE	KNNW	区间	[0.019,0.025]	[0.020,0.024]	[0.018,0.023]	[0.030,0.041]	[0.033,0.042]	[0.027,0.032]
		长度	0.006	0.005	0.004	0.011	0.009	0.005
	MF	区间	[0.027,0.034]	[0.027,0.033]	[0.026,0.030]	[0.042,0.052]	[0.040,0.048]	[0.036,0.040]
		长度	0.007	0.006	0.005	0.010	0.008	0.005
	KNNWMF	区间	[0.019,0.024]	[0.019,0.023]	[0.018,0.023]	[0.031,0.042]	[0.033,0.042]	[0.025,0.029]
		长度	0.005	0.005	0.004	0.011	0.009	0.005

根据表 4-3-4 的结果显示：第一，在不同的缺失率前提下，依据 KNNWMF 所得 MAE、RMSE、MAPE 结果对应的 95%置信区间长度始终优于 KNNW，这代表经过 MF 算法校准后的 KNNW 稳定性会显著提升；第二，当缺失率 $p=0.05,0.1,0.15$ 时，KNNWMF 在 MAE、MAPE 评价准则下的置信区间长度始终保持最佳，当缺失率 $p=0.2,0.25,0.3$ 时，KNNWMF 在 MAPE 评价准则下的置信区间长度与 MF 几乎保持相同，而当缺失率 $p=0.2,0.25,0.3$ 时，KNNWMF 在 MAE 评价准则下的置信区间长度高于 MF 算法，这证明了在随机缺失前提下 KNNWMF 仍然具备优良的稳定性；第三，从置信区间的起止

位置来看，在 MAE、MAPE 评价准则下，基于不同的缺失率，KNNWMF 的区间起点几乎总能小于 KNNW 和 MF，而 KNNWMF 的区间终点也出现一样的规律，尽管当缺失率 $p=0.2,0.25,0.3$ 时，KNNWMF 并不能保证获得最窄的置信区间，但结合起始点的优势表明 KNNWMF 在保证填补精度的同时也具备了最佳的填补效果；第四，随着缺失率的增大，三种算法对应置信区间的起点和终点均有不同程度的增大现象，且与之对应的区间长度也出现逐步变大的趋势，且当 $p\geq 0.25$ 时不再具备明显的特征，由此可以看出，随着缺失率的增大，三种填补算法的填补精度确有降低，且稳定性变差。

4.3.3 非随机缺失

同 MCAR 和 MAR 的实验模拟过程一样，依据 NMAR 的定义，实验结果如下：

表 4-3-5：不同缺失率下、3 种填补算法在 NMAR 前提下填补 100 次误差结果的均值(单变量)

评价准则	填补算法	缺失率					
		5%	10%	15%	20%	25%	30%
MAE	KNNW	1.407	1.785	1.611	1.842	1.993	2.709
	MF	1.176	1.567	1.507	1.771	2.068	2.693
	KNNWMF	1.312	1.667	1.505	1.751	1.950	2.739
RMSE	KNNW	1.740	2.086	1.887	2.251	2.534	3.090
	MF	1.362	1.744	1.715	2.116	2.518	2.926
	KNNWMF	1.570	1.904	1.744	2.128	2.447	2.939
MAPE	KNNW	0.086	0.109	0.101	0.109	0.119	0.156
	MF	0.072	0.097	0.095	0.105	0.123	0.155
	KNNWMF	0.080	0.103	0.095	0.105	0.117	0.158

根据表 4-3-5 的结果显示：第一，随着缺失率 p 的增加，三种填补算法下的 $\overline{MAE}$、$\overline{RMSE}$、$\overline{MAPE}$ 结果随缺失率的增呈现出显著增大的趋势，只有当 $p=0.15$ 时，其对应的 $\overline{MAE}$、$\overline{RMSE}$、$\overline{MAPE}$ 结果出现了略微的下降，且相较于前两种缺失情况，在该情况下，三种算法的填补效果均显著性降低，这代表在 NMAR 下，三种填补算法将变得不在适用；第二，尽管如此，基于不同的缺失率，KNNWMF 的 $\overline{MAE}$、

$\overline{RMSE}$、$\overline{MAPE}$ 结果依然能够优于 KNNW，仅在缺失率 $p=0.3$ 时，其对应的 $\overline{MAE}$、$\overline{MAPE}$ 值略高于 KNNW，这代表 KNNWMF 在 NMAR 下相较于 KNNW，其填补精度和稳定性依然有效。

同样需要使用单变量 NAMR 下三种算法的每一次 MAE、RMSE、MAPE 评价结果绘制箱线图，结果如图 4-3-5 所示：

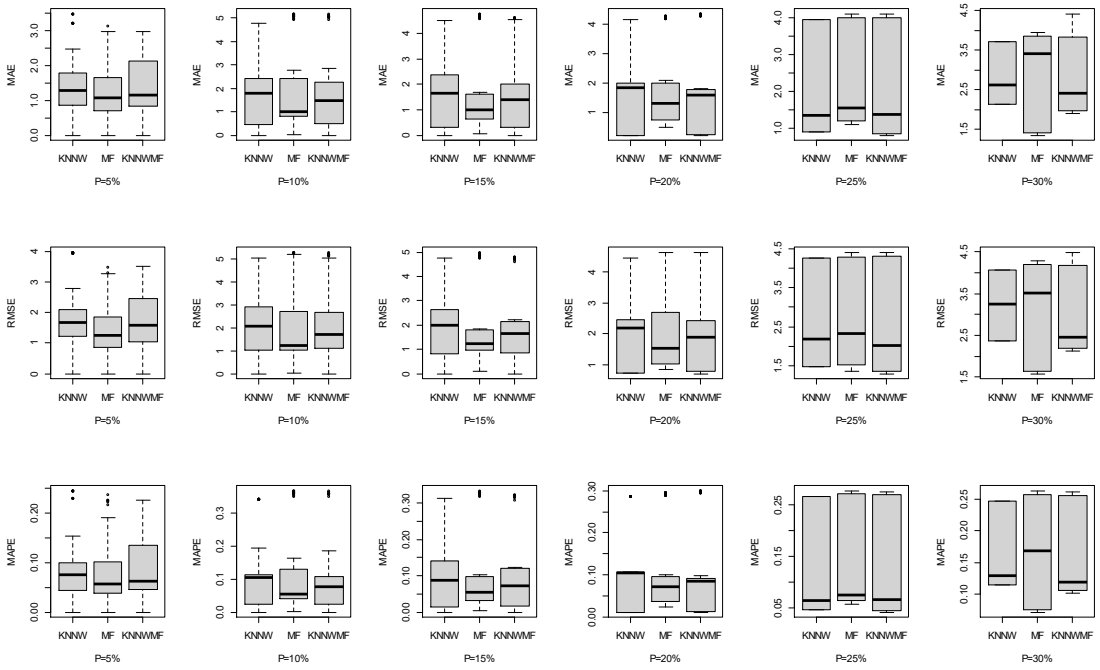

图 4-3-5 不同缺失率下、三种算法在 NMAR 前提下填补 100 次的误差结果箱线图（单变量）

根据图 4-3-5 的实验结果显示，第一，相较于前两种缺失机制，基于不同的缺失率，KNNWMF 在不同评价准则下的结果中位数均小于于 KNNW，且随着缺失率的增大，该现象愈发明显；第二，基于不同的缺失率和评价准则，三种算法对应的最大值、中位数、四分位数几乎全部出现增大现象，且对应的四分间距之间也无明显规律，这表明在 NAMR 下，三种算法几乎完全失效。综上，尽管在单变量非随机缺失下三种算法均面临失效，但 KNNWMF 依然具备一定的优势。下面对三种算法的所对应的置信区间及区间长度进行构造，结果如表 4-3-6 所示：

根据表 4-3-6 的结果显示：第一，在不同的缺失率前提下，三种填补算法所对应的 MAE、RMSE、MAPE 区间长度相较于完全随机缺失和随机缺失出现了明显的增大现象，这代表 NMAR 下算法不适用所带来的影响；第二，结合置信区间的起始位置及区间长度来看，只有在 $p\le 0.15$ 时，三种算法的 MAE、RMSE、MAPE 结果仍

能显示出与前文类似的规律特征。综上，在 NMAR 下，本文提到的三种算法将全部失效，需要考虑新的填补方法或改进措施来应对该缺失情况。

表 4-3-6：不同缺失率下、3 种填补算法在 NMAR 前提下填补 100 次误差结果的置信区间及其长度(单变量)

评价准则	填补算法	结果类型	缺失率					
			5%	10%	15%	20%	25%	30%
MAE	KNNW	区间	[1.211,1.602]	[1.506,2.063]	[1.306,1.917]	[1.588,2.097]	[1.753,2.233]	[2.581,2.837]
		长度	0.391	0.557	0.611	0.509	0.480	0.257
	MF	区间	[1.007,1.345]	[1.276,1.858]	[1.209,1.805]	[1.527,2.016]	[1.831,2.306]	[2.467,2.920]
		长度	0.337	0.582	0.596	0.489	0.475	0.453
	KNNWMF	区间	[1.136,1.489]	[1.375,1.960]	[1.201,1.810]	[1.486,2.016]	[1.695,2.205]	[2.576,2.902]
		长度	0.353	0.585	0.609	0.530	0.510	0.326
RMSE	KNNW	区间	[1.514,1.966]	[1.801,2.370]	[1.579,2.196]	[2.009,2.492]	[2.320,2.748]	[2.951,3.229]
		长度	0.452	0.569	0.617	0.483	0.428	0.278
	MF	区间	[1.178,1.546]	[1.455,2.033]	[1.414,2.016]	[1.865,2.367]	[2.289,2.748]	[2.696,3.155]
		长度	0.367	0.578	0.602	0.503	0.459	0.459
	KNNWMF	区间	[1.373,1.767]	[1.613,2.195]	[1.442,2.045]	[1.875,2.382]	[2.211,2.682]	[2.768,3.111]
		长度	0.394	0.582	0.603	0.506	0.471	0.344
MAPE	KNNW	区间	[0.072,0.099]	[0.090,0.129]	[0.080,0.122]	[0.091,0.127]	[0.101,0.137]	[0.145,0.167]
		长度	0.027	0.040	0.043	0.036	0.036	0.022
	MF	区间	[0.060,0.084]	[0.076,0.118]	[0.073,0.116]	[0.087,0.123]	[0.105,0.141]	[0.139,0.170]
		长度	0.024	0.042	0.043	0.036	0.036	0.030
	KNNWMF	区间	[0.068,0.092]	[0.082,0.124]	[0.074,0.117]	[0.086,0.123]	[0.099,0.136]	[0.145,0.170]
		长度	0.025	0.042	0.043	0.038	0.038	0.026

4.4 基于多变量缺失的实证分析

数据集中仅有某一个观测变量存在缺失值往往不符合实际情况，因此，为了验证算法的有效性，只在单变量缺失的前提下，从不同缺失机制、不同缺失率的角度，基于 MAE、RMSE、MAPE 三种评价准则去判定算法的填补效果是远远不够的。令 $J=3,9,11$，采用前文提到的多变量完全随机缺失、多变量随机缺失、多变量非

随机缺失的方式对 Boston 数据集对应列进行挖空处理，然后使用单变量数据缺失的实验方法重新对对 KNNW、MF、KNNWMF 进行评估。

4.4.1 完全随机缺失

先考虑完全随机缺失机制下多变量数据缺失的情况，根据不同的缺失率大小，每一次程序运行所得到的含有缺失值的数据集均是随机产生的，因此，多变量缺失的情形更加符合实际问题，但也相对增加了填补过程中的不确定因素，这更有利于评估算法的优良性，实验结果如表 4-4-1 所示：

表 4-4-1：不同缺失率下、3 种填补算法在 MCAR 前提下填补 100 次误差结果的均值(多变量)

评价准则	填补算法	缺失率					
		5%	10%	15%	20%	25%	30%
MAE	KNNW	0.541	0.501	0.520	0.571	0.579	0.592
	MF	0.778	0.763	0.773	0.778	0.778	0.785
	KNNWMF	0.535	0.486	0.507	0.541	0.546	0.551
RMSE	KNNW	1.392	1.317	1.400	1.557	1.535	1.551
	MF	1.415	1.381	1.396	1.405	1.392	1.409
	KNNWMF	1.401	1.299	1.400	1.520	1.490	1.507
MAPE	KNNW	0.122	0.118	0.121	0.131	0.138	0.134
	MF	0.149	0.147	0.149	0.152	0.155	0.154
	KNNWMF	0.122	0.117	0.120	0.127	0.134	0.130

根据表 4-4-1 的实验结果显示，第一，在不同的评价准则下，随着缺失率的增大，三种填补算法的填补效果均有不同程度的下降趋势，这也表明，如果在缺失率较大的情形下，使用填补算法进行缺失值填补，并不一定得到符合预期的填补结果，也从侧面印证了数据收集过程中数据本身质量的重要性；第二，在 MAE、MAPE 评价准则下，KNNWMF 在不同缺失率下均优于 KNNW；第三，在 RMSE 评价准则下，当缺失率 $p \geq 0.15$ 时，KNNWMF 的 $\overline{RMSE}$ 结果仍优于 KNNW，且该值仅次于 MF，而当缺失率 $p \leq 0.1$ 时，由于缺失率较小，且实验带有随机性，因此 KNNWMF 在 RMSE 评价准则下并无明显优势。

同单变量完全随机缺失情形一样，本部分仍采用不同缺失率下 KNNW、MF、

KNNWMF 填补的每一次 MAE、RMSE、MAPE 结果进行箱线图绘制，结果如图 4-4-1 所示：

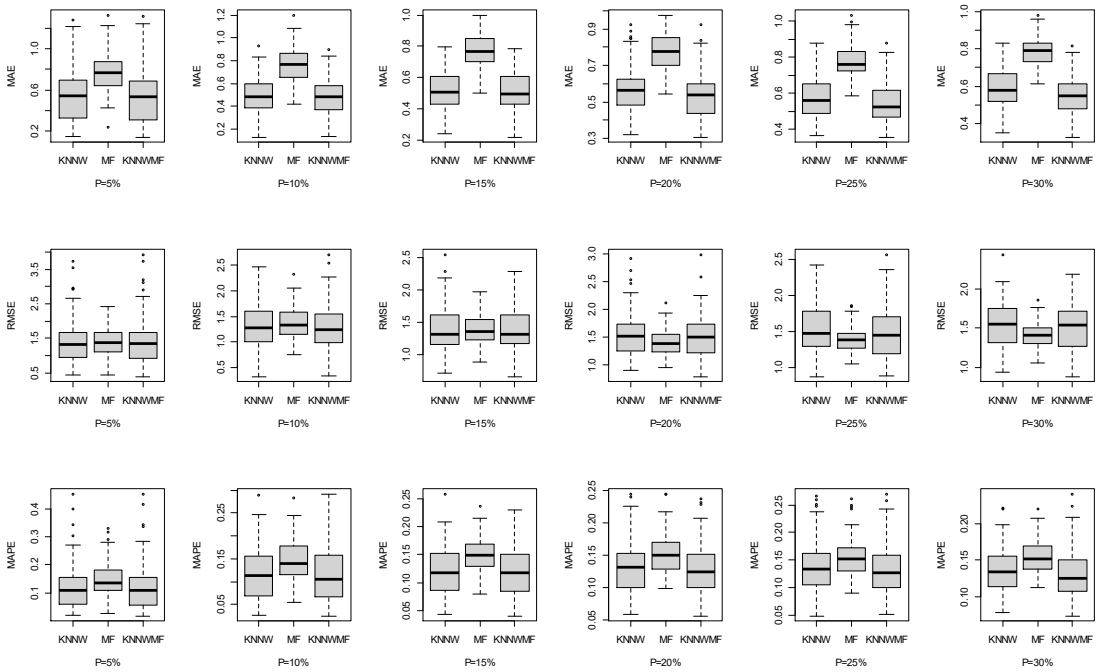

图 4-4-1 不同缺失率下、三种算法在 MCAR 前提下填补 100 次的误差结果箱线图（多变量）

图 4-4-1 的实验结果表明，第一，在不同的评价准则和缺失率前提下，MF 的四分间距均显著小于 KNNW、KNNWMF，这与单变量完全随机缺失中图 4-3-4 的第一点结论一致；第二，在不同的评价准则和缺失率前提下，三种算法对应的 100 次填补评价结果大多出现异常点，同样表明使用单次评价结果对算法的优良性进行评估是不合理的；第三，基于不同的缺失率和评价准则，采用 KNNWMF 所得的评价结果的最大值、最小值、下四分位数、中位数均小于 KNNW，且随着缺失率的增大，该现象愈发显著，这点与单变量完全随机缺失中图 4-3-4 的第三点结论一致。综上，在 MCAR 前提下，由单变量缺失推广至多变量缺失时，KNNWMF 填补效果保持稳定。

为了进一步探究三种算法在 MAE、RMSE、MAPE 评价准则下的整体优良性，同前文所述一致，分别对 MAE、RMSE、MAPE 求取不同缺失率下的 95%置信区间，实验结果如表 4-4-2。

通过分析表 4-4-2 的实验结果可以发现，第一，随着缺失率的增大，三种算法基于 MAE、RMSE、MAPE 得到的 95%置信区间长度有整体变窄的趋势，但观察各自对应的置信区间起始点发现，不同缺失率、不同评价准则下的区间起点和终点

表 4-4-2：不同缺失率下、3 种填补算法在 MCAR 前提下填补 100 次误差结果的置信区间及其长度(多变量)

评价准则	填补算法	结果类型	缺失率					
			5%	10%	15%	20%	25%	30%
MAE	KNNW	区间	[0.491,0.591]	[0.470,0.532]	[0.496,0.544]	[0.546,0.596]	[0.556,0.601]	[0.572,0.612]
		长度	0.099	0.062	0.048	0.050	0.045	0.040
	MF	区间	[0.739,0.816]	[0.735,0.790]	[0.752,0.793]	[0.759,0.797]	[0.761,0.796]	[0.770,0.799]
		长度	0.076	0.055	0.041	0.038	0.035	0.028
	KNNWMF	区间	[0.483,0.586]	[0.455,0.517]	[0.484,0.530]	[0.517,0.565]	[0.523,0.568]	[0.532,0.570]
		长度	0.103	0.062	0.047	0.047	0.045	0.039
RMSE	KNNW	区间	[1.262,1.521]	[1.231,1.403]	[1.327,1.472]	[1.479,1.635]	[1.471,1.598]	[1.494,1.608]
		长度	0.259	0.172	0.145	0.156	0.127	0.114
	MF	区间	[1.338,1.491]	[1.322,1.441]	[1.352,1.440]	[1.361,1.449]	[1.356,1.428]	[1.379,1.440]
		长度	0.154	0.119	0.088	0.088	0.072	0.061
	KNNWMF	区间	[1.266,1.536]	[1.211,1.388]	[1.329,1.471]	[1.444,1.596]	[1.423,1.558]	[1.451,1.562]
		长度	0.270	0.176	0.142	0.152	0.134	0.111
MAPE	KNNW	区间	[0.106,0.139]	[0.107,0.130]	[0.112,0.129]	[0.123,0.138]	[0.129,0.147]	[0.128,0.140]
		长度	0.032	0.023	0.017	0.016	0.018	0.013
	MF	区间	[0.137,0.160]	[0.138,0.156]	[0.143,0.155]	[0.146,0.157]	[0.148,0.161]	[0.150,0.158]
		长度	0.023	0.018	0.012	0.011	0.013	0.009
	KNNWMF	区间	[0.105,0.139]	[0.105,0.129]	[0.111,0.128]	[0.120,0.135]	[0.125,0.144]	[0.123,0.136]
		长度	0.033	0.023	0.016	0.015	0.019	0.013

会随着缺失率的增大呈现出而出现区间右移现象，综合两点进行考虑发现，随着缺失率的增大，三种算法的填补效果出现不同程度的下降，这印证了表 4-1 结论分析的第一点结论；第二，当缺失率 $p \geq 0.15$ 时，KNNWMF 基于三种评价准则得到的置信区间长度始终短于 KNNW，而当缺失率 $p \leq 0.1$ 时，则恰好相反，然而两种算法的区间长度并不存在显著差异，此外，MF 算法在不同缺失率、不同评价准则下所得的置信区间始终最窄；第三，针对不同的缺失率来看，在 MAE、MAPE 评价准则下，KNNWMF 的置信区间起点始终优于 KNNW，并显著优于 MF，其置信区间的终

点也存在相似的规律，而在 RMSE 评价准则下，当 $p \geq 0.1$ 时，该现象依然成立，只有在缺失率 $p = 0.05$ 时出现了相反结果，然而区别并不明显；综合第二、第三点结论发现，经过 MF 校准后的 KNNWMF 在不影响填补准确性的前提下继承了 MF 算法的稳定性。

4.4.2 随机缺失

通过营造不同的缺失情形来评价算法的优良性至关重要，表 4-4-3 显示了在 MAR 下，多变量数据缺失情形中三种填补算法的优良性，如下：

表 4-4-3：不同缺失率下、3 种填补算法在 MAR 前提下填补 100 次误差结果的均值(多变量)

评价准则	填补算法	缺失率					
		5%	10%	15%	20%	25%	30%
MAE	KNNW	0.631	0.655	0.690	0.873	1.032	0.898
	MF	0.791	0.838	0.817	0.875	0.913	0.877
	KNNWMF	0.595	0.590	0.613	0.666	0.728	0.687
RMSE	KNNW	1.526	1.561	1.744	2.062	2.358	2.213
	MF	1.390	1.400	1.458	1.503	1.538	1.570
	KNNWMF	1.440	1.426	1.510	1.575	1.632	1.637
MAPE	KNNW	0.143	0.148	0.151	0.219	0.226	0.170
	MF	0.152	0.165	0.156	0.180	0.177	0.160
	KNNWMF	0.137	0.139	0.138	0.172	0.165	0.143

表 4-4-3 的实验结果显示，第一，随着缺失率的增大，三种填补算法的 $\overline{MAE}$、$\overline{RMSE}$、$\overline{MAPE}$ 结果依然没有表现出严格的递增趋势，但却具备单变量随机缺失情形中的第一点结论，并在 $p = 0.25$ 时，三种算法 $\overline{MAE}$、$\overline{MAPE}$ 结果达到顶峰，而 MF、KNNWMF 的 $\overline{RMSE}$ 结果却出现了严格的递增趋势，结合计算机模拟实验的随机性，这同样表明，随着缺失率的增大，算法在逐步失效；第二，基于不同的缺失率，在 MAE、RMSE、MAPE 评价准则下，KNNWMF 的评价结果均值始终大幅度优于 KNNW，且随着缺失率的增大，这种优势更加明显；第三，由于 MF 具备优良的稳定性，且对数据的要求甚少，因此在 RMSE 评价准则下 MF 获得了最佳的结果，而

KNNWMF 仅次于 MF，通过分析发现，随着缺失率的增加，KNNWMF 的 $\overline{RMSE}$ 会逐步向 MF 结果靠拢。

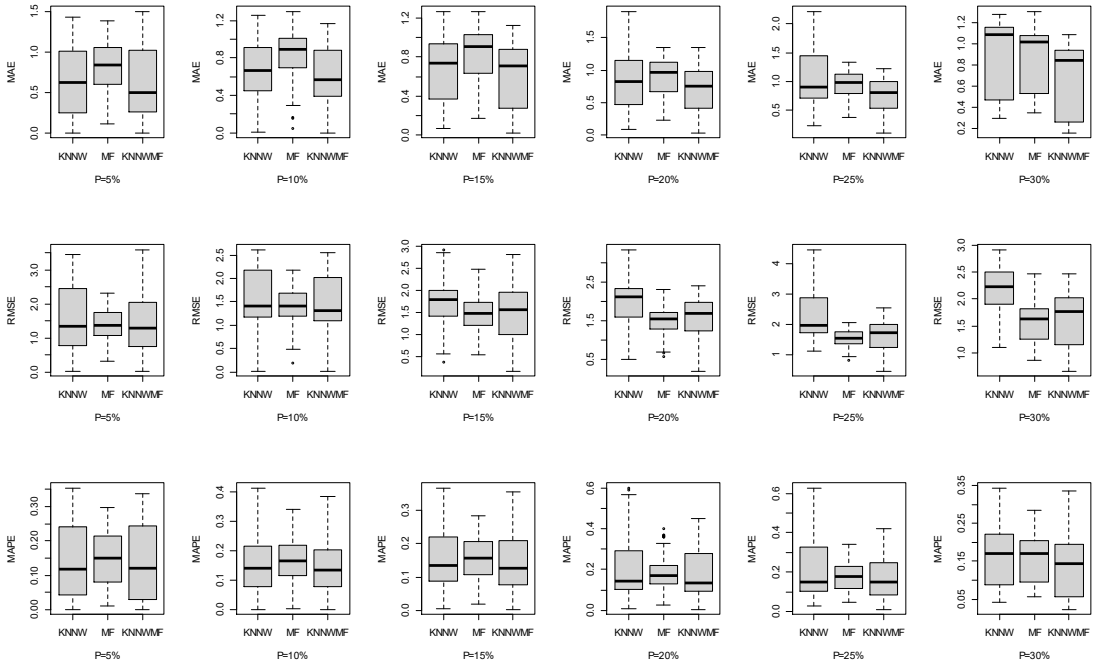

图 4-4-2 不同缺失率下、三种算法在 MAR 前提下填补 100 次的误差结果箱线图（多变量）

图 4-4-2 是多变量随机缺失前提下，基于不同的评价准则和缺失率，采用 KNNW、MF、KNNWM 填补的每一次 MAE、RMSE、MAPE 结果绘制的箱线图，图 4-3-5 的实验结果表明，第一，相较于单变量随机缺失情形，在不同的缺失率前提和评价准则下，三种算法所得结果的异常值很少，三种算法基于 MAE、RMSE、MAPE 评价结果绘制的箱线图均出现最大值、中位数、四分位数、最小值几乎不会出现重合现象，这是由于多变量缺失导致的不确定性因素增加，从而导致每次实验模拟所用的非完整数据集不完全一致所导致的，这样的结果更具有代表性；第二，基于不同缺失率和评价准则，KNNWMF 对应的评价结果最大值、中位数、四分位数、最小值均要小于或等于 KNNW，且随着缺失率的增大，该现象愈发明显；第三，在不同的缺失率和评价准则下，MF 的四分间距始终小于 KNNW 和 KNNWMF，且在缺失率 $P \geq 0.1$ 时，MF 在 RMSE 评价准则下的中位数结果达到最优，KNNWMF 的中位数始终仅次于 MF。综上，在多变量随机缺失机制下，KNNW、KNNWMF 的准确性优于 MF，而 MF 的稳定性表现相对较优，且 KNNWMF 继承了 KNNW 的准确性和 MF 的稳定性，尤其在缺失率较大时，填补性能表现优异。

在多变量随机缺失情形下，同样考虑不同缺失率前提下，三种评价结果 95%

置信区间及区间长度，实验结果如表 4-4-4：

表 4-4-4：不同缺失率下、3 种填补算法在 MAR 前提下填补 100 次误差结果的置信区间及其长度(多变量)

评价准则	填补算法	结果类型	缺失率					
			5%	10%	15%	20%	25%	30%
MAE	KNNW	区间	[0.548,0.715]	[0.593,0.716]	[0.627,0.752]	[0.771,0.976]	[0.926,1.139]	[0.831,0.965]
		长度	0.167	0.123	0.125	0.205	0.213	0.134
	MF	区间	[0.729,0.853]	[0.788,0.888]	[0.761,0.872]	[0.817,0.933]	[0.861,0.964]	[0.823,0.930]
		长度	0.124	0.100	0.111	0.115	0.103	0.107
	KNNWMF	区间	[0.510,0.680]	[0.531,0.650]	[0.549,0.678]	[0.597,0.736]	[0.663,0.793]	[0.623,0.750]
		长度	0.171	0.119	0.129	0.140	0.130	0.127
RMSE	KNNW	区间	[1.334,1.718]	[1.436,1.686]	[1.644,1.844]	[1.930,2.195]	[2.180,2.536]	[2.133,2.294]
		长度	0.384	0.250	0.199	0.265	0.356	0.161
	MF	区间	[1.302,1.479]	[1.327,1.473]	[1.382,1.534]	[1.436,1.569]	[1.483,1.593]	[1.500,1.640]
		长度	0.178	0.147	0.152	0.133	0.110	0.140
	KNNWMF	区间	[1.248,1.633]	[1.298,1.555]	[1.384,1.637]	[1.464,1.685]	[1.535,1.729]	[1.537,1.738]
		长度	0.385	0.257	0.253	0.220	0.194	0.201
MAPE	KNNW	区间	[0.122,0.164]	[0.131,0.166]	[0.133,0.169]	[0.185,0.253]	[0.190,0.261]	[0.154,0.186]
		长度	0.042	0.035	0.036	0.069	0.071	0.033
	MF	区间	[0.137,0.168]	[0.150,0.180]	[0.142,0.170]	[0.161,0.199]	[0.162,0.192]	[0.148,0.172]
		长度	0.031	0.030	0.028	0.037	0.030	0.024
	KNNWMF	区间	[0.116,0.158]	[0.122,0.156]	[0.120,0.156]	[0.148,0.197]	[0.144,0.186]	[0.126,0.160]
		长度	0.042	0.034	0.035	0.049	0.042	0.034

根据表 4-4-4 的实验结果，第一，在不同缺失率前提下，基于三种评价结果所得到的置信区间长度整体不存在递增或递减的趋势，然而，随着缺失率的变化，MF 的置信区间长度相对稳定，而 KNNW 的置信区间长度波动严重，KNNWMF 的置信区间长度波动情况优于 KNNW；第二，在三种评价准则下，KNNWMF 的置信区间起点值和终点值始终优于 KNNW，且缺失率越大，该现象越明显，才外，随着缺失率的增大，KNNWMF 的置信区间长度明显小于 KNNW；综合这两点结论，经过 MF 校准后的 KNNWMF，由于采用迭代法进行逐步填充的缘故，不仅使该算法兼具 MF、KNNW

的优点，且在缺失率较大时，这种优势会更加明显，在随机缺失机制下依然适用于缺失率较大的样本集。

4.4.3 非随机缺失

表 4-4-5 显示了在多变量非随机缺失情形下 KNNW、MF、KNNWMF 的三种评价结果均值，以此来评估三种算法在非随机缺失机制下的填补效果。

表 4-4-5：不同缺失率下、3 种填补算法在 NMAR 前提下填补 100 次误差结果的均值(多变量)

评价准则	填补算法	缺失率					
		5%	10%	15%	20%	25%	30%
MAE	KNNW	1.263	1.608	1.713	1.815	2.177	2.223
	MF	1.211	1.299	1.529	1.642	1.885	1.923
	KNNWMF	1.145	1.411	1.498	1.547	1.856	1.862
RMSE	KNNW	2.100	2.698	2.699	2.910	3.345	3.434
	MF	1.711	1.892	2.173	2.331	2.615	2.691
	KNNWMF	1.863	2.349	2.334	2.400	2.803	2.885
MAPE	KNNW	0.223	0.338	0.321	0.379	0.440	0.400
	MF	0.220	0.276	0.280	0.325	0.383	0.350
	KNNWMF	0.214	0.302	0.285	0.327	0.381	0.344

根据表 4-4-5 的实验结果，第一，同单变量非随机缺失一样，在多变量非随机缺失情形下，三种填补算法基于不同评价准则下的结果均值出现了显著变大的情况，尤其是基于 MAPE 评价准则的结果均值在缺失率较高时达到了 0.4 的水平，这代表 KNNW、MF、KNNWMF 在随机缺失机制下几乎完全失效；第二，在情形下讨论三种算法间的填补优势依然有必要，根据不同的缺失率所得的 $\overline{MAE}$、$\overline{RMSE}$、$\overline{MAPE}$ 结果会随着缺失率的增大依然呈现出递增趋势，因随机模拟的缘故，在缺失率较小时，如 $p=0.1,0.15$，可能会出现类似于 $\overline{MAPE}$ 结果抖动的现象发生，但整体的递增趋势保持不变，这表明，填补算法的效果会随着缺失率的增加而逐步变差，这点结论与完全随机缺失和随机缺失的结论一致；第三，在不同的缺失率前提下，KNNWMF 的 $\overline{MAE}$、$\overline{RMSE}$、$\overline{MAPE}$ 结果始终小于 KNNW，且随着缺失率的增加，这种优势会更加明显，这表明，KNNW 经过 MF 校准以及迭代法改进后的 KNNWMF

所具备的稳定性不会因为缺失机制的改变而发生变化，再次表明了该算法在结构上的稳定性。

现在使用多变量非随机缺失机制下三种算法的每一次 MAE、RMSE、MAPE 评价结果绘制箱线图，结果如图 4-4-3：

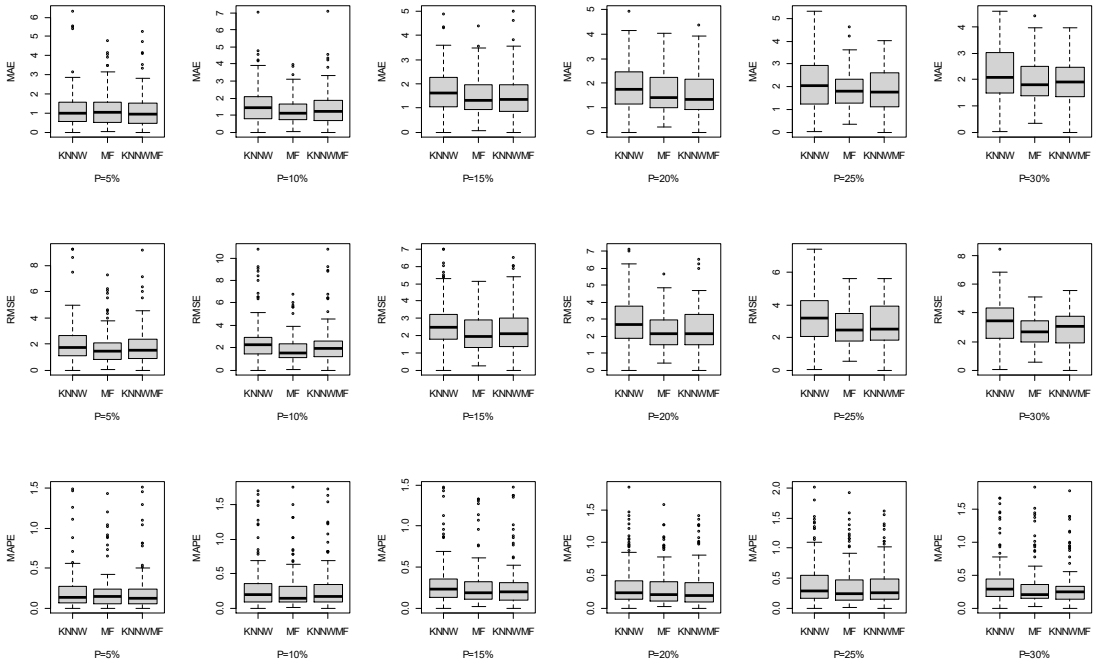

图 4-4-3 不同缺失率下、三种算法在 MAR 前提下填补 100 次的误差结果箱线图（多变量）

根据图 4-4-3 的实验结果显示，第一，基于不同的缺失率和评价准则，KNNWMF 的最大值、中位数、四分位数、最小值均小于 KNNW，且箱体长度也接近或小于 KNNW，而在 RMSE 评价准则下，MF 的最大值、中位数、四分位数以及四分间距均小于其他两种算法，且随着缺失率的增大，上述现象愈发明显；第二，基于不同的缺失率和评价准则，三种算法对应的最大值、中位数、四分位数相较于 MCAR、MAR 几乎全部出现增大现象，且出现大量的异常点，这表明在 NMAR 下，三种算法几乎完全失效。综上，尽管在多变量非随机缺失下三种算法均面临失效，但 KNNWMF 在由单变量缺失推广至多变量缺失时依然具备一定的算法优势。

同前文所述一致，考虑在多变量非随机缺失下三种算法基于不同评价准则的 95%置信区间，结果如表 4-4-6：

表 4-4-6 的实验结果显示，第一，不同缺失率下，三种算法所得的 MAE、RMSE、MAPE 评价结果的置信区间长度没有明显的规律特征，这可能与非随机缺失下 KNNW、MF、KNNWMF 不适用有关，尽管如此，MF 在不同的缺失率前提下依然获得了较为稳

定的置信区间长度；第二，观察不同算法在不同评价准则下，基于不同缺失率前提所获得的置信区间起点和终点来看，KNNWMF 依然具备优良的填补效果，结果与多变量完全随机缺失、多变量随机缺失两种情形下基本一致，此处不在赘述。

表 4-4-6：不同缺失率下、3 种填补算法在 NMAR 前提下填补 100 次误差结果的置信区间及其长度(多变量)

评价准则	填补算法	结果类型	缺失率					
			5%	10%	15%	20%	25%	30%
MAE	KNNW	区间	[1.035,1.491]	[1.370,1.845]	[1.513,1.913]	[1.625,2.005]	[1.944,2.410]	[2.019,2.428]
		长度	0.456	0.475	0.400	0.380	0.466	0.410
	MF	区间	[1.024,1.398]	[1.136,1.462]	[1.369,1.688]	[1.480,1.804]	[1.714,2.056]	[1.763,2.084]
		长度	0.374	0.326	0.318	0.324	0.343	0.320
	KNNWMF	区间	[0.946,1.345]	[1.189,1.632]	[1.310,1.687]	[1.373,1.720]	[1.661,2.052]	[1.700,2.024]
		长度	0.399	0.443	0.376	0.347	0.392	0.324
RMSE	KNNW	区间	[1.756,2.444]	[2.280,3.116]	[2.402,2.997]	[2.612,3.208]	[3.022,3.669]	[3.132,3.736]
		长度	0.689	0.836	0.594	0.597	0.647	0.603
	MF	区间	[1.434,1.988]	[1.636,2.148]	[1.954,2.392]	[2.111,2.551]	[2.402,2.829]	[2.491,2.890]
		长度	0.554	0.512	0.439	0.441	0.427	0.399
	KNNWMF	区间	[1.559,2.166]	[1.958,2.739]	[2.053,2.615]	[2.148,2.653]	[2.547,3.060]	[2.655,3.114]
		长度	0.608	0.781	0.563	0.504	0.513	0.459
MAPE	KNNW	区间	[0.168,0.278]	[0.260,0.415]	[0.256,0.385]	[0.303,0.456]	[0.356,0.525]	[0.325,0.475]
		长度	0.110	0.155	0.128	0.153	0.169	0.150
	MF	区间	[0.166,0.274]	[0.210,0.341]	[0.222,0.339]	[0.262,0.389]	[0.305,0.460]	[0.278,0.423]
		长度	0.108	0.131	0.117	0.128	0.154	0.145
	KNNWMF	区间	[0.156,0.273]	[0.230,0.375]	[0.223,0.346]	[0.258,0.396]	[0.308,0.454]	[0.276,0.413]
		长度	0.117	0.145	0.122	0.138	0.146	0.137

4.5 本章小结

针对加权 K 近邻填补因算法结构本身的特点而导致的填补效果欠佳问题，本章提出混合迭代填补的方式对非完整数据集进行填补处理，尝试使该算法同时继承加权 K 近邻和缺失森林两种算法各自的优点，实证分析结果显示，第一，

KNNWMF 在完全随机缺失和随机缺失机制下，基于不同的缺失率前提，均发现，在填补效果方面继承了加权 K 近邻算法，而在稳定性方面继承了缺失森林算法，填补效果明显优于 KNNW；第二，随着缺失率的增大，三种填补算法的有效性均有所下降，然而在缺失率较大的情形下，KNNWMF 填补优势更加显著；第三，在非随机缺失前提下，基于不同的缺失率，三种填补方法虽然面临失效，但 KNNWMF 所继承的有效性和稳定性依然存在。

5 总结与展望

5.1 工作总结及主要创新

在当今社会，日常生活中数据所占据的主导地位已愈发凸显，数据的质量和完整性将是统计分析的基石，也是决策者做出正确判断的前提条件。而数据缺失问题又广泛存在社会调查、医学等领域，如何有效的解决缺失值填补问题，从而使样本数据的价值最大化的呈现出来是人们关注的焦点，也是业界、学界研究的重点方向之一。

本文基于三种缺失机制，在不同的缺失率下，针对单变量缺失和多变量缺失两种情形展开研究，通过实证分析的方式进行实验模拟，以求最大限度还原现实生活中复杂的问题情形。更为具体的是，本文先对学者提及的三种缺失机制进行详细的数学描述之后，在此基础上，对不同缺失机制下的不同缺失情形做了深入探讨（详见第二章）；针对不同的缺失率，选用KNN和KNNW进行填补处理，进而在三种评价准则下评估算法的优良性，并引入均值填补、中位数填补两种传统填补法进行对比分析（详见第三章）；更进一步的，针对KNNW算法结构自身缺陷进行改进，提出混合迭代填补的思想，使之继承父辈算法的所有优势（详见第四章）。

本文的特点在于考虑到真实的填补问题具有随机性，而在研究过程中，为了摒弃随机性所带来的分析不确定性问题，采用公共数据集在对每一种算法基于不同的评价准则进行评估时，将每一种缺失机制、每一个缺失率前提下的单变量缺失情形和多变量缺失情形重复运行多次，并对运行次数的多少进行了简要分析，然后使用多次程序运行得到的评价准则的均值、方差、置信区间等作为算法优劣的对比依据，其主要创新点如下：

（1）KNN中采用距离公式度量待分析样本点与每一个完全观测样本点的距离之后，针对k值的大小，本文采用交叉验证法对不同缺失率前提下的单变量缺失情形和多变量缺失情形做了深入分析，根据具体观测值分析相对最优解，避免了算法默认k值或固定k值对当前缺失数据填补所带来的影响；

（2）尽管如此，k值优化仍然无法有效解决过大或过小所带来的问题，KNNW虽然在一定程度上缓解了该问题，但面对实际问题中每个含缺失值样本点到其他完全观测样本点距离集合存在差异的情况，为k个近邻采用固定的权重分配方案显然不合理，本文针对此问题提出动态调参法，为每一个含缺失值样本点的k个

近邻提供最合适的权重分配方案，且该方法针对不同的样本集和权重函数有一定的通用性。

（3）针对KNNW过分依赖最邻近样本点从而导致稳定性下降的问题，本文提出采用MF对其每一个填补值进行校准,而面对填补算法随缺失率增大而出现的填补效果存在不同程度降低的问题，本文提出迭代法进行处理，并使非完整数据集的缺失率逐步降低至0%，基于这两种方法的混合填补思想是一个通用框架，为今后的填补算法研究工作提供了新的思路。

5.2 未来研究展望

面对现实生活中复杂的缺失情形及不同特征的数据形式，在未来的研究中，依托于填补精度和稳定性两个方面，进一步开展以下几个方面的工作：

（1）针对NMAR下填补算法全面失效的问题，尝试采用其他填补算法进行处理，或是考虑缺失变量的先验分布信息，结合本文提到的填补算法和改进思想对填补值进行调整，从而应对NMAR下数据缺失所带来的影响。

（2）针对MF的参数优化问题，本文没有进行深入讨论，在后续的研究中，尝试对其核心参数的进行优化，探索该算法中新的参数优化途径；针对KNNWMF中迭代次数与填补效果之间的关系，本文没有做更进一步的分析，而面对迭代次数越多，算法时间开销越大的问题，本文也没有进行相应的优化，在后续的研究工作中应予以解决。

（3）本文所采用的数据集为公共数据集，在实验过程中没有对其进行中心化或标准化处理，这对现实生活中多样化的数据形式研究略显不足，在后续研究中，应弱化数据本身的内在含义，并采用不同数据集或不同模拟数据集对算法的填补效果进行分析，并进一步总结不同填补算法的应用场景，以此为真实问题提供理论参考。

参考文献

[1]帅平,李晓松,周晓华,刘玉萍.缺失数据统计处理方法的研究进展[J].中国卫生统计,2013,30(01):135-139+142.
[2]Allison P D,Missing data.[J].Thousand Oaks Ca Sage Quantitative Applications in the Social Sciences,2009,17(9):285-314.
[3]Schafer,Joseph L,Graham,John W.Missing data: Our view of the state of the art.[J].Psychological Methods,2002,7(2):147-177.
[4]鲍晓蕾,高辉,胡良平.多种填补方法在纵向缺失数据中的比较研究[J].中国卫生统计,2016,33(01):45-48.
[5]李琳,杨红梅,杨日东,胡珊,张学良,周毅.基于临床数据集的缺失值处理方法比较[J].中国数字医学,2018,13(04):8-10+80.
[6]邓建新,单路宝,贺德强,唐锐.缺失数据的处理方法及其发展趋势[J].统计与决策,2019,35(23):28-34.
[7]Molenberghs G,Kenward M.Missing Data in Clinical Studies[J].Journal of the Royal Statistical Society,2010,171(4):1039-1040.
[8]Linder J A,Ma J,Bates D W.Electronic health record use and the quality of ambulatory care in the United States[J].Archives of internal medicine,2007,167(13):1400-1405.
[9]宋亮,万建洲.缺失数据插补方法的比较研究[J].统计与决策,2020,36(18):10-14.
[10]刘佳星,张宏烈,刘艳菊,刘彦忠.基于缺失率的不完整数据填补算法[J].统计与决策,2021,37(02):39-41.
[11]孙玲莉,董世杰,杨贵军.常用多重插补法的插补重数选择[J].统计与决策,2019,35(23):5-10.
[12]Graham J W.Missing Data: Analysis and Design[M].Heidelberg: Springer,2012.
[13]杨贵军,李小峰,王清.双无回答层抽样的三重抽样比率估计量[J].应用数学学报,2015,38(02):366-378.
[14]Zhang Shaodian,Kang Tian,Zhang Xingting,Wen Dong,Elhadad Noemie,Lei Jianbo.Speculation Detection for Chinese Clinical Notes:Impacts of Word Segmentation and Embedding Models[J].Journal of biomedical informatics,2016(60):334-341.
[15]Little R,Rubin D.Statistical Analysis With Missing Data[M].New York:Wiley and Sons Inc,1987.
[16]Little R,Rubin D.Statistical Analysis With Missing Data[M].New York:John Wiley and Sons,2002.
[17]李春林,高玉鹏,李圣瑜.不完全数据多重插补的Bootstrap方差估计[J].统计与决策,2017(18):74-76.
[18]Little R,Roderick J A,Rubin D.Statistical analysis with missing data[J].Technometrics,2002,45(4):364-365.
[19]Enders C K.Applied missing data analysis[M]. England: Guilford press,2010.
[20]S Chevret,S Seaman,M Resche-Rigon.Multiple imputation:a mature

approach to dealing with missing data[J].Intensive care medicine, 2015, 41(2):348-350.
[21]N J Horton;N M Laird.Maximum likelihood analysis of generalized linear models with missing covariates[J].Statistical methods in medical research,1999,8(1):37-50.
[22]PAUL D.ALLISON.Multiple Imputation for Missing Data:A Cautionary Tale[J].Sociological Methods and Research,2000,28(3):301-309.
[23]严洁.缺失数据的多重插补[M].重庆: 重庆大学出版社,2017.
[24]张成萍.残缺数据的填补[D].长沙: 中南大学, 2006.
[25]杨晓倩.缺失数据插补方法的选择研究[D].兰州财经大学,2016.
[26]金勇进,邵君.缺失数据的统计处理[M].北京：中国统计出版社, 2009.
[27]庞新生.缺失数据插补处理方法的比较研究[J].统计与决策,2012(24):18-22.
[28]O Troyanskaya,M Cantor,G Sherlock,P Brown,T Hastie,R Tibshirani,D Botstein,R B Altman. Missing value estimation methods for DNA microarrays [J].Bioinformatics,2001,17(6):520- 525.
[29]R Malarvizhi,Antony Selvadoss Thanamani.K-Nearest Neighbor in Missing Data Imputation[J].International Journal of Engineering Research and Development,2012,5(1).
[30]Jianli Xiao.SVM and KNN ensemble learning for traffic incident detection[J].Physica A: Statistical Mechanics and its Applications, 2019, 517(0):29-35.
[31]Purwar,Archana,Singh Sandeep Kumar.Hybrid prediction model with missing value imputation for medical data[J].Expert Systems with Applications,2015,42(13):5621-5631.
[32]郑智泉,王孟孟,田维琦.基于加权K近邻算法的缺失数据填补研究[J].智能计算机与应用,2021,11(11):31-33+42.
[33]Tutz G, Ramzan S. Improved methods for the imputation of missing data by nearest neighbor methods[J]. Computational Statistics & Data Analysis, 2015,90:84-99.
[34]De Silva H, Perera A S. Missing data imputation using Evolutionary k-Nearest neighbor algorithm for gene expression data[C].2016 Sixteenth International Conference on Advances in ICT for Emerging Regions (ICTer), IEEE,2016:141-146.
[35]杨日东,李琳,陈秋源,周毅.LKNNI:一种局部K近邻插补算法[J].中国卫生统计,2019,36(05):780-783.
[36]Ma Liyao,Destercke Sebastien,Wang Yong.Online active learning of decision trees with evidential data[J].Pattern Recognition,2016,52(C): 33-45.
[37]Dietterich,Thomas G.An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging,Boosting,and Randomi -zation[J].Machine Learning, 2000, 40(2):139-157.
[38]L B Statistics,L Breiman.Random forests[J].Machine Learning, 2001, 45(1):5-32.
[39]方匡南,吴见彬,朱建平,谢邦昌.随机森林方法研究综述[J].统计与信息论

坛,2011,26(03):32-38.
[40]沈琳,胡国清,陈立章,谭红专.缺失森林算法在缺失值填补中的应用[J].中国卫生统计,2014,31(05):774-776.
[41]武晓岩,李康.随机森林方法在基因表达数据分析中的应用及研究进展[J].中国卫生统计,2009,26(04):437-440.
[42]Stekhoven,Daniel J,Buehlmann,Peter.Missforest-Non-parametric miss-ing value imputation for mixed-type data[J].Bioinformatics,2012,28(1):112-118.
[43]陈婉娇.缺失数据插补方法及其在医学领域的应用研究[D].华南理工大学,2019.
[44]孟杰,李春林.基于随机森林模型的分类数据缺失值插补[J].统计与信息论坛,2014,29(09):86-90.
[45]Schafer,Joseph L,Graham,John W.Missing data: Our view of the state of the art[J].Psychological Methods,2002,7(2):147-177.
[46]Rubin,Donald B.Multiple Imputation for Nonresponse in Surveys[M].New York: JohnWiley&Sones,2002.
[47]Rubin, Donald B.Multiple Imputation After 18+ Years[J].Journal of the American Statistical Association,1996,91(434):473-489.
[48]Lin T H.A comparison of multiple imputation with EM algorithm and MCMC method for quality of life missing data[J].Quality and Quantity, 2010, 44(2): 277-287.
[49]Siddiqui O. MMRM versus MI in dealing with missing data—a comparison based on 25 NDA data sets[J].Journal of biopharmaceutical statistics, 2011, 21(3): 423-436.
[50]Mattei A. Estimating and using propensity score in presence of missing background data: an application to assess the impact of childbearing on wellbeing[J]. Statistical Methods and Applications,2009,18(2):257-273.
[51]Little R.Missing-data adjustments in large surveys[J].Journal of Business and Economic Statistics,1988,6(3):287-296.
[52]杨贵军,李静华.基于PMM多重插补法的线性模型系数估计量的模拟研究[J].数量经济技术经济研究,2014,31(10):139-150.
[53]J Honaker,Gary King.What to do about missing values in time-series cross-section data[J].American Journal of Political Science,2010,54(2):561-581.
[54]杨贵军,孙玲莉,孟杰.基于EMB多重插补法的线性模型系数估计量的模拟研究[J].数量经济技术经济研究,2016,33(10):128-141.
[55]杨贵军,骆新珍.基于DA插补法的线性回归模型系数估计值的模拟研究[J].统计与信息论坛,2014,29(03):3-8.
[56]A Dempster,N Laird,Rubin.Maximum Likelihood From Incomplete Data Via The EM algorithm[J].Journal of the Royal Statistical Society Series B (Statistical Methodology),1977,39(1):1-38.
[57]孙玲莉,董世杰,杨贵军.常用多重插补法的插补重数选择[J].统计与决策,2019,35(23):5-10.

致　谢

时光荏苒，岁月如梭，转眼间我在贵州民族大学数据科学与信息工程学院的硕士研究生学习已接近尾声。在这三年当中，我不仅学到了本专业的知识，而且有缘结识了一批良师益友。在此，请允许我表达我深深的谢意。

首先感谢我的导师黄介武教授，他渊博的专业知识和严谨的科学态度深深感染了我，不仅教会我怎样去发现问题和解决问题，而且指引我如何去思考，更重要的是，在日常生活方面体现出的苦恼与困惑，黄老师仍然会为我传道、授业、解惑。感谢李荣老师兢兢业业为学院研究生所付出的一切，在您的引领下，我们才能一步步顺利完成学业。

感谢贵州民族大学数据科学与信息工程学院的各位领导和老师们的教诲，感谢在中期检查、预答辩等研究生培养环节中提出宝贵意见的答辩专家。

感谢同门王孟孟、田维琦两位同学，不管是小组讨论时给出的意见，还是闲聊谈心时的突发奇想，亦或是学习生活上的互相帮助，无数的生活琐事构成了三年的点点滴滴，三年的同门友谊有太多可以值得去回忆的地方。还要感谢同门的师兄师姐、师弟师妹以及所有研究生同学们，因为有你们，研究生阶段的学习才充斥着欢声与笑语。

最后还要感谢我的家人们，感谢你们一直以来对我的支持和理解。

仅以此文献给关爱、支持、包容和鼓励我的亲人、师长、同学和朋友们！

在校期间科研成果

郑智泉，2019 级统计学研究生，主要研究方向为统计模拟与统计计算，在校期间发表论文如下：

[1]郑智泉,王孟孟,田维琦.基于加权 K 近邻算法的缺失数据填补研究[J].智能计算机与应用,2021,11(11):31-33+42.

[2]郑智泉,杨楠.智能革命下数据驱动的智慧图书馆建设分析[J].智能计算机与应用,2020,10(08):183-185+191.

[3]郑智泉.基于不同缺失率的数据填补算法稳定性研究[J]统计与决策，2021 年 12 月已录用。

附　录

本文中涉及到的部分变量符号说明：

M：待分析样本集

M_o：进行随机排序后的 M

$M_{ij}, i \in [1,n], j \in [1,m]$：样本集中第 i 行第 j 列元素，n 代表样本点数量，m 代表变量总数

$M_{\zeta}^{mis}, \zeta \in [1, n_{mis}]$：样本集 M 中含有缺失值的样本点集合的第 ζ 个样本点

$M_{\gamma}^{obs}, \gamma \in [1, n - n_{mis}]$：样本集 M 中不含有缺失值的样本点集合的第 γ 个样本点

$D_{\left(M_{\zeta}^{mis}, M_{\gamma}^{obs}\right)}$：代表 M_{ζ}^{mis} 和 M_{γ}^{obs} 两个样本点之间的距离大小

D_{γ}：代表当前 M_{ζ}^{mis} 到所有 M_{γ}^{obs} 之间的距离集合

$D_l, l \in [1,k]$：代表当前 M_{ζ}^{mis} 到 k 个近邻 M_l^{obs} 之间的距离集合，且 $D_1 \le D_2 \le \cdots \le D_k$

X_j：样本集 M 中第 j 个观测变量

p：代表缺失率

n_{obs}：代表不含缺失值的样本点数量

n_{mis}：代表含有缺失值的样本点数量

Φ：代表样本集中缺失值的总数量

i_{mis}：代表 M 中含有缺失值的样本点的 i 值集合

i_{obs}：代表 M 中不含缺失值的样本点的 i 值集合

$J, J \subseteq [1,m]$：代表 M 中含有缺失值的观测变量的 j 值集合

X_J：样本集 M 中含有缺失值的观测变量的集合

$X_J^{\omega}, \omega \in (1, h(J))$：为 X_J 中的第 ω 个含缺失值的变量

y_J^{ω}：变量 X_J^{ω} 中的所有观测值

n_{mis}^{ω}：变量 X_J^{ω} 中含缺失值的个数

n_{mis}^{σ}：代表 n_{mis}^{ω} 中最小的数值

p^{ω}：变量 X_J^{ω} 中含缺失值的占比

W^{ω}：代表变量 x_j^{ω} 中所含缺失值的个数占 n_{mis} 的比重

$i_{mis}^{\zeta}, \zeta \in [1, n_{mis}]$：$i_{mis}^{\zeta}$ 代表 i_{mis} 中的第 ζ 个值

τ：代表当前样本点中含缺失值的观测变量对应的 j 值集合

Ω：代表除 M 中观测变量以外与 M 中观测变量有关的还未观测到的变量集合

X_t：X_t 为 Ω 中的其中一个变量

y_t：代表 X_t 中所有观测值的一个子集

$y_t^{\zeta}, \zeta \in [1, n_{mis}]$：代表 y_t 中的第 ζ 个元素

$X_{\varphi}, \varphi \in [[1, m] - J]$：其中 φ 代表不包含缺失值的变量对应的 j 值

y_j：代表 X_j 中的所有观测值

$g(\cdot)$：代表取整函数

$f(\cdot)$：代表等概率抽样函数

$h(\cdot)$：代表计算向量长度函数

$\sigma(\cdot)$：代表取最小值函数

i_o：代表对 X_{φ} 中观测值进行排序并按序逐一取出该观测值排序前在 M 中所对应的 i 值

$i_o^{\beta}, \beta \in [1, z]$：代表 i_o 的第 β 个子集

n^{β}：表示 i_o^{β} 中元素个数

n_{obs}^{β}：表示 i_o^{β} 中未被抽取的元素个数

i_{mis}^{β}：表示 i_o^{β} 中被抽取的元素集合

n_{mis}^{β}：表示 i_{mis}^{β} 中未被抽取的元素个数

$\hat{y} = \{\hat{y}_1, \hat{y}_2, \hat{y}_3, \ldots, \hat{y}_{\Phi}\}$：代表算法填补值

$y = \{y_1, y_2, y_3, \ldots, y_{\Phi}\}$：代表真实值

N：代表实验重复执行次数

MAE：代表平均绝对误差

$RMSE$：代表均方根误差

$MAPE$：代表平均绝对误差百分比

MAE_{λ}：代表第 λ 次实验所得的 MAE 值

$RMSE_{\lambda}$：代表第 λ 次实验所得的 $RMSE$ 值

$MAPE_{\lambda}$：代表第 λ 次实验所得的 $MAPE$ 值

$\overline{MAE}$：代表实验重复执行所得 N 次 MAE 结果的均值

$\overline{RMSE}$：代表实验重复执行所得 N 次 $RMSE$ 结果的均值

$\overline{MAPE}$：代表实验重复执行所得 N 次 $MAPE$ 结果的均值

$w_l, l \in [1, k]$：代表第 l 个近邻点的权重值