Statistics and Application 统计学与应用, 2018, 7(2), 210-220
Published Online April 2018 in Hans. http://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2018.72025

Hans汉斯

An Empirical Study on Maximum Likelihood Estimation of Missing Data by EM Algorithm

Lei Li, Aixiang Chen, Zanjie Yao

College of Statistics and Mathematics, Guangdong University of Finance & Economics, Guangzhou Guangdong
Email: 943770426@qq.com

Received: Apr. 1st, 2018; accepted: Apr. 21st, 2018; published: Apr. 28th, 2018

Abstract

The maximum likelihood estimation of missing data by EM is a basic method to deal with missing data. In this paper based on the maximum likelihood estimation method based on EM, was filled with the EM method for a class of random missing data sets; the results show that in 10%, 20%, 30%, three different loss rate, the relative error of EM method to fill up the missing data is less than 0.1, showing a high accuracy under the condition of low the loss rate. Further, the EM method is applied to the missing data in the actual questionnaire survey, and the influence of the survey results before and after the complement is analyzed.

Keywords

EM Algorithm, Maximum Likelihood Estimation, Missing Data, Random Deletion

EM算法对缺失数据极大似然估计的实证研究

黎　镭，陈蔼祥，姚赞杰

广东财经大学，统计与数学学院，广东　广州
Email: 943770426@qq.com

收稿日期：2018年4月1日；录用日期：2018年4月21日；发布日期：2018年4月28日

摘　要

用EM方法对缺失数据进行极大似然估计是处理缺失数据的一种基本方法。本文在介绍基于EM的极大似然估计方法基础上，对一类随机缺失数据集用EM方法进行补齐，结果表明在10%、20%、30%三种不同缺失率下，EM方法进行缺失数据补齐的相对误差均小于0.1，呈现出低缺失率情况下的高准确性的特

文章引用：黎镭，陈蔼祥，姚赞杰．EM 算法对缺失数据极大似然估计的实证研究[J]．统计学与应用，2018，7(2): 210-220. DOI: 10.12677/sa.2018.72025

点。进一步将EM方法应用于对实际问卷调查中的缺失数据进行补齐，对补齐前后调查结果的影响进行了分析。

关键词

EM算法，极大似然估计，缺失数据，随机缺失

1. 引言

数据缺失是数据分析处理中的常见问题，广泛存在于社会科学、流行病学、统计学、生物学、计算机科学等领域。一个典型的数据缺失的例子是机器学习领域的推荐系统[1] [2] [3]：要求从不完全的用户评价数据中为给定用户自动推荐感兴趣的商品列表。

大多数算法和统计学方法对数据的推断是建立在训练数据无缺失的情况下,如果将缺失数据直接作用于这些方法并不合适。因此对缺失数据的处理的一种常见的做法是将有缺失的记录丢弃，但这会减少可能本来就不多的样本数,从而达不到相应的估计精度。此外,数据中有缺失的记录和无缺失的记录可能彼此之间存在显著差异,这意味着丢弃缺失的记录会直接影响到统计推断的结果,使估计出现严重偏差或无效,甚至得到相反的结果。

对缺失数据的处理,目前已有不少的研究[4] [5]。其中基于 EM 的缺失数据处理方法[6]是一种通用的不完全数据下的极大似然估计(简称 EM 方法，下同)，该方法具有良好的收敛性以及每次迭代都能使似然函数值单调不减的良好性质。并且在满足随机缺失的假定下，该方法被证明可以从不完全数据中得到无偏估计[7]。因此大多数与缺失数据有关的问题中都会结合 EM 算法来分析。

国内对 EM 算法的理论研究较少，主要集中在应用研究方面：李顺静运用 EM 算法补齐重庆市居民的交通起止点调查表中的缺失数据，很好的展现了该算法的价值[8]；谷海彤等利用 EM 插补算法计算缺失值的插补值，并作为多重插补的初始值，得到的贝叶斯线性回归的 DA 多重插补结果比 EM 插补误差更低[9]。

本文工作主要分成两部分：首先利用 EM 方法对缺失率分别为 10%、20%、30%的随机缺失数据进行自动补齐，在此基础上，将该方法应用于实际调查数据-康华医院妇产中心调查问卷中的缺失数据进行插补,并分析了当二级指标的权重不同时，补齐数据相对于未补齐数据的总体满意度变化情况。

下文内容安排如下，第 1 节介绍了 EM 方法基本原理和计算步骤，并给出了个算例。第 2 节介绍 EM 算法补齐实验室缺失数据中的应用，分析比较了 EM 算法在不同随机缺失率下的性能,并将 EM 算法应用于实际调查数据的处理，最后第 3 节给出了论文结论及下一步工作。

2. EM 算法

EM 算法由 Dempster 等于 1977 年提出，是期望最大化法的简称，通过迭代来计算极大似然估计或者后验概率分布。由于 EM 算法具有良好的收敛性和每次迭代都能使似然函数值单调不减的优良性质，所以许多与缺失数据有关的问题中都会结合 EM 算法来分析[10]。EM 算法过程首先用缺失值由估计值替代，之后对完整数据进行参数估计，然后根据上述的参数估计值反过来再估计缺失值。EM 算法包含两部分，

E 步是求期望，即在给定观测数据的条件下求缺失值的条件期望，并用计算出的条件期望对缺失数据进行插补；M 步是做极大化估计，对 M 步之后的完整数据集的参数进行极大似然估计[11]。具体如下：设 $X_1, X_2, \cdots, X_n$ 为 n 维正态分布总体 $N_n(\mu, \Sigma)$ 的随机样本，完全数据的两个充分统计量为：

$$T_1 = \sum_{j=1}^{n} X_j \tag{1}$$

$$T_2 = \sum_{j=1}^{n} X_j X_j' = (n-1)S + X_j X_j' \tag{2}$$

假定 μ 和 Σ 均为未知，则缺失的估计需要进行如下计算：

1) 给出 μ 和 Σ 的估计初始值。可以通过计算未缺失数据的列均值来作为每列缺失数据的初始值[12]，然后通过总体协方差公式求得 Σ，即作为 Σ 的极大似然估计初始值；

2) E 步：通过计算条件数学期望来求含缺失值的每一个向量 X_j。若用 $X_j^{(1)}$ 表示缺失的分量，$X_j^{(2)}$ 表示已知的分量，对 $\bar{\mu}$ 和 $\bar{\Sigma}$ 进行相应的分块变化，则 $X_j^{(1)}$ 的条件正态分布的均值：

$$\overline{X_j^{(1)}} = E\left(X_j^{(1)} \middle| X_j^{(2)}; \bar{\mu}, \bar{\Sigma}\right) = \bar{\mu}^{(1)} + \bar{\Sigma}_{12}\bar{\Sigma}_{22}^{-1}\left(X_j^{(2)} - \bar{\mu}^{(2)}\right) \tag{3}$$

且 $X_j^{(1)}X_j^{(1)\prime}$ 和 $X_j^{(1)}X_j^{(2)\prime}$ 的条件正态分布的均值分别是：

$$\overline{X_j^{(1)}X_j^{(1)\prime}} = E\left(X_j^{(1)}X_j^{(1)\prime} \middle| X_j^{(2)}; \bar{\mu}, \bar{\Sigma}\right) = \bar{\Sigma}_{11} - \bar{\Sigma}_{12}\bar{\Sigma}_{22}^{-1}\bar{\Sigma}_{21} + \bar{X}_j^{(1)}\bar{X}_j^{(1)\prime} \tag{4}$$

$$\overline{X_j^{(1)}X_j^{(2)\prime}} = E\left(X_j^{(1)}X_j^{(2)\prime} \middle| X_j^{(2)}; \bar{\mu}, \bar{\Sigma}\right) = \bar{X}_j^{(1)}\bar{X}_j^{(2)\prime} \tag{5}$$

并用计算出的结果，求出充分的估计量 $\bar{T}_1, \bar{T}_2$。

3) M 步：计算 μ, Σ 的最大似然估计校正值 $\bar{\mu}, \bar{\Sigma}$：

$$\bar{\mu} = \frac{\bar{T}_1}{n} \tag{6}$$

$$\bar{\Sigma} = \frac{1}{n}\bar{T}_2 - \bar{\mu}\bar{\mu}' \tag{7}$$

4) 重复以上的 2~3 步，直至 $\bar{\mu}, \bar{\Sigma}$ 收敛为止。

EM 算法使用于大样本，在现实生活中有很高的使用价值，该方法能够通过在已获得的数据条件下，能稳定、可靠的找到最优值。缺点是计算复杂难度较大。利用 EM 算法来补齐如下 4 × 3 的矩阵，其中缺失值数据位于第一行第一列以及第四行第一、二列。

$$X = \begin{bmatrix} NA & 0 & 3 \\ 7 & 2 & 6 \\ 5 & 1 & 2 \\ NA & NA & 5 \end{bmatrix}$$

经过上述 EM 迭代算法步骤，经 R 语言编程可得：

$$\bar{\mu} = \begin{bmatrix} 6.053 \\ 1.115 \\ 4.000 \end{bmatrix}, \quad \bar{\Sigma} = \begin{bmatrix} 0.597 & 0.388 & 1.216 \\ 0.388 & 0.540 & 0.865 \\ 1.216 & 0.865 & 2.500 \end{bmatrix}$$

迭代算法补入的两个数据：第一行第一列为 5.672，第四行第一、二列数据为别为 6.539，1.461。

3. EM 算法的应用案例

本文实证研究分为两部分：第一部分是比较多变量在 10%、20%、30%三种不同缺失率下，用 EM 算法补齐三层神经网络训练参数的随机缺失数据，计算其相对误差。第二部分分别用多重插补以及 EM 算法对康华医院妇产中心调查问卷中的缺失数据进行插补，并给出未补齐数据、用多重插补补齐数据以及用 EM 算法补齐数据的患者满意度的计算公式，探讨当二级指标的权重不同时，用两种算法补齐数据相对于未补齐数据的总体满意度增加率。

3.1. 案例一——关于网络训练参数的随机缺失补齐

本案例选用的数据集为：用 6000 个样本数据，构造三层神经网络，隐层有 120 个节点，学习率为 0.0005，传递函数为双曲正切函数，共训练 300 轮而得到的 4 × 3 的不含缺失值的矩阵，其中 40 代表测试次数，三个属性依次为最小训练误差(x_1)、测试误差(x_2)、训练时间(x_3)。为验证不同缺失率下 EM 算法补齐数据的精度，通过随机缺失的方式来构造 10%缺失率、20%缺失率以及 30%缺失率的三个不同矩阵，记为 X_1, X_2, X_3 其三个矩阵含缺失值的数目分别为 12 个，24 个和 36 个，缺失位置如图 1。

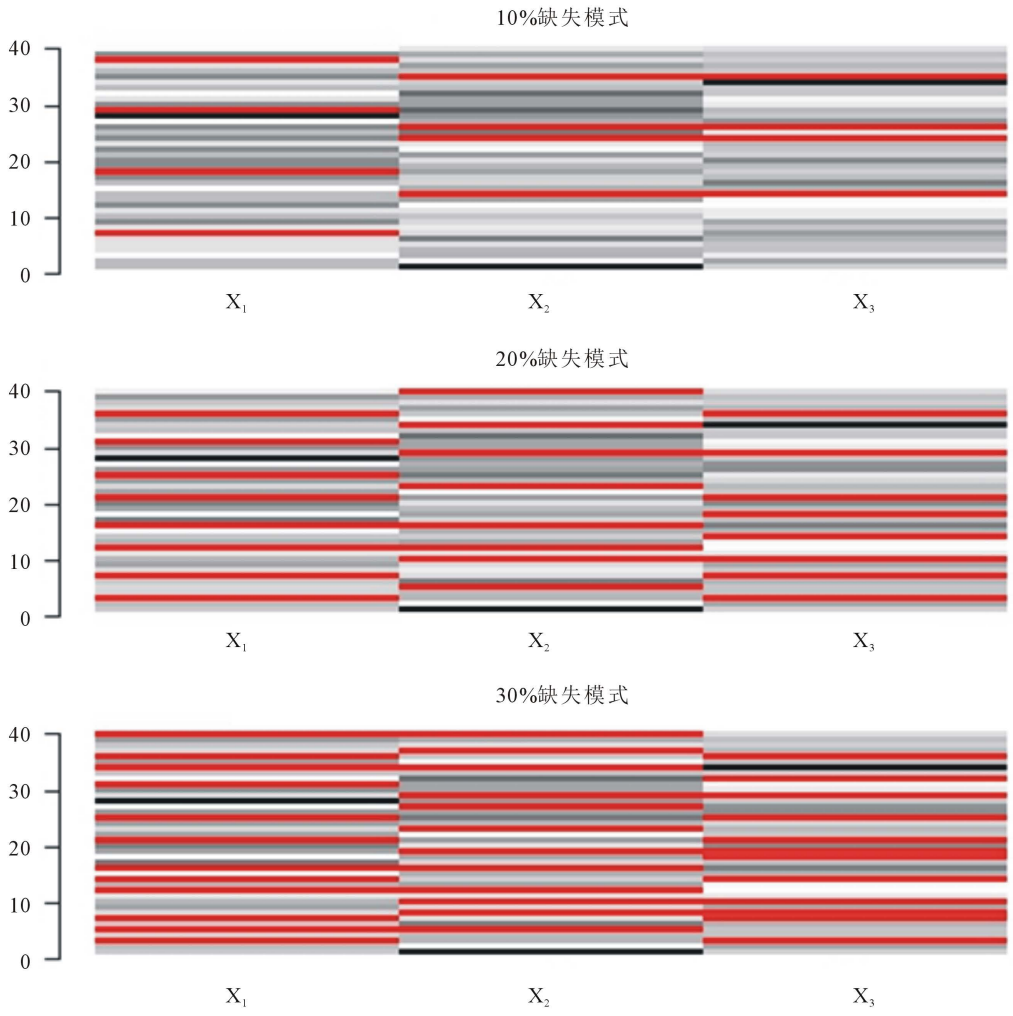

Figure 1. Missing location of different missing ratios
图 1. 不同缺失比率的缺失位置图

多元正态性检验

在做 EM 算法补齐数据之前，需对各变量进行正态分布检验。面对多元正态性检验问题，由于多元正态随机向量 $P \sim N(\mu, \Sigma)$，那么 P 与 μ 的马氏距离($Mahalanobis\ D^2$)的平方服从自由度为 p 的卡方分布。

$$Mahalanobis\ D^2 = (P-\mu)' \Sigma^{-1} (P-\mu) \tag{8}$$

首先通过多元正态性检验的 Q-Q 图来判断变量是否服从正态分布，用 R 中 qqplot 函数分别画出 X, X_1, X_2, X_3 各自的 QQ-plot，如图 2 中的四张图所示，发现点大部分都落在斜率为 1，截距项为 0 的直线附近，故认为其数据近似服从多元正态分布。

进一步验证变量的正态性，采用了 R 中的 mshapiro.test 函数，通过对 X, X_1, X_2, X_3 数据集中的三个变量做正态性检验，得到的各自 P 值如表 1，结果显示 p 值都大于 0.05，故认为其四个数据集中的变量均服从正态分布。

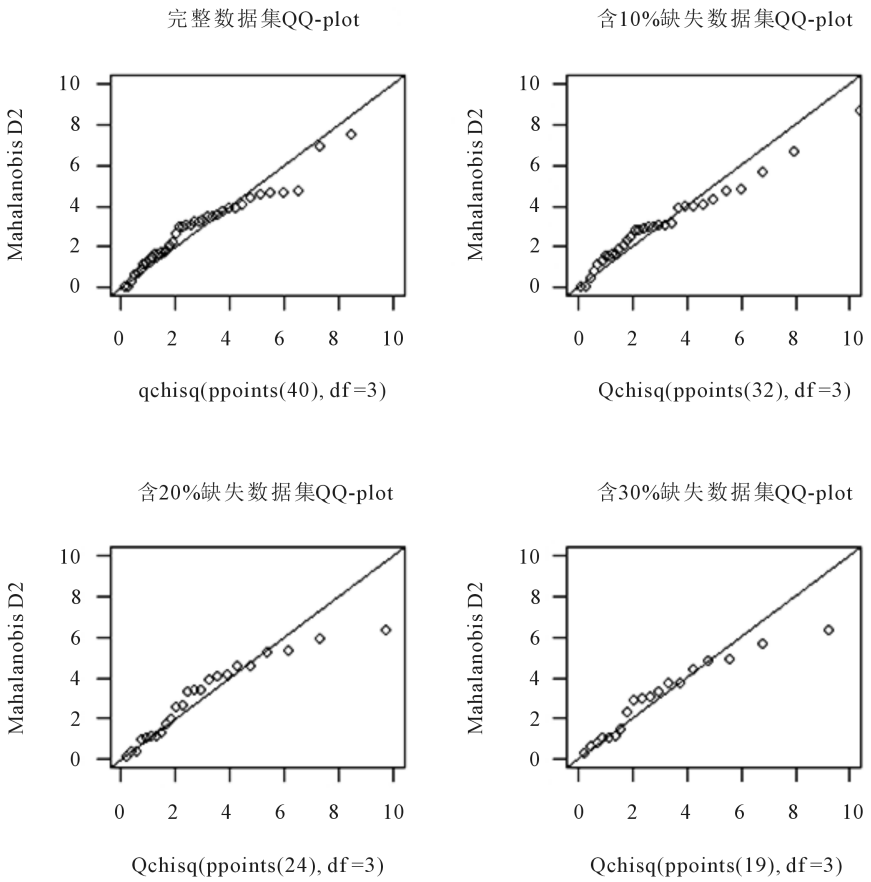

Figure 2. Normality test Q-Q diagram
图 2. 正态性检验 Q-Q 图

Table 1. Normal test p-value table
表 1. 正态性检验 p 值表

数据集	X	X_1	X_2	X_3
p-value	0.16	0.18	0.85	0.73

通过 EM 迭代算法的原理，在 R 语言进行编程，首先计算不含缺失值的列均值来作为初始数据来补齐，并得到 Σ 的极大似然估计初始值，然后计算条件数学期望来求含缺失值的每一个向量 X_j，并计算充分的估计量 $\overline{T}_1,\overline{T}_2$，通过极大似然估计法不断校正 $\mu,\sum$，直到收敛。迭代算法补入的三个矩阵的数据与其本来的数据相对误差 $\delta=\frac{|\Delta|}{L}$，$\Delta=l'-l$ 表示误差 l' 表示插补的值，l 表示真值，计算出当数据缺失率为 10%，20%，30%时[13]，计算的相对误差均值为别为 0.0947,0.0371,0.0413，均小于 0.1，可见 EM 算法在处理低缺失率的数据集的效果比较稳定，且精度比较好，相对误差的箱线图如图 3。

3.2. 案例二——康华医院患者满意度调查数据补齐

3.2.1. 调查背景与目的

EM 算法常用于医学研究中，尤其是临床医学中很常见，因为在临床医学需要对同一试验单位进行多次重复观测，而在这个过程中由于各种原因经常导致试验观测数据缺失，如动物的意外死亡，记录仪器发生故障，或是被调查者拒绝回答相关调查项目等。对于缺失值通常的做法是删去具有缺失的观察记录，但这样会造成信息的损失以及分析结果的偏性[14]。为了保证信息的完整，必须对缺失数据进行适当的处理。EM 算法相比别的算法的优势在于，以“补缺”的方式将含有缺失值的“不完全资料”转化为“完全资料”，从而提高了估计精度。

东莞康华医院由东莞康华集团投资建成，是目前全国首家最大规模的民营三甲医院，2017 年，医院为进一步提高医院患者满意度，严抓医疗服务质量，开展第三方患者满意度调查，以检验与提升医院各项服务工作。医院患者满意度调查具体目的如下：

1) 了解医院患者总体满意度水平；
2) 了解医院各环节服务满意度水平；
3) 了解医院各科室服务满意度水平；
4) 收集医院患者反馈的意见与建议；
5) 发现医院服务存在的问题与不足；
6) 为进一步提升医院服务提出相应建议。

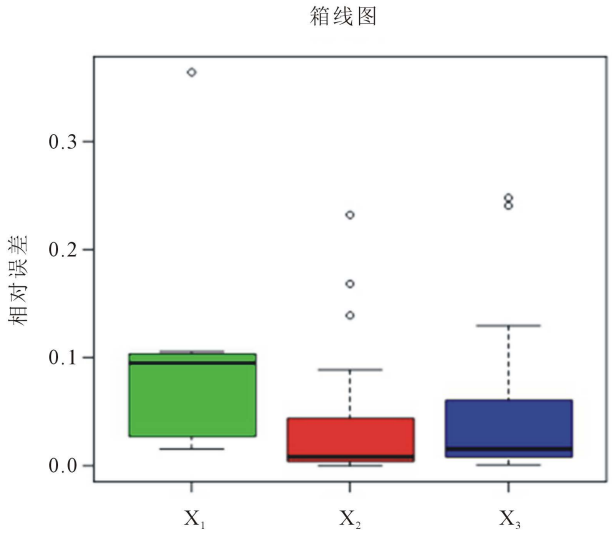

Figure 3. Box plot of relative error for different missing rate datasets
图 3. 不同缺失率数据集相对误差的箱线图

3.2.2. 调查对象与方法

本次调查对象为调研期间到达康华医院就诊的患者，包括门急诊患者、住院患者和出院患者[15]。具体为：门急诊患者(调查当天在医院接受门诊或急诊服务的患者)，住院患者(调查期间在医院接触住院服务，且住院时间在 1 天以上的住院患者)，出院患者(调查前三个月内接受过医院服务的住院患者，且调查时已出院)。

根据医院患者满意度调查对象，项目调查采取定量问卷调查的形式，并分别采用现场访问、病房访问以及电话访问三种方法进行。具体如表 2。

3.2.3. 调查内容与说明

本项目调查内容采用医院节点服务满意度调查，即从患者进入医院到离开医院期间所接触的服务出发，了解患者对其所接触服务点的满意度评价[16]，从而为医院整改提供具体化的建议。其中门诊整体各节点服务的一级指标为总体满意度，二级指标分别为：就医环境、窗口服务、接诊服务、医技服务和护理服务。住院整体各节点服务的一级指标为总体满意度，二级指标分别为：环境后勤、医生服务、护理服务和医技检查。根据这些指标体系以及各个科室的不同服务，建立了包含健康管理中心、门诊患者、血透中心患者、重症医学科患者、住院妇产中心以及住院患者 6 种不同的调查问卷。

3.2.4. 数据描述与整理

总的数据集包含健康管理中心、门诊患者、血透中心患者、重症医学科患者、住院妇产中心以及住院患者在内的 6 个科室的现场访问的样本数据以及包含住院妇产中心和住院患者在内的 2 个电话访问的样本数据。为了充分利用好现场访问和电话访问这两个类型的样本数据，了解患者在医院的整个满意度调查[17]，本案例只挑选了住院妇产中心(电访)这个数据集来补齐数据。

住院妇产中心(电访)包含 41 个样本，问卷包含 25 个能体现打分的有效问题，其中有 9 个问题的回答缺失率超过 30%，而这些问题如，第 5 题(3、您对医院订餐、送餐和饭菜情况是否满意？)，第 8 题(1、您对护士响应呼叫与帮助的及时性是否满意？)这种很多患者没有体验过，所以没有评价，对于这种大比例缺失的情况填补价值不大，故对这 9 个问题予以删除。删除过 9 个问题后，得到了一个 41 行，16 列的数据集，16 个问题用 $A_j\left(j=1,2,3,\cdots,16\right)$ 来表示，如表 3，此时每个问题的缺失情况如图 4。

3.2.5. 缺失数据处理方法——多重插补和 EM 算法

康华医院患者满意度调查数据中，由于被调查者未能体验到医院的某些服务，导致数据缺失，与其他任何观测变量或未观测变量都不相关。对数据做了 MCAR 检验，p 值为 0.059，故接受原假设，认为该数据为完全随机缺失。

对于 MCAR 的缺失模式，运用多重插补的方法，利用 SPSS 对缺失变量进行插补。通过对每个缺失值都构造 20 个插补值，产生了 20 个完全数据集，计算了每个数据集的均值和方差的点估计值，并求出了各个变量用于评价多重插补的指标 ∂, r, RE 的值，详细的归因模型如表 4。

Table 2. Summary of sampling methods at Kanghua Hospital
表 2. 康华医院抽样方法汇总

序号	调查对象	调查方法	抽样方式
1	门诊患者	现场拦截访问	随机抽样
2	住院患者	病房访问	等距抽样
3	出院患者	电话访问	系统随机抽样

Table 3. Questionnaires and labels for part of the questionnaires
表 3. 妇产中心部分问卷题目及标号

A_1	第 5 题(1、您对病房环境卫生（如走道、地板、浴厕等）是否满意？)
A_2	第 5 题(2、您对病房便民设施（如病床、柜、电器等）是否满意？)
A_3	第 5 题(4、您对病区安全保卫工作是否满意？)
A_4	第 5 题(5、您对援助中心工作人员带患者检查服务是否满意？)
A_5	第 5 题(6、您对住院部收费处人员的服务态度是否满意？)
A_6	第 6 题(1、您对主管医生病房巡视服务（巡视次数、检查等）是否满意？)
A_7	第 6 题(2、您对主管医生病情沟通与解答是否满意？)
A_8	第 6 题(3、您对主管医生的服务态度是否满意？)
A_9	第 6 题(4、您对主管医生的技术水平是否满意？)
A_{10}	第 9 题(1、您对 B 超室人员服务是否满意？)
A_{11}	第 9 题(2、您对心电图人员服务是否满意？)
A_{12}	第 9 题(3、您对手术室工作人员服务是否满意？)
A_{13}	第 10 题(1、您对本院整体服务的满意度感知如何？)
A_{14}	第 10 题(2、您对医院向患者提供信息查询或提供费用清单评价如何？)
A_{15}	第 10 题(3、您对医院信息公开方式和公开内容评价如何？)
A_{16}	第 10 题(4、您对医院就诊时间、医生出诊时间、患者须知等信息评价如何？)

Table 4. Multiple Imputation attribution model
表 4. 多重插补的归因模型

缺失变量	类型	缺失值	归因值	分数缺失值(∂)	相对增加方差(r)	相对效率(RE)
A_4	Logistic 回归	1	20	0.225	0.283	0.989
A_5	Logistic 回归	1	20	0.238	0.305	0.988
A_9	Logistic 回归	2	20	0.052	0.055	0.997
A_{10}	Logistic 回归	2	20	0.057	0.06	0.997
A_{11}	Logistic 回归	4	20	0.262	0.346	0.987
A_{12}	Logistic 回归	6	20	0.201	0.247	0.99
A_{14}	Logistic 回归	9	20	0.032	0.033	0.998
A_{15}	Logistic 回归	11	20	0.056	0.059	0.997

未补齐之前的初始数据的含缺失值的各变量的平均均值和平均标准差为 4.299，0.875，而构造了 20 个完整插补数据集后，通过比较每个完整数据集的平均均值和平均标准差，按照平均均值和平均标准差越接近未补之前的平均均值和平均标准差的原则，从而选取了第 15 个数据集作为多重插补的最终结果，其平均均值和平均标准差分别为 4.23，0.90。将多重插补的结果与用 EM 算法编程得到的结果相比如表 5。

3.2.6. 妇产中心总体满意度评价

回到医院做调查问卷的本意，是希望通过问卷的形式了解患者对医院各项服务的满意度情况，以一级指标为总体满意度为因变量 S，自变量为二级指标：环境后勤 B_1、医生服务 B_2、护理服务 B_3 和医技检查 B_4，且 $\sum_{i=1}^{4} c_i = 1$，则 S 与 $B_i\left(i=1,2,3,4\right)$ 的关系可表示为：

$$S = c_1B_1 + c_2B_2 + c_3B_3 + 8c_4B_4 \tag{9}$$

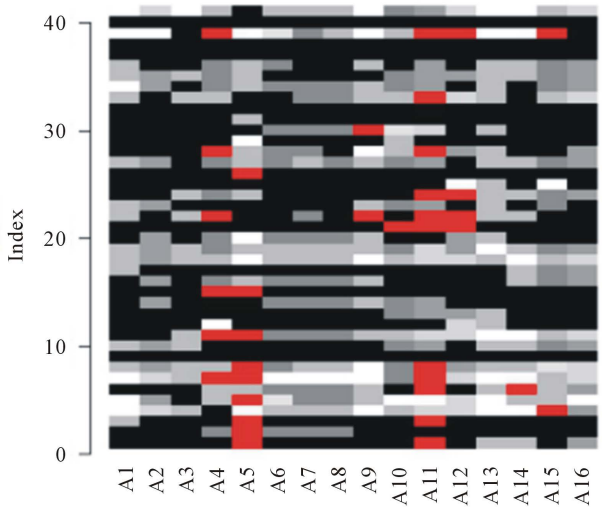

Figure 4. Missing locations for different variables
图 4. 不同变量缺失位置

Table 5. Multiple imputation and EM algorithm complete data results
表 5. 多重插补和 EM 算法补齐数据结果

	A_4		A_5		A_9		A_{10}		A_{11}		A_{12}		A_{14}		A_{15}	
样本标号	EM算法	多重插补	EM算法	多重插补	EM算法	多重插补	EM算法	多重插补	EM算法	多重插补	EM算法	多重插补	EM算法	多重插补	EM算法	多重插补
1			4	4					3	4						
2			5	4												
3			5	3					5	3						
4															5	2
5			5	3												
6									5	2			5	5		
7	3	4	4	4					5	4						
8			4	3					3	4						
11	3	1	4	4												
15	5	4	4	4												
21							5	5	3	5	4	5				
22	5	4			5	4			5	3	5	5				
24									5	5	5	5				
26			5	3												
28	4	4							5	4						
30					4	3										
33									4	2						
39	4	4							5	5	4	4			4	5

其中 c_i 表示各自变量前面的系数。

由于问卷调查数据的补齐是无监督信号的，所以不能通过对比补齐的数据和标准答案的相对误差精度来比较方法的优劣。可利用上文对缺失数据的处理的基础，用三级指标的均值来作为二级指标数据，得到未补齐数据、用 EM 算法补齐数据、用多重插补补齐数据各自的二级指标数值，并得到对应的总体满意度公式：

未补齐数据的公式：

$$S_1 = 4.41c_1 + 4.31c_2 + 4.40c_3 + 4.26c_4 \tag{10}$$

用 EM 算法补齐数据后的公式：

$$S_2 = 4.44c_1 + 4.31c_2 + 4.40c_3 + 4.29c_4 \tag{11}$$

用多重插补补齐数据后的公式：

$$S_3 = 4.39c_1 + 4.29c_2 + 4.40c_3 + 4.23c_4 \tag{12}$$

EM 算法补齐数据相对于未补齐数据的总体满意度增加率：

$$L_1 = \frac{|S_2 - S_1|}{S_1} \times 100\% \tag{13}$$

以及多重插补补齐数据相对于未补齐数据的总体满意度增加率：

$$L_2 = \frac{|S_3 - S_1|}{S_1} \times 100\% \tag{14}$$

当自变量前的系数及各二级指标的权重为 $c_1 = c_2 = c_3 = c_4$ 时，计算出 L_1 为 0.356%，即 EM 算法补齐数据后的总体满意度相对于未补齐数据的总体满意度增加了 0.356%，而 L_2 为 0.253%，即多重插补算法补齐数据后的总体满意度相对于未补齐数据的总体满意度增加了 0.253%，EM 算法补齐数据后的增加率比多重插补算法补齐数据后的增加率提高了 0.103%。

当自变量前的系数及各二级指标的权重为 $2c_1 = c_2 = 2c_3 = 2c_4$ 时，计算出 L_1 为 0.174%，即 EM 算法补齐数据后的总体满意度相对于未补齐数据的总体满意度增加了 0.174%，而 L_2 为 0.172%，即多重插补算法补齐数据后的总体满意度相对于未补齐数据的总体满意度增加了 0.172%，此时 EM 算法补齐数据后的增加率和多重插补算法补齐数据后的增加率相差无几。

当自变量前的系数及各二级指标的权重为 $5c_1 = c_2 = \frac{5}{3}c_3 = 5c_4$ 时，计算出 L_1 为 0.162%，即 EM 算法补齐数据后的总体满意度相对于未补齐数据的总体满意度增加了 0.162%，而 L_2 为 0.216%，即多重插补算法补齐数据后的总体满意度相对于未补齐数据的总体满意度增加了 0.216%，EM 算法补齐数据后的增加率比多重插补算法补齐数据后的增加率降低了 0.054%。

4. 小结

本文是基于 EM 算法实证分析，第一个案例通过比较多变量在 10%、20%、30%三种不同缺失率下，用 EM 算法补齐三层神经网络训练参数的随机缺失数据，得到三者的相对误差均小于 0.1，验证了 EM 算法的在低缺失率情况下的高准确性。第二个案例分别用多重插补(基于 logistic 回归)以及 EM 算法对康华医院妇产中心调查问卷中的缺失数据进行插补，并给出未补齐数据、用多重插补补齐数据以及用 EM 算法补齐数据的患者满意度的计算公式，探讨了当二级指标的权重不同时，用两种算法补齐数据相对于未补齐数据的总体满意度增加率。从实证结果可得到：1) 当权重不同时，EM 算法补齐数据后的增加率相

比于多重插补算法补齐数据后的增加率可增可减；2) 用 EM 算法补齐数据后的增加率和多重插补算法补齐数据后的增加率都不大，原因可能有对于二级指标护理服务下面的 8 个问题，由于每个问题的缺失率超过了 30%，故没有用算法补齐，导致计算 S_1, S_2, S_3 中的 c_3 系数都一样，还有本调查问卷的打分体系是 0~5 分，所以用算法补齐后的数据与均值相差不大。

参考文献

[1] Marlin, B.M. and Zemel, R.S. (2009) Collaborative Prediction and Ranking with Non-Random Missing Data. *Proceedings of the Third ACM Conference on Recommender Systems*, 5-12. https://doi.org/10.1145/1639714.1639717

[2] Marlin, B.M., Zemel, R.S., Roweis, S. and Slaney, M. (2007) Collaborative Filtering and the Missing at Random Assumption. *UAI*.

[3] Marlin, B.M., Zemel, R.S., Roweis, S.T. and Slaney, M. (2011) Recommender Systems: Missing Data and Statistical Model Estimation. *IJCAI*.

[4] Buhi, E.R., Goodson, P. and Neilands, T.B. (2008) Out of Sight, Not Out of Mind: Strategies for Handling Missing Data. *American Journal of Health Behavior*, **32**, 83-92. https://doi.org/10.5993/AJHB.32.1.8

[5] Rubin, D.B. (1996) Multiple Imputation after 18+ Years. *Journal of the American Statistical Association*, **91**, 473-489. https://doi.org/10.1080/01621459.1996.10476908

[6] Dempster, A.P., Laird, N.M. and Rubin, D.B. (1997) Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society Series B Statistical Methodology*.

[7] Little, R.J.A. and Rubin, D.B. (2002) Statistical Analysis with Missing Data. Wiley. https://doi.org/10.1002/9781119013563

[8] 李顺静. 基于 EM 算法的缺失数据的统计分析及应用[D]: [硕士学位论文]. 重庆: 重庆工商大学, 2015.

[9] 谷海彤, 陈邵华, 等. DA 多重插补法在电网电能量数据缺失处理中的应用[J]. 广西科技大学学报, 2017(6): 104-106.

[10] 邹薇, 王会进. 基于朴素贝叶斯的 EM 缺失数据填充算法[J]. 微型机与应用, 2011(16): 75-77.

[11] 吕涛. 市场调查中样本数据缺失值问题研究[J]. 商场现代化, 2014(12): 70-71.

[12] 杨基栋. EM 算法理论及其应用[J]. 安庆师范学院学报(自然科学版), 2009, 15(4): 30-35.

[13] 谭宏卫, 曾捷. Logistic 回归模型的影响分析[J]. 数理统计与管理, 2013, 32(3): 476-485.

[14] 游晓锋, 丁树良, 刘红云. 缺失数据的估计方法及应用[J]. 江西师范大学学报(自然科学版), 2011, 35(3): 325-330.

[15] 王建军. 影响医患关系和谐的因素及对策研究[J]. 江苏卫生事业管理, 2011, 22(5): 118-120.

[16] 兰烯, 刘国恩, 李林. 医疗机构产权性质对医疗服务质量的影响——基于全国试点城市微观数据的实证分析[J]. 中国经济问题, 2014(2): 67-78.

[17] 陈娜. 加强医德医风建设工作的思考[J]. 管理观察, 2014(7): 185-186.