arXiv:2305.01794v1 [stat.ME] 2 May 2023

MISNN：基于半参数神经网络的多重插补

布志奇*、戴宗宇*、张毅良*和龙琦{zbu,daizy,zylthu14}@sas.upenn.edu, qlong@pennmedicine.upenn.edu

摘要。 多重插补（MI）已广泛应用于生物医学、社会科学和计量经济学研究中的缺失值问题，以避免下游数据分析中出现不恰当的推断。在存在高维数据的情况下，包含特征选择的插补模型，尤其是 ℓ_1 正则化回归（如 Lasso、自适应Lasso 和弹性网络），是防止模型欠定的常见选择。然而，结合特征选择进行多重插补具有一定难度：现有方法通常计算效率低下且性能不佳。我们提出了 MISNN，这是一种将特征选择融入多重插补的新颖高效算法。借助神经网络的逼近能力，MISNN 构建了一个通用且灵活的框架，能够兼容任意特征选择方法、任意神经网络架构、高维或低维数据以及一般缺失模式。通过实证实验，MISNN 在插补准确性、统计一致性和计算速度方面均展现出相较于最先进的插补方法（如贝叶斯 Lasso 和矩阵补全）的显著优势。

关键词： 缺失值、插补、半监督学习

1 引言

1.1 缺失值机制与插补方法

数据分析中经常会遇到缺失数据。众所周知，未能妥善处理缺失数据会导致有偏差的研究结果、不恰当的统计推断 [11,37]以及较差的预测表现。缺失数据插补是有效的解决方法之一。现有插补方法主要分为单次插补（SI）和多重插补（MI） [26]。前者仅对缺失值插补一次，而后者则从某种分布中多次生成插补值。在金融和医学研究等领域，人们通常更倾向于使用线性模型，因为不仅需要准确预测，还需要解释预测的不确定性及特征的影响。考虑到统计推断，包括本文提出的 MISNN 在内的多重插补方法更为适用，因为它们能够充分考虑插补的不确定性，并提供恰当的推断。

一般而言，插补的表现与产生缺失值的机制密切相关，这些机制可分为三类：缺失

*共同贡献

完全随机缺失（MCAR）、随机缺失（MAR）和非随机缺失（MNAR）。如果所有条目缺失的概率都相同，则称缺失数据为完全随机缺失；随机缺失是指缺失概率仅取决于已观测值；非随机缺失是指缺失概率取决于未观测到的缺失值。直观而言，在完全随机缺失机制下进行插补更容易，因为缺失概率只是一个（未知的）常数，因此大多数方法都是为完全随机缺失情形设计的。不过，随机缺失和非随机缺失通常更难处理，能够在这些问题上表现良好的方法也较少。

1.2 插补模型中的特征选择

在许多应用中，包括基因表达和金融时间序列研究，我们需要分析特征数量远大于样本数量的高维数据。在这种情况下，多重插补需要估计缺失数据的（条件）分布，但由于特征数量过多，估计可能不准确。已有研究 [11,37] 提出在构建插补模型之前使用正则化线性模型进行特征选择。一些代表性模型包括 Lasso [31], SLOPE [4], 弹性网络[39], 自适应Lasso [38], 稀疏组Lasso [13,27], 等。

虽然正则化线性模型成功减少了特征数量，但在构建插补模型时，它们依赖所选特征之间的线性关系，并且会丢失未选特征中的信息，因此往往无法捕捉缺失数据的真实分布。因此，相应的推断可能存在显著偏差。本文提出的 MISNN 通过半参数神经网络克服了这一缺陷。概括而言，MISNN 是一种基于神经网络的半参数模型，它将预测变量分为两组：第一组用于构建线性模型，另一组用于构建神经网络，而神经网络通常被视为非参数模型。我们强调，MISNN 的优越表现既得益于其神经网络部分，也得益于线性部分。神经网络能够有效捕捉插补模型中的非线性关系；线性模型除了捕捉线性关系外，还能通过对回归参数进行最大似然估计，实现高效的多重插补。

1.3 我们的贡献

本文有两项贡献。首先，我们提出了 MISNN，这是一种新颖的插补方法，在插补准确性、统计一致性和计算速度方面均优于最先进的插补方法。MISNN 易于调参、可解释，并且能够稳健地处理高缺失率和高维特征。其次，MISNN 是一个灵活的插补框架，可与任何合适的特征选择方法结合使用，例如 Lasso 和前向选择。此外，MISNN 兼容任何神经网络，包括参数不足或参数过多的网络、卷积神经网络、残差网络、丢弃法等。

2 相关工作

在缺失数据插补方面，早在多重插补（MI）概念提出之前，单一插补方法（SI）就已存在很长时间 [26]，其中一种有代表性的方法是均值插补。

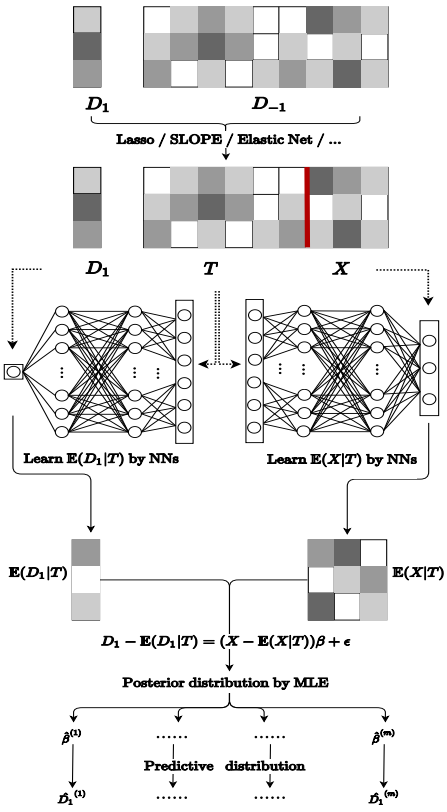


图 1. MISNN框架。

近年来，单一插补方法（SI）中出现了矩阵补全方法，将插补转化为优化问题。现有方法如 SoftImpute[23] 和 MMMF（最大间隔矩阵分解） [28] ，已被证明在完全随机缺失机制下有效。与此同时，越来越多的多重插补方法也受到研究：MICE [32,5] 通过链式方程插补缺失值；MissForest [30] 则通过对多棵树进行自助聚合来插补缺失值。研究人员还提出了用于插补的深度生成模型 [34,14,22,19,9]，包括生成对抗式插补网络（GAIN）。值得注意的是，现有方法大多仅在完全随机缺失（MCAR）情况下有理论保证（尽管有些方法在随机缺失（MAR）情况下在实证中表现良好）。

针对高维数据的多重插补，已有研究提出了正则化线性模型。贝叶斯 Lasso [24,16] 估计系数的后验分布，而其他方法 [37,11] 则对正则化线性回归的估计量进行去偏处理。具体而言，这些方法包括直接使用正则化回归（DURR）和间接使用正则化回归（IURR）。然而，线性插补模型无法捕捉缺失数据条件分布中潜在的非线性关系。MISNN 属于这一研究方向，其计算效率高于贝叶斯Lasso，并能在插补过程中捕捉非线性关系。

近期研究凸显了缺失数据插补中可信度的重要性，其中，保护隐私的 [18,10,8] 插补模型和关注公平性的 [21,36,6,35]插补模型受到广泛关注。MISNN 具有很强的可解释性，有助于更好地理解插补过程，并增强对结果的信任。

3 数据设置

将数据矩阵记为 $\mathbf{D} \in \mathbb{R}^{n \times p}$，其中 n 是样本/案例的数量，p 是特征/变量的数量。我们将第 j个特征记为 $\mathbf{D}_j$，其补集特征记为 $\mathbf{D}_{-j} := \mathbf{D}_{2:p}$，其中 $j \in [p]$。存在缺失数据时，可将 $\mathbf{D}$ 分为两个子矩阵 $\mathbf{D}_{\text{cc}}$ 和 $\mathbf{D}_{\text{ic}}$，其中 $\mathbf{D}_{\text{cc}}$ 表示所有完整案例（即所有特征均已观测到），$\mathbf{D}_{\text{ic}}$ 表示所有不完整案例。我们分别用 $\mathbf{D}_{\text{cc},j}$ 和 $\mathbf{D}_{\text{ic},j}$ 表示完整案例和不完整案例的第 j个特征。我们还定义 $\mathbf{D}_{\text{miss}}$，即 $\mathbf{D}$ 中缺失特征的集合，以及 $\mathbf{D}_{\text{obs}}$，即 $\mathbf{D}$ 中样本的观测特征集合。简言之，为插补缺失值，我们使用 $\mathbf{D}_{\text{obs}}$拟合插补模型 g，并将 $\hat{} \mathbf{D}_{\text{ic,obs}}$ 作为输入，以得到插补结果 $\mathbf{D}_{\text{miss}}$。为便于说明，我们先从单个特征缺失的情形入手，此时 $\mathbf{D}$ 中只有第一列（即 $\mathbf{D}_1$）包含缺失值。随后，我们将在第 5 节讨论多个特征缺失的一般缺失模式。

3.1 多重插补框架

这里简要介绍多重插补的一般框架，MISNN 也采用了该框架。在上述数据设定下，多重插补方法估计条件分布 $\rho(\mathbf{D}_{\text{miss}}|\mathbf{D}_{\text{obs}})$，并从中多次抽取插补值。假设 $\mathbf{D}$ 的分布由未知参数 $\boldsymbol{\xi}$ 表征，则

$$\rho(\mathbf{D}_{\text{miss}}|\mathbf{D}_{\text{obs}}) = \int \rho_{\text{miss}}(\mathbf{D}_{\text{miss}}|\mathbf{D}_{\text{obs}}, \boldsymbol{\xi}) \rho_2(\boldsymbol{\xi}|\mathbf{D}_{\text{obs}}) d\boldsymbol{\xi}$$

其中，ρ、ρ_1、ρ_2 是三个条件分布。对于第 m 次插补，我们从 $\boldsymbol{\xi}$ 的后验分布中随机抽取 $\boldsymbol{\xi}^{(m)}$，即 $\rho_2(\boldsymbol{\xi}|\mathbf{D}_{\text{obs}})$；然后从预测分布 $\rho_1(\mathbf{D}_{\text{miss}}|\mathbf{D}_{\text{obs}}、\boldsymbol{\xi})$ 中生成第 m 个插补数据 $\mathbf{D}^{(m)}{}_{\text{miss}}$。借助多个插补数据集，可以利用 Rubin 法则 [20,26] 进行进一步的分析和推断。附录 A 对 Rubin 法则进行了详细介绍。

4 基于半参数神经网络的多重插补（MISNN）

在高层次上，MISNN 通过部分线性模型（PLM）对每一列中的缺失数据进行插补，其形式为

$$\hat{\mathbf{D}}_1 = \mathbf{X}\hat{\beta} + \hat{f}(\mathbf{T})$$

其中，经过特征选择确定的（$\mathbf{X}$，$\mathbf{T}$）是其余$\hat{} \hat{}p-1$列的一个划分。虽然 β和 f（$\mathbf{T}$）可以任意选择

方式，但我们采用偏回归剔除方法 [25] （在 [7]中也称为正交化），如果真实模型具有$\mathbf{D}_1 = \mathbf{X}\beta + f（\mathbf{T}） + \boldsymbol{\epsilon}$的形式，就能提供一致的参数估计。为此，我们对$\mathbf{T}$取条件期望，并假设 $\mathbb{E}（\boldsymbol{\epsilon}|\mathbf{T}） = 0$:

$$\begin{aligned} \mathbf{D}_1 &= \mathbf{X}\boldsymbol{\beta} + f(\mathbf{T}) + \boldsymbol{\epsilon} \\ \mathbb{E}(\mathbf{D}_1|\mathbf{T}) &= \mathbb{E}(\mathbf{X}|\mathbf{T})\boldsymbol{\beta} + f(\mathbf{T}) \\ \mathbf{D}_1 - \mathbb{E}(\mathbf{D}_1|\mathbf{T}) &= (\mathbf{X} - \mathbb{E}(\mathbf{X}|\mathbf{T}))\,\boldsymbol{\beta} + \boldsymbol{\epsilon} \end{aligned} \tag{1}$$

令 $\mathcal{S}$ 表示所选特征的集合。注意，$\mathbf{T} := \mathbf{D}_{-1} \setminus \mathbf{D}_{\mathcal{S}}$ 在最后一个方程中被显式移除。因此，只要能够控制所选特征的数量（即，$|\mathcal{S}|$ 较小），我们就可以得到一个低维线性模型（因为 $\mathbf{X} - \mathbb{E}\mathbf{(X|T)} \in \mathbb{R}^{n\times|\mathcal{S}|}$），前提是能够适当地估计映射 $\mathbb{E}\mathbf{(D}_1\mathbf{|T)}$ 和 $\mathbb{E}\mathbf{(X|T)}$。为实现上述方法，MISNN 算法采取三个关键步骤:

– 特征选择： 在对每个缺失特征进行插补时，MISNN 会执行特征选择，最多选择 n 个特征。所选特征 $\mathbf{X}$ 应与缺失特征具有显著的线性相关性，之后将在一个线性模型（例如最小二乘法）中对其进行拟合。**– 拟合部分线性模型：** 假设选择后剩余的特征表示为$\mathbf{T}$，MISNN 拟合两个神经网络，以学习 $\mathbb{E}\mathbf{(D}_{\text{miss}}\mathbf{|T)}$ 和 $\mathbb{E}\mathbf{(X|T)}$，从而推导出低维普通线性模型（1）；**– 多重插补：** MISNN 使用最大似然估计（1）中的^参数，然后从 $\widehat{\boldsymbol{\beta}}$ 的后验分布中抽取 M 次样本，并进一步从预测分布中抽取$\mathbf{D}_{\text{miss}}$。

请注意，前两个步骤结合起来与 DebiNet [33],密切相关，尽管我们没有进一步采用过度参数化的神经网络，并且使用两个神经网络来学习 $\mathbf{(}\mathbb{E}\mathbf{(D}_{\text{miss}}\mathbf{|T)}\ \mathbb{E}\mathbf{(X|T))}$。下面，我们介绍算法 1 中的 MISNN，并严格验证 MISNN 的过程。这里假设缺失特征是连续的。对于非连续特征，需要对算法进行一些修改。详情见附录B。

注 1 如果只关注预测而非推断，可以在算法 *1* 中进行单次插补。具体而言，普通最小二乘法可以求解步骤 *(4)* 中的线性模型，我们通过以下方式进行插补

$$\widehat{\mathbf{D}}_{ic,1} = (\mathbf{X}_{cc} - \mathbb{E}(\mathbf{X}_{cc}|\mathbf{T}_{cc}))\,\hat{\boldsymbol{\beta}} + \mathbb{E}(\mathbf{D}_{cc,1}|\mathbf{T}_{cc})$$

我们将该插补算法命名为 *SISNN*（见附录*D*中的算法*4*）。

4.1 从后验分布和预测分布中抽样

要在 MISNN 中进行多重插补，我们需要从后验分布中抽取参数 $\rho_2\left(\boldsymbol{\beta}, \sigma^2 \middle| \mathbf{D}_{\text{obs,1}}, \mathbf{X}_{\text{obs}}, \mathbf{T}_{\text{obs}}\right)$ 以及预测

noend 1 通过半参数神经网络进行多重插补（MISNN）**输入：** 不完整数据 $\mathbf{D}$, 插补次数 M

1: 通过惩罚函数 P 拟合一个正则化回归 $\mathbf{D}_{\mathrm{cc},1} \sim \mathbf{D}_{\mathrm{cc},-1}$，方法是

$$(\widehat{\boldsymbol{\alpha}}, \hat{\alpha}_0) := \underset{(a,a_0)}{\operatorname{argmin}} \frac{1}{2}\|\mathbf{D}_{\mathrm{cc},1} - \mathbf{D}_{\mathrm{cc},-1}\boldsymbol{a} - a_0\|^2 + P(\boldsymbol{a}).$$

2: 获取活动集 $\mathcal{S} := \{i : \hat{\alpha}_i \neq 0\}$，并将 $\mathbf{D}_{-1}$ 拆分为子矩阵 $\mathbf{X} = [\mathbf{D}_{-1}]_{\mathcal{S}}$ 和 $\mathbf{T} = \mathbf{D}_{-1} \backslash \mathbf{X}$。

3: 给定训练数据 $\{\mathbf{T}_{\mathrm{cc}}, \mathbf{D}_{\mathrm{cc},\ 1}, \mathbf{X}_{\mathrm{cc}}\}$，训练神经网络以学习

$$\eta_D(\mathbf{T}) := \mathbb{E}(\mathbf{D}_1|\mathbf{T}), \eta_X(\mathbf{T}) := \mathbb{E}(\mathbf{X}|\mathbf{T})$$

4: 将标准最大似然技术应用于

$$\mathbf{D}_{\mathrm{cc},1} - \mathbb{E}(\mathbf{D}_{\mathrm{cc},1}|\mathbf{T}_{\mathrm{cc}}) = (\mathbf{X}_{\mathrm{cc}} - \mathbb{E}(\mathbf{X}_{\mathrm{cc}}|\mathbf{T}_{\mathrm{cc}}))\,\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

其中 $\boldsymbol{\epsilon} \sim \mathcal{N}\ 0\ \sigma^2$ $\rho_2\left(\boldsymbol{\beta}, \sigma\middle|\ _{\mathrm{obs}\ 1}\quad _{\mathrm{obs}}\quad _{\mathrm{obs}}\right)(,\ )$

5 $m \in \{1, \ldots, M\}$

并近似分布 $\mathbf{D}_{,}$, $\mathbf{X}$, $\mathbf{T}$: **对于 do**

6: 随机抽取 $\hat{\boldsymbol{\beta}}^{(m)}$, $\hat{\sigma}^{(m)}$ 从条件分布 $\rho_2\left(\boldsymbol{\beta}, \sigma\middle|\mathbf{D}_{\mathrm{cc},\ 1}, \mathbf{X}_{\mathrm{cc}}, \mathbf{T}_{\mathrm{cc}}\right)$ 中抽取。随后，从预测分布 $\rho_1\left(\mathbf{D}_{\mathrm{ic},\ 1}|\mathbf{X}_{\mathrm{ic}}, \mathbf{T}_{\mathrm{ic}}, \hat{\boldsymbol{\beta}}^{(m)}, \hat{\boldsymbol{\sigma}}^{(m)^2}\right)$ 中随机抽取，以 $\widehat{\ }\mathbf{D}^{(m)}{}_{\mathrm{ic},\ 1}$ 插补 $\mathbf{D}_{\mathrm{ic},\ 1}$

在 MISNN 中，分布 $\rho_1\left(\mathbf{D}_{\mathrm{miss},\ 1}|\mathbf{X}_{\mathrm{miss}}, \mathbf{T}_{\mathrm{miss}}, \hat{\boldsymbol{\beta}}^{(m)}, \hat{\boldsymbol{\sigma}}^{(m)^2}\right)$（参见算法 1）。通过偏回归剔除，我们在步骤 (4) 中拟合线性回归，

$$\mathbf{D}_{\mathrm{obs},1} - \mathbb{E}(\mathbf{D}_{\mathrm{obs},1}|\mathbf{T}_{\mathrm{obs}}) = (\mathbf{X}_{\mathrm{obs}} - \mathbb{E}(\mathbf{X}_{\mathrm{obs}}|\mathbf{T}_{\mathrm{obs}}))\,\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

我们使用以下方法近似 $\boldsymbol{\beta}$, σ 的后验分布

$$\rho_2\left(\boldsymbol{\beta}, \sigma^2\middle|\mathbf{D}_{\mathrm{obs},1}, \mathbf{X}_{\mathrm{obs}}, \mathbf{T}_{\mathrm{obs}}\right) = f_1\left(\boldsymbol{\beta}\middle|\mathbf{D}_{\mathrm{obs},1}, \mathbf{X}_{\mathrm{obs}}, \mathbf{T}_{\mathrm{obs}}\right) \times f_2\left(\sigma^2\middle|\mathbf{D}_{\mathrm{obs},1}, \mathbf{X}_{\mathrm{obs}}, \mathbf{T}_{\mathrm{obs}}\right)$$

¯假设 $\boldsymbol{\beta}$ 的普通最小二乘估计及其方差分别为 $\boldsymbol{\beta}$ 和 $\Sigma_{\boldsymbol{\beta}}$。我们可以用正态分布近似 $\boldsymbol{\beta}$ 的分布：

$$f_1\left(\boldsymbol{\beta}\middle|\mathbf{D}_{\mathrm{obs},1}, \mathbf{X}_{\mathrm{obs}}, \mathbf{T}_{\mathrm{obs}}\right) \sim \mathcal{N}\left(\bar{\boldsymbol{\beta}}, \Sigma_{\boldsymbol{\beta}}\right)$$

其中参数定义如下：

$$\bar{\boldsymbol{\beta}} = \operatorname{argmin}_b \|\mathbf{D}_{\mathrm{obs},1} - \eta_D(\mathbf{T}_{\mathrm{obs}}) - [\mathbf{X}_{\mathrm{obs}} - \eta_X(\mathbf{T}_{\mathrm{obs}})]\boldsymbol{b}\|^2$$

$$\Sigma_{\boldsymbol{\beta}} = \bar{\sigma}^2\left((\mathbf{X}_{\mathrm{obs}} - \eta_X(\mathbf{T}_{\mathrm{obs}})^\top(\mathbf{X}_{\mathrm{obs}} - \eta_X(\mathbf{T}_{\mathrm{obs}}))\right)^{-1}$$

此处 $\bar{\sigma}^2$ 可估计为残差平方的均值：

$$f_2\left(\sigma^2\middle|\mathbf{D}_{\mathrm{obs},1}, \mathbf{X}_{\mathrm{obs}}, \mathbf{T}_{\mathrm{obs}}\right) = \left\|\mathbf{D}_{\mathrm{obs},1} - \eta_D(\mathbf{T}_{\mathrm{obs}}) - (\mathbf{X}_{\mathrm{obs}} - \eta_X(\mathbf{T}_{\mathrm{obs}}))\bar{\boldsymbol{\beta}}\right\|^2 / n_{\mathrm{obs}}$$

至于从预测分布中抽样，我们从 f_2 计算出 $\hat{}\sigma^{(m)}$ （将$\bar{\beta}$ 替换为$\hat{}\boldsymbol{\beta}^{(m)}$）。最后，我们可以从中抽取 $\widehat{\ }\mathbf{D}^{(m)}$ 缺失，1

$$\rho_1\left(\mathbf{D}_{\text{miss},1}|\mathbf{X}_{\text{miss}},\mathbf{T}_{\text{miss}},\hat{\boldsymbol{\beta}}^{(m)},\hat{\boldsymbol{\sigma}}^{(m)^2}\right)=\eta_D(\mathbf{T}_{\text{miss}})+(\mathbf{X}_{\text{miss}}-\eta_X(\mathbf{T}_{\text{miss}}))\hat{\boldsymbol{\beta}}^{(m)}+\mathcal{N}(0,\hat{\boldsymbol{\sigma}}^{(m)^2})$$

4.2 MISNN 框架的灵活性

再次强调，MISNN 框架在两个方面具有灵活性：它可以在插补过程中采用任意特征选择方法和任意神经网络模型。

MISNN 可以采用任意特征选择方法。这里，我们采用 Lasso 选择特征 $\mathbf{X}=\mathbf{D}_{\mathcal{S}}$ 和 $\mathbf{T}=\mathbf{D}_{-1}\setminus\mathbf{D}_{\mathcal{S}}$，其中 $\mathcal{S}=\{i>0:\hat{}\alpha_i\neq 0\}$来自 Lasso 估计量的非零部分

$$(\hat{\boldsymbol{\alpha}},\hat{\alpha}_0)=\text{argmin}_{\boldsymbol{a},a_0}\frac{1}{2}\|\mathbf{D}_{\text{cc},1}-\mathbf{D}_{\text{cc},-1}\boldsymbol{a}-a_0\|_2^2+\lambda\|\boldsymbol{a}\|_1$$

MISNN 与所有类型的网络都兼容。特别是，使用过参数化神经网络时，MISNN 可以借鉴 DebiNet [33, 定理 1&2]，从而得出 $\sqrt{n}$-一致性和指数级快速收敛的结论。

在实践中，MISNN 可以使用比神经切核区域 [12,1]理论支持的丰富得多的神经网络类别。这包括欠参数化、适度宽且深的神经网络。实证实验表明，由这类神经网络学习得到的 PLM 具有很强的预测准确性和选择后推断能力（见表 2）。

4.3 MISNN 的其他性质

除了前文提到的框架灵活性、 $\boldsymbol{\beta}$ 的一致估计以及 PLM 的快速训练之外，本节还将讨论 MISNN 所具备的一些性质。数值证据见第 5 节。

可训练性： 与贝叶斯Lasso（可能需要耗时较长的预热期，见表 3）、自助法方法（如 DURR，需要大量自助抽样子样本才能获得准确结果）或 MICE（逐一迭代拟合每个特征，在高维情况下可能较慢）相比，MISNN 可以高效地使用现有优化器进行训练。

稳健性： 从经验上看，MISNN 对超参数调优具有稳健性（例如，隐藏层的宽度对性能影响不大）。从数据角度来看，即使特征维度高、缺失率高（例如，与 DURR、IURR 和 GAIN 相比），MISNN 仍能取得不错的效果。

4.4 MISNN 适用于一般缺失模式

插补过程可以自然地扩展到一般缺失模式的情形，其伪代码见附录D中的算法 3。假设$\mathbf{D}$中前 K 列缺失，记为$\mathbf{D}_{\text{full},\ [K]}$而第 k列记为$\mathbf{D}_{\text{full},\ k}$。集合 $-[K]$ 表示除 $[K]$中的列之外的所有其他列。与单列缺失的情形类似，要构建部分线性模型，我们需要将数据划分为$\mathbf{X}$ 和$\mathbf{T}$。

我们对每个存在缺失值的 K 列分别拟合正则化线性回归，得到 K 个活跃集。随后，我们提出使用交集或并集将这些集合合并为一个集合，并将其作为$\mathbf{X}$。为估计参数 β，在每次插补过程中，对于第 k列，我们考虑一个以$\mathbf{D}_{\text{full, }[K]}$ 为回归变量、以第 k列为响应变量的普通最小二乘模型。采用最大似然法生成回归系数 $\boldsymbol{\beta}_k$。

我们指出，也可以采用其他合适的特征选择方法和集合合并规则，替代本文使用的方法。我们也可以借鉴 MICE 的思路，采用迭代方法进行按列插补。借助广义部分线性模型，可以推广到离散缺失值的情形，这与附录B中的讨论类似。

5 数值结果

我们在各种合成数据集和真实世界数据集上，将 MISNN 与其他最先进的方法进行了比较。为建立基线，我们纳入了完整数据分析、完整案例分析和列均值插补。我们还评估了两种在高维情形下采用正则化线性模型进行特征选择的多重插补方法：MICE-DURR 和 MICE-IURR。此外，我们还纳入了 MissForest（一种使用随机森林作为回归模型的 MICE 方法）、GAIN（一种基于深度学习的插补方法）以及两种矩阵补全方法：SoftImpute 和 MMMF。有关实验设置和结果的更多详细信息，请参见附录C。

方法	风格	Bias	插补均方误差	覆盖率	秒	SE	SD
完整数据	-	**0.0027**	-	**0.954**	-	0.1126	0.1150
完整案例	-	0.1333	-	0.854	-	0.1556	0.1605
均值填补	SI	0.1508	12.6215	**0.994**	**0.005**	0.3268	0.1933
MISNN-wide（Lasso）	多重插补	**-0.0184**	**4.2382**	**0.902**	**0.324**	0.1438	0.1713
MISNN-wide（弹性网络）	多重插补	**-0.0134**	**4.2191**	**0.924**	**0.286**	0.1431	0.1641
MISNN-narrow（Lasso）	多重插补	**-0.0251**	6.2666	**0.944**	**0.370**	0.1816	0.1755
MISNN-narrow（弹性网络）	多重插补	**-0.0246**	6.2550	**0.956**	**0.344**	0.1818	0.1647
MICE-DURR（Lasso）	MI	0.1815	12.6704	**0.978**	1.266	0.2275	0.1196
MICE-DURR（弹性网络）	MI	0.1314	10.8060	**0.990**	0.633	0.2241	0.1219
MICE-IURR（Lasso）	MI	0.2527	15.7803	0.886	1.483	0.2136	0.1150
MICE-IURR（弹性网络）	MI	0.2445	15.3266	0.892	0.566	0.2153	0.1399
MissForest	MI	0.0579	9.6174	**0.962**	69.948	0.2851	0.2609
GAIN	SI	0.7578	27.3505	0.289	14.812	0.2869	0.4314
SoftImpute	SI	-0.1432	**4.6206**	0.842	**0.019**	0.1804	0.2005
MMMF	SI	-0.1239	**4.0956**	0.782	3.385	0.1491	0.1869

表1. 合成数据中 500 个蒙特卡洛数据集的多特征缺失模式。偏差：$\hat{}\ \beta_1 - \beta_1$ 的平均偏差；插补均方误差：$\|\widehat{\mathbf{D}}_{\text{miss, }1:3} - \mathbf{D}_{\text{miss, }1:3}\|^2/n_{\text{miss}}$；覆盖率：$\beta_1$ 的 95% 置信区间的覆盖概率；秒：实际插补时间；标准误：$\hat{}\ \beta_1$ 的平均标准误；标准差：$\hat{}\ \beta_1$ 的蒙特卡洛标准差。模型设置见第 5.1 节，数据生成见附录 C.2。

除了评估插补准确性外，我们还评估基于插补数据集进行统计推断时插补模型的表现。在所有

实验中，我们在数据矩阵 $\hat{\mathbf{D}} = (\mathbf{Z}\ y)$ 中指定一组预测变量和一个响应变量。使用插补数据集拟合线性回归 $\hat{y} = \mathbf{Z}\boldsymbol{\theta}$，以预测 $\hat{y}$，并记录回归参数 $\boldsymbol{\theta}$。在合成数据集中，我们可以获得真实值 $\boldsymbol{\theta}$，因此重点关注推断性能。在真实数据分析中，我们无法获得真实的 $\boldsymbol{\theta}$，因此转而关注预测误差。

5.1 统计推断的观点

在统计推断方面，我们考察四个统计量：$\hat{\boldsymbol{\theta}}$ 的偏差、$\boldsymbol{\theta}$ 的 95% 置信区间覆盖率（CR）、$\boldsymbol{\theta}$ 的平均标准误（SE）以及 $\boldsymbol{\theta}$ 的蒙特卡洛标准差（SD）。我们还比较插补均方误差（MSE）。我们研究 MISNN 在一般缺失模式下的表现，在这种模式下，数据集中的多个列（特征）都可能包含缺失值。我们采用与 [11] 中类似的实验设置，并在 500 个蒙特卡洛数据集上评估性能。详细的实验说明见附录C。

方法	风格	估计量	插补均方误差	秒	SE	预测均方误差
完整数据	-	**0.0532**	-	-	0.0676	**0.8695**
完整案例	-	0.1278	-	-	0.1392	1.3376
均值填补	SI	-0.0374	1.3464	**0.006**	0.0686	0.8938
MISNN（Lasso）	MI	**0.0545**	0.6620	**1.501**	0.0681	**0.8780**
MISNN（弹性网络）	MI	**0.0521**	**0.5140**	**0.861**	0.0716	**0.8789**
MICE-DURR（Lasso）	MI	**0.0504**	1.8256	3.946	0.0508	**0.8755**
MICE-DURR（弹性网络）	MI	0.0426	1.6998	2.709	0.0552	0.8817
MICE-IURR（Lasso）	MI	0.0474	2.0404	4.093	0.0476	**0.8747**
MICE-IURR（弹性网络）	MI	0.0318	2.0219	2.620	0.0484	0.8803
GAIN	SI	0.0304	0.9902	67.432	0.0504	**0.8749**
SoftImpute	SI	**0.0533**	0.6667	**0.0344**	0.0763	0.8808
MMMF	SI	0.0833	**0.3051**	5.0261	0.0838	**0.8755**

表2. ADNI数据集中经100次重复的多特征缺失模式。估计量：通过普通最小二乘法（OLS）使用前5个特征作为回归变量得到的$\hat{}$estimated β_1；插补均方误差：插补均方误差 $\|\hat{\mathbf{D}}_{\text{miss},1:3} - \mathbf{D}_{\text{miss},1:3}\|^2/n_{\text{miss}}$；秒：实际插补时间；标准误：$\hat{}$ β_1的平均标准误；预测均方误差：$\mathbf{A}\hat{\boldsymbol{\theta}}$与$\mathbf{y}$之间的均方误差。模型设置见第5.2节，数据生成见附录C.3。MissForest速度过慢（每个数据集超过5分钟），因此不予考虑。

原则上，对于单列缺失，也可以将 MICE 与 MISNN 结合起来。不过，我们通过在附录D中提出算法3来避免这样做；该算法以并行计算的方式，对一般缺失模式采用不同的处理方法。在实验中，我们在算法3的步骤（3）中使用不同的网络结构：MISNN-wide 使用两个宽度为500的隐藏层，每层后接 ReLU激活、批归一化层 [17] 和 Dropout层 [29]，丢弃率为0.1。MISNN-narrow 中的神经网络与 MISNN-wide 中的相同，但隐藏层的宽度为50。

结果汇总于表1。值得指出的是，与其他插补方法相比，所有 MISNN 方法的估计偏差都最小。MISNN 的插补均方误差、统计覆盖率和计算速

度也令人满意。相比之下，两种矩阵补全方法的插补均方误差相当，但其覆盖率远低于多重插补方法。

值得注意的是，与 MISNN-narrow 相比，MISNN-wide 的插补均方误差和估计偏差往往更小。然而，前者的覆盖率不如后者，主要原因是标准误较小。我们建议，在实践中，如果主要关注插补准确性或参数估计，应采用具有宽隐藏层的 MISNN。如果更强调对感兴趣参数进行统计推断，则应为 MISNN 配备窄隐藏层。

5.2 预测视角

我们将 MISNN 应用于阿尔茨海默病神经影像学计划（ADNI）基因数据集 *。该数据集包含 649 位患者的逾 1.9 万个基因组特征，以及一个取值范围为 $[0.4, 0.6]$ 的响应变量——VBM右侧海马体积。我们选取了与响应变量相关性最高的前 1000 个特征，并重点研究响应变量与前 5 个特征之间的线性分析模型。由于无法获取线性模型中的真实系数，我们研究了完整数据分析所得估计系数与插补数据集所得估计系数之间的差异。我们在与响应变量相关性最高的前 3 个特征中人为生成随机缺失值，缺失率约为 65%。我们使用了包含一个宽度为 500 的隐藏层和一个批归一化层的 MISNN，并拟合响应变量$\mathbf{y}$ 与前五个特征 $\mathbf{D}_1 \sim \mathbf{D}_5$ 之间的线性回归，以进行下游预测。

我们的结果总结于表2，显示MISNN以较高的计算效率实现了较小的插补均方误差和预测均方误差，尤其是与其他多重插补方法相比时。此外，MISNN（以及SoftImpute）得到的估计量最接近完整数据分析的金标准。更多实验详情请参阅附录C。

6 讨论

在这项工作中，我们提出了 MISNN，这是一种用于表格/矩阵数据缺失值多重插补的新型基于深度学习的方法。我们证明，MISNN 能够灵活地适配任意特征选择和任意神经网络架构。MISNN 可使用现成优化器进行训练，计算速度快，并能为插补模型提供可解释性，同时对数据维度和缺失率具有鲁棒性。通过合成数据集和真实世界数据集开展的多项实验表明，MISNN 显著优于最先进的插补模型。

尽管 MISNN 适用于广泛的分析模型，但我们仅讨论了使用偏回归剔除处理连续型缺失值的情形。通过考虑广义部分线性模型（GPLM，详见第 4.1 节），我们可以很容易地将 MISNN 扩展到离散型缺失值问题。然而，

*完整的ADNI致谢请见 http://adni.loni.usc.edu/ wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf。

偏回归剔除技术通常对 GPLM 无效。因此，可能需要包括回代拟合在内的迭代方法来学习 MISNN，而这可能会比较缓慢。

7 致谢

本研究部分得到了美国国立卫生研究院资助项目 R01GM124111 的支持。本文内容完全由作者负责，并不一定代表美国国立卫生研究院的官方观点。

参考文献

1. Allen-Zhu, Z., Li, Y., Song, Z.：过参数化下深度学习的收敛理论。载于：国际机器学习大会。第242–252页。PMLR（2019）2. Barnard, J., Rubin, D.B.：杂录：多重插补下的小样本自由度。《生物统计学》**86**(4), 948–955（1999）3. Belloni, A., Chernozhukov, V., 等：高维稀疏模型中模型选择后的最小二乘法。《伯努利》**19**(2), 521–547（2013）4. Bogdan, M., Van Den Berg, E., Sabatti, C., Su, W., Cand`es, E.J.：SLOPE——通过凸优化实现自适应变量选择。《应用统计年鉴》**9**(3), 1103（2015）5. Buuren, S.v., Groothuis-Oudshoorn, K.：mice：R语言中的链式方程多变量插补。《统计软件杂志》第1–68页（2010）6. Caton, S., Malisetty, S., Haas, C.：插补策略对机器学习公平性的影响。《人工智能研究杂志》**74**, 1011–1035（2022）7. Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., Robins, J.：用于处理效应和结构参数的双重/去偏机器学习。《计量经济学杂志》**21**(1), C1–C68（2018）8. Clifton, C., Hanson, E.J., Merrill, K., Merrill, S.：差分隐私k近邻缺失数据插补。《ACM隐私与安全汇刊》**25**(3), 1–23（2022）9. Dai, Z., Bu, Z., Long, Q.：利用生成对抗网络解决高维分块缺失值问题的多重插补。载于：2021年第20届IEEE国际机器学习与应用大会（ICMLA）。第791–798页。IEEE（2021）10. Das, S., Drechsler, J., Merrill, K., Merrill, S.：差分隐私下的插补。arXiv预印本 arXiv:2206.15063（2022）11. Deng, Y., Chang, C., Ido, M.S., Long, Q.：高维数据存在时一般缺失数据模式的多重插补。《科学报告》**6**(1), 1–10（2016）12. Du, S., Lee, J., Li, H., Wang, L., Zhai, X.：梯度下降能够找到深度神经网络的全局极小值。载于：国际机器学习大会。第1675–1685页。PMLR（2019）13. Friedman, J., Hastie, T., Tibshirani, R.：关于组套索和稀疏组Lasso的说明。arXiv预印本 arXiv:1001.0736（2010）14. Gondara, L., Wang, K.：Mida：使用去噪自编码器进行多重插补。载于：亚太知识发现与数据挖掘会议。第260–272页。施普林格（2018）15. Gramacy, R.B.：monomvn：对具有单调缺失的数据进行多元正态分布和学生t分布估计。R软件包版本，第1–8页（2010）

16. HANS, C.：贝叶斯Lasso回归。《生物统计学》pp. 1–11（2009） 17. Ioffe, S., Szegedy, C.：批量归一化：通过减少内部协变量偏移加速深度网络训练。载于：国际机器学习大会。pp. 448–456。PMLR（2015） 18. Jagannathan, G., Wright, R.N.：缺失数据的隐私保护插补。《数据与知识工程》**65**(1), 40–56（2008） 19. Li, S.C.X., Jiang, B., Marlin, B.：Misgan：利用生成对抗网络从不完整数据中学习。arXiv预印本 arXiv:1902.09599（2019） 20. Little, R.J., Rubin, D.B.：《缺失数据的统计分析》。Wiley（2002） 21. Mart´ınez-Plumed, F., Ferri, C., Nieves, D., Hern´andez-Orallo, J.：公平性与缺失值。arXiv预印本 arXiv:1905.12728（2019） 22. Mattei, P.A., Frellsen, J.：Miwae：不完整数据的深度生成建模与插补。arXiv预印本 arXiv:1812.02633（2018） 23. Mazumder, R., Hastie, T., Tibshirani, R.：用于学习大型不完整矩阵的谱正则化算法。《机器学习研究杂志》**11**, 2287–2322（2010） 24. Park, T., Casella, G.：贝叶斯Lasso。《美国统计学会杂志》**103**(482), 681–686（2008） 25. Robinson, P.M.：根号 n 一致的半参数回归。《计量经济学：计量经济学会杂志》pp. 931–954（1988） 26. Rubin, D.B.：调查中无响应的多重插补，第81卷。John Wiley & Sons（2004） 27. Simon, N., Friedman, J., Hastie, T., Tibshirani, R.：稀疏组Lasso。《计算与图形统计学杂志》**22**(2), 231–245（2013） 28. Srebro, N., Rennie, J., Jaakkola, T.：最大间隔矩阵分解。《神经信息处理系统进展》**17**, 1329–1336（2004） 29. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.：丢弃法：一种防止神经网络过拟合的简单方法。《机器学习研究杂志》**15**(1), 1929–1958（2014） 30. Stekhoven, D.J., B¨uhlmann, P.：Missforest——用于混合类型数据的非参数缺失值插补。《生物信息学》**28**(1), 112–118（2012） 31. Tibshirani, R.：通过Lasso进行回归收缩与变量选择。《英国皇家统计学会杂志：B辑（方法论）》**58**(1), 267–288（1996） 32. Van Buuren, S.：通过完全条件设定对离散和连续数据进行多重插补。《医学研究统计方法》**16**(3), 219–242（2007） 33. Xu, S.：Debinet：利用非线性过参数化神经网络对线性模型进行去偏。arXiv预印本 arXiv:2011.00417（2020） 34. Yoon, J., Jordon, J., Van Der Schaar, M.：Gain：使用生成对抗网络进行缺失数据插补。arXiv预印本 arXiv:1806.02920（2018） 35. Zhang, Y., Long, Q.：公平性感知的缺失数据插补。载于：NeurIPS 2022可信且具社会责任的机器学习研讨会 36. Zhang, Y., Long, Q.：缺失数据插补中的公平性。arXiv预印本 arXiv:2110.12002（2021） 37. Zhao, Y., Long, Q.：高维数据情形下的多重插补。《医学研究统计方法》**25**(5), 2021–2035（2016） 38. Zou, H.：自适应Lasso及其预言机性质。《美国统计学会杂志》**101**(476), 1418–1429（2006） 39. Zou, H., Hastie, T.：通过弹性网络进行正则化与变量选择。《英国皇家统计学会杂志：B辑（统计方法论）》**67**(2), 301–320（2005）

A 分析模型与 Rubin 法则

本节将说明 Rubin 法则的用途及其理论基础。假设估计目标 $\boldsymbol{\theta}$ 是我们最终感兴趣的量，并且如果观测到完整数据集，就可以计算出它。收集到 M 插补¯数据集后，我们利用 Rubin 法则推断 $\boldsymbol{\theta}$ ，并谨慎量化模型不确定性。Rubin 法则的具体实现见算法 2。

无末尾 2 用于计算置信区间的 Rubin 法则**输入：**

M 个插补数据集，显著性水平 s

(1) 根据分析模型，计算第 m 个插补数据集中的 $\hat{}\boldsymbol{\theta}^{(m)}$ 和 $\mathrm{SE}\left(\hat{}\boldsymbol{\theta}^{(m)}\right)$ 。(2) 通过 $\bar{\boldsymbol{\theta}} = \sum_{m=1}^{M} \hat{}\boldsymbol{\theta}^{(m)}$ 计算合并均值。(3) 通过以下方式计算合并方差：

$$
\begin{aligned}
\mathrm{Var}_{\mathrm{within}} &= \frac{\sum_{m=1}^{M} \mathrm{SE}\left(\hat{\boldsymbol{\theta}}^{(m)}\right)^2}{M} \\
\mathrm{Var}_{\mathrm{between}} &= \frac{\sum_{m=1}^{M} (\hat{\boldsymbol{\theta}}^{(m)} - \bar{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}}^{(m)} - \bar{\boldsymbol{\theta}})^{\top}}{M-1} \\
\mathrm{Var}_{\mathrm{total}} &= \mathrm{Var}_{\mathrm{within}} + \left(1 + \frac{1}{M}\right) \mathrm{Var}_{\mathrm{between}}
\end{aligned}
$$

(4) 构造端点如下的 $(1-s)$ 置信区间：其中 Φ 是标准正态分布的累积分布函数，

$$
\bar{\boldsymbol{\theta}} \pm \Phi^{-1}\left(1 - \frac{s}{2}\right) \sqrt{\mathrm{Var}_{\mathrm{total}}}
$$

注 2 给定合并均值和合并方差后，还可以用其他方法构造置信区间。例如，可以将高斯分位数 $\Phi^{-1}\left(1-\frac{s}{2}\right)$ 替换为 t 分布分位数 $t_{\mathrm{df},\ 1-\frac{s}{2}}$，自由度（自由度）的常见取值为 自由度$_{old}$ *[2]* 或 自由度$_{adjusted}$ *[26]*。 ¯

这里我们解释 Rubin 法则。根据全期望定律：

$$
\mathbb{E}(\boldsymbol{\theta}|\mathbf{D}_{\mathrm{obs}}) = \mathbb{E}(\mathbb{E}[\boldsymbol{\theta}|\mathbf{D}_{\mathrm{obs}}, \mathbf{D}_{\mathrm{miss}}]|\mathbf{D}_{\mathrm{obs}})
$$

这个方程说明，在合并多重插补的结果时，应采用 Rubin 法则。假设 $\hat{}\boldsymbol{\theta}^{(m)}$ 是第 m 次插补的估计值，则合并均值为

$$
\bar{\boldsymbol{\theta}} = \sum_{m=1}^{M} \hat{\boldsymbol{\theta}}^{(m)}
$$

根据全方差公式， $\boldsymbol{\theta}$ 基于观测数据的后验方差来自两个方面：

$$\begin{aligned}\mathrm{Var}(\boldsymbol{\theta}|\mathbf{D}_{\mathrm{obs}}) = \mathbb{E}[\mathrm{Var}(\boldsymbol{\theta}|\mathbf{D}_{\mathrm{obs}}, \mathbf{D}_{\mathrm{miss}})|\mathbf{D}_{\mathrm{obs}}] \\ + \mathrm{Var}[\mathbb{E}(\boldsymbol{\theta}|\mathbf{D}_{\mathrm{obs}}, \mathbf{D}_{\mathrm{miss}})|\mathbf{D}_{\mathrm{obs}}]\end{aligned}$$

第一部分是 $\boldsymbol{\theta}$ 在插补数据集中的后验方差均值，称为“组内方差”。第二部分是 $\boldsymbol{\theta}$ 后验均值之间的方差，称为“组间方差”。给定 M 个插补数据集，组内方差可估计为

$$\mathrm{Var}_{\mathrm{within}} = \frac{\sum_{m=1}^{M} \mathrm{Var}\left(\hat{\boldsymbol{\theta}}^{(m)}\right)}{M}$$

组间方差可以通过以下方法估计。

$$\mathrm{Var}_{\mathrm{between}} = \frac{\sum_{m=1}^{M} (\hat{\boldsymbol{\theta}}^{(m)} - \bar{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}}^{(m)} - \bar{\boldsymbol{\theta}})^{\top}}{M-1}$$

¯其中 $\boldsymbol{\theta}$ 按上述方式计算。

我们强调，直接将 $\mathrm{Var}_{\mathrm{within}}$ 和 $\mathrm{Var}_{\mathrm{between}}$ 相加以推导¯总方差是不正确的，会低估方差。因为 $\boldsymbol{\theta}$ 是通过有限数量的 M 数据集估计的，而不是通过 $M \to \infty$估计的。有限估计与无限估计之间的差异可以用 $\frac{1}{M}\mathrm{Var}_{\mathrm{between}}$[26]来表征。因此，总而言之， $\boldsymbol{\theta}$ 的后验方差按照 Rubin 法则估计为

$$\mathrm{Var}_{\mathrm{total}} = \mathrm{Var}_{\mathrm{within}} + \left(1 + \frac{1}{M}\right)\mathrm{Var}_{\mathrm{between}}$$

B 非高斯缺失值

当缺失值不服从高斯分布时，我们在算法 1 的步骤 (3) 中拟合广义部分线性模型（广义部分线性模型），从此处开始，该算法与算法 1 不同。

$$\begin{aligned}\mathbb{E}(\mathbf{D}_1|\mathbf{X}, \mathbf{T}) &= h^{-1}\left(\alpha_0 + \mathbf{D}_{\mathrm{full},\mathcal{S}}\boldsymbol{\beta} + \overline{\mathbf{D}_{\mathrm{full},\mathcal{S}}}\boldsymbol{\gamma}\right) \\ &= h^{-1}\left(\alpha_0 + \mathbf{X}\boldsymbol{\beta} + f(\mathbf{T})\right)\end{aligned}$$

例如，当缺失值为二元变量时，我们拟合逻辑部分线性模型，其中 h 为 logit 函数。

然而，我们不估计条件期望 $\mathbb{E}(\mathbf{D}_1|\mathbf{T})$ 和 $\mathbb{E}(\mathbf{X}|\mathbf{T})$，因为广义部分线性模型不适用偏回归剔除技术，这是由于 h 的非线性。相反，我们使用回代拟合等方法来近似 $\boldsymbol{\beta}$ 的后验分布以及广义部分线性模型中的函数 f。

将任何拟合方法视为学习广义部分线性模型的黑箱，我们给出两个多重插补的示例，其中

– $\mathbf{D}_1$ 服从伯努利分布，则 $h(z) = \log(\frac{z}{1-z})$。我们从 $\boldsymbol{\beta}$ 的后验分布中抽取 $\hat{\boldsymbol{\beta}}^{(m)}$，并进一步从中抽取二元插补值

$$\begin{aligned}\widehat{\mathbf{D}}_{i,1}^{(m)} &\sim \text{Bernoulli}(\pi_i^{(m)}) \\ \pi_i^{(m)} &= \frac{1}{1+\exp(-\mathbf{X}_i\hat{\boldsymbol{\beta}}^{(m)} - \hat{f}(\mathbf{T}_i))}\end{aligned}$$

– $\mathbf{D}_1$ 服从泊松分布，则 $h(z) = \log(z)$。我们从 $\boldsymbol{\beta}$ 的后验分布中抽取 $\hat{\boldsymbol{\beta}}^{(m)}$，并进一步抽取正值且离散的插补值。

$$\begin{aligned}\widehat{\mathbf{D}}_{i,1}^{(m)} &\sim \text{Poisson}(\mu_i^{(m)}) \\ \mu_i^{(m)} &= \exp(\mathbf{X}_i\hat{\boldsymbol{\beta}}^{(m)} + \hat{f}(\mathbf{T}_i))\end{aligned}$$

随后，Rubin 法则以与算法 2 相同的方式应用。

C 实验详情

在介绍实验细节之前，我们希望简要介绍其他采用特征选择的多重插补方法。

通过贝叶斯*Lasso*进行多重插补 贝叶斯Lasso [24]通过对回归系数 $\boldsymbol{\alpha}$赋予双指数（拉普拉斯）先验分布，为Lasso构建了一个分层贝叶斯模型：

$$\rho\left(\boldsymbol{\alpha} \mid \sigma^2, \lambda, \rho\right) = \prod_{j=1}^{p-1} \frac{\lambda}{2\sqrt{\sigma^2}} \exp \frac{-\lambda\left|\alpha_j\right|}{\sqrt{\sigma^2}} \tag{2}$$

这里 a，b，r，s 是预先设定的超参数，用于控制方差 σ^2 和惩罚参数 λ的先验分布：$\sigma^2 \sim$ 逆伽马分布(a，b) 和$\lambda \sim$ 伽马分布(r，s)。抽样过程通常通过马尔可夫链蒙特卡罗（MCMC）进行。然而，在高维情形下，抽样过程极其耗时，使得贝叶斯Lasso虽然准确，却不切实际。

备注 3（贝叶斯Lasso需要预热）为了从贝叶斯*Lasso*中获得令人满意的结果，抽样过程中必须进行大量预热迭代。因此，计算成本可能很高。

提出了两种替代的多重插补（MI）方法：DURR 和 IURR [37,11], ，这两种方法的计算效率均高于贝叶斯Lasso，且适用于 Lasso 以外的其他正则化方法。

预热迭代 偏差		插补均方误差
50	-0.7752	2.9925
500	-0.4774	1.6328
5000	-0.1986	0.9249
10000	-0.1979	0.9244

表3. 表4所设定的情形下，贝叶斯Lasso在不同预热期长度下的表现。

通过直接使用正则化回归进行多重插补（*DURR*）在第 m次插补中，DURR $\langle 1\rangle$ 通过从原始数据集$\mathbf{D}$中有放回抽样，生成大小为 n 的自助法数据集$\mathbf{D}^{(m)}$， $\langle 2\rangle$ 使用正则化回归获得估计值 $\boldsymbol{\alpha}^{(m)}$， $\langle 3\rangle$ 使用包含$\boldsymbol{\alpha}^{(m)}$的预测分布插补缺失值。值得注意的是，DURR是唯一使用自助法近似 $\boldsymbol{\alpha}$后验分布的方法。

通过间接使用正则化回归进行多重插补（*IURR*）在第 m次插补中，IURR $\langle 1\rangle$ 拟合正则化回归并识别活动集 $\mathcal{S}$， $\langle 2\rangle$ 使用最大似然近似 $\boldsymbol{\alpha}$的后验分布， $\langle 3\rangle$ 根据 $\boldsymbol{\alpha}^{(m)}$的预测分布插补缺失值。IURR 的思路是只对选定（重要）的条目进行推断，而忽略其他条目。前两个步骤合起来称为 OLS 后套索 [3] ，与 MISNN 不同的是，我们使用 PLM，它还利用了未选中特征中的信息。

C.1 合成缺失数据：单列缺失

在本小节中，我们将 MISNN 与其他最先进的插补方法在单变量缺失模式的合成数据上的表现进行比较。我们研究与 [37]中类似的设定。在所有模拟中，我们固定样本量 $n = 100$ 和特征数量 $p = 1000$。每个模拟数据集包含标签 y 和预测变量集合 $\mathrm{D} \in \mathbb{R}^{n\times p}$，其中 D_1 包含缺失值。$(\mathrm{D}2, \ldots, \mathrm{D}p)$ 服从均值为 0、具有 0.5 自相关的一阶自回归协方差矩阵的多元正态分布。D_1 服从方差为 1、均值为$\alpha_0 + \mathrm{D}_S\boldsymbol{\alpha}$的正态分布，其中 $\mathrm{D}_S = \{\mathrm{D}2, \ldots, \mathrm{D}_{11}, \mathrm{D}50, \ldots, \mathrm{D}_{59}\}$ 且 $\alpha_j = \sqrt{0.2}$。标签 y 服从均值为 $\beta_0 + \beta_1\mathrm{D}_1 + \beta_2\mathrm{D}_2 + \beta_3\mathrm{D}_3$ $(\beta_j = 1, \mathrm{A} = [\mathrm{D}_1, \mathrm{D}_2, \mathrm{D}_3])$、方差为 1 的正态分布。

为生成随机缺失（MAR）数据，缺失值指示变量遵循 $\mathrm{logit}[\Pr(\mathbf{D}_1\text{缺失}|\mathbf{D}_{-1}, \mathbf{y})] = 3 - 0.1\mathbf{D}_2 + 3\mathbf{D}_3 - 2\mathbf{y}$, 因此约有 50% 的$\mathbf{D}_1$缺失。我们在表4中总结了不同插补方法的表现。显然，MISNN 能够稳健地应对如此高维和高缺失率的情况：MISNN 的速度显著快于其他方法，例如贝叶斯Lasso；这些方法的覆盖率合理且偏差较小。尽管矩阵补全方法和 Lasso 的插补均方误差最小，但这些单一插补方法会导致不恰当的推断，估计值 β_1 存在偏差，且覆盖率较低。值得注意的是，GAIN 的表现最差，因为它只适用于完全随机缺失机制。

对于所有需要进行特征选择的 方法，我们始终使用相同的 Lasso 惩罚（$\lambda = 0.1$）。对于多重插补方法，我们生成 30 个插补数据集。我们使用 R 实现贝叶斯Lasso（R 软件包 monomvn [15]），并进行 10000 次预热迭代；在先验模型 (2) 中使用 $(a, b, r, s) = (0.1, 0.1, 0.01, 0.01)$，并使用 TensorFlow 实现 GAIN [34]。我们使用 Python/pythorch 实现其他方法。

MISNN 详情：我们采用两层全连接神经网络作为 MISNN 的结构。隐藏层宽度为 500。激活函数为 ReLU，并应用了批量归一化。我们使用 Adam 优化器，学习率为 10^{-3}。我们采用早停机制，耐心值为 1。

方法	风格	Bias	插补均方误差	覆盖率	秒	SE	SD
完整数据	-	**-0.0057**	-	**0.951**	-	0.1412	0.1419
完整案例	-	0.2052	-	0.754	-	0.1823	0.1977
均值填补	SI	0.8086	2.9364	0.021	**0.014**	0.2145	0.1833
多重插补 MISNN-Lasso	MI	**0.0894**	**0.8431**	**0.938**	**0.114**	0.1844	0.1690
贝叶斯Lasso	MI	-0.1979	**0.9244**	**0.840**	82.72	0.2035	0.2708
DURR-Lasso	MI	0.4244	1.5518	0.318	**0.022**	0.1798	0.1305
IURR-Lasso	MI	0.4542	1.7734	0.146	**0.038**	0.1551	0.1099
GAIN	SI	0.6277	3.7016	0.430	89.75	0.1500	0.4970
SoftImpute	SI	-0.4199	**0.4306**	0.320	**0.043**	0.1599	0.2369
MMMF	SI	-0.4384	**0.4242**	0.302	2.070	0.1668	0.1999
Lasso	SI	-0.5350	**0.4030**	0.132	**0.032**	0.1521	0.2007

表4. 在500个蒙特卡洛数据集上的合成数据单特征缺失模式。偏差：平均偏差 $\hat{}\ \beta_1 - \beta_1$；插补均方误差：插补均方误差 $\|\widehat{\mathbf{D}}_{\text{miss},\,1} - \mathbf{D}_{\text{miss},\,1}\|^2 / n_{\text{miss}}$；覆盖率：$\beta_1$的95%置信区间覆盖概率 ^；秒：实际插补时间；标准误：β_1的平均标准误；标准差：$\hat{}\ \beta_1$的蒙特卡洛标准差。数据生成和模型设置见附录 C.1。

C.2 合成缺失数据：多个缺失列

数据生成：这里我们采用了 [11] 中的数据生成流程，其中数据集的三列包含缺失值。与单列缺失模式的设置类似，每个模拟数据集的样本量为 $n = 200$，并包含 y。每个观测有 $p = 1000$ 个特征，前三个特征包含缺失值。我们首先从均值为 0、相关性为 0.5 的一阶自回归协方差矩阵所对应的多元正态分布中生成 $(\mathrm{D}4, \ldots, \mathrm{D}p)$。给定 $(\mathrm{D}4, \ldots, \mathrm{D}p)$，从正态分布 $\mathcal{N}(0, \mathrm{D}_S\boldsymbol{\alpha})$ 中独立生成 $(\mathrm{D}_1, \mathrm{D}_2, \mathrm{D}_3)$，其中 $\mathrm{D}_S = \{\mathrm{D}2, \ldots, \mathrm{D}_{11}, \mathrm{D}50, \ldots, \mathrm{D}_{59}\}$ 且 $\alpha_j = \sqrt{0.2}$。标签 y 由均值为 $\beta_0 + \beta_1\mathrm{D}_1 + \beta_2\mathrm{D}_2 + \beta_3\mathrm{D}_3 + \beta_4\mathrm{D}_4 + \beta_5\mathrm{D}_5 (\beta_j = 1)$、方差为 6 的正态分布生成。

通过逻辑回归模型生成 $(\mathbf{D}_1, \mathbf{D}_2, \mathbf{D}_3)$ 中的缺失值。假设缺失指示变量分别为 $\delta_1, \delta_2, \delta_3$。那么逻辑回归模型为 $\text{logit}(\delta_1 = 1) = -1 - \mathbf{D}_4 + 2\mathbf{D}_5 - \mathbf{y}$，$\text{logit}(\delta_2 = 1) = -1 - \mathbf{D}_4 + 2\mathbf{D}_{51} - \mathbf{y}$，$\text{logit}(\delta_3 = 1) = -1 - \mathbf{D}_{50} + 2\mathbf{D}_{51} - \mathbf{y}$。

在插补步骤中，我们采用了两种特征选择方法：Lasso（$\lambda = 0.2$）和弹性网络（$\lambda = 1.0$，惩罚比率为 ℓ_1 0.5）。所有需要使用这些方法的方法都采用相同的惩罚项。以 Adam 作为优化器时，MISNN-wide 的学习率设为 0.01，MISNN-narrow 的学习率设为 0.001。MISNN-wide 训练 5 个训练轮次，MISNN-narrow 训练 15 个训练轮次。我们还为两个网络都加入了学习率衰减，衰减率为 0.6，每两步衰减一次，每一步在处理一个批次后进行。

C.3 阿尔茨海默病数据 udy

在本小节中，我们将详细介绍 ADNI数据集 的实验。原始数据包含 649 名受试者；每名受试者有 19822 个特征和一个连续响应变量 y。我们首先通过去除均值并缩放至单位方差，对所有特征和响应变量进行标准化。然后选出与响应变量相关性最高的 1000 个特征，并按降序排列。将这些特征记为 $\mathbf{A} = [\mathbf{A}1, \ldots, \mathbf{A}_{1000}]$。根据逻辑回归模型 $\text{logit}(\delta_1 = 1) = -1 + 2\mathbf{D}_4 + \mathbf{D}_5 + 2\mathbf{y}$、$\text{logit}(\delta_2 = 1) = -1 + \mathbf{D}_4 + 2\mathbf{D}_{51} + 3\mathbf{y}$、$\text{logit}(\delta_3 = 1) = -1 - \mathbf{D}_{50} + 2\mathbf{D}_{51} + \mathbf{y}$，在 $[\mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3]$ 中按随机缺失机制生成缺失值。之后，通过合并特征矩阵 A 和响应变量 y，得到预处理数据集 D。对于所有需要特征选择的方法，均使用预先指定的惩罚项。对于 Lasso，惩罚项为 $\lambda = 0.1$；对于弹性网络，惩罚项为 $\lambda = 0.8$，惩罚比率为 ℓ_1 0.5。MISNN 的结构与附录 C.1 中相同。

D MISNN 变体

D.1 用于单次插补的 MISNN

假设只关注插补和预测任务，那么在 MISNN 中无需从后验分布和预测分布中进行抽样。我们在算法4中展示了处理单变量缺失模式时 MISNN 的单次插补版本。如附录B和第4.4节所述，这种单次插补算法可以直接扩展到包含离散缺失值或一般缺失模式的情形。

noend 3 具有多个缺失列的 MISNN**输入：** 不完整数据 $\mathbf{D}$，插补次数 M

(1) **针对** $k \in \{1, \ldots, K\}$ **进行**拟合正则化回归 $\mathbf{D}_{\text{obs},k} \sim \mathbf{D}_{\text{obs},-[K]}$ by

$$(\widehat{\boldsymbol{\alpha}}_k, \hat{\alpha}_{0,k}) := \underset{(a,a_0)}{\operatorname{argmin}} \frac{1}{2} \|\mathbf{D}_{\text{obs},k} - \mathbf{D}_{\text{obs},-[K]}\boldsymbol{a} - a_0\|^2 + P(\boldsymbol{a})$$

其中 P 是惩罚函数。

获得活动集 $\mathcal{S}_k := \{i : \hat{\alpha}_{k,i} \neq 0\}$

通过取交集 $\mathcal{S} = \cap_{k=1}^{K} \mathcal{S}_k$ 或并集 $\mathcal{S} = \cup_{k=1}^{K} \mathcal{S}_k$，将 K 个活动集合并为一个活动集。

(2) 将 $\mathbf{D}_{\text{full},-[K]}$ 拆分为子矩阵 $\mathbf{X} = [\mathbf{D}_{\text{full},-[K]}]_{\mathcal{S}}$ 和 $\mathbf{T} = \mathbf{D}_{\text{full},-[K]} \backslash \mathbf{X}$。

(3) 给定训练数据 $\{\mathbf{T}_{\text{obs}}、\mathbf{D}_{\text{obs},[K]}、\mathbf{X}_{\text{obs}}\}$，训练神经网络以学习

$$\eta_D(\mathbf{T}) := \mathbb{E}(\mathbf{D}_{\text{full},[K]}|\mathbf{T}), \eta_X(\mathbf{T}) := \mathbb{E}(\mathbf{X}|\mathbf{T})$$

(4)**对** $k \in \{1, \ldots, K\}$ **执行**

应用标准最大似然技术于

$$\mathbf{D}_{\text{obs},k} - \mathbb{E}(\mathbf{D}_{\text{obs},k}|\mathbf{T}_{\text{obs}}) = (\mathbf{X}_{\text{obs}} - \mathbb{E}(\mathbf{X}_{\text{obs}}|\mathbf{T}_{\text{obs}}))\,\boldsymbol{\beta}_k + \boldsymbol{\epsilon}$$

其中 $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma_k^2)$，并近似后验分布

$$\rho_2\left(\boldsymbol{\beta}_k, \sigma_k \middle| \mathbf{D}_{\text{obs},k} - \eta_D(\mathbf{T}_{\text{obs}}), \mathbf{X}_{\text{obs}} - \eta_X(\mathbf{T}_{\text{obs}})\right)$$

(5)
for $m \in \{1, \ldots, M\}$ **do**
  for $k \in \{1, \ldots, K\}$ **do**
    Randomly draw $\hat{\boldsymbol{\beta}}_k^{(m)}, \hat{\sigma}_k^{(m)}$ from posterior distribution $\rho_2\left(\boldsymbol{\beta}_k, \sigma_k \middle| \mathbf{D}_{\text{obs},k}, \mathbf{X}_{\text{obs}}, \mathbf{T}_{\text{obs}}\right)$.
    Subsequently, impute $\mathbf{D}_{\text{miss},k}$ with $\widehat{\mathbf{D}}_{\text{miss},k}^{(m)}$ by drawing randomly from the predictive distribution $\rho_1\left(\mathbf{D}_{\text{miss},k}|\mathbf{X}_{\text{miss}}, \mathbf{T}_{\text{miss}}, \hat{\boldsymbol{\beta}}_k^{(m)}, \hat{\boldsymbol{\sigma}}_k^{(m)^2}\right)$

noend 4 基于半参数神经网络的单次插补 (SISNN)**输入：** 不完整数据 $\mathbf{D}$，插补次数 M

(1) 拟合正则化回归 $\mathbf{D}_{\text{obs, }1}$ ~$\mathbf{D}_{\text{obs, }-1}$，方法是

$$(\widehat{\boldsymbol{\alpha}}, \hat{\alpha}_0) := \underset{(a,a_0)}{\operatorname{argmin}} \frac{1}{2} \|\mathbf{D}_{\text{obs},1} - \mathbf{D}_{\text{obs},-1}\boldsymbol{a} - a_0\|^2 + P(\boldsymbol{a})$$

其中，P 是惩罚函数。

(2) 获取活动集 $\mathcal{S} := \{i : \hat{}\alpha_i \neq 0\}$，并将 $\mathbf{D}_{-1}$ 拆分为子矩阵 $\mathbf{X} = [\mathbf{D}_{-1}]_{\mathcal{S}}$ 和 $\mathbf{T} = \mathbf{D}_{-1} \backslash \mathbf{X}$。

(3) 给定训练数据 $\{\mathbf{T}_{\text{obs}}, \mathbf{D}_{\text{obs},1}, \mathbf{X}_{\text{obs}}\}$，训练神经网络以学习

$$\eta_D(\mathbf{T}) := \mathbb{E}(\mathbf{D}_1|\mathbf{T}), \eta_X(\mathbf{T}) := \mathbb{E}(\mathbf{X}|\mathbf{T})$$

(4) 使用普通最小二乘法在.....上推导 $\hat{} \boldsymbol{\beta}$

$$\mathbf{D}_{\text{obs},1} - \mathbb{E}(\mathbf{D}_{\text{obs},1}|\mathbf{T}_{\text{obs}}) = (\mathbf{X}_{\text{obs}} - \mathbb{E}(\mathbf{X}_{\text{obs}}|\mathbf{T}_{\text{obs}}))\,\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

(5) 使用.....进行插补

$$\mathbf{D}_{\text{miss},1} = [\mathbf{X}_{\text{miss}} - \eta_X(\mathbf{T}_{\text{miss}})]\,\hat{\boldsymbol{\beta}} + \eta_D(\mathbf{T}_{\text{miss}})$$