arXiv:2601.00517v1 [stat.ML] 2 Jan 2026

生成式条件缺失值插补网络

George Sun[1] 和 Yi-Hui Zhou[1, 2*]

[1]生物信息学研究中心，北卡罗来纳州立大学，Lampe Drive 1号，罗利，27695，北卡罗来纳州，美国。[2*]生物科学与统计学系，北卡罗来纳州立大学，Lampe Drive 1号，罗利，27695，北卡罗来纳州，美国

*通讯作者。电子邮箱：yihui zhou@ncsu.edu；贡献者：gsun4@ncsu.edu；

摘要

本研究提出了一种先进的生成式条件策略，旨在对数据集中的缺失值进行插补；这在统计分析中具有重要意义。具体而言，我们首先阐述生成式条件缺失值插补网络（GCMI）的理论基础，并证明其在完全随机缺失（MCAR）和随机缺失（MAR）机制下具有稳健的性质。随后，我们通过结合采用链式方程法的多重插补框架，增强了 GCMI 的稳健性和准确性。这一创新有助于提升模型稳定性，并显著改善插补效果。最后，通过一系列严谨的模拟，以及利用基准数据集进行的实证评估，我们证明，与目前可用的其他领先插补技术相比，所提出的方法效果更优。这项全面评估不仅凸显了 GCMI 的实用性，也进一步证实了它作为统计数据分析领域前沿工具的潜力。

关键词：生成式神经网络；缺失值插补；链式方程多重插补

1 引言

在处理现实世界的数据时，缺失数据是一个普遍存在的问题，常常阻碍下游分析中统计方法的应用与验证。处理缺失数据时，缺失机制是首先需要考虑的关键因素，因为选择合适的缺失值插补方法取决于能否找出数据背后的正确缺失模式 [1]。缺失机制可归纳为三类：完全随机缺失（MCAR）、随机缺失（MAR）和非随机缺失（MNAR）。完全随机缺失指数据缺失的可能性与数据本身无关，无论数据是观测到的还是未观测到的。形式上，当满足以下条件时，完全随机缺失成立：$\Pr(R \mid X) = \Pr(R)$。这里，“R”和“X”分别表示缺失性和数据（包括观测数据与缺失数据）。在完全随机缺失的情况下，数据与缺失值的出现之间不存在明显联系。[2]。对于缺失数据，更灵活且更符合现实的假设是随机缺失：在这种情况下，缺失的可能性取决于观测数据，但在考虑观测信息后，与未观测数据无关。数学上，当满足以下条件时，随机缺失成立：$\Pr(R \mid X) = \Pr(R \mid X_{obs})$。缺失性的最后一类是非随机缺失，其缺失的可能性与观测数据和未观测数据都有关。数学上，非随机缺失可由以下方程描述：$\Pr(R \mid X) = \Pr(R \mid X_{obs}, X_{mis})$ 这表明，即使考虑了观测信息，缺失性与数据之间的关系仍然存在。

对缺失数据处理不当的传统方法通常会导致样本量和统计功效下降、数据表示信息丢失以及结果偏倚 [3]。例如，完整案例分析假设数据完全随机缺失；若在给定模型中变量的条件下，缺失性与感兴趣的结局相互独立，则在其他缺失机制下会引入偏倚。由于未使用被删除的数据，这种方法总会导致统计功效损失。用于纵向研究的末次观测值结转法通常会产生偏倚 [4]。单一插补法会填补缺失值，通常采用回归预测模型，但它未能将插补中的不确定性纳入考虑，从而导致假阳性发现增加，并使下游分析产生偏倚 [5]。相较之下，链式方程多重插补（MICE）等多重插补（MI）方法 [6, 7] 更稳健且能提供更多信息，因为它们能够更准确地捕捉各变量的潜在分布，并提供不确定性估计。

近期深度生成方法的发展展现出良好的性能，能够捕捉复杂高维数据的潜在结构、变量间相关性和表示 [8]。例如，Yoon 等人 [9] 提出了生成对抗网络（GAIN），用于学习目标分布，并为判别器提供有关原始缺失性的信息。Nazabal 等人 [10]设计了变分自编码器，用于拟合不完整异质数据（HI-VAE）。Li、Jiang 和 Marlin 在 2019 [11] 提出的 MISGAN，着重于利用 GAN 框架中的生成模型提高插补准确性。Ghalebikesabi 等人提出了“深度生成式缺失模式集混合

模型”[12]，通过深度生成技术应对缺失数据模式的复杂性，为插补质量开辟了新的维度。《不完整数据集的深度生成建模与插补》（MIWAE）[13]致力于通过创新的生成技术提高插补数据的质量。MIWAE 模型的扩展——《缺失非随机数据的深度生成建模》（not-MIWAE）[14], 进一步展示了深度生成建模在复杂真实数据集中的应用，并解决数据非随机缺失这一棘手问题。Ma 和 Zhang 于 2021 年通过其关于“可识别生成模型”的研究为该领域作出了贡献[15]，这些模型旨在插补非随机缺失的数据，从而应对更复杂的缺失数据情形。上述研究凸显了人们对使用生成对抗网络进行缺失数据插补日益增长的兴趣，以及该领域不断涌现的创新。GAIN 只能处理同质数据类型，而 HI-VAE 对异质数据采用不同的似然模型。本文提出一种生成式条件方法，用于插补给定数据集中的缺失值。该方法旨在弥补现有文献中的关键空白，重点是在不施加分布假设的情况下，灵活处理连续数据和分类数据。具体而言，我们首先证明，生成式条件缺失值插补网络（GCMI）在 MCAR 和随机缺失机制下具有理论性质；随后结合基于链式方程的方法进行多重插补，以提高模型的稳定性和插补性能。最后，我们在基准数据集上开展了广泛的模拟实验和真实数据实验，展示所提方法相较于其他最先进的缺失插补方法具有更优的性能。我们的方法旨在提升插补数据的质量和准确性，使其成为处理各领域不完整数据集的实用工具。

2 问题表述

考虑一个包含 N 个样本/对象和 P 个特征/变量的数据矩阵。将数据矩阵记为 $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}2, \ldots, \mathbf{x}_P) \in \mathbb{R}^{N \times P}$，其中 $\mathbf{x}_j = (x_{1j}, \ldots, x_{Nj})^T$ 是一个特征向量，包含第 j 个特征的 N 个属性；$i = 1, \ldots, N$ 是行索引。每个 N 个属性都可能缺失或已观测。当 $\mathbf{x}_j$ 既非完全缺失也非完全观测时，可根据每个特征向量 $\mathbf{x}_j$ 的缺失值，将数据矩阵分为四部分：1. 变量 $\mathbf{x}_j$ 的已观测属性，记为 $\mathbf{x}_j^{obs}$；2. 变量 $\mathbf{x}_j$ 的缺失属性，记为 $\mathbf{x}_j^{mis}$；3. 特征矩阵中对应于已观测 $\mathbf{x}_j$ 的其余部分，记为 $\mathbf{X}_{(-j)}^{obs}$；4. 特征矩阵中对应于 $\mathbf{x}_j$ 缺失值的其余部分，记为 $\mathbf{X}_{(-j)}^{mis}$。此外，我们使用索引集 miss(j)、obs(j)分别表示第 j 个特征中缺失和已观测的行，并用 n_j 表示 obs(j) 中元素的数量。目标是在不对观测数据过拟合的前提下，为每个特征 j 插补缺失值。此外，缺失值插补通常还伴随结果预测等下游任务。例如，研究人员可能希望补全电子健康记录中的各类患者级特征，如人口统计信息、诊断、药物、生命体征和实验室数据，以预测不良事件或疾病

结果。当特征得到正确插补时，有望提升后续性能。

2.1 用于缺失值插补的生成式条件网络（GCMI）

对于每个变量 j，目标是根据$\mathbf{x}_j^{mis}$的观测值，生成缺失属性；这些观测值来自$\mathbf{X}_{(-j)}^{mis}$，可从特征 j 的条件分布$p(x_j|x_{\neg j})$中采样。因此，GCMI 是由 P 对条件生成器和判别器组成的生成式条件网络集成。

生成器 G_j：第 j 个生成器 G_j 旨在对第 j 列中的缺失值进行插补。设 $\mathbf{z}$ 表示一个来自 $\mathcal{N}(\mathbf{0},\ \mathbf{I})$ 的独立 k 维噪声，则生成器 G_j：$\mathbb{R}^{P-1} \times \mathbb{R}^k \to \mathbb{R}^1$ 是一个将 $\mathbf{X}_{(-j)}$ 和噪声 $\mathbf{z}$ 映射为插补向量 $\hat{\mathbf{x}}_j = G_j(\mathbf{X}_{(-j)},\ \mathbf{z})$ 的函数。在训练阶段，生成器使用$\mathbf{X}_{(-j)}^{obs}$ 和 $\mathbf{x}_j^{obs}$来学习$p_{(\mathbf{x}_{(j)}|\mathbf{X}_{(-j)})}$的条件分布，然后我们将 G_j应用于 $\mathbf{X}_{(-j)}^{mis}$，以生成 $\mathbf{x}_j^{mis}$的插补值。

判别器 D_j：第 j 个判别器 D_j：$\mathbb{R}^{P-1} \times \mathbb{R}^1 \to (0,\ 1)$ 是一个函数，它根据关于 $x_{\neg j}$的信息，试图区分x_j 中的属性是观测值还是插补值。

我们将 G_j, D_j 的损失函数定义为：

$$\min_{D_j} L_D(D_j) = \frac{1}{2}\mathbb{E}_{\boldsymbol{x}\sim p_{\mathbf{x}_j}(\boldsymbol{x})}\left[(D_j(\mathbf{X}_{(-j)}, \boldsymbol{x}) - 2)^2\right] + \frac{1}{2}\mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z})}\left[(D_j(\mathbf{X}_{(-j)}, G_j(\mathbf{X}_{(-j)}, \boldsymbol{z})))^2\right]$$

$$\min_{G_j} L_G(G_j) = \frac{1}{2}\mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z})}\left[(D_j(\mathbf{X}_{(-j)}, G_j(\mathbf{X}_{(-j)}, \boldsymbol{z})) - 1)^2\right],$$

2.1.1 GCMI 的理论性质

为确保插补值的可靠性，我们建立了 GCMI 在完全随机缺失（MCAR）和随机缺失（MAR）机制下的理论性质。我们的目标是找到一个生成器 $G_j(x_{-j}, \mathbf{z})$，使其尽可能接近条件分布 $P_{\mathbf{x}_j|\mathbf{X}_{(-j)}=x_{-j}}$。对于给定的 $x_{-j} \in \mathbf{X}_{(-j)}$，将 $G_j(x_{-j}, \mathbf{z})$ 的条件分布与 $P_{\mathbf{x}_j|\mathbf{X}_{(-j)}=x_{-j}}$ 匹配，等价于匹配 $(\mathbf{X}_{(-j)}, G_j(\mathbf{X}_{(-j)}, z))$ 的联合分布与 $(\mathbf{X}_{(-j)}, \mathbf{x}_j)$ 的联合分布，前提是二者涉及相同的 $\mathbf{X}_{(-j)}$ 边缘分布 [16]。为此，我们首先介绍 f-散度的概念，它用于衡量两个概率函数之间的分布差异。假设 P 和 Q 是定义在 $\mathbb{R}^d$ 上、密度分别为 p 和 q 的两个概率密度函数，且 Q 关于 P 绝对连续。Q 相对于 P 的 f- 散度 [17] 定义为

$$\mathbb{D}_f(q\|p) = \int f\left(\frac{q(z)}{p(z)}\right) p(z)dz,$$

其中，f 是一个非负凸函数，在 $f(1) = 0$处取得最小值。根据詹森不等式，对于每个 q，p都有 $\mathbb{D}_f(q\|p) \geq 0$，且当且仅当 $q = p$时，$\mathbb{D}_f(q\|p) = 0$。

我们将说明，通过定义上述损失函数，实际上使用的是 χ^2散度，即 f-散度的一种特定形式。

$$
\begin{aligned}
L_D(D_j) &= \frac{1}{2}\mathbb{E}_{\boldsymbol{x}\sim p_{\mathbf{x}_j}(\boldsymbol{x})}\left[(D_j(\mathbf{X}_{(-j)},\boldsymbol{x})-2)^2\right] + \frac{1}{2}\mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z})}\left[(D_j(\mathbf{X}_{(-j)},G_j(\mathbf{X}_{(-j)},\boldsymbol{z})))^2\right] \\
&= \frac{1}{2}\int_{\mathcal{X}_j} p_{\mathbf{x}_j}(\boldsymbol{x})\left[(D_j(\mathbf{X}_{(-j)},\boldsymbol{x})-2)^2\right]\mathrm{d}\boldsymbol{x} + \frac{1}{2}\int_{\mathcal{Z}} p_{\boldsymbol{z}}(\boldsymbol{z})\left[(D_j(\mathbf{X}_{(-j)},G_j(\mathbf{X}_{(-j)},\boldsymbol{z})))^2\right]\mathrm{d}\boldsymbol{z} \\
&= \frac{1}{2}\int_{\mathcal{X}_j} p_{\mathbf{x}_j}(\boldsymbol{x})\left[(D_j(\mathbf{X}_{(-j)},\boldsymbol{x})-2)^2\right]\mathrm{d}\boldsymbol{x} + \frac{1}{2}\int_{\mathcal{X}_j} p_{G_j}(\boldsymbol{x})\left[(D_j(\mathbf{X}_{(-j)},\boldsymbol{x}))^2\right]\mathrm{d}\boldsymbol{x} \\
&= \frac{1}{2}\int_{\mathcal{X}_j} \{p_{\mathbf{x}_j}(\boldsymbol{x})\left[(D_j(\mathbf{X}_{(-j)},\boldsymbol{x})-2)^2\right] + p_{G_j}(\boldsymbol{x})\left[(D_j(\mathbf{X}_{(-j)},\boldsymbol{x}))^2\right]\}\mathrm{d}\boldsymbol{x}
\end{aligned}
$$

为了找到使 $L_D(D_j)$ 最小的 D_j，令

$$
\frac{\partial\{p_{\mathbf{x}_j}(\boldsymbol{x})\left[(D_j(\mathbf{X}_{(-j)},\boldsymbol{x})-2)^2\right] + p_{G_j}(\boldsymbol{x})\left[(D_j(\mathbf{X}_{(-j)},\boldsymbol{x}))^2\right]\}}{\partial D_j(\mathbf{X}_{(-j)},\boldsymbol{x})} = 0,
$$

那么，对于固定的 G_j，我们有最优判别器 D_j：

$$
D_j^*(\mathbf{X}_{(-j)},\boldsymbol{x}) = \frac{2p_{\mathbf{x}_j}(\boldsymbol{x})}{p_{\mathbf{x}_j}(\boldsymbol{x}) + p_{G_j}(\boldsymbol{x})}
$$

然后，我们可以将 $L_G(G_j)$ 重新表述为:

$$
\begin{aligned}
L_G(G_j) &= \frac{1}{2}\mathbb{E}_{\boldsymbol{x}\sim p_{\mathbf{x}_j}(\boldsymbol{x})}\left[(D_j^*(\mathbf{X}_{(-j)},\boldsymbol{x})-1)^2\right] + \frac{1}{2}\mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z})}\left[(D_j^*(\mathbf{X}_{(-j)},G_j(\mathbf{X}_{(-j)},\boldsymbol{z}))-1)^2\right] \\
&= \frac{1}{2}\int_{\mathcal{X}_j} p_{\mathbf{x}_j}(\boldsymbol{x})\left[(D_j^*(\mathbf{X}_{(-j)},\boldsymbol{x})-1)^2\right]\mathrm{d}\boldsymbol{x} + \frac{1}{2}\int_{\mathcal{Z}} p_{\boldsymbol{z}}(\boldsymbol{z})\left[(D_j^*(\mathbf{X}_{(-j)},G_j(\mathbf{X}_{(-j)},\boldsymbol{z}))-1)^2\right]\mathrm{d}\boldsymbol{z} \\
&= \frac{1}{2}\int_{\mathcal{X}_j} p_{\mathbf{x}_j}(\boldsymbol{x})\left[(D_j^*(\mathbf{X}_{(-j)},\boldsymbol{x})-1)^2\right]\mathrm{d}\boldsymbol{x} + \frac{1}{2}\int_{\mathcal{X}_j} p_{G_j}(\boldsymbol{x})\left[(D_j^*(\mathbf{X}_{(-j)},\boldsymbol{x})-1)^2\right]\mathrm{d}\boldsymbol{x} \\
&= \frac{1}{2}\int_{\mathcal{X}_j} p_{\mathbf{x}_j}(\boldsymbol{x})\left[(\frac{2p_{\mathbf{x}_j}(\boldsymbol{x})}{p_{\mathbf{x}_j}(\boldsymbol{x})+p_{G_j}(\boldsymbol{x})}-1)^2\right]\mathrm{d}\boldsymbol{x} + \frac{1}{2}\int_{\mathcal{X}_j} p_{G_j}(\boldsymbol{x})\left[(\frac{2p_{\mathbf{x}_j}(\boldsymbol{x})}{p_{\mathbf{x}_j}(\boldsymbol{x})+p_{G_j}(\boldsymbol{x})}-1)^2\right]\mathrm{d}\boldsymbol{x} \\
&= \frac{1}{2}\int_{\mathcal{X}_j} \frac{\left((p_{\mathbf{x}_j}(\boldsymbol{x})+p_{G_j}(\boldsymbol{x})) - 2p_{G_j}(\boldsymbol{x})\right)^2}{p_{\mathbf{x}_j}(\boldsymbol{x})+p_{G_j}(\boldsymbol{x})}\mathrm{d}x \\
&= \frac{1}{2}\mathbb{D}_{\chi^2}(p_{\mathbf{x}_j}+p_{G_j}\|2p_{G_j})
\end{aligned}
$$

其中，χ^2_{Pearson} 是 Pearson χ^2 散度。因此，最小化上述方程得到 lds 最小化 $p_{\mathbf{x}_j}+p_{G_j}$ 与 $2p_{G_j}$ 之间的 Pearson χ^2 散度。

因此，我们证明了损失函数 $L_D(D_j)$ 和 $L_G(G_j)$ 当且仅当 $D_j^*(\mathbf{X}^{(-j)},\ \boldsymbol{x}) = \frac{2p_{\mathbf{x}_j}(\boldsymbol{x})}{p_{\mathbf{x}_j}(\boldsymbol{x})+p_{G_j}(\boldsymbol{x})}$ 以及 $p_{\mathbf{x}_j}=p_{G_j}$ 时达到全局最小值

定理 1（GCMI 理论基础）

GCMI 损失函数 $L_D(D_j)$和 $L_G(G_j)$当且仅当生成器分布与真实条件分布一致时达到全局最小值，即 $p_{G_j} = p_{x_j}$。

具体而言：
1. **最优判别器：** 对于任意固定的生成器 G_j，最优判别器为：$D_j^*(X_{(-j)}, x) = \frac{p_{x_j}(x)}{2p_{x_j}(x) + p_{G_j}(x)}$ 2. ~~χ^2~~ **散度最小化：** 生成器损失等价于最小化 Pearson χ^2 散度：$L_G(G_j) = \frac{1}{2}\mathbb{D}_{\chi^2}(p_{x_j} + p_{G_j} \| 2p_{G_j})$ 3. **全局最优条件：** 当 $D_j^* = 1$ 和 $p_{x_j} = p_{G_j}$ 时，两个损失函数均达到全局最小值。

实际意义： *GCMI* 可证明能够学习真实的条件分布 $p(x_j|X_{(-j)})$，从而确保插补值保持与观测数据相同的统计性质。

在训练过程中，我们考虑 $L_D(D_j)$ 和 $L_G(G_j)$ 的经验版本：

$$\widehat{L}_D(D_j) = \frac{1}{2n}\sum_{i=1}^{n}(D_j(\mathbf{X}_{i,(-j)}, \mathbf{x}_{i,j}) - 2)^2 + \frac{1}{2n}\sum_{i=1}^{n}(D_j(\mathbf{X}_{i,(-j)}, G_j(\mathbf{X}_{i,(-j)}, z_i)))^2$$

$$\widehat{L}_G(G_j) = \frac{1}{2n}\sum_{i=1}^{n}(D_j(\mathbf{X}_{i,(-j)}, G_j(\mathbf{X}_{i,(-j)}, z_i)) - 1)^2$$

为进一步提高生成器准确性，我们在生成器损失函数中加入以下惩罚项：

$$L_{acc}(x, \hat{x}) = \begin{cases} (\hat{x} - x)^2, & \text{if } x \text{ is continuous} \\ -x\log(\hat{x}) - (1-x)\log(1-\hat{x}), & \text{if } x \text{ is binary} \end{cases}$$

3 算法

通过迭代过程可以最小化损失函数。首先，使用均值插补或其他方法填补$\mathbf{X}$ 中的缺失值。其次，根据$\mathbf{x}j, j = 1, \ldots, P$中的缺失比例，按升序排列变量的列。第三，对于每个变量$\mathbf{x}j$，首先以$\mathbf{x}_j^{obs}$为响应变量、以$\mathbf{X}_{(-j)}^{obs}$为预测变量训练 GCIN；然后，将训练好的 GCMI 应用于$\mathbf{X}_{(-j)}^{mis}$，预测$\mathbf{x}_j^{mis}$中的缺失值。

算法 GCIN 可视化如图 1 所示，由两个关键组件构成：条件生成器 (G) 和条件判别器 (D)。G 和 D 均使用修正线性单元 (ReLU) 激活函数。

G 的作用是为缺失位置生成插补值。它根据给定的条件 $X_{(-j)}$，将随机噪声向量 z 作为输入，并生成 $X_{(j)}$ 的插补值作为输出。生成器中采用了 ReLU 激活函数，以引入非线性并增强模型的表达能力。

D 负责评估其接收的数据样本是否真实，旨在区分真实值与插补值，同时考虑给定的条件 $X_{(-j)}$。判别器还采用 ReLU 激活函数，在计算中引入非线性。

训练期间，G 和 D 的参数交替更新。在每次迭代中，生成器生成合成样本，然后将这些样本与训练集中的真实样本一同输入判别器。判别器评估这些样本，并向 G 和自身提供反馈。通过迭代更新 G 和 D 的参数，GCIN 旨在达到一种平衡：生成器产生高质量的插补值，使其无法与真实样本区分，而判别器则越来越擅长区分真实值和插补值。

随后，我们整合一种基于链式方程的多重插补策略，以增强模型的稳定性和插补性能。也就是说，GCIN 插补过程会重复进行，直至满足停止准则。GCMI 中链式方程插补的最大迭代次数为 20。多重插补在处理缺失数据时具有若干重要优势。首先，多重插补通过生成多个完整数据集来减轻缺失数据带来的偏差，每个数据集都包含一组不同的插补值。这降低了随机性对插补的影响，并有助于更有效地恢复真实的数据分布。其次，多重插补能够实现有效的方差量化。有效的方差估计对于准确评估参数估计值和假设检验相关的不确定性至关重要。该过程将每个插补数据集内部的变异性（插补内方差）与各插补数据集之间的变异性（插补间方差）结合起来，从而得到总体方差估计。按照 Rubin 法则，汇总方差通常通过组合插补内方差与插补间方差来计算 [18]。这些法则以恰当考虑缺失数据所致不确定性的方式合并方差，从而得到有效且稳健的参数方差估计。图 2所示的 GCMI 基于 GCIN 和多重插补来插补缺失值。其实现细节见算法 1。

我们的 GCMI 链式插补循环采用双重收敛准则，同时监测数值变量和分类变量：对于数值变量，$\gamma_{\text{num}} = \frac{\sum(X_{\text{new}} - X_{\text{old}})^2}{\sum X_{\text{new}}^2}$。对于分类变量，$\gamma_{\text{cat}} = \frac{\sum(X_{\text{new}} \neq X_{\text{old}})}{n_{\text{missing}}}$。只要任一准则显示出改进（$\gamma_{\text{num,new}} < \gamma_{\text{num,old}}$ 或 $\gamma_{\text{cat,new}} < \gamma_{\text{cat,old}}$），算法就会继续运行；当两项指标均趋于稳定，或达到最多 100 次迭代后，算法便会终止。这种双重方法通过追踪连续变量的归一化均方差以及分类变量的变化比例，确保混合数据类型具有稳健的收敛性。

Algorithm 1: Make an initial guess for the missing values in $\mathbf{X}$ using a chosen imputation method. Then, sort the variables $\mathbf{x}_j, j = 1, \ldots, P$ based on the missing proportion starting with the lowest amount.

while *imputation stopping criterion not converged* **do**

for $j \in \{1, \ldots, P\}$ **do**

- Split data into fours parts as $\mathbf{x}_j^{obs}, \mathbf{x}_j^{mis}, \mathbf{X}_{(-j)}^{obs}, \mathbf{X}_{(-j)}^{mis}$.
- Fit a generative conditional imputation neural network (GCIN) model $\mathbf{x}_j^{obs} \sim G_j(\mathbf{X}_{(-j)}^{obs}, \mathbf{z})$, which includes Generator G_j with standard normal initial weights $\mathbf{W}$, and Discriminator D_j with standard normal initial weights $\mathbf{U}$.

while *GCIN not converged* **do**

- Update D_j by descending its stochastic gradient:

$$g_{\mathbf{U}} = \nabla_{\mathbf{U}} \frac{1}{B} \left[\sum_{i=1}^{B/2} \left(D_j(\mathbf{X}_{i,(-j)}^{obs}, \mathbf{X}_{i,j}^{obs}) - 2 \right)^2 + \sum_{i=B/2+1}^{B} \left(D_j(\mathbf{X}_{i,(-j)}^{obs}, G_j(\mathbf{X}_{i,(-j)}^{obs}, \mathbf{z_i}) \right)^2 + \alpha_1 \|\mathbf{U}\|^2 \right.$$

- Randomly select $c_i \in \mathbf{obs}(j)$ for $i = 1, 2, \ldots, B$,
- Update G_j by descending its stochastic gradient:

$$g_{\mathbf{W}} = \nabla_{\mathbf{W}} \frac{1}{B} \sum_{i=1}^{B} \left(D_j(\mathbf{X}_{c_i,(-j)}^{obs}, G_j(\mathbf{X}_{c_i,(-j)}^{obs}, \mathbf{z}_{\mathbf{c_i}})) - 1 \right)^2 + \alpha_{21} \|\mathbf{W}\|^2 + \alpha_{22} L_{acc} \left(\mathbf{x}_{c_i,j}^{obs}, G_j(\mathbf{X}_{c_i,(-j)}^{obs}, \mathbf{z}_{\mathbf{c_i}}) \right)$$

- Update weights parameters $[\mathbf{W}, \mathbf{U}] \leftarrow [\mathbf{W} - \eta_1 g_{\mathbf{W}}, \mathbf{U} - \eta_2 g_{\mathbf{U}}]$
- Check GCIN convergence criterion

end

Update $\mathbf{X}$: $\mathbf{x}_j^{mis} \leftarrow G_j(\mathbf{X}_{(-j)}^{mis}, \mathbf{z})$

end

Check imputation convergence criterion

end

4 实验

缺失数据插补方法通常通过基于合成数据和真实数据的模拟进行评估。在本节中，我们针对不同的缺失机制、缺失比例和样本量，验证 GCMI 在协变量插补准确性方面的表现。当主要目标是插补缺失的协变量时，我们将 GCMI 与几种传统及先进的缺失数据插补方法进行比较：

- **均值插补**：使用每个连续变量的均值和每个分类变量的众数来插补缺失值。

- **MICE**：一种迭代式多重插补方法，根据其他已观测变量对缺失数据进行建模。我们使用 *mice* Python 包，并采用默认超参数设置。

- **矩阵插补**：通过核范数正则化，对含有缺失值的矩阵进行低秩矩阵逼近。

- **MissForest**：一种迭代式随机森林插补算法，在观测数据上进行训练，以预测缺失数据。我们使用带有默认超参数设置的 *MissForest* Python 软件包。

• **GAIN**：一种生成对抗插补方法，通过使用额外的提示矩阵对生成器和判别器进行对抗训练来插补缺失数据，从而揭示缺失性。

4.1 合成数据

我们生成 $N = 2000$ 个完整观测，其中包含 $P = 15$ 个独立连续变量，这些变量服从多元正态分布

$$\boldsymbol{X}_i = (X_{i1}, \ldots, X_{iP})^T \sim \mathcal{N}\left(0, \sigma^2 \left[(1-\rho)I_P + \rho 1_P 1_P^T\right]\right)$$

其中， $\sigma^2 = 1$ 为边际方差， $\rho = 0.3$ 控制协变量之间的相关性。连续响应变量 Y_i 由线性回归模型生成

$$Y_i = X_i^T \alpha + \epsilon_i, \quad \alpha \sim \text{Unif}\left([-1,1]^P\right), \epsilon_i \sim N(0,1), \tag{1}$$

其中， ϵ_i 表示个体层面的随机噪声，回归系数 α 从均匀分布中随机模拟得到，其实际取值为 (0.542, - 0.769, 0.298, -0.156, 0.778, -0.391, -0.629, 0.311, 0.913, -0.025, -0.676, 0.512, 0.840, -0.265, -0.678)。

4.2 真实数据

除合成数据外，我们还使用三个真实世界基准数据集评估 GCMI 与最先进插补方法的性能：两个重症监护病房数据集（MIMIC-III 和 eICU），以及一个 UCI 数据库数据集。由于不规则采样或运行限制，临床重症监护病房数据往往存在缺失值，因此稳健的插补对于下游分析至关重要。下文将详细介绍各重症监护病房数据集的特征及其缺失性模式。

4.2.1 MIMIC-III 实验室数据

医疗信息市场重症监护数据库 III（MIMIC-III） [19]是一个公开的重症监护数据库，收录了 2008 年至 2012 年间入住美国马萨诸塞州波士顿贝斯以色列女执事医疗中心重症监护病房的所有患者信息。该数据库汇集了来自不同数据源的 3.5 亿余条记录，涵盖患者的人口统计信息、诊断代码、实验室检查、生命体征和临床事件。

MIMIC-III数据集中的实验室表所涵盖的时间范围不仅限于患者在重症监护病房的住院期间，还包括其整个住院过程，其中也可能包含门诊记录。这个全面的数据集为患者在入住重症监护病房之前预测其死亡率提供了机会。然而，许多实验室特征的缺失率超过 50%，有些甚至高达 80%。图 3 展示了 36 项常见实验室检查的缺失比例。如此高的缺失数据发生率可能会对某些预测方法的准确性产生重大影响。

对于重症监护病房患者，实验室检查的开具缺乏统一的标准化规范，导致实验室特征空间庞大且高度稀疏。Frassica 先前开展的研究在 2005 [20] 中考察了多达 45,188 项实验室检查及其特征

，涵盖三个不同的重症监护病房，结果显示，这些重症监护病房中 80% 的检查及其特征可以通过少于 25 项检查充分表示。在此基础上，Sharafoddini[21] 根据重症监护医学信息集市 III 数据库对实验室检验项目进行了调整，并选取了 36 项实验室检验项目用于预测院内死亡。受这些既往分析的启发，我们提出假设：增加实验室检验项目的数量未必会提高预测准确性。这是由于三维患者-实验室检验-日数组的固有特征所致，该数组具有极高的稀疏性和很高的缺失数据比例。因此，在可获得的 726 个不同实验室检验特征中，我们的分析聚焦于患者首次住院入院后记录的项目。随后，我们按日对这些数值进行汇总，以降低稀疏性。如果患者在某一天接受了同一实验室检查的多次检测，我们计算其平均值。对这些数值取平均有助于减小噪声的影响，并对潜在变量提供更稳定、更具代表性的估计。如果某项实验室检查在特定日期未对患者进行，则将其视为缺失值。检查流程具有信息量，这一点显而易见，因为它同时涵盖了是否进行检查以及检查频率。具体而言，决定让某些个体接受检查本身就包含有价值的信息。相反，未接受检查的患者，其协变量分布可能存在差异，这或许表明其健康状况不同——具体而言，可能病情更重，也可能病情较轻，取决于所考察的具体检查。在完成初步汇总后，我们将实验室数据组织为三维患者-实验室检验数组，其不同维度分别表示患者 ID、日期和实验室检验。尽管我们处理了实验室检验的稀疏性问题，但高缺失率的挑战仍然存在。我们选择对文献中报道的、用于预测死亡的顶级特征进行插补 [22]。其中，最关键的特征包括血小板计数、红细胞分布宽度、丙氨酸氨基转移酶和血尿素氮。

由于我们没有未观测实验室值的真实值，为了评估缺失值插补性能，我们随机丢弃观测数据，并额外设置10%至50%的缺失率，然后根据此类丢弃数据的已知值及其插补值计算插补误差。

4.2.2 eICU 协作研究数据库

eICU 协作研究数据库是开展重症医学领域深入研究的一项庞大且宝贵的资源。该数据库由飞利浦医疗与麻省理工学院计算生理学实验室合作创建，旨在为研究人员提供来自多样化且规模庞大的患者群体的全面临床数据集。

eICU 协作研究数据库收录了美国 200 多家医院超过 20 万例重症监护病房住院记录。该数据集包含 13.9 万多名患者的资料，以及数千个涵盖生命体征、实验室检查结果、药物、诊断和干预措施的变量，全面呈现了患者护理情况。数据涵盖数年，并以很高的时间粒度采集，使研究人员能够从全国范围和时间维度探索重症监护实践。这些数量充分体现了

eICU数据库的广度和深度，使其成为研究工作的重要资源，包括探讨重症监护环境中的缺失数据插补技术。

eICU数据库以改善患者护理和预后为首要目标，包含丰富的信息，涵盖患者人口统计学信息、生命体征、实验室检查结果、药物和治疗干预等多个医学领域。通过提供在美国重症监护病房收治患者的护理和预后信息，eICU数据库促进了临床决策支持、预测建模、流行病学和质量改进等领域的研究。

在处理 eICU 协作研究数据库时，我们采用了与 MIMIC-III数据集分析相似的数据预处理流程。该流程生成了一个包含 40 个关键特征的数据集，用于插补。由于未观测实验室值缺乏真实值，我们采用一致的评估策略来衡量缺失数据插补的表现。为此，我们在观测数据中额外引入缺失，随机丢弃 10% 至 50% 的数据点。随后，我们通过计算插补误差，将插补值与被丢弃数据的已知值进行比较，以评估插补技术的质量。

4.3 实验设置

分别在完全随机缺失、随机缺失和非随机缺失机制下生成缺失数据集。

•完全随机缺失：以恒定概率 $p \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6\}$. 随机删除原始完整协变量矩阵 $\boldsymbol{X}$ 中的每个元素
• 随机缺失：根据缺失倾向得分模型 logit$[M_j = 1] = \boldsymbol{X}_{\boldsymbol{c}}^T \beta_j$，，在 $\boldsymbol{X}_{\boldsymbol{m}} = (X_5, \ldots\ldots, X_p)^T$ 中生成缺失值，其中 $\boldsymbol{X}_{\boldsymbol{c}} = (X_1, \ldots\ldots, X_4)^T$，$\beta_j \sim$Unif $\left([-1, 1]^P\right)$，$j = 1, 2, \ldots\ldots, p$. M_j 是列 j 的缺失指示向量。
•非随机缺失：以恒定概率 $p = b_0 + b_1 * X_{ij}$ 随机删除原始完整协变量矩阵 $\boldsymbol{X}$ 中的每个元素，其中 $b_0 \in \{-4, -3.1, -2.5, -2, -1.5, -1\}$，$b_1 = 3$。

模拟中使用的神经网络设定如下。训练数据的批大小设为 256（当 $n_{train} = 5000$）。训练轮数上限为 10000。GCMI 中链式方程插补的最大迭代次数为 20。

5 结果

所有方法均根据原始完整协变量矩阵 X 与插补后的协变量矩阵 $\hat{X}$ 之间的均方根误差（均方根误差）进行评估。在每种模拟情景下，我们在 100 个蒙特卡洛（MC）数据集上重复实验，并报告 100 次重复的平均均方根误差。

5.1 合成数据插补

图 4 展示了仅进行协变量矩阵插补时，在不同缺失机制和缺失率下缺失数据插补的均方根误差。当缺失

机制为随机缺失时，均值插补的均方根误差最高。这是预期之中的，因为均值插补不假设任何预测模型，因此无法考虑协变量之间的相关性，也无法利用这些相关性进行预测。GAIN 和矩阵填补方法产生了次差的结果，随后是 MissForest 和 MICE。我们提出的 GCMI 在各种缺失率下始终具有最高的预测准确率和最低的均方根误差。结果表明，与单次填补方法（均值插补、矩阵填补和 GAIN）相比，基于多重插补的方法（MissForest、MICE 和 GCMI）具有更高的精确度，也更加稳定。在完全随机缺失和非随机缺失机制下也观察到了类似趋势。

5.2 真实 ICU 数据集上的插补性能

我们在两个重症监护数据集——MIMIC-III 和 eICUCollaborativeResearch 数据库——上评估了所提出的 GCMI 方法与最先进的替代方法在随机缺失（MAR）条件下的表现，特征缺失率范围为 10% 至 50%。性能采用均方根误差（RMSE，均值 ±标准误）进行量化。

5.2.1 MIMIC-III 重症监护数据

对 MIMIC-III 实验室数据集（表1）的分析揭示了三项主要发现。随着缺失性增加，传统方法的性能如预期般下降：均值插补提供了稳定但有限的基线（RMSE 0.$073 \pm 0.$002），而 MissForest（RMSE 范围：0.065– 0.089）和 GAIN（0.061–0.070）等先进方法则带来了适度但不稳定的提升（相较均值插补，误差降低 12%–18%）。矩阵补全方法表现出显著局限——Soft Impute 的结果波动较大（0.080– 0.098），其掩码变体在缺失率超过 40% 后变得不可靠。最重要的是，GCMI 的性能在统计上显著优于其他方法（配对 t 检验： $p < 0.01$在所有缺失率水平下），同时保持了稳健的准确性（0.058–0.068），相较最佳基线方法 GAIN，平均误差降低了 22%。这种稳定性表明，该方法特别适用于临床实验室数据填补。

5.2.2 eICU 多中心验证

eICU 数据集上的结果（表 2）证实了 GCMI 在不同机构间的泛化能力。我们的方法始终优于其他方法（均方根误差：0.060–0.081），其误差比矩阵补全方法低 15%–25%，同时展现出对极端缺失性（$p < 0.05$在 50% 缺失率下）的显著稳健性。这项多中心验证结合 MIMIC-III 的结果，确立了 GCMI 在多样化重症监护病房数据环境中的准确性与可靠性——这是临床实施的一项关键要求。

6 实践实施说明

我们提供了 GCMI 可复现性及实际应用的关键实现细节。

6.1 超参数配置

学习率配置为生成器使用 $\eta_G = 0.001$，判别器使用 $\eta_D = 0.0005$；这种非对称设置可防止训练过程中判别器占据主导地位。我们采用 L2 正则化（$\alpha = 0.0001$）和 Adam 优化（动量 $= 0.9$），并且每个周期执行 50 次生成器迭代、10 次判别器迭代，以保持对抗训练的稳定性。架构规模采用适应数据集的策略：对于样本数为 ≤ 20K 的数据集，使用单个隐藏层（100 个单元）；对于中等规模的数据集（20K–30K 个样本），使用两个隐藏层（200、100 个单元）；对于样本数为 ≥ 30K 且具有 ≥ 50 个特征的数据集，则使用更大的架构（400、200 个单元）。最多迭代 100 次，并使用容差 $\epsilon = 10^{-4}$，依据标准 $\gamma_{\text{new}} < \gamma_{\text{old}}$ 进行收敛。

6.2 计算需求

GCMI 的计算复杂度随 $O(P \times N \times I)$ 增长，其中 P是特征数量，N是样本量，I 表示训练迭代次数。内存需求从小型数据集（$<$ 50K 个样本、$<$ 100 个特征）所需的 8GB RAM，到中等规模应用所需的 16GB 不等。在标准硬件（Intel i7、16GB RAM）上，小型数据集的训练时间为 5–15 分钟，大型数据集则为 2–6 小时。特征级并行化在多核系统上可实现 2–4$\times$ 倍加速。

开源 Python GCMI 软件包将在论文被接收后于 Github 上公开。

7 结论

总之，本文提出的生成式条件缺失值插补（GCMI）方法，是对数据插补领域的一项重要贡献。GCMI 利用条件 GAN 架构的稳健性，擅长处理不同机制下的数据缺失性，在插补准确性方面表现出竞争力，并在多种情境下优于现有方法。尽管基于 GAN 的算法（如 GCMI）训练起来可能颇具挑战，尤其是在样本量较小时，其优势仍显而易见。GCMI 在训练期间适应不完整数据集的能力，以及保持变量间相关性的特点，都是增强其适用性的关键优势。基于真实数据集的实证验证证实了 GCMI 作为插补方法工具库中强大工具的潜力，即使面对信息缺失，也能提高预测质量。

参考文献

[1] Gelman, A., Hill, J.:《使用回归和多层级/层次模型进行数据分析》第1卷。剑桥大学出版社，美国纽约州纽约，???（2007）[2] Baraldi, A.N., Enders, C.K.: 现代缺失数据分析导论。《学校心理学杂志》48(1)，5–37（2010）

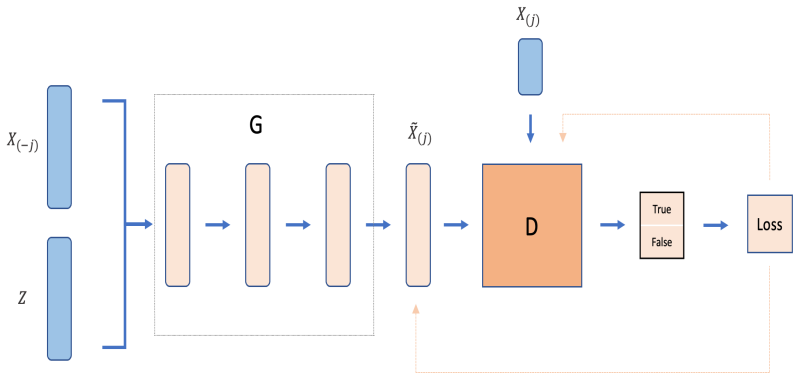


图 1：GCIN： G 将随机噪声向量 z 作为输入，在给定条件 $X_{(-j)}$的约束下，为 $X_{(j)}$ 生成插补值作为输出。 D 旨在评估输入数据样本，并根据给定条件 $X_{(-j)}$区分真实值和插补值。生成器中采用 ReLU 激活函数，以引入非线性并增强模型的表达能力。

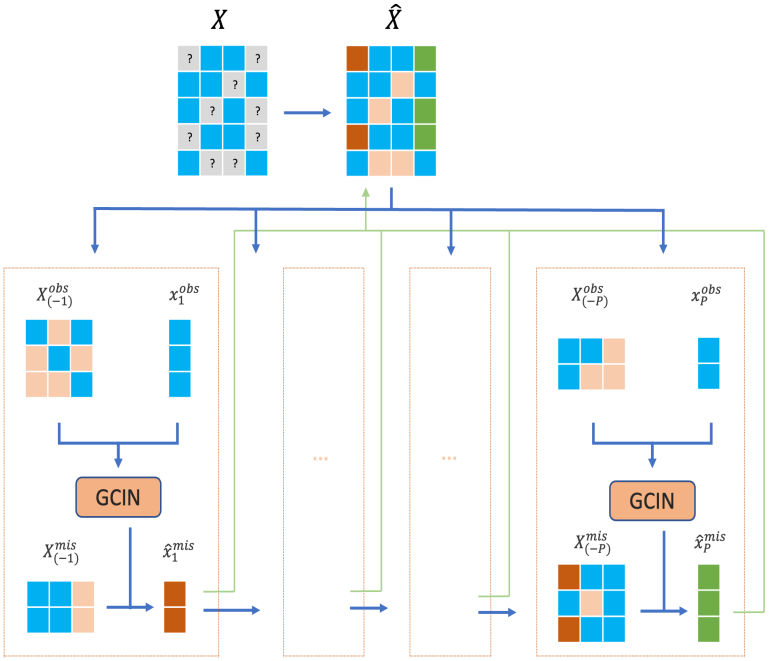


图 2：GCMI：使用选定的方法对 X 中的缺失值进行初始插补，并根据缺失比例对变量排序。遍历每一列 j ，并在$\mathbf{x}_j^{obs} \sim G_j(\mathbf{X}_{(-j)}^{obs}, \mathbf{z})$上拟合生成式条件插补神经网络（GCIN）。GCIN 算法收敛后，使用$\mathbf{X}$ 更新为$\mathbf{x}_j^{mis} \leftarrow G_j(\mathbf{X}_{(-j)}^{mis}, \mathbf{z})$，并检查插补是否收敛。

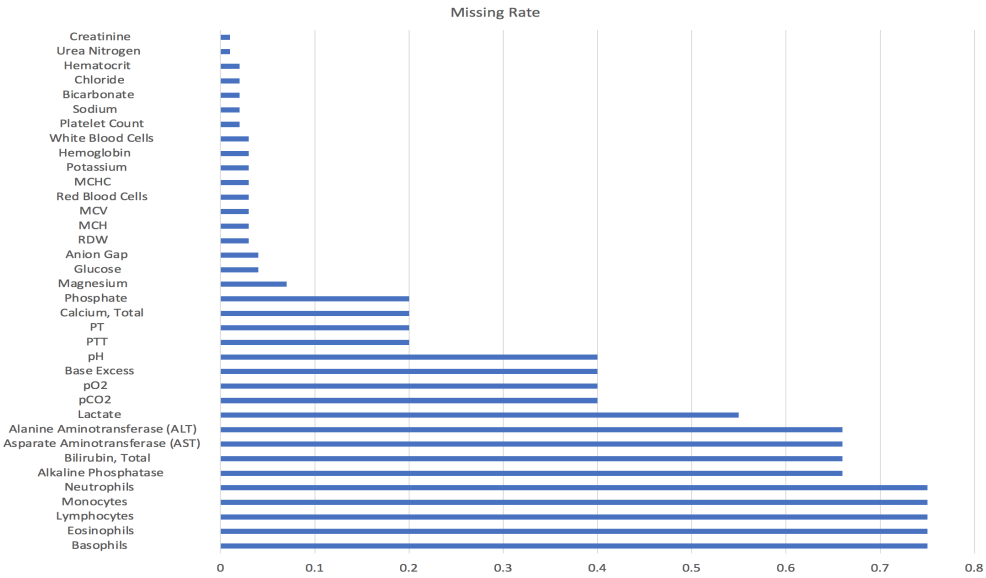
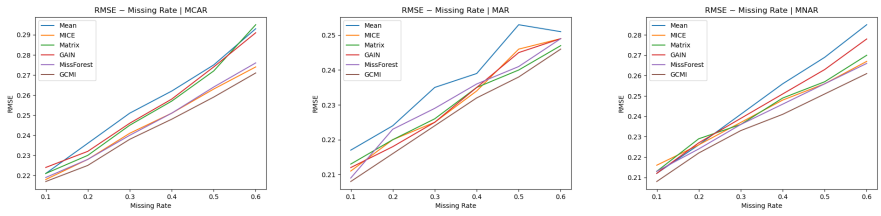


图 3：MIMIC-III 中 36 项常见实验室检查的缺失比例。



(a) 完全随机缺失　(b) 随机缺失　(c) 非随机缺失

图 4：在三种缺失机制和不同缺失率范围下，以均方根误差衡量的协变量矩阵缺失值插补误差

表 1：MIMIC-III 实验室数据——缺失值插补：在随机缺失情况下，使用 $n = 5000$ 和 $p = 24$ 进行块状缺失数据模拟。特征的缺失率范围为 10% 至 50%。性能结果以均值 ±标准差表示。

缺失率	0.1	0.2	0.3	0.4	0.5
均值插补	0.073 ± 0.001	0.073 ± 0.001	0.073 ± 0.001	0.073 ± 0.001	0.073 ± 0.001
多重插补	0.061 ± 0.001	0.063 ± 0.001	0.065 ± 0.000	0.067 ± 0.001	0.069 ± 0.001
MI-KNN	0.064 ± 0.001	0.066 ± 0.001	0.068 ± 0.001	0.070 ± 0.001	0.071 ± 0.001
MissForest	0.059 ± 0.001	0.061 ± 0.001	0.064 ± 0.001	0.066 ± 0.001	0.069 ± 0.001
Soft	0.080 ± 0.002	0.094 ± 0.001	0.098 ± 0.006	0.095 ± 0.006	0.092 ± 0.004
Soft Mask	0.097 ± 0.003	0.091 ± 0.003	0.097 ± 0.004	0.106 ± 0.005	0.125 ± 0.007
GAIN	0.061 ± 0.001	0.063 ± 0.001	0.065 ± 0.001	0.068 ± 0.001	0.070 ± 0.001
GCMI	$\mathbf{0.058 \pm 0.001}$	$\mathbf{0.061 \pm 0.001}$	$\mathbf{0.063 \pm 0.001}$	$\mathbf{0.066 \pm 0.001}$	$\mathbf{0.068 \pm 0.002}$

表2：eICU 协作研究数据库 - 缺失值插补：在随机缺失（MAR）条件下，使用 $n = 5000$ 和 $p = 40$ 进行块状缺失数据模拟。特征的缺失率范围为 10% 至 50%。性能结果以均值 ± 标准差表示。

缺失率	0.1	0.2	0.3	0.4	0.5
均值插补	0.097 ± 0.001	0.097 ± 0.001	0.097 ± 0.001	0.097 ± 0.001	0.097 ± 0.002
多重插补	0.065 ± 0.001	0.070 ± 0.001	0.073 ± 0.001	0.078 ± 0.001	0.083 ± 0.002
MI-KNN	0.073 ± 0.001	0.076 ± 0.001	0.080 ± 0.001	0.084 ± 0.001	0.088 ± 0.002
MissForest	0.062 ± 0.001	0.067 ± 0.001	0.073 ± 0.001	0.078 ± 0.001	0.083 ± 0.002
Soft	0.072 ± 0.001	0.076 ± 0.001	0.079 ± 0.001	0.084 ± 0.001	0.090 ± 0.002
Soft Mask	0.100 ± 0.001	0.107 ± 0.002	0.112 ± 0.002	0.114 ± 0.002	0.113 ± 0.002
GAIN	0.076 ± 0.001	0.079 ± 0.001	0.082 ± 0.001	0.086 ± 0.001	0.112 ± 0.002
GCMI	$\mathbf{0.060 \pm 0.001}$	$\mathbf{0.065 \pm 0.002}$	$\mathbf{0.070 \pm 0.001}$	$\mathbf{0.076 \pm 0.001}$	$\mathbf{0.081 \pm 0.002}$

[3] Little, R.J., Rubin, D.B.：《缺失数据的统计分析》第793卷。John Wiley& Sons,???(2019年)

[4]Cook, R.J., Zeng, L., Yi, G.Y.：不完整纵向二元数据的边际分析：关于末次观测值结转插补的警示。《生物计量学》**60**(3), 820–828 (2004)

[5] Donders, A.R.T., Van Der Heijden, G.J., Stijnen, T., Moons, K.G.：缺失值插补简介。《临床流行病学杂志》**59**(10),1087–1091 (2006)

[6] Van Buuren, S., Brand, J.P., Groothuis-Oudshoorn, C.G., Rubin, D.B.：多变量插补中的完全条件设定。《统计计算与模拟杂志》**76**(12), 1049–1064 (2006)

[7]Van Buuren, S.、Groothuis-Oudshoorn, K.：mice：R 中基于链式方程的多变量插补。《统计软件杂志》**45**，1–67（2011）

[8] Yoon, S.、Sull, S.：GAMIN：针对高缺失率数据的生成对抗多重插补网络。载于：IEEE/CVF计算机视觉与模式识别会议论文集，第8456–8464页（2020）[9] Yoon, J.、Jordon, J.、Schaar, M.：GAIN：使用生成对抗网络进行缺失数据插补。载于：国际机器学习会议，第5689–5698页。PMLR，???（2018）

[10]Nazabal, A.、Olmos, P.M.、Ghahramani, Z.、Valera, I.：使用变分自编码器处理不完整异质数据。《模式识别》**107**，107501（2020）

[11] 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年）[12] Ghalebikesabi, S.,Cornish, R., Holmes, C., Kelly, L.: 深度生成式缺失性

模式集混合模型。载于：国际人工智能与统计会议，第3727–3735页。PMLR，???（2021年）[13] Mattei, P.-A., Frellsen, J.: MIWAE：不完整数据集的深度生成建模与插补。载于：第36届国际机器学习会议论文集，第4413–4423页。PMLR 97，???（2019年）[14] Ipsen, N.B., Mattei, P.-A., Frellsen, J.: not-MIWAE：缺失非随机数据的深度生成建模。载于：第9届国际学习表征会议（2021年）[15] Ma, C., Zhang, C.: 非随机缺失数据插补的可识别生成模型。载于：《神经信息处理系统进展》第34卷，第27645–27658页（2021年）[16] Zhou, X., Jiao, Y., Liu, J., Huang, J.: 一种用于条件采样的深度生成方法。《美国统计学会会刊》，第1–12页（2022年）[17]Ali, S.M., Silvey, S.D.: 一个分布相对于另一个分布的散度系数的一般类。《英国皇家统计学会会刊：B辑（方法论）》**28**（1），131–142（1966年）[18] Rubin, D.B.: 调查中的无应答多重插补。Wiley **81**（394），543–603（1987年）https://doi.org/10.2307/2289753[19] Johnson, A.E., Pollard, T.J., Shen, L., Li-Wei, H.L., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L.A., Mark, R.G.: MIMIC-III：一个可免费访问的重症监护数据库。《科学数据》**3**（1），1–9（2016年）[20] Frassica, J.J.: 重症监护病房实验室检测使用情况及其对大规模数据收集工作的影响。《美国医学信息学会会刊》**12**（2），229–233（2005年）[21] Sharafoddini, A., Dubin, J.A., Maslove, D.M., Lee, J.: 深入了解重症监护病房患者档案中的缺失数据：观察性研究。《JMIR医学信息学》**7**（1），11605（2019年）[22] Sun, Y., Zhou, Y.-H.: 用于重症监护病房死亡率预测的机器学习流程。《国际数字健康期刊》**2**（1），3（2022年）[23]Rosenbaum, P.R., Rubin, D.B.: 倾向评分在观察性研究因果效应中的核心作用。Biometrika **70**（1），41–55（1983年）[24] Li, L., Shen, C., Li, X., Robins, J.M.: 关于缺失数据的加权方法。《医学研究统计方法》**22**（1），14–30（2013年）[25] Seaman, S.R., White, I.R.: 关于用于处理缺失数据的逆概率加权的综述。《医学研究统计方法》**22**（3），278–295（2013年）

[26] Austin, P.C.：观察性研究中减少混杂影响的倾向评分方法简介。《多变量行为研究》**46(3)**，399–424（2011）[27] Shimodaira, H.：通过对数似然函数加权，改善协变量偏移下的预测推断。《统计规划与推断杂志》**90(2)**，227–244（2000）[28] Cortes, C.、Mansour, Y.、Mohri, M.：重要性加权的学习界。《神经信息处理系统进展》**23** （2010）[29] Fernandes, K.、Vinagre, P.、Cortez, P.：用于预测在线新闻受欢迎程度的主动式智能决策支持系统。载于：葡萄牙人工智能会议，第535–546页（2015）。施普林格[30] Asuncion, A.、Newman, D.：UCI机器学习库。美国加利福尼亚州欧文（2007）