Pure Mathematics 理论数学, 2026, 16(11), 34-47
Published Online November 2026 in Hans. https://www.hanspub.org/journal/pm
https://doi.org/10.12677/pm.2026.1611232

Hans汉斯

异质缺失数据的重要性加权变分插补

史俊歌[1,2]，廖剑晖[2*]

[1]华北水利水电大学数学与统计学院，河南 郑州
[2]河南省科学院数学研究所，河南 郑州

收稿日期：2026年9月29日；录用日期：2026年10月28日；发布日期：2026年11月9日

摘　要

针对异质表格数据中的缺失值插补问题，本文构建重要性加权变分插补框架H-MIWAE (Heterogeneous Missing-data Importance-Weighted Autoencoder)。该框架在混合离散－连续潜变量结构下引入类型特异观测分布，并结合多样本重要性加权目标与自归一化重要性采样，实现对异质属性的统一建模与条件预测。为区分训练阶段与插补阶段的重要性采样作用，本文分别以 K_{tr} 和 L_{imp} 表示两个阶段使用的重要性样本数，并研究二者对不同属性类型插补性能的联合影响。在合成数据集和真实异质数据集上的实验表明，H-MIWAE能够对多类异质属性进行统一插补，并在不同数据条件下表现出较强的竞争力。结果表明，多样本训练所带来的插补收益依赖于后续的重要性采样，且这种依赖在不同属性类型上表现并不一致。

关键词

异质数据，缺失数据，变分推断，重要性加权，数据插补

Importance-Weighted Variational Imputation for Heterogeneous Missing Data

Junge Shi[1,2], Jianhui Liao[2*]

[1]College of Mathematics and Statistics, North China University of Water Resources and Electric Power, Zhengzhou Henan
[2]Institute of Mathematics HNAS, Henan Academy of Sciences, Zhengzhou Henan

Received: September 29, 2026; accepted: October 28, 2026; published: November 9, 2026

Abstract

This paper develops the Heterogeneous Missing-data Importance-Weighted Autoencoder (H-MIWAE),

*通讯作者。

文章引用: 史俊歌, 廖剑晖. 异质缺失数据的重要性加权变分插补[J]. 理论数学, 2026, 16(11): 34-47.
DOI: 10.12677/pm.2026.1611232

an importance-weighted variational imputation framework for missing-value imputation in heterogeneous tabular data. Under a mixed discrete-continuous latent-variable structure, the framework incorporates type-specific observation distributions and combines a multi-sample importance-weighted objective with self-normalized importance sampling to achieve unified modeling and conditional prediction for heterogeneous attributes. To distinguish the roles of importance sampling in the training and imputation stages, K_{tr} and L_{imp} denote the numbers of importance samples used in the two stages, respectively, and their joint effects on imputation performance across different attribute types are investigated. Experiments on synthetic and real-world heterogeneous data sets show that H-MIWAE can perform unified imputation for multiple types of heterogeneous attributes and achieves competitive performance under different data settings. The results further show that the imputation gains brought by multi-sample training depend on the subsequent importance sampling, and that this dependence varies across attribute types.

Keywords

Heterogeneous Data, Missing Data, Variational Inference, Importance Weighting, Data Imputation

Copyright © 2026 by author(s) and Hans Publishers Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/
Open Access

1. 引言

在实际数据分析中，数据缺失是一类普遍存在的问题，其产生可能与数据采集、测量过程及个体响应等多种因素有关[1] [2]。与此同时，表格数据通常由统计性质不同的多类变量共同构成，不同属性在取值空间和分布特征上可能存在明显差异[3]。数据的不完整性与变量类型的多样性共同增加了数据建模与分析的复杂性。对这类不完整异质数据进行合理分析，有助于更好地理解数据特征，并为缺失值处理及后续统计分析提供基础。

针对不完整表格数据，已有研究提出了多种缺失值处理方法，例如多重插补、近邻插补和随机森林插补等[4]-[7]。这些方法从不同角度利用已观测信息对缺失部分进行估计。进一步而言，概率模型通过刻画数据的联合分布，可以在给定已观测部分的条件下对未观测部分进行概率推断，为缺失数据分析提供了另一类建模思路。

为灵活刻画联合分布中的复杂非线性关系，以变分自编码器(variational autoencoder, VAE)为代表的深度潜变量模型，利用神经网络参数化观测分布与变分近似分布，并通过变分推断学习模型参数[8] [9]。VAE 的早期典型应用通常建立在完整观测数据上，例如对二值图像采用伯努利观测分布，对实值数据采用高斯观测分布[8]。针对不完整数据，生成对抗插补网络(generative adversarial imputation nets, GAIN) [10]、任意条件变分自编码器(variational autoencoder with arbitrary conditioning, VAEAC) [11]等方法分别从对抗式插补和任意条件生成等方向进行了拓展，相关研究还涉及异质观测建模及缺失条件下的变分推断[12] [13]，推动了深度生成模型在缺失数据分析中的应用。

对于同时包含连续型、计数型、类别型和有序型等多种统计类型属性的不完整表格数据，上述研究在异质观测分布的刻画、缺失条件下的多样本变分学习以及训练后的条件预测等方面，通常具有不同的关注点[11]-[13]。其中，重要性加权方法从识别分布抽取潜变量样本，利用相应的重要性权重构造训练阶段的多样本变分目标，并在训练后通过自归一化重要性采样近似缺失属性的条件预测量[13]。训练重要性样本数 K_{tr} 决定多样本目标的构造，插补重要性样本数 L_{imp} 则控制固定模型下条件预测量的 Monte Carlo

近似。已有研究指出，更紧的变分下界并不保证识别网络的学习效果同步改善[14]；缺失数据重要性加权自编码器(missing-data importance-weighted autoencoder, MIWAE)的实验也观察到，不同训练设置的插补表现对插补重要性样本数具有不同的依赖[13]。

针对上述问题，本文构建面向不完整异质表格数据的重要性加权变分插补模型 H-MIWAE (Heterogeneous Missing-data Importance-Weighted Autoencoder)，研究训练与插补重要性样本数对不同属性类型插补性能的联合影响。本文主要工作如下：

(1) 建立统一的异质缺失数据潜变量生成框架，针对不同统计类型属性设置相应的概率观测模型，并基于观测数据实现重要性加权变分学习与缺失属性条件预测。

(2) 区分重要性采样在训练阶段和插补阶段的不同作用，分别考察训练重要性样本数 K_{tr} 与插补重要性样本数 L_{imp} 对不同类型属性插补性能的影响，并分析二者之间的作用关系。

(3) 在合成数据和真实异质表格数据上开展系统实验，从类型特异插补误差、相同评价重要性样本数下的观测数据重要性加权下界估计，以及归一化有效样本量等方面，评价模型性能及推断特性。

2. H-MIWAE 模型与重要性加权插补

H-MIWAE 的整体建模、训练与插补流程如图 1 所示。模型由不完整观测构造潜变量的识别分布，并分别以 K_{tr} 和 L_{imp} 控制训练阶段与插补阶段的重要性采样，最终得到不同类型缺失属性的条件预测结果。

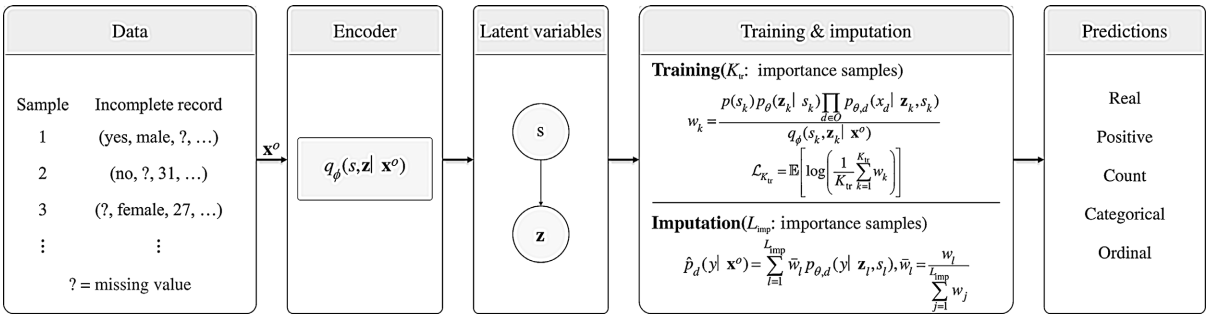


Figure 1. Overall framework of H-MIWAE for heterogeneous missing-data imputation
图 1. H-MIWAE 异质缺失数据建模、训练与插补的整体框架

2.1. 异质缺失数据的概率模型

设数据集包含 N 个样本，每个样本具有 D 个属性。第 n 个完整数据向量记为

$$\boldsymbol{x}_n=(x_{n1},\cdots,x_{nD})\in\mathcal{X}_1\times\cdots\times\mathcal{X}_D,\qquad n=1,\cdots,N$$

其中，$\mathcal{X}_d$ 表示第 d 个属性的取值空间。根据属性的取值空间与统计结构，本文考虑实值型、正值型、计数型、无序分类型和有序分类型五类属性。

对于第 n 个样本，记其观测属性索引集合和缺失属性索引集合分别为 O_n 和 M_n，满足

$$\mathrm{O}_n\cup\mathrm{M}_n=\{1,\cdots,D\},\qquad \mathrm{O}_n\cap\mathrm{M}_n=\varnothing$$

相应地，其观测子向量和缺失子向量分别记为

$$\boldsymbol{x}_n^{\mathrm{o}}=(x_{nd})_{d\in\mathrm{O}_n},\qquad \boldsymbol{x}_n^{\mathrm{m}}=(x_{nd})_{d\in\mathrm{M}_n}$$

为简化记号，以下针对单个样本进行推导时省略样本下标 n，并分别记为 $\boldsymbol{x}$, $\boldsymbol{x}^{\mathrm{o}}$, $\boldsymbol{x}^{\mathrm{m}}$, O, M 。

为刻画潜变量空间中的多成分结构，本文引入离散成分变量 $s\in\{1,\cdots,S\}$ 和连续潜变量 $\boldsymbol{z}\in\mathbb{R}^q$，其中

s 用于指示混合成分。参考高斯混合潜变量建模方式[12]，对各成分采用等权先验，并设给定成分后的连续潜变量服从高斯条件先验：

$$p(s=j)=\frac{1}{S},\qquad p_\theta(\boldsymbol{z}\mid s=j)=\mathcal{N}\left(\boldsymbol{z}\mid \boldsymbol{\mu}_j,\boldsymbol{I}_q\right),\qquad j=1,\cdots,S \tag{1}$$

其中，$\boldsymbol{\mu}_j$ 为第 j 个潜在成分的均值参数，$\boldsymbol{I}_q$ 为 q 维单位矩阵。

给定 $(s,\boldsymbol{z})$ 后，假设各属性条件独立，则潜变量与完整数据的联合分布可分解为

$$p_\theta(s,\boldsymbol{z},\boldsymbol{x})=p(s)\,p_\theta(\boldsymbol{z}\mid s)\prod_{d=1}^{D}p_{\theta,d}\left(x_d\mid \boldsymbol{z},s\right) \tag{2}$$

其中，$p_{\theta,d}\left(x_d\mid \boldsymbol{z},s\right)$ 表示给定潜变量后第 d 个属性的条件观测分布。模型基于 $(s,\boldsymbol{z})$ 通过共享解码网络获得中间表示，并通过与属性类型对应的输出层生成各属性条件观测分布的参数。根据属性的取值空间与统计结构，实值型、正值型和计数型属性分别采用高斯分布、对数正态分布和泊松分布；无序分类型属性采用多项 logit 参数化的类别分布，有序分类型属性采用累积 logit 模型[12]。

对于有序分类型属性，还需刻画不同等级之间的顺序关系。设第 d 个有序属性共有 C_d 个等级，并以 $c\in\{1,\cdots,C_d\}$ 表示等级。其中，logistic 函数定义为

$$\Lambda(t)=\frac{1}{1+\exp(-t)}$$

$\eta_d(\boldsymbol{z},s)$ 表示给定潜变量后的潜在得分，$\tau_{d,c}(s)$ 表示第 c 个内部阈值。定义第 c 个等级对应的累积概率为

$$F_{\theta,d,c}(\boldsymbol{z},s)=\Pr_\theta\left(X_d\le c\mid \boldsymbol{z},s\right)=\Lambda\left(\tau_{d,c}(s)-\eta_d(\boldsymbol{z},s)\right),\qquad c=1,\cdots,C_d-1 \tag{3}$$

约定

$$F_{\theta,d,0}(\boldsymbol{z},s)=0,\qquad F_{\theta,d,C_d}(\boldsymbol{z},s)=1$$

则第个等级的条件概率为

$$p_{\theta,d}\left(X_d=c\mid \boldsymbol{z},s\right)=F_{\theta,d,c}(\boldsymbol{z},s)-F_{\theta,d,c-1}(\boldsymbol{z},s),\qquad c=1,\cdots,C_d \tag{4}$$

为保证累积概率随等级单调增加，各内部阈值满足

$$\tau_{d,1}(s)<\tau_{d,2}(s)<\cdots<\tau_{d,C_d-1}(s)$$

根据样本的观测部分 $\boldsymbol{x}^{\mathrm{o}}$，构造联合潜变量 $(s,\boldsymbol{z})$ 的变分识别分布，并分解为

$$q_\phi\left(s,\boldsymbol{z}\mid \boldsymbol{x}^{\mathrm{o}}\right)=q_\phi\left(s\mid \boldsymbol{x}^{\mathrm{o}}\right)q_\phi\left(\boldsymbol{z}\mid \boldsymbol{x}^{\mathrm{o}},s\right) \tag{5}$$

其中，离散潜变量采用类别变分分布 c

$$q_\phi\left(s\mid \boldsymbol{x}^{\mathrm{o}}\right)=\mathrm{Cat}\left(s\mid \boldsymbol{\pi}_\phi\left(\boldsymbol{x}^{\mathrm{o}}\right)\right) \tag{6}$$

其中，Cat 表示类别分布。给定 s 后，连续潜变量采用对角高斯变分分布

$$q_\phi\left(\boldsymbol{z}\mid \boldsymbol{x}^{\mathrm{o}},s\right)=\mathcal{N}\left(\boldsymbol{z}\mid \boldsymbol{\mu}_\phi\left(\boldsymbol{x}^{\mathrm{o}},s\right),\mathrm{diag}\left[\boldsymbol{\sigma}_\phi^2\left(\boldsymbol{x}^{\mathrm{o}},s\right)\right]\right) \tag{7}$$

实际编码时，对缺失位置进行与属性类型相匹配的填充，并将填充后的输入与缺失掩码共同作为识别网络的输入。为简化记号，式(5)~(7)仍以 $\boldsymbol{x}^{\mathrm{o}}$ 表示识别网络所利用的不完整观测及其缺失模式信息。

上述生成模型与识别模型共同构成本文异质缺失数据建模的概率基础。

2.2. 多样本观测数据目标

对于不完整观测，由式(2)可得观测数据的边缘似然

$$p_\theta\left(\boldsymbol{x}^{\mathrm{o}}\right)=\sum_{s=1}^{S}\int p(s)\,p_\theta\left(\boldsymbol{z}\mid s\right)\prod_{d\in\mathrm{O}}p_{\theta,d}\left(x_d\mid \boldsymbol{z},s\right)\mathrm{d}\boldsymbol{z} \tag{8}$$

式(8)中涉及对潜变量的求和与积分，一般难以直接计算。因此，将识别分布 $q_\phi\left(s,\boldsymbol{z}\mid\boldsymbol{x}^{\mathrm{o}}\right)$ 作为重要性采样的提议分布，可将观测数据边缘似然改写为

$$\begin{aligned}p_\theta\left(\boldsymbol{x}^{\mathrm{o}}\right)&=\sum_{s=1}^{S}\int q_\phi\left(s,\boldsymbol{z}\mid\boldsymbol{x}^{\mathrm{o}}\right)\frac{p(s)\,p_\theta\left(\boldsymbol{z}\mid s\right)\prod_{d\in\mathrm{O}}p_{\theta,d}\left(x_d\mid\boldsymbol{z},s\right)}{q_\phi\left(s,\boldsymbol{z}\mid\boldsymbol{x}^{\mathrm{o}}\right)}\mathrm{d}\boldsymbol{z}\\&=\mathbb{E}_{q_\phi\left(s,\boldsymbol{z}\mid\boldsymbol{x}^{\mathrm{o}}\right)}\left[w\left(s,\boldsymbol{z};\boldsymbol{x}^{\mathrm{o}}\right)\right]\end{aligned} \tag{9}$$

其中，重要性权重定义为

$$w\left(s,\boldsymbol{z};\boldsymbol{x}^{\mathrm{o}}\right)=\frac{p(s)\,p_\theta\left(\boldsymbol{z}\mid s\right)\prod_{d\in\mathrm{O}}p_{\theta,d}\left(x_d\mid\boldsymbol{z},s\right)}{q_\phi\left(s\mid\boldsymbol{x}^{\mathrm{o}}\right)q_\phi\left(\boldsymbol{z}\mid\boldsymbol{x}^{\mathrm{o}},s\right)} \tag{10}$$

设 $K_{\mathrm{tr}}\in\mathbb{N}_+$ 表示训练阶段构造多样本目标时使用的重要性样本数。从识别分布 $q_\phi\left(s,\boldsymbol{z}\mid\boldsymbol{x}^{\mathrm{o}}\right)$ 中独立同分布地抽取

$$\left(s_k,\boldsymbol{z}_k\right),\qquad k=1,\cdots,K_{\mathrm{tr}}$$

并记相应的重要性权重为

$$w_k=w\left(s_k,\boldsymbol{z}_k;\boldsymbol{x}^{\mathrm{o}}\right)$$

参考重要性加权自编码器(importance-weighted autoencoder, IWAE)和 MIWAE 的多样本重要性加权目标构造[13] [15]，定义单个观测样本的训练目标为

$$\mathcal{L}_{K_{\mathrm{tr}}}\left(\theta,\phi;\boldsymbol{x}^{\mathrm{o}}\right)=\mathbb{E}_{\prod_{k=1}^{K_{\mathrm{tr}}}q_\phi\left(s_k,\boldsymbol{z}_k\mid\boldsymbol{x}^{\mathrm{o}}\right)}\left[\log\left(\frac{1}{K_{\mathrm{tr}}}\sum_{k=1}^{K_{\mathrm{tr}}}w_k\right)\right] \tag{11}$$

由 Jensen 不等式，式(11)构成观测数据对数似然 $\log p_\theta\left(\boldsymbol{x}^{\mathrm{o}}\right)$ 的下界。

当 $K_{\mathrm{tr}}=1$ 时，式(11)化为相应的单样本观测数据变分下界。将重要性权重展开，可得

$$\begin{aligned}\mathcal{L}_1=\;&\mathbb{E}_{q_\phi\left(s,\boldsymbol{z}\mid\boldsymbol{x}^{\mathrm{o}}\right)}\left[\sum_{d\in\mathrm{O}}\log p_{\theta,d}\left(x_d\mid\boldsymbol{z},s\right)\right]-\mathrm{KL}\left(q_\phi\left(s\mid\boldsymbol{x}^{\mathrm{o}}\right)\|\,p(s)\right)\\&-\mathbb{E}_{q_\phi\left(s\mid\boldsymbol{x}^{\mathrm{o}}\right)}\left[\mathrm{KL}\left(q_\phi\left(\boldsymbol{z}\mid\boldsymbol{x}^{\mathrm{o}},s\right)\|\,p_\theta\left(\boldsymbol{z}\mid s\right)\right)\right]\end{aligned} \tag{12}$$

因此，$K_{\mathrm{tr}}=1$ 对应本文模型的单样本训练情形；当 $K_{\mathrm{tr}}>1$ 时，采用多样本重要性加权训练目标。不同 K_{tr} 表示训练阶段使用不同数量的重要性样本，而生成模型与识别模型的概率结构保持不变。

对于包含 N 个样本的数据集，总体训练目标定义为各样本训练目标之和：

$$\mathcal{J}_{K_{\mathrm{tr}}}\left(\theta,\phi\right)=\sum_{n=1}^{N}\mathcal{L}_{K_{\mathrm{tr}}}\left(\theta,\phi;\boldsymbol{x}_n^{\mathrm{o}}\right) \tag{13}$$

2.3. 缺失属性的条件插补

训练完成后，固定模型参数 $\left(\hat{\theta},\hat{\phi}\right)$。对于任一缺失属性 $d\in\mathrm{M}$，其给定观测部分 $\boldsymbol{x}^{\mathrm{o}}$ 的条件预测分布

可表示为

$$p_{\hat{\theta},d}\left(y\mid \boldsymbol{x}^{\mathrm{o}}\right)=\sum_{s=1}^{S}\int p_{\hat{\theta},d}\left(y\mid \boldsymbol{z},s\right)p_{\hat{\theta}}\left(s,\boldsymbol{z}\mid \boldsymbol{x}^{\mathrm{o}}\right)\mathrm{d}\boldsymbol{z} \tag{14}$$

式(14)涉及潜变量的条件后验 $p_{\hat{\theta}}\left(s,\boldsymbol{z}\mid \boldsymbol{x}^{\mathrm{o}}\right)$，通常难以直接计算。因此，利用识别分布 $q_{\hat{\phi}}\left(s,\boldsymbol{z}\mid \boldsymbol{x}^{\mathrm{o}}\right)$ 进行重要性采样近似。

具体地，设 $L_{\mathrm{imp}}\in\mathbb{N}_{+}$ 表示插补阶段近似条件预测分布时使用的重要性样本数，从识别分布 $q_{\hat{\phi}}\left(s,\boldsymbol{z}\mid \boldsymbol{x}^{\mathrm{o}}\right)$ 中独立同分布地抽取

$$\left(s_l,\boldsymbol{z}_l\right),\qquad l=1,\cdots,L_{\mathrm{imp}}$$

利用式(10)，在训练后参数 $\left(\hat{\theta},\hat{\phi}\right)$ 下计算各样本对应的未归一化重要性权重

$$w_l=w\left(s_l,\boldsymbol{z}_l;\boldsymbol{x}^{\mathrm{o}}\right),\qquad l=1,\cdots,L_{\mathrm{imp}}$$

并归一化为

$$\bar{w}_l=\frac{w_l}{\sum_{r=1}^{L_{\mathrm{imp}}}w_r},\qquad l=1,\cdots,L_{\mathrm{imp}} \tag{15}$$

采用自归一化重要性采样[13]，第 d 个缺失属性的条件预测分布可近似为

$$\hat{p}_d\left(y\mid \boldsymbol{x}^{\mathrm{o}}\right)=\sum_{l=1}^{L_{\mathrm{imp}}}\bar{w}_l\,p_{\hat{\theta},d}\left(y\mid \boldsymbol{z}_l,s_l\right) \tag{16}$$

式(16)对不同统计类型的缺失属性给出了统一的预测分布近似。在此基础上，点插补值根据属性类型及相应损失函数确定。

对于实值型属性，若第 l 个重要性样本对应的条件高斯均值为 μ_{dl}，则在平方损失下，其点插补值为

$$\hat{x}_d=\sum_{l=1}^{L_{\mathrm{imp}}}\bar{w}_l\mu_{dl} \tag{17}$$

对于正值型属性，若

$$X_d\mid \boldsymbol{z}_l,s_l\sim\mathrm{LogNormal}\left(\mu_{dl},\sigma_{dl}^2\right)$$

则其原始尺度上的条件均值为

$$\exp\left(\mu_{dl}+\frac{\sigma_{dl}^2}{2}\right)$$

因此在平方损失下，点插补值为

$$\hat{x}_d=\sum_{l=1}^{L_{\mathrm{imp}}}\bar{w}_l\exp\left(\mu_{dl}+\frac{\sigma_{dl}^2}{2}\right) \tag{18}$$

对于计数型属性，若第 l 个重要性样本对应的条件泊松率为 λ_{dl}，则在平方损失下，点插补值为

$$\hat{x}_d=\sum_{l=1}^{L_{\mathrm{imp}}}\bar{w}_l\lambda_{dl} \tag{19}$$

对于具有 C_d 个类别的无序类别型属性，记第 l 个重要性样本的类别概率为 π_{dlc}，则重要性加权后的类别概率为

$$\hat{p}_{dc} = \sum_{l=1}^{L_{\text{imp}}} \bar{w}_l \pi_{dlc}, \qquad c = 1, \cdots, C_d \tag{20}$$

在 0~1 损失下，相应的点插补为

$$\hat{c}_d = \arg\max_{1 \le c \le C_d} \hat{p}_{dc} \tag{21}$$

对于有序类别型属性，记第 l 个重要性样本在第 c 个等级上的条件概率为

$$p_{dlc} = p_{\hat{\theta},d}\left(X_d = c \mid \boldsymbol{z}_l, s_l\right)$$

则重要性加权后的等级概率为

$$\hat{p}_{dc} = \sum_{l=1}^{L_{\text{imp}}} \bar{w}_l p_{dlc}, \qquad c = 1, \cdots, C_d \tag{22}$$

在绝对等级损失下，采用上述混合预测分布的中位等级作为点插补：

$$\hat{c}_d = \min\left\{ c : \sum_{r=1}^{c} \hat{p}_{dr} \ge \frac{1}{2} \right\} \tag{23}$$

由此，L_{imp} 个联合潜变量样本及其归一化重要性权重共同构成缺失属性条件预测分布的有限样本近似；不同统计类型的点插补均由该统一预测分布进一步确定。

3. 数值实验

3.1. 实验设置

实验采用 Synthetic-HET5、Adult 和 Default Credit 三个数据集，其样本规模与属性类型构成见表 1。各数据集按 70%、15%和 15%的比例划分为训练集、验证集和测试集，并设置完全随机缺失(missing completely at random, MCAR)和随机缺失(missing at random, MAR)两种缺失机制。其中，MCAR 在原有观测单元中分别随机遮蔽 20%和 40%的数据；MAR 的缺失概率由始终观测属性决定，对其余属性中原有可观测单元设置 20%的目标缺失率。插补误差仅在人工遮蔽且真值已知的测试单元上计算。

Table 1. Datasets and attribute-type composition
表 1. 数据集及属性类型构成

数据集	样本数	属性数	实值型	正值型	计数型	无序分类	有序分类
Synthetic-HET5	10,000	10	2	2	2	2	2
Adult	32,561	12	2	1	2	6	1
Default Credit	30,000	24	12	1	1	10	0

为保证不同训练重要性样本数设置下的模型结构与参数规模一致，本文在所有实验中固定潜在成分数 $S = 10$ 和连续潜变量维数 $q = 10$。编码器与共享解码网络均采用两个包含 128 个隐藏单元的全连接层，并使用 ReLU 激活函数。模型采用 Adam 优化器，学习率设为 10^{-3}，批量大小设为 256。连续潜变量采用重参数化梯度估计；离散潜变量在 $K_{\text{tr}} = 1$ 时采用带基线的得分函数估计，在 $K_{\text{tr}} > 1$ 时采用 VIMCO (Variational Inference for Monte Carlo Objectives)梯度估计[16]。

训练阶段和插补阶段使用的重要性样本数分别设为

$$K_{\text{tr}} \in \{1, 5, 20\}, \qquad L_{\text{imp}} \in \{1, 5, 20, 100, 500, 2000\} \tag{24}$$

在相同 L_{imp} 下比较不同 K_{tr} 训练所得模型，考察 L_{imp} 的影响时保持已训练模型不变；方法间的主比较统一采用 $L_{\text{imp}}=2000$。

对比方法包括相同 H-MIWAE 模型结构下 $K_{\text{tr}}=1$ 的单样本训练模型、统计插补、混合类型迭代随机森林和 GAIN [10]。针对不同属性类型，本文采用与其取值结构相匹配的插补误差指标进行评价：对实值型、正值型和计数型属性，采用标准差归一化均方根误差(standard-deviation-normalized root mean squared error, NRMSE)。记 $\mathcal{T}_d$ 为属性 d 中人工遮蔽且真值已知的测试单元集合，$s_{d,\text{tr,obs}}$ 为该属性训练集可观测值的样本标准差，则

$$\text{NRMSE}_d=\frac{\left[\dfrac{1}{|\mathcal{T}_d|}\sum_{n\in\mathcal{T}_d}\left(x_{nd}-\hat{x}_{nd}\right)^2\right]^{1/2}}{s_{d,\text{tr,obs}}} \tag{25}$$

对无序分类型属性采用分类错误率，对有序分类型属性采用归一化等级位移误差：

$$\begin{aligned}\text{Err}_d&=\frac{1}{|\mathcal{T}_d|}\sum_{n\in\mathcal{T}_d}\mathbf{1}\{c_{nd}\neq\hat{c}_{nd}\}\\ \text{Disp}_d&=\frac{1}{(C_d-1)|\mathcal{T}_d|}\sum_{n\in\mathcal{T}_d}\left|c_{nd}-\hat{c}_{nd}\right|\end{aligned} \tag{26}$$

其中，c_{nd} 和 $\hat{c}_{nd}$ 分别表示真实类别(或等级)与预测类别(或等级)，C_d 为第 d 个有序分类型属性的等级数。

为在相同的重要性样本数下比较不同 K_{tr} 训练所得模型，本文以 K_{eval} 表示评价阶段计算观测数据重要性加权下界时使用的重要性样本数，并固定 $K_{\text{eval}}=1000$。对于测试样本 $\boldsymbol{x}_n^{\text{o}}$，记第 k 个重要性样本对应的权重为 w_{nk}，则其观测数据重要性加权下界的 Monte Carlo 估计为

$$\hat{\mathcal{L}}_{K_{\text{eval}}}\left(\boldsymbol{x}_n^{\text{o}}\right)=\log\left(\frac{1}{K_{\text{eval}}}\sum_{k=1}^{K_{\text{eval}}}w_{nk}\right) \tag{27}$$

测试集上的评价结果由各测试样本的 $\hat{\mathcal{L}}_{K_{\text{eval}}}\left(\boldsymbol{x}_n^{\text{o}}\right)$ 取平均得到。

同时，采用归一化有效样本量(normalized effective sample size, nESS)刻画重要性权重的集中程度。设

$$\bar{w}_{nk}=\frac{w_{nk}}{\sum_{\ell=1}^{K_{\text{eval}}}w_{n\ell}} \tag{28}$$

为测试样本 n 的归一化重要性权重，则

$$\text{nESS}_n=\frac{1}{K_{\text{eval}}\sum_{k=1}^{K_{\text{eval}}}\bar{w}_{nk}^2} \tag{29}$$

nESS 越大，表示重要性权重越均衡；nESS 越小，表示权重越集中。测试集上的 nESS 由各测试样本的 nESS_n 取平均得到。

主实验采用四组配对重复，结果报告四组重复的均值和标准差。

3.2. 异质属性的插补性能

表 2 汇总了 $L_{\text{imp}}=2000$ 时不同方法在不同数据集、缺失率和属性类型下的插补误差。

总体来看，H-MIWAE 在异质缺失数据插补中表现出较强的竞争力。在 Synthetic-HET5 上，三种 K_{tr} 设置在两档 MCAR 缺失率和五类属性下均取得了低于统计插补、迭代随机森林和 GAIN 的平均插补

Table 2. Type-specific imputation errors at $L_{imp} = 2000$

表 2. $L_{imp} = 2000$ 时不同方法的类型特异插补误差

缺失率	属性类型	统计插补	迭代随机森林	GAIN	H-MIWAE		
					$K_{tr} = 1$	$K_{tr} = 5$	$K_{tr} = 20$
(a)完全随机缺失(MCAR)							
Synthetic-HET5							
20%	实值	0.999 (0.013)	0.932 (0.007)	1.080 (0.057)	0.896 (0.007)	**0.893 (0.006)**	0.894 (0.007)
	正值	0.969 (0.069)	0.708 (0.048)	0.949 (0.039)	0.684 (0.035)	0.673 (0.037)	**0.670 (0.037)**
	计数	0.988 (0.028)	0.821 (0.041)	0.980 (0.057)	0.801 (0.031)	0.790 (0.030)	**0.788 (0.030)**
	无序类别	0.726 (0.019)	0.546 (0.017)	0.749 (0.023)	0.529 (0.021)	0.519 (0.020)	**0.518 (0.021)**
	有序类别	0.321 (0.005)	0.277 (0.009)	0.387 (0.020)	**0.267 (0.003)**	0.268 (0.004)	0.270 (0.004)
40%	实值	0.994 (0.012)	0.988 (0.023)	1.103 (0.109)	0.914 (0.016)	0.898 (0.017)	**0.898 (0.017)**
	正值	0.987 (0.033)	0.826 (0.035)	1.022 (0.068)	0.735 (0.029)	0.724 (0.026)	**0.721 (0.025)**
	计数	0.986 (0.012)	0.918 (0.013)	1.042 (0.030)	0.833 (0.014)	0.820 (0.014)	**0.820 (0.012)**
	无序类别	0.712 (0.014)	0.568 (0.028)	0.749 (0.034)	0.552 (0.004)	**0.527 (0.010)**	0.528 (0.010)
	有序类别	0.318 (0.002)	0.290 (0.003)	0.419 (0.014)	0.289 (0.007)	0.274 (0.004)	**0.273 (0.005)**
Adult							
20%	实值	**1.016 (0.048)**	1.059 (0.035)	1.033 (0.050)	1.032 (0.050)	1.037 (0.037)	1.052 (0.033)
	正值	0.981 (0.023)	1.022 (0.015)	1.038 (0.054)	**0.980 (0.023)**	0.982 (0.025)	0.983 (0.022)
	计数	1.003 (0.010)	0.875 (0.013)	1.014 (0.042)	0.852 (0.017)	0.829 (0.011)	**0.827 (0.011)**
	无序类别	0.471 (0.007)	0.299 (0.005)	0.398 (0.025)	0.332 (0.011)	0.300 (0.010)	**0.292 (0.005)**
	有序类别	0.125 (0.002)	**0.097 (0.005)**	0.132 (0.002)	0.119 (0.004)	0.102 (0.007)	0.099 (0.006)
40%	实值	**0.981 (0.047)**	1.150 (0.027)	0.991 (0.047)	1.017 (0.054)	1.009 (0.042)	1.007 (0.020)
	正值	0.992 (0.029)	1.087 (0.032)	1.291 (0.183)	**0.992 (0.028)**	0.994 (0.029)	0.993 (0.028)
	计数	0.997 (0.013)	0.970 (0.004)	1.138 (0.121)	0.899 (0.040)	0.867 (0.013)	**0.857 (0.010)**
	无序类别	0.471 (0.004)	0.335 (0.002)	0.460 (0.024)	0.371 (0.021)	0.338 (0.006)	**0.329 (0.004)**
	有序类别	0.124 (0.002)	**0.106 (0.004)**	0.167 (0.023)	0.121 (0.004)	0.108 (0.003)	0.106 (0.004)
Default Credit							
20%	实值	1.027 (0.102)	**0.491 (0.075)**	0.939 (0.063)	0.628 (0.088)	0.570 (0.102)	0.534 (0.069)
	正值	1.000 (0.052)	**0.750 (0.025)**	0.970 (0.051)	0.788 (0.039)	0.771 (0.032)	0.770 (0.032)
	计数	1.005 (0.035)	**0.924 (0.027)**	1.004 (0.040)	0.990 (0.034)	0.962 (0.042)	0.955 (0.025)
	无序类别	0.445 (0.0004)	**0.193 (0.002)**	0.373 (0.026)	0.267 (0.005)	0.231 (0.005)	0.223 (0.004)
40%	实值	1.013 (0.050)	**0.596 (0.054)**	1.543 (0.374)	0.655 (0.044)	0.607 (0.046)	0.600 (0.047)
	正值	1.008 (0.022)	0.800 (0.033)	1.137 (0.142)	3.656 (0.054)	0.790 (0.020)	**0.784 (0.022)**
	计数	1.014 (0.011)	1.000 (0.010)	1.210 (0.145)	1.007 (0.007)	**0.972 (0.007)**	0.977 (0.008)
	无序类别	0.448 (0.002)	**0.226 (0.003)**	0.387 (0.026)	0.276 (0.002)	0.252 (0.004)	0.248 (0.002)

续表

(b)随机缺失(MAR，目标缺失率 20%)							
Synthetic-HET5							
20%	实值	1.002 (0.025)	0.777 (0.031)	1.057 (0.075)	0.760 (0.026)	**0.749 (0.025)**	0.750 (0.022)
	正值	0.842 (0.044)	0.600 (0.010)	0.768 (0.072)	0.580 (0.021)	**0.561 (0.017)**	0.565 (0.017)
	计数	1.040 (0.012)	0.955 (0.023)	1.047 (0.008)	0.923 (0.015)	0.924 (0.016)	**0.923 (0.016)**
	无序类别	0.683 (0.023)	0.516 (0.027)	0.659 (0.056)	0.494 (0.015)	**0.477 (0.015)**	0.478 (0.018)
	有序类别	0.326 (0.004)	0.278 (0.005)	0.398 (0.052)	0.272 (0.004)	**0.270 (0.005)**	0.271 (0.003)
Adult							
20%	实值	**0.973 (0.081)**	1.020 (0.073)	0.983 (0.088)	0.987 (0.082)	0.973 (0.090)	0.983 (0.069)
	正值	0.977 (0.023)	1.008 (0.013)	1.077 (0.058)	**0.973 (0.022)**	0.976 (0.021)	0.975 (0.021)
	计数	1.039 (0.017)	0.946 (0.029)	1.392 (0.215)	0.926 (0.014)	**0.912 (0.013)**	0.916 (0.011)
	无序类别	0.500 (0.004)	0.314 (0.005)	0.480 (0.057)	0.341 (0.008)	0.320 (0.007)	**0.314 (0.006)**
	有序类别	0.121 (0.004)	**0.095 (0.002)**	0.143 (0.047)	0.110 (0.005)	0.100 (0.005)	0.099 (0.006)
Default Credit							
20%	实值	1.327 (0.110)	**0.635 (0.103)**	1.830 (0.262)	0.772 (0.116)	0.676 (0.114)	0.671 (0.101)
	正值	1.147 (0.023)	**0.746 (0.015)**	1.148 (0.041)	0.775 (0.024)	1.339 (0.014)	2.133 (0.273)
	计数	1.014 (0.025)	**0.930 (0.033)**	1.018 (0.013)	1.001 (0.023)	0.963 (0.010)	0.969 (0.001)
	无序类别	0.441 (0.004)	**0.165 (0.005)**	0.296 (0.019)	0.244 (0.006)	0.207 (0.006)	0.199 (0.006)

误差，在 MAR 条件下亦呈现类似结果。

在真实数据集上，H-MIWAE 的竞争力主要体现在部分属性类型。在 Adult 数据集上，其在 MCAR 和 MAR 条件下的计数型属性均取得了各对比方法中最低的平均插补误差，在无序类别型属性上也表现出较强的竞争力。在 Default Credit 数据集上，H-MIWAE 在实值型和无序类别型属性上的平均插补误差普遍低于 GAIN，并随训练重要性样本数的增加而逐渐降低。总体来看，H-MIWAE 能够有效处理真实数据中的多种异质属性，但不同属性类型的性能收益存在差异。

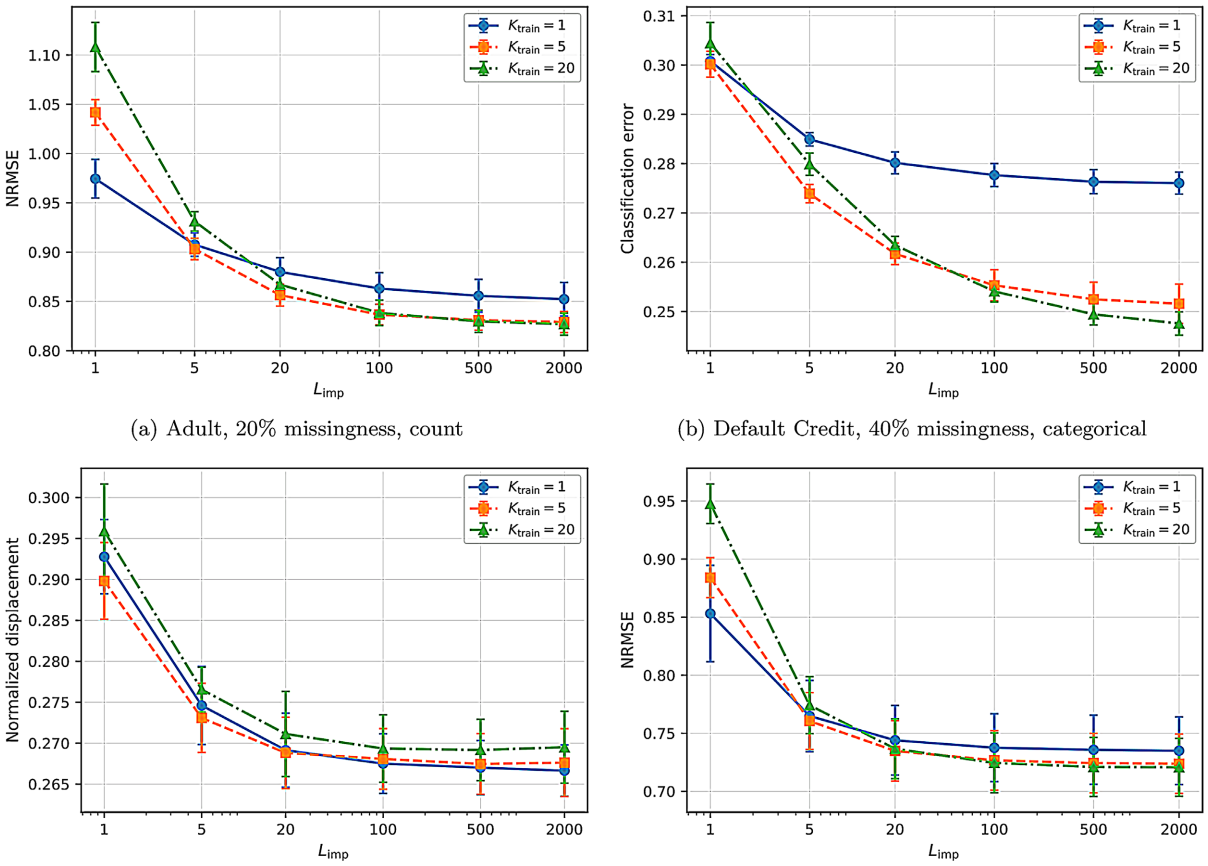
进一步比较 H-MIWAE 的不同训练设置可以发现，较大的 K_{tr} 在多数实验条件下对应更低的插补误差，但表现最优的 K_{tr} 并不固定，其相对表现具有明显的数据集和属性类型差异。

3.3. 训练与插补重要性样本数的联合影响

在 MCAR 条件下，为考察训练重要性样本数和插补重要性样本数对插补性能的联合影响，本文固定各 K_{tr} 设置下已训练模型的参数，改变 L_{imp}，并比较不同 K_{tr} 下插补误差随 L_{imp} 的变化。图 2 展示了四个具有不同变化模式的实验条件下，插补误差随 L_{imp} 的变化，完整结果见附录。

图 2 及附录中的完整结果表明，较大的 K_{tr} 往往需要较大的 L_{imp} 才能体现其插补优势，且这一现象具有明显的属性类型差异。其中，计数型属性上的规律最为稳定，无序类别型属性中也较为明显，而其余属性类型则表现出更强的数据依赖性。

不同训练设置的相对表现并非保持一致。在 Adult 数据集 20%缺失率的计数型属性和 Default Credit 数据集 40%缺失率的无序类别型属性中，$K_{\mathrm{tr}}=20$ 在较小 L_{imp} 下的插补误差较高，但随着 L_{imp} 增大，其误差下降更为明显，并最终实现相对排序的反转。相比之下，Synthetic-HET5 在 20%缺失率下的有序类别型



(a) Adult, 20% missingness, count

(b) Default Credit, 40% missingness, categorical

(c) Synthetic-HET5, 20% missingness, ordinal

(d) Synthetic-HET5, 40% missingness, positive

注：点和误差线分别表示四组配对重复的均值和标准差。

Figure 2. Imputation error versus L_{imp} under different training importance-sample settings

图 2. 不同训练重要性样本数下插补误差随 L_{imp} 的变化

属性未出现相同的变化模式。

统一采用 $K_{eval} = 1000$ 进行评价后，不同 K_{tr} 训练所得模型在观测数据重要性加权下界和 nESS 上呈现出不同特征，具体结果见表 3。其中，下界估计随 K_{tr} 的变化因数据集和缺失率而异，而 nESS 则随 K_{tr} 增大而持续降低，对应的重要性权重分布逐渐趋于集中。

Table 3. Importance-weighted lower bound and nESS at $K_{eval} = 1000$

表 3. $K_{eval} = 1000$ 时的重要性加权下界与 nESS

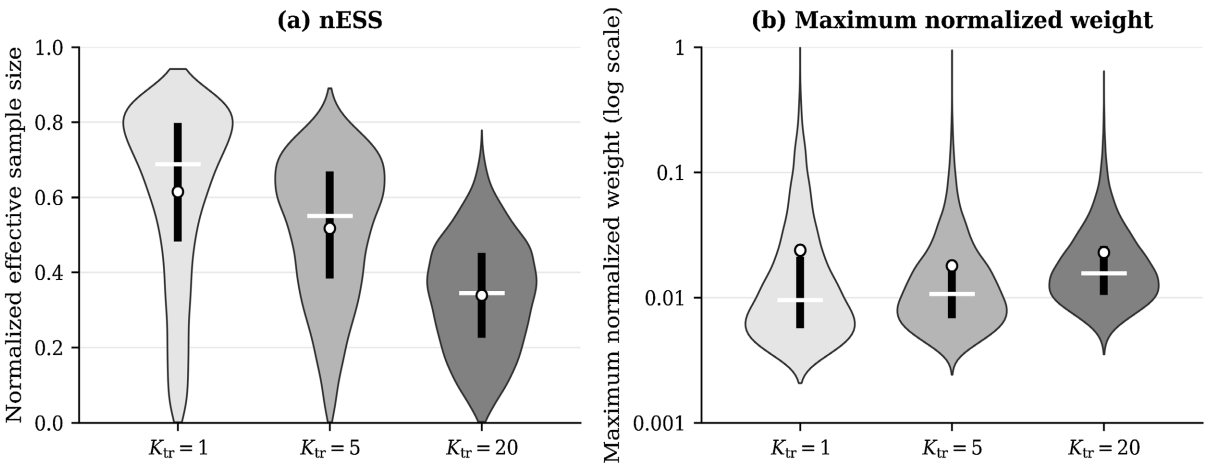
数据集	缺失率	诊断量	$K_{tr} = 1$	$K_{tr} = 5$	$K_{tr} = 20$
Synthetic-HET5	20%	下界	−12.1634 ± 0.0262	−12.1138 ± 0.0207	−12.1183 ± 0.0177
		nESS	0.8933 ± 0.0075	0.7582 ± 0.0076	0.5183 ± 0.0217
	40%	下界	−9.3134 ± 0.0520	−9.2110 ± 0.0355	−9.2138 ± 0.0334
		nESS	0.8919 ± 0.0070	0.7885 ± 0.0216	0.5547 ± 0.0245
Adult	20%	下界	−26.7832 ± 0.6111	−26.0004 ± 0.0401	−25.9103 ± 0.1572
		nESS	0.6150 ± 0.0220	0.5173 ± 0.0261	0.3394 ± 0.0254

续表

	40%	下界	−20.6121 ± 0.2623	−19.7317 ± 0.0494	−19.6685 ± 0.0372
		nESS	0.6672 ± 0.0262	0.6092 ± 0.0130	0.3803 ± 0.0347
Default Credit	20%	下界	−111.5428 ± 0.2487	−107.7468 ± 0.4478	−106.3956 ± 0.1212
		nESS	0.1410 ± 0.0168	0.0559 ± 0.0027	0.0249 ± 0.0022
	40%	下界	−85.9698 ± 2.0902	−110.0845 ± 5.4602	−85.6637 ± 7.9902
		nESS	0.1712 ± 0.0090	0.0866 ± 0.0031	0.0413 ± 0.0039

注：数值为四组配对重复的均值 ± 标准差。

为进一步考察不同训练设置下的重要性采样特征，图 3 展示了 Adult 数据集在 20% MCAR 条件下的样本级 nESS 和最大归一化权重分布。随着 K_{tr} 从 1 增至 20，nESS 的中位数由 0.6881 降至 0.3451，最大归一化权重的中位数则由 0.0095 升至 0.0156，表明较大的训练重要性样本数对应更加集中的权重分配。在自归一化重要性采样中，权重集中可能使有限粒子估计对高权重粒子的抽取更加敏感。结合图 2(a)的计数型属性插补结果，$K_{tr} = 20$ 的模型在较大的 L_{imp} 下才体现出插补优势，这与上述有限样本重要性采样的特征相符，提示插补预算可能在不同训练设置的性能差异中发挥作用。



注：(a) 固定 $K_{eval} = 1000$，每条测试记录的指标先对 5 次独立评价取平均；白色圆点、白色横线和黑色粗线分别表示均值、中位数和四分位区间；(b) 采用对数纵轴。

Figure 3. Sample-level importance-weight diagnostics on Adult under 20% MCAR
图 3. Adult 数据集 20% MCAR 条件下的样本级重要性权重诊断

3.4. 计算效率分析

重要性样本数的选择不仅影响插补性能，还关系到计算开销。为分析不同采样预算下的计算成本，本文在 20% MCAR 条件下，分别考察训练重要性样本数(K_{tr})对训练优化耗时的影响，以及插补重要性样本数(L_{imp})对插补耗时的影响，结果如表 4 所示。

由表 4 可见，在所考察的三个数据集上，随着 K_{tr} 从 1 增至 20，训练优化耗时仅小幅增长，而插补耗时随 L_{imp} 增大明显增加。以 Adult 数据集为例，当 L_{imp} 从 20 增至 2000 时，插补耗时由 0.083 s 增至 1.106 s。结合前述插补性能分析，较大的插补重要性样本数能够在部分属性类型上降低插补误差，但也伴随着更高的计算开销。

Table 4. Training optimization and imputation times under 20% MCAR
表 4. 20% MCAR 条件下的训练优化与插补计算时间

数据集	训练优化耗时(min)			独立插补耗时(s)		
	$K = 1$	$K = 5$	$K = 20$	$L = 20$	$L = 500$	$L = 2000$
Synthetic-HET5	16.01 (1.36)	16.27 (1.12)	16.68 (1.52)	0.081 (0.002)	0.283 (0.012)	1.058 (0.034)
Adult	55.93 (2.05)	56.56 (2.19)	59.28 (3.81)	0.083 (0.006)	0.287 (0.012)	1.106 (0.051)
Default Credit	81.12 (2.87)	81.93 (2.82)	82.99 (2.45)	0.141 (0.008)	0.497 (0.012)	1.931 (0.051)

注：数值为四组重复实验的均值(标准差)。训练耗时为参数优化步骤的累计时间；插补耗时基于($K_{tr} = 5$)的模型，每次处理 256 条测试记录。

4. 讨论

较大的 K_{tr} 往往需要足够的 L_{imp} 才能充分体现其插补优势。在多样本训练目标下，重要性重加权形成的有效变分近似可以比原始识别分布更加灵活，但这并不意味着 $q_\phi\left(s, z \mid x^o\right)$ 会随 K_{tr} 增大而同步改善[13] [14]。对于有序分类型属性插补阶段仍以该识别分布作为重要性采样的提议分布，因而不同训练设置对 L_{imp} 的依赖也会有所不同。

在 H-MIWAE 中，这种依赖进一步受到异质预测结构的影响。相同的潜变量重要性权重需要用于估计不同属性类型的条件预测量，而不同属性类型由条件预测分布得到点插补值的规则并不相同。识别分布与后验之间的近似差异，因而会在有限 L_{imp} 下以不同方式传递到各类属性的预测风险中。这使得 K_{tr} 与 L_{imp} 之间的关系呈现出明显的类型依赖，而非对所有属性一致的变化模式。

多样本训练带来的收益并不必然体现为提议分布采样效率的同步提高。重要性重加权允许模型在原始识别分布仍存在近似偏差时，获得较好的观测数据目标；而在缺失值预测中，同一识别分布又继续承担重要性采样提议分布的作用。由此，训练目标的优化与后续有限样本推断之间存在一定不一致性：前者可以从重加权中获益，后者仍受到提议分布质量的制约。观测数据重要性加权下界与 nESS 的差异正体现了这种不一致性，也说明最终插补性能取决于模型训练与后续重要性采样的共同作用。

本文在 MCAR 和 MAR 缺失条件下考察了 H-MIWAE 的异质属性插补性能，并重点分析了 MCAR 条件下训练与插补重要性样本数的联合影响。然而，现有研究仍限于特定的异质表格数据及所构造的缺失机制。后续工作可进一步考察非随机缺失(missing not at random, MNAR)等更复杂的缺失情形，并将 K_{tr} 与 L_{imp} 的联合分析扩展至更丰富的数据结构与条件预测任务。

5. 结论

本文围绕异质缺失数据的重要性加权变分插补问题，构建 H-MIWAE，并将五类属性的概率建模、观测数据训练与重要性加权条件预测统一于同一框架。通过明确区分训练重要性样本数 K_{tr} 和插补重要性样本数 L_{imp}，本文进一步将多样本训练与有限样本插补作为两个相互关联的阶段进行分析，并从属性类型层面考察二者在异质缺失数据中的作用。由此形成了从异质概率建模、多样本训练到类型特异插补的统一分析框架，为研究重要性加权方法在异质缺失数据中的训练与预测行为提供了基础。

基金项目

河南省科学技术厅河南省科学院科技研发计划联合基金(235200810046)，国家自然科学基金资助项目(12571462)。

参考文献

[1] Josse, J. and Reiter, J.P. (2018) Introduction to the Special Section on Missing Data. *Statistical Science*, **33**, 139-141. https://doi.org/10.1214/18-sts332in

[2] Little, R.J.A. and Rubin, D.B. (2002) Statistical Analysis with Missing Data. 2nd Edition, Wiley. https://doi.org/10.1002/9781119013563

[3] Valera, I. and Ghahramani, Z. (2017) Automatic Discovery of the Statistical Types of Variables in a Dataset. *Proceedings of the* 34*th International Conference on Machine Learning*, **70**, 3521-3529.

[4] Buuren, S.V. and Groothuis-Oudshoorn, K. (2011) Mice: Multivariate Imputation by Chained Equations in *R*. *Journal of Statistical Software*, **45**, 1-67. https://doi.org/10.18637/jss.v045.i03

[5] Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., *et al*. (2001) Missing Value Estimation Methods for DNA Microarrays. *Bioinformatics*, **17**, 520-525. https://doi.org/10.1093/bioinformatics/17.6.520

[6] Stekhoven, D.J. and Bühlmann, P. (2012) MissForest-Non-Parametric Missing Value Imputation for Mixed-Type Data. *Bioinformatics*, **28**, 112-118. https://doi.org/10.1093/bioinformatics/btr597

[7] Murray, J.S. (2018) Multiple Imputation: A Review of Practical and Theoretical Findings. *Statistical Science*, **33**, 142-159. https://doi.org/10.1214/18-sts644

[8] Kingma, D.P. and Welling, M. (2014) Auto-Encoding Variational Bayes. arXiv: 1312.6114.

[9] Rezende, D.J., Mohamed, S. and Wierstra, D. (2014) Stochastic Backpropagation and Approximate Inference in Deep Generative Models. *Proceedings of the* 31*st International Conference on Machine Learning*, Beijing, 21-26 June 2014, 1278-1286.

[10] Yoon, J., Jordon, J. and van der Schaar, M. (2018) GAIN: Missing Data Imputation Using Generative Adversarial Nets. *Proceedings of the* 35*th International Conference on Machine Learning*, Stockholm, 10-15 July 2018, 5689-5698.

[11] Ivanov, O., Figurnov, M. and Vetrov, D. (2019) Variational Autoencoder with Arbitrary Conditioning. arXiv: 1806.02382.

[12] Nazábal, A., Olmos, P.M., Ghahramani, Z. and Valera, I. (2020) Handling Incomplete Heterogeneous Data Using Vaes. *Pattern Recognition*, **107**, Article ID: 107501. https://doi.org/10.1016/j.patcog.2020.107501

[13] Mattei, P.A. and Frellsen, J. (2019) MIWAE: Deep Generative Modelling and Imputation of Incomplete Data Sets. *Proceedings of the* 36*th International Conference on Machine Learning*, Vancouver,13-19 July 2025, 4413-4423.

[14] Rainforth, T., Kosiorek, A., Le, T.A., Maddison, C., Igl, M., Wood, F. and Teh, Y.W. (2018) Tighter Variational Bounds Are Not Necessarily Better. *Proceedings of the* 35*th International Conference on Machine Learning*, Stockholm, 10-15 July 2018, 4277-4285.

[15] Burda, Y., Grosse, R. and Salakhutdinov, R. (2016) Importance Weighted Autoencoders. arXiv: 1509.00519.

[16] Mnih, A. and Rezende, D.J. (2016) Variational Inference for Monte Carlo Objectives. *Proceedings of the* 33*rd International Conference on Machine Learning*, Vancouver, 13-19 July 2025, 2188-2196.