Rethinking the Diffusion Models for Missing Data Imputation: A Gradient Flow Perspective

NeurIPS 2024, Main Track, Poster

Presenter： Zhichao Chen

Authors： Zhichao Chen, Haoxuan Li, Faingyikang Wang, Odin Zhang, Hu Xu, Xiaoyu Jiang, Zhihuan Song, and Hao Wang

Nov. 08. 2024

Outline

1. **Background**

2. **Motivation**

3. **Proposed Approach**

4. **Experimental Results**

Outline

1. Background

2. Motivation

3. Proposed Approach

4. Experimental Results

Background Introduction

1.1 Missing Data Imputation (MDI) Task

① Suppose we have an **ideal tabular data**: $\boldsymbol{X}^{(\text{ideal})} \in \mathbb{R}^{\mathrm{N}\times\mathrm{D}}$.

② However, at hand, we have an **observational data** : $\boldsymbol{X}^{(\text{obs})} = \boldsymbol{X}^{(\text{ideal})} \odot \boldsymbol{M} + \text{NaN} \odot (\mathbf{1}_{\mathrm{N}\times\mathrm{D}} - \boldsymbol{M})$.

③ Where NaN is the abbreviation of **not a number**, $\boldsymbol{M} \in \{0,1\}^{\mathrm{N}\times\mathrm{D}}$ is **mask matrix**, and $\mathbf{1}_{\mathrm{N}\times\mathrm{D}}$ is the **matrix of ones**.

④ We should **recover** $\boldsymbol{X}^{(\text{ideal})}$ **by imputation matrix** $\boldsymbol{X}^{(\text{imp})}$ **as follows:** $\widehat{\boldsymbol{X}} = \boldsymbol{X}^{(\text{ideal})} \odot \boldsymbol{M} + \boldsymbol{X}^{(\text{imp})} \odot (\mathbf{1}_{\mathrm{N}\times\mathrm{D}} - \boldsymbol{M})$.

NeurIPS 2024, Main Track, Submission ID: #1850

Background Introduction

1.2 Diffusion Model for Missing Data Imputation

① Suppose we have a **score function**: $\nabla_{\boldsymbol{X}} \log p(\boldsymbol{X})$

② Diffusion models generate samples by simulating the SDE: $\mathrm{d}\boldsymbol{X}_\tau = f(\boldsymbol{X}_\tau)d\tau + g_\tau dW_\tau$

③ Where τ is the time, $f(\boldsymbol{X}_\tau)$ is **drift term,** which is **concerned with score function**, g_τ is the **volatility term**. The **density** $r(\boldsymbol{X}_\tau)$ is **governed by**: $\frac{\partial r(\boldsymbol{X}_\tau)}{\partial \tau} = -\nabla \cdot \big(r(\boldsymbol{X}_\tau) f(\boldsymbol{X}_\tau)\big) + \frac{1}{2} g_\tau^2 \nabla \cdot \nabla r(\boldsymbol{X}_\tau)$

④ Diffusion-Model-based MDI treats the MDI problem as a conditional generative problem, which aims to generate samples from **conditional score function**: $\nabla_{\boldsymbol{X}^{(\mathrm{miss})}} \log p\big(\boldsymbol{X}^{(\mathrm{miss})} \big| \boldsymbol{X}^{(\mathrm{obs})}\big)$

⑤ In practice, ground-truth missing values are unavailable, thus, we should **mask part of data** to construct the score function: $\nabla_{\boldsymbol{X}^{(\mathrm{miss})}} \log p\big(\boldsymbol{X}^{(\mathrm{miss})} \big| \boldsymbol{X}^{(\mathrm{obs})}\big)$.

Background Introduction

1.3 Wasserstein Gradient Flow

① Suppose we want to optimize a **cost functional**: $\mathcal{F}_{\text{cost}}: \mathcal{P}_2(\mathbb{R}^{\mathrm{D}}) \to \mathbb{R}$

② Wasserstein Gradient Flow is an absolute continuous trajectory $(q_\tau)_{\tau\geq 0}$, that descend $\mathcal{F}_{\text{cost}}$ **as effective as possible**.

③ The trajectory in Wasserstein Gradient Flow is governed by the **continuity equation**: $\frac{\partial q_\tau}{\partial \tau} = -\nabla \cdot (u_\tau q_\tau)$

④ **Velocity field** u_τ is given by $u_\tau = -\nabla_X \frac{\delta \mathcal{F}_{\text{cost}}}{\delta q_\tau}$.

⑤ Based on this, the evolution of $X \in \mathbb{R}^{\mathrm{D}}$ can be **delineated by the ODE** $\frac{dX_\tau}{d\tau} = u_\tau$

Outline

1. Background

2. **Motivation**

3. Proposed Approach

4. Experimental Results

Motivation

2.1 The Task for MDI: An Optimization Perspective

Based on the **Maximum Likelihood Estimation principle**, we can obtain the following optimization problem:

$$\boldsymbol{X}^{(\mathrm{imp})} = \operatorname{argmax}_{\boldsymbol{X}^{(\mathrm{miss})}} \log \hat{p}\left(\boldsymbol{X}^{(\mathrm{miss})} \middle| \boldsymbol{X}^{(\mathrm{obs})}\right).$$

From the perspective of **probabilistic machine learning**, we can reframe the following cost functional:

$$\operatorname{argmax}_{r(\boldsymbol{X}^{(\mathrm{miss})})} \mathbb{E}_{r(\boldsymbol{X}^{(\mathrm{miss})})}\left[\log \hat{p}\left(\boldsymbol{X}^{(\mathrm{miss})} \middle| \boldsymbol{X}^{(\mathrm{obs})}\right)\right],$$

where we assume that $\boldsymbol{X}^{(\mathrm{miss})}$ **comes from a proposal distribution** $r(\boldsymbol{X}^{(\mathrm{miss})})$, optimizing the sample $\boldsymbol{X}^{(\mathrm{miss})}$ is optimizing the distribution.

Motivation

2.2 A Toy Case for DM-based Optimization

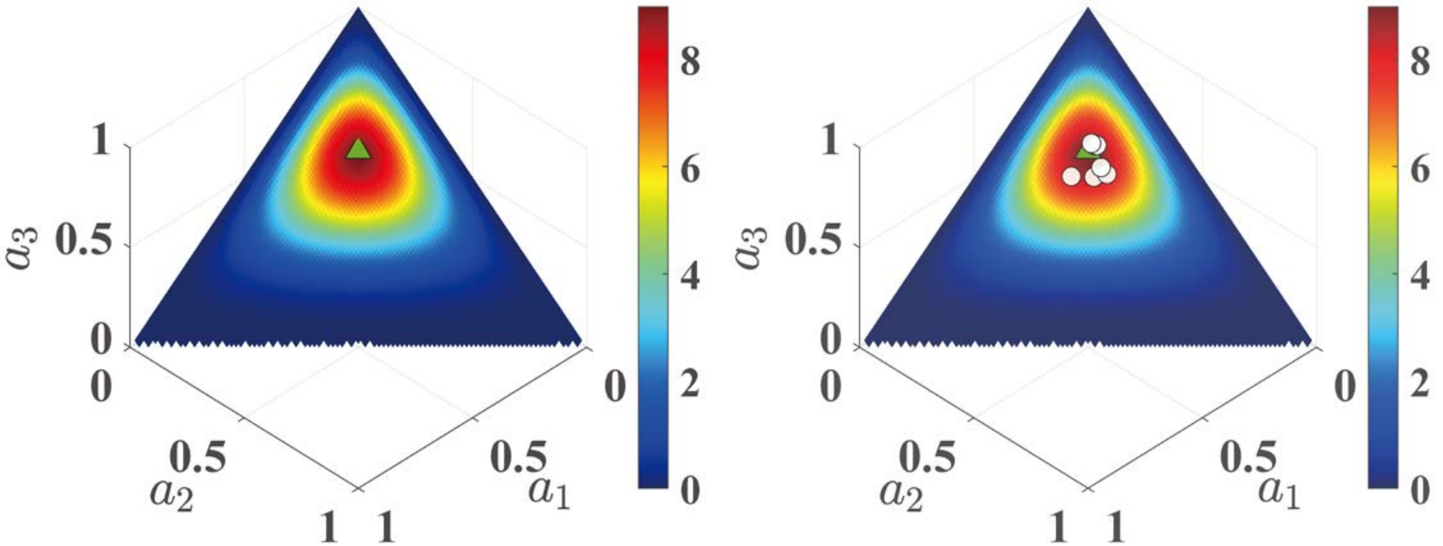
Suppose we have a Dirichlet distribution supports on Δ^2, and we want to optimize the functional defined as follows:

$$\text{argmax}_{\boldsymbol{a}_h \in \Delta^2} \sum_{h=1}^{\mathrm{H}} \{\log\left(\frac{\Gamma\left(\sum_{k=1}^{3} \rho_k\right)}{\prod_{k=1}^{3} \Gamma(\rho_k)}\right) + \sum_{k=1}^{3} (\rho_k - 1)\log \boldsymbol{a}_{k,h}\} \ ,$$

where $\boldsymbol{a}_h$ is the variable, $\rho_k|_{k=1}^{3} = [2.5,2.5,5.0]$ is concentration parameter, and H is the sample number.

Motivation

2.2 A Toy Case for DM-based Optimization



Expected Optimal Results Results by Diffusion Models

- The results tend to **cluster around** the expected optimal results
- There might be something **implicitly optimized** during DMs
- And this **implicitly optimized** term may result in **diversity**

Outline

1. Background

2. Motivation

3. Proposed Approach

4. Experimental Results

Proposed Approach

3.1 What makes a diversified imputation result?

① $\mathrm{d}\boldsymbol{X}_\tau = f(\boldsymbol{X}_\tau)d\tau + g_\tau dW_\tau$ is governed by $\frac{\partial r(\boldsymbol{X}_\tau)}{\partial \tau} = -\nabla \cdot \big(r(\boldsymbol{X}_\tau) f(\boldsymbol{X}_\tau)\big) + \frac{1}{2} g_\tau^2 \nabla \cdot \nabla r(\boldsymbol{X}_\tau)$.

② The trajectory in Wasserstein Gradient Flow is governed by the **continuity equation**: $\frac{\partial q_\tau}{\partial \tau} = -\nabla \cdot (u_\tau q_\tau)$

Let us **analyze and improve** the diffusion model-based MDI within the Wassersetin gradient flow framework!

Proposed Approach

3.1 What makes a diversified imputation result?

For diffusion model-based MDI, we can find that they are optimizing the following cost functional:

$$\mathrm{argmax}_{r(\boldsymbol{X}^{(\mathrm{miss})})}\ \mathbb{E}_{r(\boldsymbol{X}^{(\mathrm{miss})})}\left[\log \hat{p}\left(\boldsymbol{X}^{(\mathrm{miss})} \middle| \boldsymbol{X}^{(\mathrm{obs})}\right)\right] + \psi\left(\boldsymbol{X}^{(\mathrm{miss})}\right) + \mathrm{const}$$

- **VP-SDE:** $\psi\left(\boldsymbol{X}^{(\mathrm{miss})}\right) = \frac{1}{2}\mathbb{H}\left[r\left(\boldsymbol{X}^{(\mathrm{miss})}\right)\right] + \frac{1}{4}\mathbb{E}_{r(\boldsymbol{X}^{(\mathrm{miss})})}\left\{\left[\boldsymbol{X}^{(\mathrm{miss})}\right]^{\top}\left[\boldsymbol{X}^{(\mathrm{miss})}\right]\right\} \geq 0$
- **VE-SDE:** $\psi\left(\boldsymbol{X}^{(\mathrm{miss})}\right) = \frac{1}{2}\mathbb{H}\left[r\left(\boldsymbol{X}^{(\mathrm{miss})}\right)\right] \geq 0$
- **sub-VP-SDE:** $\psi\left(\boldsymbol{X}^{(\mathrm{miss})}\right) = \frac{1}{2}\mathbb{H}\left[r\left(\boldsymbol{X}^{(\mathrm{miss})}\right)\right] + \frac{1}{4\gamma_\tau}\mathbb{E}_{r(\boldsymbol{X}^{(\mathrm{miss})})}\left\{\left[\boldsymbol{X}^{(\mathrm{miss})}\right]^{\top}\left[\boldsymbol{X}^{(\mathrm{miss})}\right]\right\} \geq 0$

➢ $\psi\left(\boldsymbol{X}^{(\mathrm{miss})}\right)$ consistently **greater than 0**.

➢ **Entropy** term $\frac{1}{2}\mathbb{H}\left[r\left(\boldsymbol{X}^{(\mathrm{miss})}\right)\right]$ results in **diversity**.

Proposed Approach

3.2 How to eliminate the diversity?

- $\psi\left(\boldsymbol{X}^{(\mathrm{miss})}\right)$ should be **smaller than 0**.
- The design regularized term should **eliminate diversity**.
- The **negative entropy** is a suitable choice:

$$\psi\left(\boldsymbol{X}^{(\mathrm{miss})}\right) = -\lambda \mathbb{H}\left[r\left(\boldsymbol{X}^{(\mathrm{miss})}\right)\right], \lambda \geq 0$$

- We can define a novel cost functional as follows:

$$\mathcal{F}_{\mathrm{NER}} = \mathbb{E}_{r\left(\boldsymbol{X}^{(\mathrm{miss})}\right)}\left[\log \hat{p}\left(\boldsymbol{X}^{(\mathrm{miss})} \middle| \boldsymbol{X}^{(\mathrm{obs})}\right)\right] - \lambda \mathbb{H}\left[r\left(\boldsymbol{X}^{(\mathrm{miss})}\right)\right]$$

- We call our approach termed '**N**egative **E**ntropy-regularized **W**asserstein Gradient Flow-based **Imp**utation', aka, **NewImp**.

Proposed Approach

3.3 How to optimize this functional?

- Within WGF framework, we can optimize the $\mathcal{F}_{\mathrm{NER}}$ with the help of the following velocity field:

$$u(\boldsymbol{X}^{(\mathrm{miss})}) = -\nabla_{\boldsymbol{X}^{(\mathrm{miss})}} \frac{\delta \mathcal{F}_{\mathrm{NER}}}{\delta r(\boldsymbol{X}^{(\mathrm{miss})})}$$

$$= \nabla_{\boldsymbol{X}^{(\mathrm{miss})}} \log \hat{p}(\boldsymbol{X}^{(\mathrm{miss})} | \boldsymbol{X}^{(\mathrm{obs})}) + \lambda \nabla_{\boldsymbol{X}^{(\mathrm{miss})}} \log r(\boldsymbol{X}^{(\mathrm{miss})})$$

However, implementing this velocity filed to obtain imputed value by $\frac{\mathrm{d}\boldsymbol{X}^{(\mathrm{miss})}}{\mathrm{d}\,\tau} = u(\boldsymbol{X}^{(\mathrm{miss})})$ requires explicitly estimating intractable density function $r(\boldsymbol{X}^{(\mathrm{miss})})$:

- Directly estimating $r(\boldsymbol{X}^{(\mathrm{miss})})$ **is intractable.**
- **Analytically solving** the continuity equation $\frac{\partial r(\boldsymbol{X}^{(\mathrm{miss})})}{\partial \tau} = -\nabla \cdot [u(\boldsymbol{X}^{(\mathrm{miss})}) r(\boldsymbol{X}^{(\mathrm{miss})})]$ is **difficult**.

NeurIPS 2024, Main Track, Submission ID: #1850

Proposed Approach

3.3 How to optimize this functional?

Fortunately, with the help of the following two conditions, we can realize the velocity filed in computer language:

① Velocity filed is restricted within the RKHS satisfies the boundary condition: $u\left(\boldsymbol{X}^{(\text{miss})}\right) \in K\left(\boldsymbol{X}^{(\text{miss})}, \widetilde{\boldsymbol{X}}^{(\text{miss})}\right)$, and the kernel function satisfies: $\lim_{\|\widetilde{\boldsymbol{X}}^{(\text{miss})}\| \to \infty} K\left(\boldsymbol{X}^{(\text{miss})}, \widetilde{\boldsymbol{X}}^{(\text{miss})}\right) = 0$.

② Density function $r(\boldsymbol{X}^{(\text{miss})})$ is bounded.

We can get:

$$u\left(\boldsymbol{X}^{(\text{miss})}\right) = \mathbb{E}_{r(\widetilde{\boldsymbol{X}}^{(\text{miss})})}\left\{\begin{array}{c} -\lambda \nabla_{\widetilde{\boldsymbol{X}}^{(\text{miss})}} K\left(\boldsymbol{X}^{(\text{miss})}, \widetilde{\boldsymbol{X}}^{(\text{miss})}\right) \\ +\left[\nabla_{\widetilde{\boldsymbol{X}}^{(\text{miss})}} \log \hat{p}\left(\widetilde{\boldsymbol{X}}^{(\text{miss})} \middle| \boldsymbol{X}^{(\text{obs})}\right)\right]^{\top} K\left(\boldsymbol{X}^{(\text{miss})}, \widetilde{\boldsymbol{X}}^{(\text{miss})}\right) \end{array}\right\}$$

Proposed Approach

3.4 Can we sidestep the mask modeling?

Interestingly, we can find another joint distribution related cost-functional:

$$\mathcal{F}_{\text{joint-NER}} = \mathbb{E}_{r(\boldsymbol{X}^{(\text{joint})})}\left[\log \hat{p}\left(\boldsymbol{X}^{(\text{joint})}\right)\right] - \lambda \mathbb{H}\left[r\left(\boldsymbol{X}^{(\text{joint})}\right)\right]$$

We can prove that:

- $\mathcal{F}_{\text{joint-NER}} = \mathcal{F}_{\text{NER}} - \text{const}$
- Within Wasserstein gradient flow framework, the **velocity filed induced by $\mathcal{F}_{\text{joint-NER}}$ is identity to** the velocity filed induced by $\mathcal{F}_{\text{NER}}$, $u\left(\boldsymbol{X}^{(\text{joint})}\right)$ satisfies: $u\left(\boldsymbol{X}^{(\text{joint})}\right) = u(\boldsymbol{X}^{(\text{miss})})$.

Proposed Approach

3.4 Can we sidestep the mask modeling?

By far, we merely need to simulate the velocity field:

$$u\left(\boldsymbol{X}^{(\text{joint})}\right) = \mathbb{E}_{r(\widetilde{\boldsymbol{X}}^{(\text{joint})})}\left\{\begin{array}{c} -\lambda \nabla_{\widetilde{\boldsymbol{X}}^{(\text{miss})}} K\left(\boldsymbol{X}^{(\text{joint})}, \widetilde{\boldsymbol{X}}^{(\text{joint})}\right) \\ +\left[\nabla_{\widetilde{\boldsymbol{X}}^{(\text{miss})}} \log \hat{p}(\widetilde{\boldsymbol{X}}^{(\text{joint})})\right]^{\top} K\left(\boldsymbol{X}^{(\text{joint})}, \widetilde{\boldsymbol{X}}^{(\text{joint})}\right) \end{array}\right\}$$

We concerning terms can be realized by:

- $K\left(\boldsymbol{X}^{(\text{joint})}, \widetilde{\boldsymbol{X}}^{(\text{joint})}\right) = \exp(-\frac{\|\boldsymbol{X}^{(\text{joint})} - \widetilde{\boldsymbol{X}}^{(\text{joint})}\|_2^2}{2h^2})$
- $\nabla_{\widetilde{\boldsymbol{X}}^{(\text{miss})}} K\left(\boldsymbol{X}^{(\text{joint})}, \widetilde{\boldsymbol{X}}^{(\text{joint})}\right) = \nabla_{\widetilde{\boldsymbol{X}}^{(\text{joint})}} K\left(\boldsymbol{X}^{(\text{joint})}, \widetilde{\boldsymbol{X}}^{(\text{joint})}\right) \odot (\mathbf{1}_{\mathrm{N}\times\mathrm{D}} - \boldsymbol{M}) + 0 \times \boldsymbol{M}$
- $\nabla_{\widetilde{\boldsymbol{X}}^{(\text{miss})}} \log \hat{p}(\widetilde{\boldsymbol{X}}^{(\text{joint})}) = \nabla_{\widetilde{\boldsymbol{X}}^{(\text{joint})}} \log \hat{p}(\widetilde{\boldsymbol{X}}^{(\text{joint})}) \odot (\mathbf{1}_{\mathrm{N}\times\mathrm{D}} - \boldsymbol{M}) + 0 \times \boldsymbol{M}$
- $\mathbb{E}_{r(\widetilde{\boldsymbol{X}}^{(\text{joint})})}$ realized by Monte Carlo approximation
- Now we **merely remain the implementation** of $\nabla_{\widetilde{\boldsymbol{X}}^{(\text{joint})}} \log \hat{p}(\widetilde{\boldsymbol{X}}^{(\text{joint})})$.

NeurIPS 2024, Main Track, Submission ID: #1850

Proposed Approach

3.4 Estimation of Joint Distribution

Up to now, our primary task is to estimate the joint distribution $\nabla_{\boldsymbol{X}^{(\text{joint})}} \log \hat{p}(\boldsymbol{X}^{(\text{joint})})$.

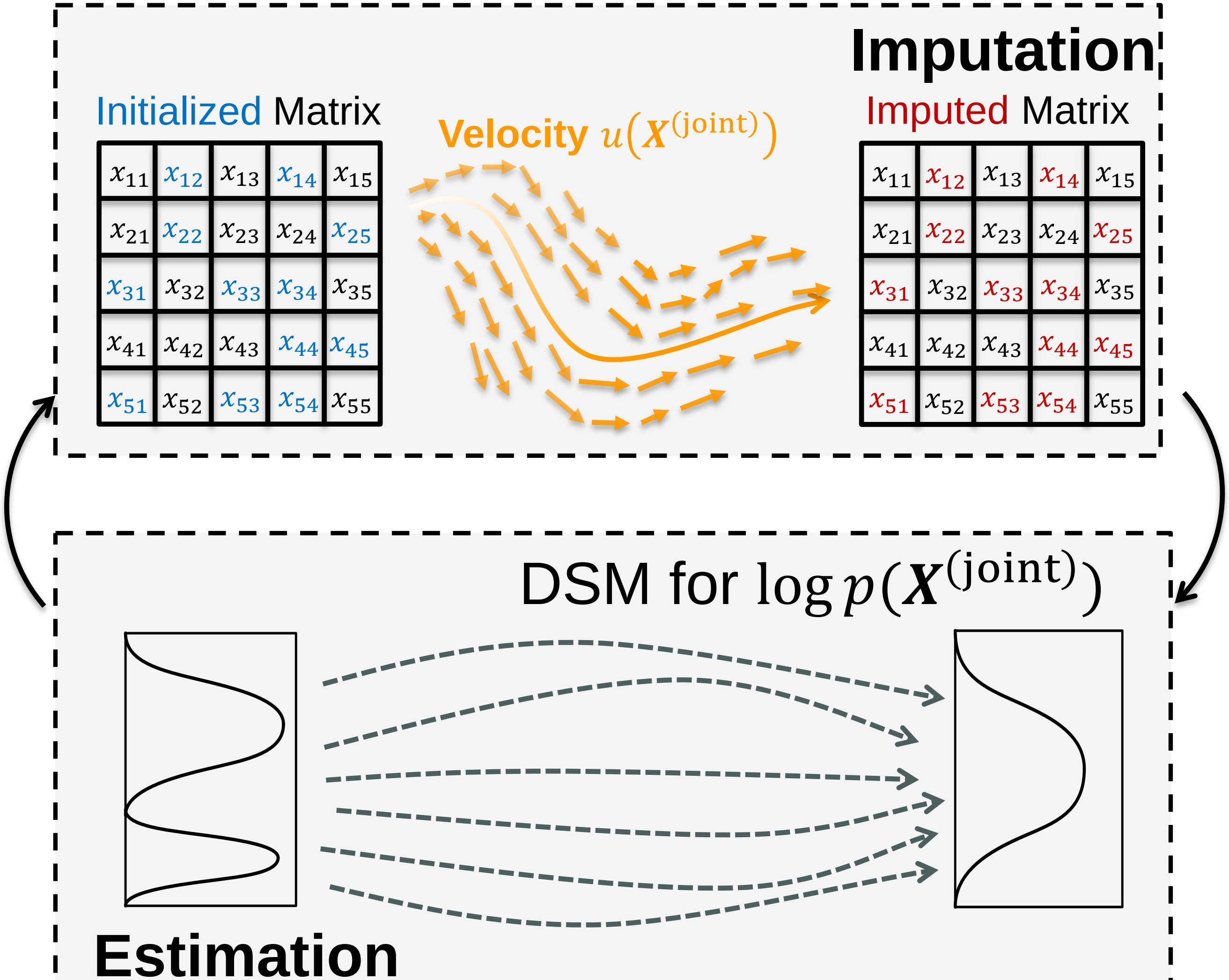
- We parameterize the score function $\nabla_{\boldsymbol{X}^{(\text{joint})}} \log \hat{p}(\boldsymbol{X}^{(\text{joint})})$ by a neural network.
- The neural network is trained by denoise score matching (DSM) by the following loss function:

$$\mathcal{L}_{\text{DSM}} = \frac{1}{2}\mathbb{E}_{q_\sigma(\widehat{\boldsymbol{X}}^{(\text{joint})}|\boldsymbol{X}^{(\text{joint})})}[||\nabla_{\widehat{\boldsymbol{X}}^{(\text{joint})}} \log \hat{p}\left(\widehat{\boldsymbol{X}}^{(\text{joint})}\right) - \nabla_{\widehat{\boldsymbol{X}}^{(\text{joint})}} \log q_\sigma\left(\widehat{\boldsymbol{X}}^{(\text{joint})}\middle|\boldsymbol{X}^{(\text{joint})}\right) ||^2]$$

- where $\widehat{\boldsymbol{X}}^{(\text{joint})} = \boldsymbol{X}^{(\text{joint})} + \epsilon, \epsilon \sim \mathcal{N}(0, \sigma^2 I)$

Proposed Approach

3.5 Overall Framework



Outline

1. Background

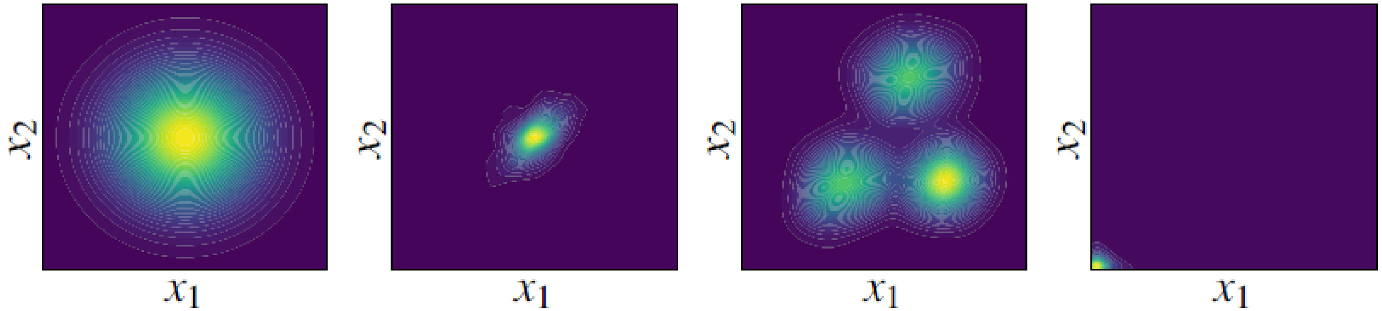
2. Motivation

3. Proposed Approach

4. Experimental Results

Experimental Results

4.1 Toy Case Study Results



(a) Standard Gaussian (b) Student's-t (c) Gaussian Mixture (d) Skewed Gaussian

Scenario	Distribution Type	MAE	WASS
MAR	Gaussian	$0.769_{\pm 0.030}$	$0.481_{\pm 0.026}$
	Student's-t	$0.737_{\pm 0.053}$	$0.513_{\pm 0.048}$
	Gaussian Mixture	$0.763_{\pm 0.097}$	$0.419_{\pm 0.104}$
	Skewed-Gaussian	$0.422_{\pm 0.253}$	$0.492_{\pm 0.025}$
MCAR	Gaussian	$0.769_{\pm 0.013}$	$0.287_{\pm 0.014}$
	Student's-t	$0.698_{\pm 0.030}$	$0.307_{\pm 0.014}$
	Gaussian Mixture	$0.824_{\pm 0.017}$	$0.391_{\pm 0.023}$
	Skewed-Gaussian	$0.417_{\pm 0.140}$	$0.210_{\pm 0.026}$
MNAR	Gaussian	$0.778_{\pm 0.034}$	$0.309_{\pm 0.030}$
	Student's-t	$0.715_{\pm 0.028}$	$0.323_{\pm 0.019}$
	Gaussian Mixture	$0.807_{\pm 0.042}$	$0.380_{\pm 0.050}$
	Skewed-Gaussian	$0.421_{\pm 0.111}$	$0.202_{\pm 0.006}$

- NewImp approach outperforms on different types of data.
- This phenomenon reflects that the NewImp approach is robust to data type like heavy-tailed, skewed, and multi-modal.

Experimental Results

4.2 Baseline Comparison

Scenario	Model	BT		BCD		CC		CBV		IS		PK		QB		WQW	
		MAE	WASS	MAE	WASS	MAE	WASS	MAE	WASS	MAE	WASS	MAE	WASS	MAE	WASS	MAE	WASS
MAR	CSDI_T	0.93 *	3.44 *	0.92 *	18.20 *	0.85 *	2.82 *	0.81 *	3.86 *	0.70 *	16.86 *	0.99 *	15.86 *	0.65 *	20.10 *	0.77 *	4.13 *
	MissDiff	0.85 *	2.20 *	0.91 *	16.53 *	0.87 *	1.59 *	0.83 *	3.87 *	0.72 *	13.25 *	0.92 *	17.07 *	0.63 *	26.25 *	0.75 *	6.88 *
	GAIN	0.75 *	0.65 *	0.54 *	1.64 *	0.75 *	0.67 *	0.68 *	0.68 *	0.56 *	1.88 *	0.59 *	1.90 *	0.65 *	5.05 *	0.68 *	0.87 *
	MIRACLE	0.62 *	0.38	0.55 *	1.92 *	0.43	**0.25**	0.55 *	0.46 *	3.39 *	35.06 *	4.14 *	34.07 *	0.46	**2.87** *	0.51 *	0.56
	MIWAE	0.64	0.53	0.52 *	1.54 *	0.76 *	0.64 *	0.82 *	0.92 *	0.50 *	1.87 *	0.65 *	1.98 *	0.55 *	5.05 *	0.62 *	0.75 *
	Sink	0.87 *	0.92 *	0.92 *	3.84 *	0.88 *	0.83 *	0.84 *	0.98 *	0.75 *	2.43 *	0.94 *	3.61 *	0.65 *	4.71 *	0.76 *	1.04 *
	TDM	0.83 *	0.89 *	0.83 *	3.47 *	0.81 *	0.73 *	0.76 *	0.85 *	0.62 *	1.96 *	0.86 *	3.36 *	0.59 *	4.46 *	0.73 *	0.99 *
	ReMasker	**0.52**	0.52	0.48 *	1.15	0.60 *	0.43 *	0.49 *	0.37 *	0.62 *	2.23 *	0.61 *	1.59 *	0.60 *	3.81	0.51 *	0.59 *
	NewImp	0.52	**0.38**	**0.34**	**0.82**	**0.35**	0.25	**0.31**	**0.20**	**0.39**	**1.31**	**0.44**	**1.21**	**0.45**	3.50	**0.46**	**0.55**
MCAR	CSDI_T	0.73 *	1.93 *	0.73 *	15.51 *	0.85 *	2.71 *	0.83 *	3.79 *	0.76 *	15.19 *	0.72 *	12.42 *	0.57 *	19.89 *	0.78 *	4.11 *
	MissDiff	0.72 *	1.62 *	0.73 *	14.39 *	0.84 *	1.23 *	0.82 *	3.31 *	0.75 *	13.01 *	0.71 *	14.12 *	0.56 *	19.67 *	0.76 *	4.95 *
	GAIN	0.72 *	0.39 *	0.38 *	1.41 *	0.78 *	0.73 *	0.72 *	0.99 *	0.57 *	3.72 *	0.46 *	1.70	0.42 *	3.62	0.73 *	1.14 *
	MIRACLE	0.52	0.15 *	0.44 *	1.94 *	0.53 *	0.35	0.61 *	0.72 *	2.99 *	52.92 *	3.38 *	42.78 *	0.35	**2.71** *	0.56 *	**0.75**
	MIWAE	0.58 *	0.24	0.50 *	2.55 *	0.76 *	0.69 *	0.83 *	1.24 *	0.64 *	4.95 *	0.51 *	2.05 *	0.48 *	5.87 *	0.67 *	0.95 *
	Sink	0.73 *	0.48 *	0.75 *	4.39 *	0.84 *	0.85 *	0.82 *	1.27 *	0.75 *	4.94 *	0.74 *	3.36 *	0.61 *	5.92 *	0.76 *	1.25 *
	TDM	0.68 *	0.42 *	0.63 *	3.57 *	0.77 *	0.75 *	0.77 *	1.15 *	0.66 *	4.20 *	0.64 *	2.89 *	0.52 *	5.34 *	0.74 *	1.20 *
	ReMasker	**0.46** *	**0.11**	0.39 *	1.69 *	0.55 *	0.37	0.56 *	0.64 *	0.54 *	4.01 *	0.48 *	1.71 *	0.45 *	3.94	0.57 *	0.76
	NewImp	0.48	0.18	**0.25**	**0.80**	**0.47**	**0.34**	**0.42**	**0.44**	**0.44**	**3.05**	**0.32**	**1.01**	**0.34**	3.66	**0.53**	0.76
MNAR	CSDI_T	0.83 *	2.29 *	0.82 *	15.68 *	0.85 *	2.78 *	0.83 *	3.83 *	0.74 *	15.54 *	0.84 *	12.20 *	0.62 *	19.77 *	0.78 *	4.09 *
	MissDiff	0.78 *	1.43 *	0.81 *	14.89 *	0.84 *	1.27 *	0.83 *	3.53 *	0.72 *	13.31 *	0.81 *	16.02 *	0.61 *	21.62 *	0.76 *	4.70 *
	GAIN	0.77 *	0.57 *	0.62 *	3.94 *	0.78 *	0.79 *	0.78 *	1.15 *	0.71 *	4.85 *	0.70 *	4.20 *	0.76 *	10.53 *	0.75 *	1.23 *
	MIRACLE	0.63	0.35	0.60 *	4.26 *	0.52 *	0.35	0.63 *	0.77 *	3.10 *	55.56 *	3.49 *	44.76 *	0.52 *	5.61	0.58 *	0.80
	MIWAE	0.66 *	0.42	0.56 *	3.31 *	0.74 *	0.68 *	0.85 *	1.30 *	0.59 *	4.33 *	0.60 *	3.06 *	0.53 *	7.21 *	0.67 *	0.97 *
	Sink	0.79 *	0.68 *	0.83 *	5.90 *	0.83 *	0.89 *	0.84 *	1.36 *	0.75 *	4.86 *	0.84 *	5.02 *	0.64 *	7.23 *	0.77 *	1.33 *
	TDM	0.76 *	0.64 *	0.74 *	5.18 *	0.76 *	0.77 *	0.79 *	1.24 *	0.64 *	4.02 *	0.76 *	4.54 *	0.57 *	6.45	0.74 *	1.23 *
	ReMasker	**0.53**	**0.28**	0.42 *	1.91 *	0.54 *	0.39 *	0.59 *	0.68 *	0.51 *	3.59 *	0.63 *	3.06 *	0.47	**5.02**	0.56	**0.77**
	NewImp	0.60	0.35	**0.32**	**1.46**	**0.44**	**0.34**	**0.46**	**0.52**	**0.40**	**2.68**	**0.39**	**1.56**	**0.42**	5.57	**0.55**	0.81

Kindly Note: The best results are **bolded** and the second best results are underliend. "*" marks the results that NewImp significantly outperform with p-value< 0.05 over paired samples t-test.

- NewImp approach outperforms most of prevalent models.

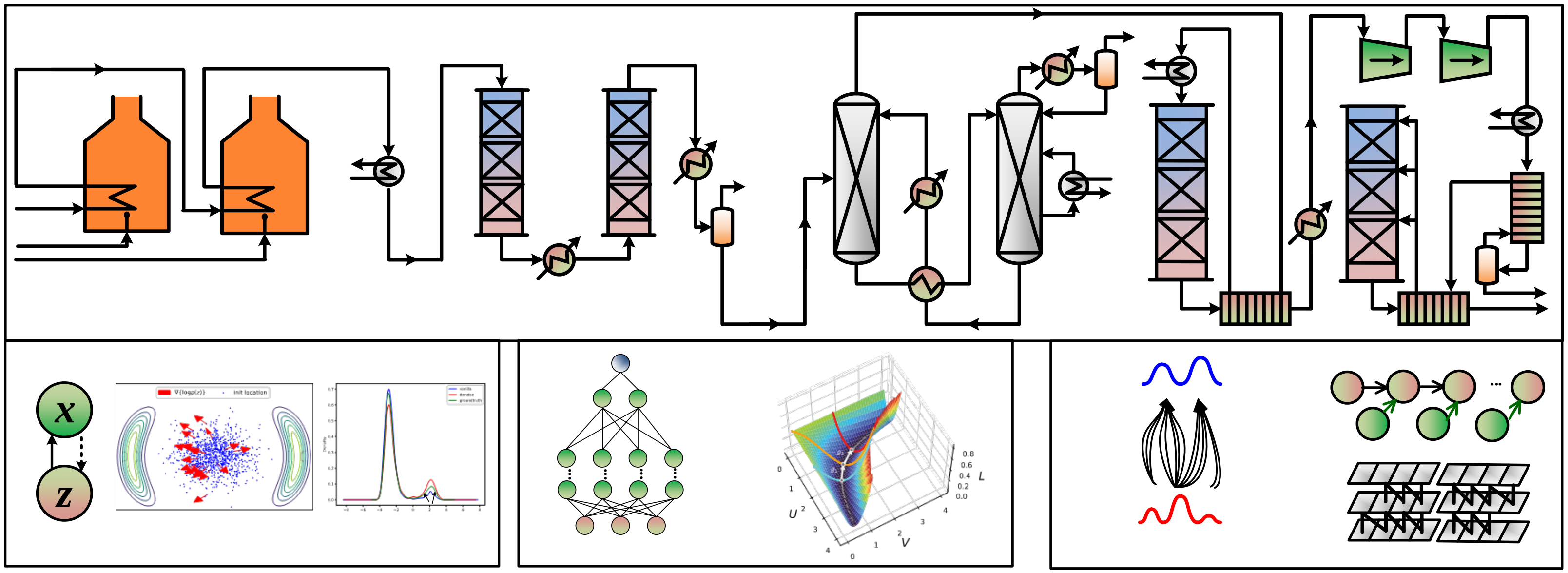
Experimental Results

4.3 Ablation Study

Scenario	NER	Joint	BT		BCD		CC		CBV		IS		PK		QB		WQW	
			MAE	WASS	MAE	WASS	MAE	WASS	MAE	WASS	MAE	WASS	MAE	WASS	MAE	WASS	MAE	WASS
MAR	✗	✗	0.96*	3.82*	1.05*	20.2*	1.04*	5.47*	0.86*	5.81*	0.67*	20.2*	1.06*	15.6*	0.72*	22.5*	0.79*	6.49*
	✗	✓	0.54	0.42	0.34	0.82	0.61*	0.40*	0.58*	0.47*	0.43*	1.34	0.46*	1.25*	0.47*	3.56*	0.55*	0.64*
	✓	✗	0.96*	3.83*	1.05*	20.3*	1.04*	5.49*	0.86*	5.83*	0.67*	20.2*	1.06*	15.6*	0.72*	22.5*	0.79*	6.51*
	✓	✓	**0.52**	**0.38**	**0.34**	**0.82**	**0.35**	**0.25**	**0.31**	**0.20**	**0.39**	**1.31**	**0.44**	**1.21**	**0.45**	**3.50**	**0.46**	**0.55**
MCAR	✗	✗	0.72*	2.11*	0.74*	16.7*	0.85*	3.72*	0.83*	5.22*	0.74*	18.4*	0.71*	12.7*	0.58*	20.1*	0.76*	5.57*
	✗	✓	0.52*	**0.17***	**0.25**	**0.79**	0.62*	0.46*	0.61*	0.71*	0.46	**3.05**	0.34	1.09	0.36*	3.74*	0.58*	0.82*
	✓	✗	0.72*	2.12*	0.73*	16.8*	0.86*	3.73*	0.83*	5.24*	0.74*	18.4*	0.71*	12.7*	0.58*	20.1*	0.76*	5.60*
	✓	✓	**0.48**	0.18	0.25	0.80	**0.47**	**0.34**	**0.42**	**0.44**	**0.44**	3.05	**0.32**	**1.01**	**0.34**	**3.66**	**0.53**	**0.76**
MNAR	✗	✗	0.81*	2.47*	0.89*	18.2*	0.87*	3.85*	0.85*	5.26*	0.69*	17.6*	0.87*	13.0*	0.64*	20.6*	0.77*	5.71*
	✗	✓	0.62	0.37	0.32	1.47	0.61*	0.47*	0.64*	0.79*	0.44	2.79	0.43*	1.88*	0.44*	5.65	0.60*	0.87*
	✓	✗	0.82*	2.57*	0.89*	18.3*	0.87*	3.86*	0.85*	5.28*	0.69*	17.7*	0.88*	13.5*	0.64*	20.7*	0.77*	5.73*
	✓	✓	**0.60**	**0.35**	**0.32**	**1.46**	**0.44**	**0.34**	**0.46**	**0.52**	**0.40**	**2.68**	**0.39**	**1.56**	**0.42**	**5.57**	**0.55**	**0.81**

Kindly Note: The best results are **bolded** and the second best results are underliend. "*" marks the results that NewImp significantly outperform with p-value<0.05 over paired samples t-test.

- Both of the negative regularization term and joint modeling strategy are effective for model performance improvement.



Thank you for listening！

All suggestions are welcomed. !