not-MIWAE: Deep Generative Modelling with Missing Not At Random Data

with sup-not-MIWAE: a Supervised Extension

Adam GASSEM Amine MAAZIZI Ewerthon MELZANI

ENSTA Paris ENS Paris-Saclay

Abstract

We study *not-MIWAE* (Ipsen et al., 2021)[1], a deep latent-variable model for learning and imputing when data are **Missing Not At Random (MNAR)**. Unlike MAR/MCAR settings where the mask mechanism can be ignored, MNAR couples the generative model and the missingness mechanism, requiring **joint modeling** of data and mask. We (i) reproduce core results on UCI datasets, (ii) extend the MNAR clipping experiment to **CelebA** images, (iii) propose a transparent supervised MNAR baseline **sup-not-MIWAE**, and (iv) release unified PyTorch implementations (MIWAE / not-MIWAE / supMIWAE / sup-not-MIWAE).

Context

Data $\boldsymbol{x} \in \mathbb{R}^p$ is split into observed and missing parts: $\boldsymbol{x} = (\boldsymbol{x}^o, \boldsymbol{x}^m)$ with mask $\boldsymbol{s} \in \{0,1\}^p$:

$$s_j = \mathbb{1}\{x_j \text{ observed}\}.$$

Missingness regimes:

- MCAR: $p(\boldsymbol{s} \mid \boldsymbol{x}) = p(\boldsymbol{s})$
- MAR: $p(\boldsymbol{s} \mid \boldsymbol{x}) = p(\boldsymbol{s} \mid \boldsymbol{x}^o)$
- MNAR: $p(\boldsymbol{s} \mid \boldsymbol{x})$ depends on $\boldsymbol{x}^m$ (non-ignorable)

Problem

Under MNAR, learning θ from only $\boldsymbol{x}^o$ is biased because the observed likelihood must integrate missing values *inside* the mask model:

$$p_{\theta,\phi}(\boldsymbol{x}^o, \boldsymbol{s}) = \int p_\theta(\boldsymbol{x}^o, \boldsymbol{x}^m)\, p_\phi(\boldsymbol{s} \mid \boldsymbol{x}^o, \boldsymbol{x}^m)\, d\boldsymbol{x}^m.$$

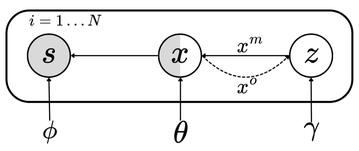
In MAR/MCAR, p_ϕ does not depend on $\boldsymbol{x}^m$ and can be factored out; **not so in MNAR**. Hence we must model the missing mechanism jointly with the data distribution.

Model: not-MIWAE

Deep latent variable model with latent $\boldsymbol{z}$ [1]:

$$p(\boldsymbol{x}^o, \boldsymbol{x}^m, \boldsymbol{s}, \boldsymbol{z}) = p(\boldsymbol{z})\, p_\theta(\boldsymbol{x}^o, \boldsymbol{x}^m \mid \boldsymbol{z})\, p_\phi(\boldsymbol{s} \mid \boldsymbol{x}^o, \boldsymbol{x}^m),$$

with conditional independence in the decoder: $p_\theta(\boldsymbol{x} \mid \boldsymbol{z}) = \prod_j p_\theta(x_j \mid \boldsymbol{z})$.



Inference: IW variational lower bound

Use variational posterior $q_\gamma(\boldsymbol{z} \mid \boldsymbol{x}^o)$ and draw K importance samples $\boldsymbol{z}_{ki} \sim q_\gamma(\boldsymbol{z} \mid \boldsymbol{x}_i^o)$, $\boldsymbol{x}_{ki}^m \sim p_\theta(\boldsymbol{x}^m \mid \boldsymbol{z}_{ki})$. Optimize [2] :

$$\mathcal{L}_K(\theta, \phi, \gamma) = \sum_{i=1}^{n} \mathbb{E}\left[\log \frac{1}{K} \sum_{k=1}^{K} w_{ki}\right], \quad w_{ki} = \frac{p_\phi(\boldsymbol{s}_i \mid \boldsymbol{x}_i^o, \boldsymbol{x}_{ki}^m)\, p_\theta(\boldsymbol{x}_i^o \mid \boldsymbol{z}_{ki})\, p(\boldsymbol{z}_{ki})}{q_\gamma(\boldsymbol{z}_{ki} \mid \boldsymbol{x}_i^o)}.$$

$\mathcal{L}_K$ is a lower bound and tightens monotonically as $K \uparrow$ [3].

Imputation & OT interpretation

Imputation is a Bayesian decision problem under $p_{\theta,\phi}(\boldsymbol{x}^m \mid \boldsymbol{x}^o, \boldsymbol{s})$:

$$\hat{\boldsymbol{x}}^m = \arg\min_{\hat{\boldsymbol{x}}^m} \mathbb{E}[L(\boldsymbol{x}^m, \hat{\boldsymbol{x}}^m) \mid \boldsymbol{x}^o, \boldsymbol{s}]\,.$$

Squared loss $\Rightarrow$ conditional mean (via SNIS), **absolute loss** $\Rightarrow$ conditional median. [1].Which is equivalent , by the Optimal transport point of view as imputations using the distance between distributions $\mathbb{W}^p$ restricted to the variational family of Dirac measures $q = \delta_{\hat{\boldsymbol{x}}^m}$, studied in details in [4], then collapsing posterior uncertainty (motivating a distributional imputation).

Using prior knowledge on the MNAR mechanism

The missing model typically factorizes as Bernoulli:

$$p_\phi(\boldsymbol{s} \mid \boldsymbol{x}) = \prod_{j=1}^{p} \pi_{\phi,j}(\boldsymbol{x})^{s_j} \big(1 - \pi_{\phi,j}(\boldsymbol{x})\big)^{1-s_j}.$$

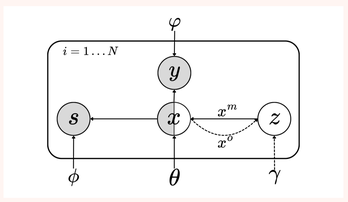
When we know a **self-masking** structure, we can constrain $\pi_{\phi,j}(\boldsymbol{x}) = \sigma(a x_j + b)$ (logistic censoring), improving identifiability and performance. [1]

sup-not-MIWAE

Goal: predict y from partially observed covariates by marginalizing missing features:

$$p(y \mid \boldsymbol{x}^o, \boldsymbol{s}) = \int p_\varphi(y \mid \boldsymbol{x}^o, \boldsymbol{x}^m)\, p_{\theta,\phi}(\boldsymbol{x}^m \mid \boldsymbol{x}^o, \boldsymbol{s})\, d\boldsymbol{x}^m.$$

Following existing supervised deep generative approaches [4], we augment not-MIWAE [1] with a predictor head $p_\varphi(y \mid x)$ (labels depend on *complete* covariates).



Training uses an IW bound with weights

$$w_{ki}^{\text{sup}} = \frac{p_\phi(\boldsymbol{s}_i \mid \boldsymbol{x}_i^o, \boldsymbol{x}_{ki}^m)\, p_\varphi(y_i \mid \boldsymbol{x}_i^o, \boldsymbol{x}_{ki}^m)\, p_\theta(\boldsymbol{x}_i^o \mid \boldsymbol{z}_{ki})\, p(\boldsymbol{z}_{ki})}{q_\gamma(\boldsymbol{z}_{ki} \mid \boldsymbol{x}_i^o)}.$$

Imputation under MNAR

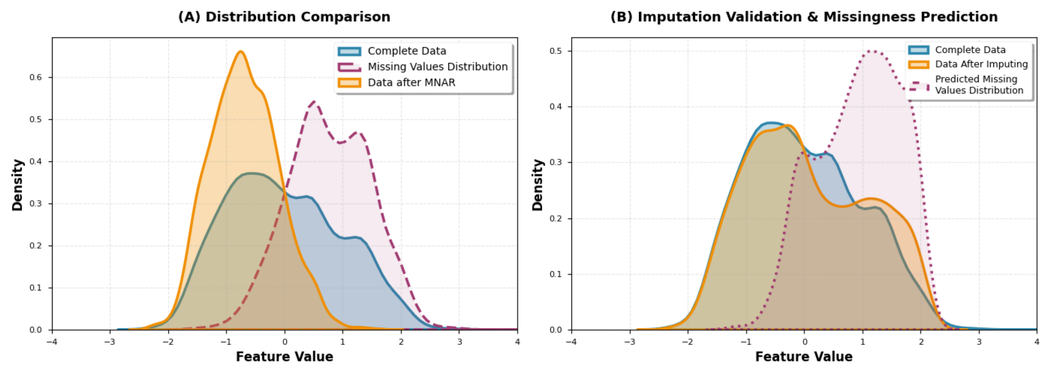
Experimental Setup:

- **Scenario:** Synthetic MNAR injected into UCI datasets and CelebA (images).
- **Baselines:** Compared against Mean, MICE, and standard MIWAE (MAR assumption).
- **Metric:** RMSE calculated against the ground truth of missing values.

Table 1. Imputation RMSE on UCI datasets under MNAR self-masking. Lower is better.

Model	Banknote (61.8%)	Concrete (50.3%)	Breast (43.6%)	White (44.6%)	CelebA (Images)
Mean	1.546	1.329	1.548	1.460	0.466
KNN	1.543	1.366	1.290	1.410	0.350
MICE	1.540	1.247	1.130	1.290	–
MIWAE	1.243	1.148	1.113	1.310	–
not-MIWAE variants					
Self-masking (Unknown parameters)	0.770	1.099	0.836	0.950	0.093
Linear	1.023	1.149	1.018	1.040	–
Non-linear	1.133	1.128	0.938	1.080	–

Quantitative Results: Table 1 shows that the proposed *not-MIWAE* (self-masking) consistently achieves the lowest RMSE, significantly outperforming standard baselines (Mean, KNN, MICE) and the MAR-based MIWAE.



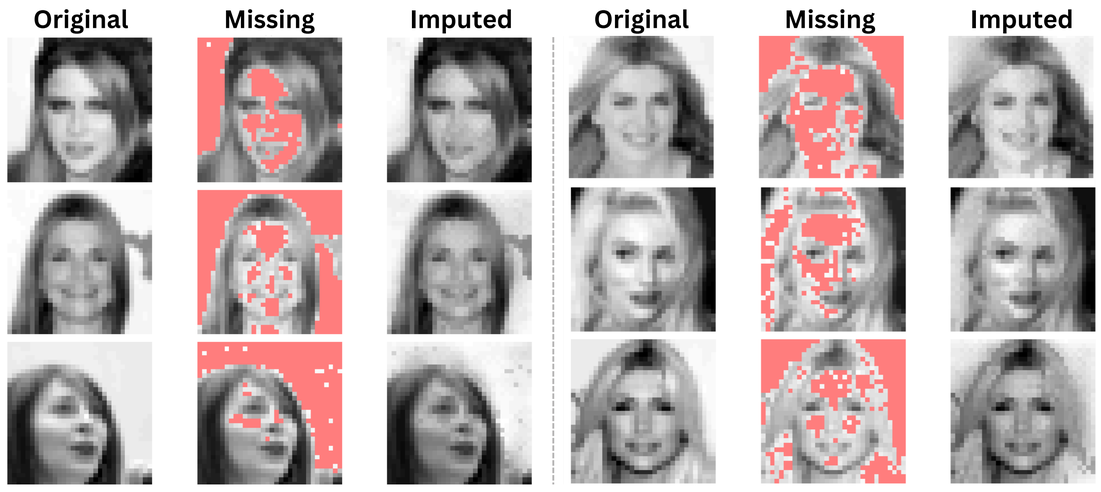
Distribution Recovery: Figure 1 illustrates this efficacy qualitatively. The model accurately predicts missing values and pushes the imputed distribution (orange) back to align with the original complete data (blue), effectively correcting the MNAR bias.

CelebA MNAR clipping

Self-masking censoring for pixels:

$$P(s_{ij} = 1 \mid x_{ij}) = \sigma(W(x_{ij} - b))\,, \quad W = -50,\ b = 0.75.$$

The model recovers missingness parameters and imputes clipped regions.



Supervised learning under MNAR

Model	Accuracy
Mean + LR	0.5405
MICE + LR	0.5425
Two-Stage (not-MIWAE + LR)	0.5445
Sup-not-MIWAE	**0.5690**
Oracle (complete data)	0.7123

Model	Test RMSE
Mean + Ridge	0.6966
MICE + Ridge	0.6713
Two-Stage (not-MIWAE + Ridge)	**0.6503**
Sup-not-MIWAE	0.6718
Oracle (complete data)	0.6173

Takeaway: For **classification** (Covertype), marginalization improves over naive baselines; for **regression** (Wine Quality), strong point imputations paired with ridge regression can be very competitive.

References

[1] Niels Bruun Ipsen, Pierre-Alexandre Mattei, and Jes Frellsen.
not-miwae: Deep generative modelling with missing not at random data, 2021.

[2] Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov.
Importance weighted autoencoders, 2016.

[3] Justin Domke and Daniel Sheldon.
Importance weighting and variational inference, 2018.

[4] Niels Bruun Ipsen, Pierre-Alexandre Mattei, and Jes Frellsen.
How to deal with missing data in supervised deep learning?
In *The First ICML Workshop on The Art of Learning with Missing Values (Artemiss)*, Vienna, Austria, July 2020.

[5] Pierre-Alexandre Mattei and Jes Frellsen.
Leveraging the exact likelihood of deep latent variable models, 2018.

[6] Geert Molenberghs, Caroline Beunckens, Cristina Sotto, and Michael Kenward.
Every missing not at random model has got a missing at random counterpart with equal fit.
Journal of the Royal Statistical Society Series B, 70:371–388, 04 2008.

[7] Gabriel Peyré and Marco Cuturi.
Computational optimal transport.
Foundations and Trends® in Machine Learning, 11(5-6):355–607, 2019.

[8] SeungHyun Kim, Hyunsu Kim, EungGu Yun, Hwangrae Lee, Jaehun Lee, and Juho Lee.
Probabilistic imputation for time-series classification with missing data, 2023.

Code: github.com/Adam-Ousse/notmiwae_pytorch **Package:** pypi.org/project/notmiwae-pytorch/