Deep Generative Modelling with Missing not at Random Data (not-MIWAE) [1]

Charles Dezons [1] Simon Blotas [1]

[1]Ecole Nationale des Ponts et Chaussées, ENPC

MATHÉMATIQUES VISION APPRENTISSAGE

Deep Generative Modeling with Missing Not at Random Data

Problem Statement: Traditional missing data approaches assume Missing At Random (MAR) mechanisms, where missingness depends only on observed data. However, real-world datasets often exhibit more complex Missing Not At Random (MNAR) patterns.

Notations and problem decomposition: We consider a dataset $\mathbf{X} = (x_1, \dots, x_n) \in \mathcal{X}^n$, where each x_i is a sample with features x_i^o (observed) and x_i^m (missing). Missingness is described by a mask $s \in \{0,1\}^p$, where $s_j = 1$ if feature j is observed, $s_j = 0$ if feature j is missing. The joint distribution $p_{\theta,\phi}(x,s)$ can be factorized as $p_{\theta,\phi}(x,s) = p_\theta(x)p_\phi(s|x)$, with three assumptions:

- **MCAR:** Missing completely at random: $p_\phi(s|x) = p_\phi(s)$
- **MAR:** Missing mechanism depending only on observed data: $p_\phi(s|x) = p_\phi(s|x^o)$
- **MNAR:** Missing mechanism dependent on both observed and missing data: $p_\phi(s|x)$ depends on x^o and x^m

IWAE Architecture and Inference in the MNAR Case

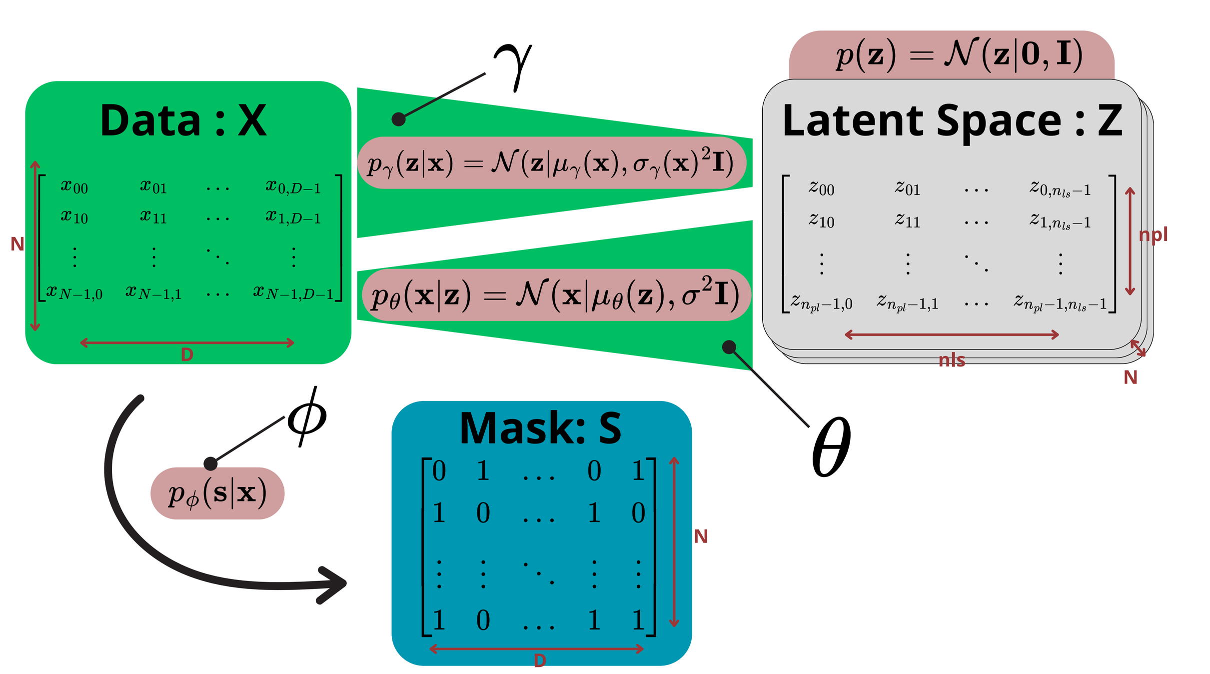


Figure 1. Schematic VAE with MNAR Missing Values

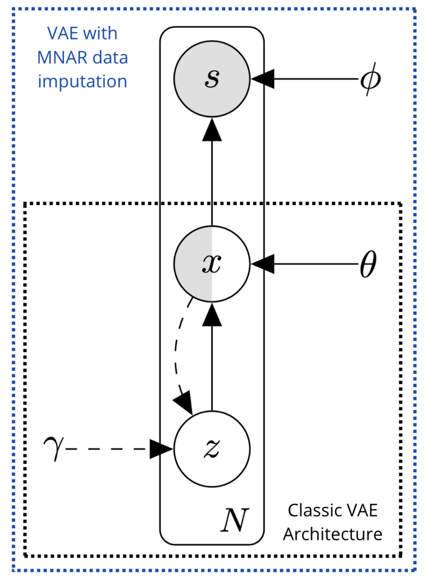


Figure 2. Probabilistic Graphical Model of not-MIWAE/VAE

MNAR case: We optimize both data-generation and missingness mechanisms by maximizing the joint log-likelihood:

$$\ell(\theta,\phi) = \sum_{i=1}^{n} \log p_{\theta,\phi}(x_i, s_i). \quad (1)$$

Direct maximum likelihood estimation is intractable due to missing and latent variables. Instead, we use a variational distribution $q_\gamma(z|x^o)$ to approximate a lower bound through importance sampling, similar to VAE and IWAE.

The contribution of data points is:

$$\log p_{\theta,\phi}(x^o, s) = \log \int p_\phi(s|x^o, x^m)\, p_\theta(x^o|z)\, p_\theta(x^m|z)\, p(z)\, dz\, dx^m \quad (2)$$

The single observation contribution is:

$$\log p_{\theta,\phi}(x^o, s) = \log \mathbb{E}_{z\sim q_\gamma(z|x^o), x^m \sim p_\theta(x^m|z)} \left[\frac{p_\phi(s|x^o, x^m)p_\theta(x^o|z)p(z)}{q_\gamma(z|x^o)}\right] \quad (3)$$

The objective is estimated using Monte Carlo sampling:

$$\mathcal{L}_K(\theta,\phi,\gamma) = \sum_{i=1}^{n} \mathbb{E}\left[\log \frac{1}{K}\sum_{k=1}^{K} w_{ki}\right] \quad (4)$$

where w_{ik} are the importance weights (see equation 6)

Once trained, the model can impute missing values by minimizing the squared error $L(x^m, \hat{x}^m)$. Optimal imputations minimize :

$$\mathbb{E}_{x^m}\left[L(x^m, \hat{x}^m)\middle|x^o, s\right]$$

resulting in:

$$\hat{x}^m = \sum_{k=1}^{K} \alpha_k \mathbb{E}[x^m|x^o, s], \quad \alpha_k = \frac{w_k}{\sum_{j=1}^{K} w_j} \quad (5)$$

The weights w_k match those used during training:

$$w_k = \frac{p_\phi(s|x^o, x_k^m)p_\theta(x^o|z_k)p(z_k)}{q_\gamma(z_k|x^o)}. \quad (6)$$

Gaussian Distribution

We generate a 2D Gaussian distribution centered at $(0,0)$, with covariance matrix $\mathbf{Cov}$.

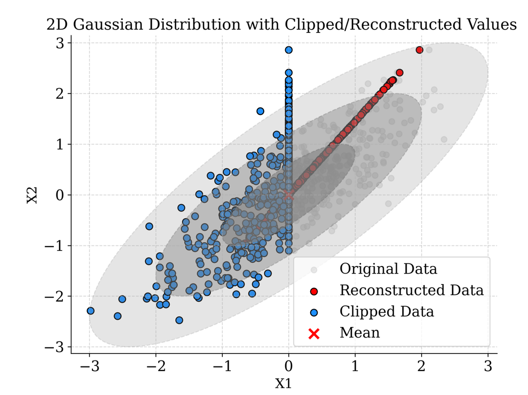
The missing values mask is defined as follows: for each sample, the x_1 component is masked if it exceeds the mean of all x_1 values in the dataset, which in this case is 0.

Ring-Shaped Distribution

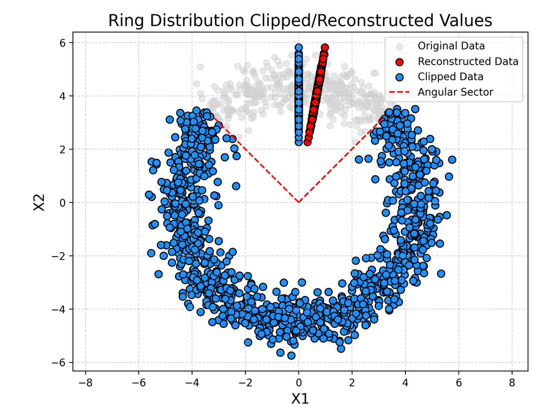
We generate a 2D ring distribution by sampling angles uniformly in $[0, 2\pi]$ and radius from a Gaussian distribution with mean $\mu = 4.5$ and variance σ^2.

The missing values mask is defined as follows: for each sample, the x_1 component is masked if the sample lies within a specific angular sector.

Reconstruction Quality



(a) Reconstructed data for Gaussian example using notMIWAE.



(b) Reconstructed data for ring example using notMIWAE.

Figure 3. Comparison of reconstructed data for Gaussian (left) and ring (right) examples using notMIWAE.

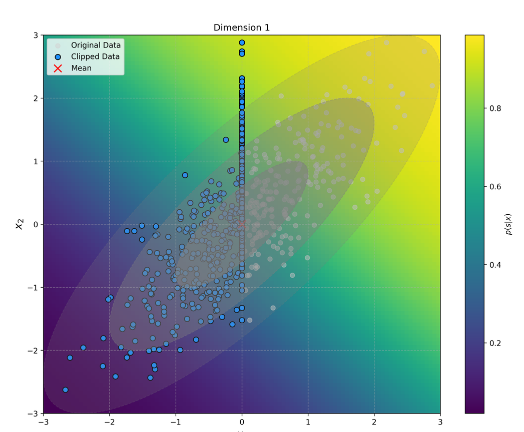
RMSE Performance

Dataset	Method	notMIWAE	MIWAE	KNN	MICE	Random Forest
Gaussian	RMSE	**1.152**	1.153	1.320	1.155	1.328
Ring	RMSE	**1.971**	1.977	3.046	1.975	4.034

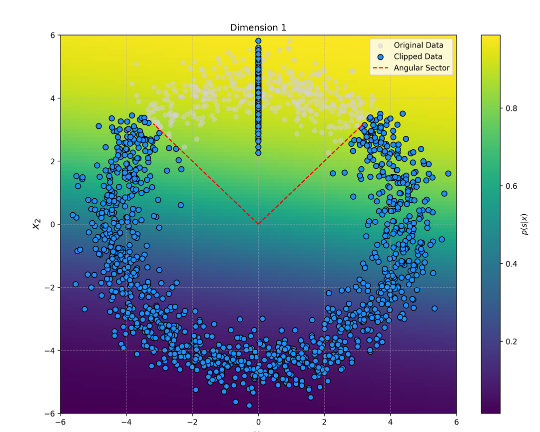
Table 1. RMSE comparison for Gaussian and ring datasets across different imputation methods.

Comparison of inferred distributions

Figure 4 illustrates the inferred conditional distributions $p(x_1 \text{ missing} \mid x)$ for both datasets, projected in 2D.



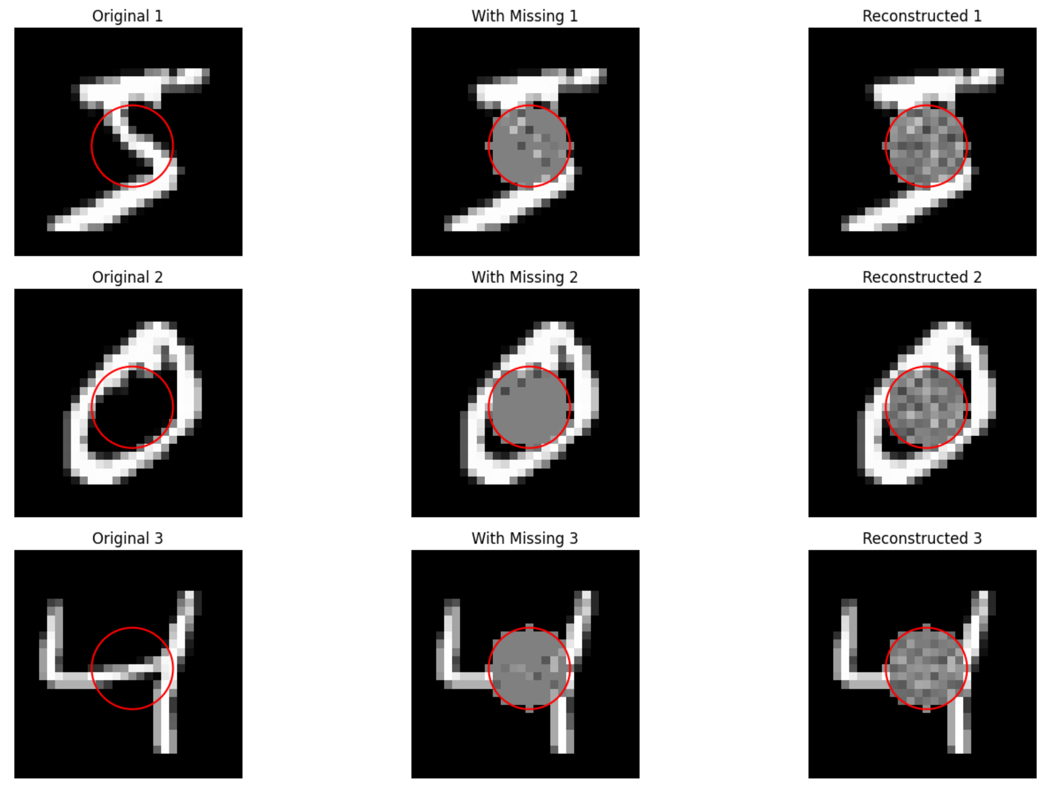
(a) Inferred $p(s|x)$ for Gaussian dataset.



(b) Inferred $p(s|x)$ for ring dataset.

Figure 4. Comparison of inferred $p(s|x)$ for Gaussian (left) and ring (right) datasets.

Experimentation with Real Dataset



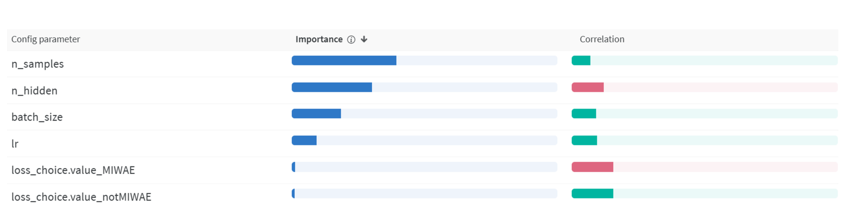
(a) Reconstructed MNIST example with imputed values.

Method	RMSE
notMIWAE	0.9301
MIWAE	0.9285
KNN	0.9353

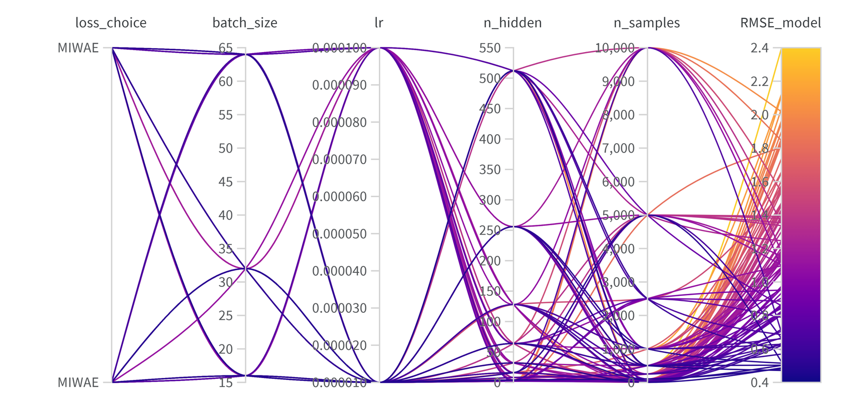
(b) RMSE comparison across imputation methods for MNIST dataset.

Figure 5. Visual and quantitative results of imputation methods. (a) Example of reconstructed MNIST data with missing values imputed. (b) RMSE comparison across different imputation methods.

Hyperparameter Optimization using W&B Sweep



(a) Correlation between hyperparameters and RMSE.



(b) Visualization of RMSE scores for different hyperparameter values.

Figure 6. Results from the W&B hyperparameter sweep.

Discussion

The notMIWAE model demonstrates robust RMSE performance, outperforming heuristic methods like KNN and Random Forest, particularly on complex datasets. Its probabilistic framework effectively utilizes observed data for imputation, but it struggles to fully capture non-linear structures and spatial dependencies, limiting its applicability to more intricate datasets. Enhancements such as deeper networks, convolutional layers, or tailored priors could address these shortcomings, enabling the model to better exploit its probabilistic design. Additionally, adopting domain-specific architectures, such as encoder-decoder frameworks for image data, may further improve performance. Despite these limitations, notMIWAE remains a competitive approach, with significant potential for optimization and broader applicability.

References

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