# Expanding Not-MIWAE : Experiments on the Not-MIWAE Model

Pierre Pauchet  Laura Choquet

MVA - ENS-Paris-Saclay

## Context

How to deal with real data set with missing values in machine learning? Our paper aims to impute missing data to enable the use of method using complete dataset.

The original paper introduces a new autoencoder model specialized in reconstructing data with a missing pattern dependent on the value of the missing data itself. This peculiar missing pattern is known as Not Missing At Random (MNAR), as opposed to Missing At Random (MAR) which may depends on the observed data but not the missing one.

## Background and Notations

Data Setting:

- Data matrix: $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^\top \in \mathcal{X}_n$ , $n$ i.i.d samples in $p$-dimensional space
- Missing values: $\mathbf{x}_i^m$  Observed values: $\mathbf{x}_i^o$
- Mask matrix: $\mathcal{S} \in \{0,1\}^{n,p}$ with $s_{ij} = 1$if $x_{ij}$ observed

Model (MNAR setting):

$$p_{\theta,\phi}(x, s) = p_\theta(x) p_\phi(s|x)$$

**Likelihood :** Marginalize over the missing data to compute the

$$p_{\theta,\phi}(\mathbf{x}^o, s) = \int p_\theta(\mathbf{x}^o, \mathbf{x}^m) p_\phi(s|\mathbf{x}^o, \mathbf{x}^m) d\mathbf{x}^m$$

## Not-MIWAE Model

**Objective:** Maximize joint log-likelihood with latent variable approach

**Initial log-likelihood:** with latent variable $z$

$$l(\theta, \phi) = \sum_{i=1}^{n} \log \int p_\phi(s|\mathbf{x}_i^o, \mathbf{x}^m) p_\theta(\mathbf{x}_i^o, \mathbf{x}^m|z) p(z) d\mathbf{x}^m dz$$

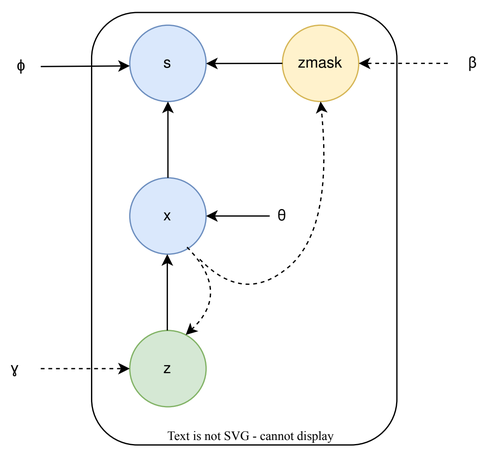
**Variational approximation:** *Using variational distribution* $q_\gamma(\mathbf{z}|\mathbf{x}^o)$ and IWAE lower bound.

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

where $z \sim q_\gamma(\mathbf{z}|\mathbf{x}^o), x^m \sim p_\theta(\mathbf{x}^m|z)$

**Optimal imputation:** $\hat{\mathbf{x}}^m$ minimizes $\mathbb{E}_{\mathbf{x}^m}[\mathcal{L}(\mathbf{x}^m, \hat{\mathbf{x}}^m)|\mathbf{x}^o, s]$. Compute using again self-normalised importance.

## Extended Model : Not-MIWAE with an additional latent variable




Model components:

- $\mathbf{x}$: data features
- $\mathbf{s}$: missingness mask
- $\mathbf{z}$: latent representation
- $\mathbf{z_{mask}}$: mask-specific latent variable

Dependencies:

- $\theta$: data generation $(\mathbf{x}|\mathbf{z})$
- $\phi$: missingness mechanism $(\mathbf{s}|\mathbf{x}, \mathbf{z_{mask}})$
- $\gamma, \beta$: variational parameters

**Additional latent variable:** $\mathbf{z}^{mask}$ for mask modeling

$$\mathbf{s} \sim p_\phi(\mathbf{s}|\mathbf{x}^o, \mathbf{x}^m, \mathbf{z}^{mask})$$

Extended likelihood:

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

Monte Carlo approximation:

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

$$\text{where } \omega_{ki} = \frac{p_\theta(\mathbf{x}_i^o|\mathbf{z}_{ki})p(\mathbf{z}_{ki})}{q_\gamma(\mathbf{z}_{ki}|\mathbf{x}_i^o)} \frac{p_\phi(s_i|\mathbf{x}_i^o, \mathbf{x}_{ki}^m, z_{ki}^{mask})p(z_{ki}^{mask})}{q_\beta(\mathbf{z}_{ki}^{mask}|\mathbf{x}_i^o)}$$

## Impact of Additional Mask Latent Variable

not-MIWAE Improvements:

- Linear setting:
  - not-MIWAE → not-MIWAE-mask :
  - White: 1.372 → **1.359**
  - Red: 1.321 → **1.250**

PPCA Improvements:

- Self-masking:
  - PPCA → PPCA-mask:
  - White: 1.002 → **0.996**
  - Red: 1.143 → **1.123**
- Known mask:
  - White: 0.998 → **0.973**
  - Red: 1.125 → **1.124**

The addition of $z_{\text{mask}}$ consistently improves RMSE performance across both datasets, with a notable **5.4% improvement** on red wine dataset.

PPCA with mask shows consistent enhancement, particularly strong in the known mask setting with a **2.5% improvement** on white win dataset.

The mask latent variable systematically improves or maintains performance across all experimental settings, suggesting its effectiveness in capturing missing data patterns.

## not-MIWAE used with supervised regression

Objective:

- Joint regressor-notMIWAE model
- MLP regressor with supervision
- Enhanced reconstruction via labels

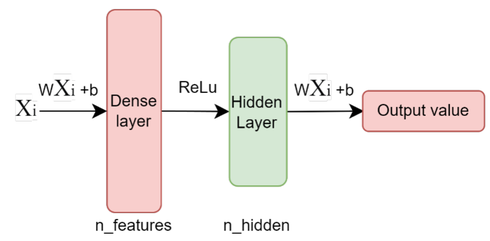
Joint Loss:

$$\log p_\Psi(y|x) + \mathbb{E}_{(\mathbf{z}_k)}\left[\log \frac{1}{K} \sum_{i=1}^{K} \frac{p_\phi(s|x)p_\theta(x|\mathbf{z}_k)p(\mathbf{z}_k)}{q_\gamma(\mathbf{z}_k|x)}\right]$$

Sum of regressor likelihood and IWAE bound

## Model Reg-Not-MIWAE

Architecture:




Model components:

- Input: features $X_i$
- Dense layer: $WX_i + b$
- Activation: ReLU
- Hidden size: n_hidden
- Output: scalar regression

Training:

- End-to-end joint optimization
- Combines imputation and prediction

## Regressor Performance Analysis

Experimental Setup:

- MLP regressor on red wine data
- Two training scenarios:
  - Frozen not-MIWAE weights
  - End-to-end joint training (Reg-not-MIWAE)

Key Findings:

- Better performance with frozen weights
- Joint training shows limitations
- Trade-off in shared parameters
- Constrained regressor learning

## Reg-not-MIWAE: Performance Analysis

Key Findings:

- Similar Test RMSE but decreased Train performance
- Comparable to baseline methods
- Limited by simplistic MLP architecture

| Metric | not-MIWAE | Reg-not-MIWAE |
|---|---|---|
| Train RMSE | 1.061 | 1.161 |
| Test RMSE | 0.866 | 0.888 |
| MICE | 1.681 | 1.683 |
| Mean | 1.838 | 1.841 |
| RF | 1.610 | 1.616 |

## Conclusions & Perspectives

Key Contributions:

- Successful reproduction of not-MIWAE results
- Enhanced model with mask latent variable
- Exploration of joint supervised learning

Main Findings:

- Improved performance with masked not-MIWAE
- Joint supervision shows limitations
- Trade-off between reconstruction and prediction

Future Work:

- Test on diverse datasets (MNIST, etc.)
- Explore complex regression architectures
- Investigate alternative joint learning approaches
- Study different MNAR missing patterns

This work demonstrates the potential for improving MNAR missing data handling, while highlighting the challenges in combining imputation with supervised learning.