Generative modelling with Missing not at Random Data: Combining DLVMs to enhance data imputation

Carlos CUEVAS VILLARMIN   Javier Alejandro LOPETEGUI GONZALEZ

ENS Paris-Saclay

Context and problem

Missing data is a common challenge in real-world problems, especially in fields like recommendation systems. Addressing it requires either adapting models to incomplete data or imputing missing values beforehand. Model performance depends heavily on *assumptions about the missing mechanism* and how accurately they reflect the true underlying process.

To address this problem, [1] proposes a solution for the specific case where the **missing data mechanism depends on the missing data themselves (MNAR)**. Specifically, their contribution is based on deep latent variable models (DLVMs), enabling flexible modeling of the conditional distribution of the missingness pattern given the data ($p(s|x)$ where s is the missingness pattern and x the data (with observed and unobserved features).

Our contribution follows this line of research. Specifically, we modify the proposed graphical model by **introducing a latent variable between the data samples and the mask**, which indicates observed and missing data. This approach assumes that the new latent space can better capture all the information needed to correctly identify complex missing data mechanism. As a result, the mechanism is expected to recover the mask depends directly from the new latent variable.

Material and methods

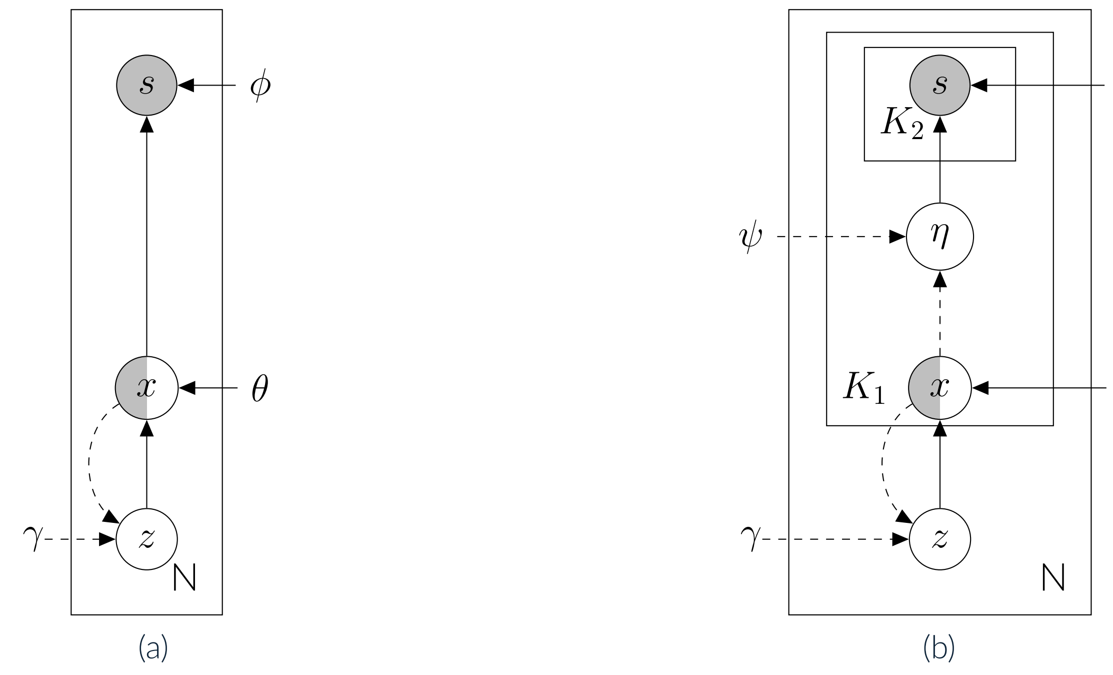

Figure 1. Graphical models of the not-MIWAE. (a) Proposed in [1]. (b) Our approach with two DLVMs.

Having the parametric model $p_{\theta,\phi}(x,s) = p_\theta(x)p_\phi(s|x)$ the general contribution of data points log$p_{\theta,\phi}(x,s)$ can be rewrite as

$$\log \int p_\phi(s|\eta)p(\eta)p_\theta(x^o|z)p_\theta(x^m|z)p(z)dzdx^m d\eta,$$

taking into account that $p(\eta|x) = p(\eta) = \mathcal{N}(\eta; 0, 1)$. The integral over missing values and latent variables is intractable. Therefore, two variational distributions are added $q_\gamma(z|x)$ and $q_\psi(\eta|x)$. So the contribution of a single observation is equal to

$$\begin{aligned} \log p_{\theta,\phi}(x,s) &= \log \int \frac{p_\phi(s|\eta)p(\eta)p_\theta(x^o)|z)p(z)}{q_\gamma(z|x^o)q_\psi(\eta|x)} q_\gamma(z|x^o)q_\psi(\eta|x)p_\theta(x^m|z)dx^m dz d\eta \qquad (1)\\ &= \log \mathbb{E}_{\substack{\eta\sim q_\psi(\eta|x)\\ z\sim q_\gamma(z|x^o)\\ x^m\sim p_\theta(x^m|z)}} \left[\frac{p_\phi(s|\eta)p(\eta)p_\theta(x^o|z)p(z)}{q_\gamma(z|x^o)q_\psi(\eta|x)}\right] \qquad (2)\end{aligned}$$

That can be translated into the objective function:

$$\begin{aligned} \mathcal{L}_{K_1,K_2}(\theta,\phi,\gamma,\psi) &= \sum_{i=1}^{N} \mathbb{E}\left[log \frac{1}{K_1K_2} \sum_{k_1=1}^{K_1} \sum_{k_2}^{K_2} w_{k_1,k_2,i}\right] \qquad (3)\\ &= \sum_{i=1}^{N} \mathbb{E}\left[log \frac{1}{K_1K_2} \sum_{k_1=1}^{K_1} \sum_{k_2}^{K_2} \frac{p_\phi(s_i|\eta_{k_1,k_2,i})p(\eta_{k_1,k_2,i})p_\theta(x_i^o|z_{k_1,i})p(z_{k_1,i})}{q_\gamma(z_{k_1,i}|x_i^o)q_\psi(\eta_{k_1,k_2,i}|x^m_{k_1,i}, x_i^o)}\right] \qquad (4)\end{aligned}$$

where $\{(z_{k_j,i}, x_{k_j,i})\}_{j=1,\ldots,K_1}$ are K_1 i.i.d. samples from $q_\gamma(x|x_i^o)$ and $p_\theta(x^m|z)$, and $\{(\eta_{k_j,k_l,i})\}_{l=1,\ldots,K_2}$ are K_2 i.i.d. samples from $q_\psi(\eta|x^m_{k_j,i}, x_i^o)$.

Following one of the experiments done in [1], we have selected four UCI datasets to test the quality of the proposed imputator. Concretely, the datasets are: Banknote authentication, Wine quality, Yeast and Breast cancer. The synthetic datasets with missing values are generated following the same missingness mechanism as in [1] (values above the feature mean are masked for half of the features).

We perform 5 runs for each model and dataset and report the RMSE mean and Standard Deviation. During training process the values of K_1 and K_2 are set to 20. In all the cases, we trained for 300 epochs. To calculate the RMSE we use 1000 importance samples in the imputator model. Notice that the masker model does not affect in this scenario. As baselines we used the three missing model approaches from [1] without PPCA. Based on their methodology we compare with mean and median imputation, missForest and MICE.

Results

	Banknote	*Wine*	*Yeast*	*Breast*
not-MIWAE				
agnostic	4.93 ± 2.03	1.86 ± 0.09	1.53 ± 0.03	1.41 ± 0.05
self-masking	1.29 ± 0.23	1.97 ± 0.27	$\mathbf{1.50 \pm 0.04}$	1.44 ± 0.05
self-masking known	1.61 ± 0.04	1.81 ± 0.04	1.52 ± 0.01	1.95 ± 0.02
missForest	1.28 ± 0.00	1.68 ± 0.00	1.72 ± 0.00	1.49 ± 0.00
MICE	1.41 ± 0.00	1.61 ± 0.00	1.76 ± 0.00	$\mathbf{0.95 \pm 0.00}$
mean	1.72 ± 0.00	1.81 ± 0.00	1.73 ± 0.00	1.82 ± 0.00
median	1.62 ± 0.00	1.76 ± 0.00	1.62 ± 0.00	1.79 ± 0.00
not-2-MIWAE (our)	$\mathbf{1.26 \pm 0.02}$	$\mathbf{1.59 \pm 0.01}$	1.72 ± 0.01	1.21 ± 0.02

Table 1. Imputation RMSE on UCI datasets affected by MNAR. Bolded numbers indicate the best performance for each dataset.

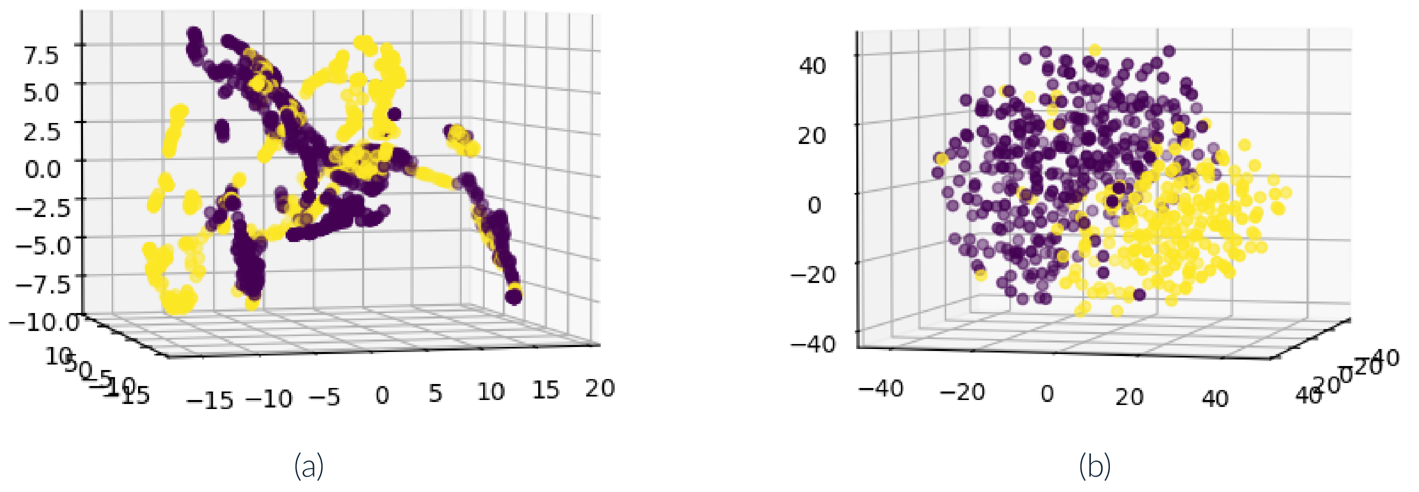

Figure 2. t-SNE in latent space z for Banknote (a) and Breast (b) datasets. Colored based on the labels y.

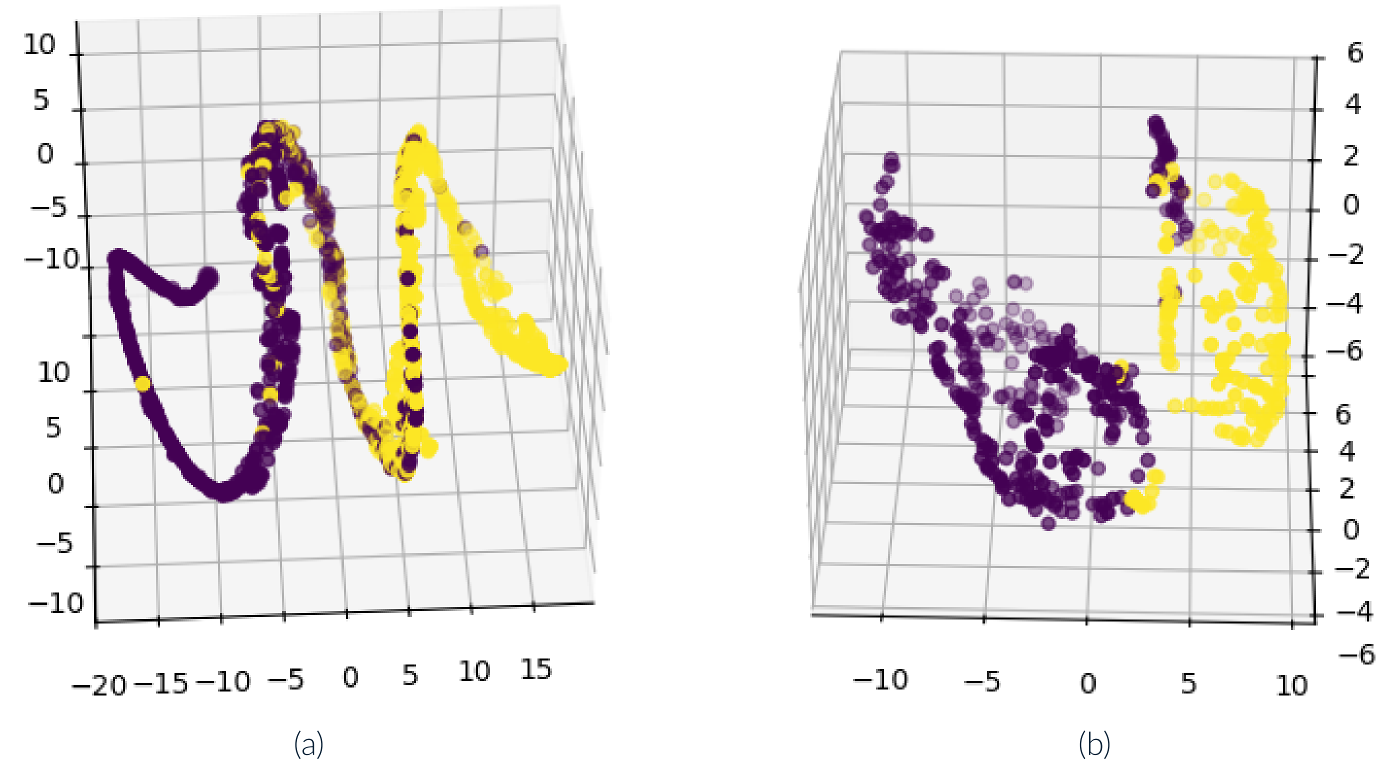

Figure 3. t-SNE in latent space η for Bankbone (a) and Breast (b) datasets. Colored based on k-means clustering done in masked vectors s.

Discussion and further work

- not-2-MIWAE outperforms for 3 out of 4 datasets the original not-MIWAE and it is the best imputation technique in Banknote and Wine quality datasets.
- Figure 2 shows that latent space z somehow makes a difference between labels to impute missing values. Furthermore, Figure 3 highlights the ability of the latent space η to capture relationships from the data to recover the missingness mechanism being able to distinguish each cluster.

It can be concluded that **each latent space seems to learn patterns that enhance the performance of the corresponding task**. Let this project serve as motivation to further work and do more specific experiments that will allow for greater explainability and confirm this hypothesis. For example, setting other prior distributions, test on real data, among others. Specifically, we consider that the **relevance of our contribution would be more notable when dealing with more complex missing patters where the MNAR mechanism does not relay just in the feature values**. Furthermore, more precise work should be done in the general pipeline in terms of hyperparameters optimization (number of epochs, number of importance samples or the layers in the model, etc.) which is left as further work too.

References

[1] Niels Bruun Ipsen, Pierre-Alexandre Mattei, and Jes Frellsen.
not-miwae: Deep generative modelling with missing not at random data.
arXiv preprint arXiv:2006.12871, 2020.