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

Carlos CUEVAS VILLARMIN
Javier Alejandro LOPETEGUI GONZALEZ
Master MVA, ENS Paris-Saclay

December 2024

1 Introduction

Missing data is a frequent issue in real-world problems, where large datasets often contain incomplete information. For example, user-item interaction matrices are typically sparse [7], as most users interact with only a small subset of available items. If not properly addressed, this incompleteness can significantly affect the performance of machine learning models by introducing biases or reducing predictive accuracy.

There are two main approaches to handling missing data: *adapting models to process incomplete input directly* or *imputing the missing values beforehand*. In the first approach, models are designed to explicitly account for missing entries during training or inference. This can include probabilistic frameworks [6] that estimate the likelihood of missing values based on observed data or neural network architectures that mask and ignore incomplete inputs. In the second approach, missing values are estimated or "imputed" before the data is used by the model. Imputation techniques can range from simple strategies, such as replacing missing values with the mean or median, to more advanced methods like multiple imputations, matrix factorization, or generative models such as Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs).

The performance of either approach depends on assumptions about the missing data mechanism. The mechanism describes why data is missing and is generally categorized into three types:

- *Missing Completely at Random (MCAR)*, where the missingness is independent of any data.
- *Missing at Random (MAR)*, where missingness depends on observed data.
- *Missing Not at Random (MNAR*), where missingness is related to the values that are missing.

Incorrect assumptions about the mechanism can lead to biased imputations or misinformed model updates, resulting in poor performance.

Ultimately, selecting an appropriate strategy to handle missing data requires understanding both the extent of the missingness and the nature of the missing mechanism. Models that effectively address these issues are better equipped to produce accurate predictions and reliable results, particularly in fields like recommendation systems, where missing data is common.

In this work, we focus on the case of Missing Not at Random (MNAR). Concretely, we will focus on the *not-missing-at-random importance-weighted autoencoder (not-MIWAE)* [3], which is inspired by [4]. Concretely, our goal is to continue the research done in [3] by combining a DLVM for data imputation (as it is done in [3]) with a DLVM to recover the missing pattern. With this variation, the missing mask s is not directly conditioned by the values of x but by a latent variable η, which we hypothesize learns deeper relationships of the data to recover the mask and therefore allows the model to learn more complex missing mechanisms.

The rest of the report consist on: a detailed presentation of not-MIWAE and our adaptation not-2-MIWAE in Section 2. Section 3 presents the pipeline followed for our experiments and the results are highlighted in 2. Finally, we will draw conclusions and discuss further work and weaknesses in Section 5.

2 not-MIWAE model and its variation

In this section, the mathematical foundations of not-MIWAE are presented and followed by the adaptation we have made to deal with the new latent variable. First of all, we set some notation that will be used in the section. Considered the data matrix $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T \in \mathcal{X}^n$, i.e., n i.i.d. copies of the random variable $\mathbf{x} \in \mathcal{X}$ being $\mathcal{X}$ a p-dimensional feature space. Taking into account that a sample $\mathbf{x}$ can have missing feature values it can be split into $\mathbf{x}_i = (\mathbf{x}_i^o, \mathbf{x}_i^m)$. The missing mechanism is defined by a mask matrix $\mathbf{S} = (\mathbf{s}_1, \ldots, \mathbf{s}_n)^T \in \{0,1\}^{n \times p}$ where $s_{ij} = 1$ if x_{ij} is observed and $s_{ij} = 0$ if x_{ij} is missing.

2.1 Single DLVM: not-MIWAE

The target parametric model is defined by $p_{\theta,\phi}(\mathbf{x}, \mathbf{s}) = p_\theta(\mathbf{x})p_\phi(\mathbf{s}|\mathbf{x})$. In this sense, the quantity of interest to maximize is the log-joint likelihood:

$$l(\theta, \phi) = \sum_{i=1}^{n} \log p_{\theta,\phi}(\mathbf{x}_i^o, \mathbf{s}_i). \tag{1}$$

The proposed model in [3] follows the graphical model that can be seen in Figure 1. The first part is a stochastic mapping parametrized by θ from a latent $\mathbf{z} \sim p(\mathbf{z})$ to the data $\mathbf{x} \sim p_\theta(\mathbf{x}|\mathbf{z})$. The second part is another stochastic mapping from the data to the missing mask $\mathbf{s} \sim p_\phi(\mathbf{s}|\mathbf{x})$. By adding the latent variable z and assuming that the observation model is fully factorized ($p_\theta(\mathbf{x}|\mathbf{z}) = \prod_j p_\theta(x_j|\mathbf{z})$), the contribution of a data point to the loss can be written as

$$\log \int p_\phi(\mathbf{s}|\mathbf{x}^o, \mathbf{x}^m)p_\theta(\mathbf{x}^o|\mathbf{z})p_\theta(\mathbf{x}^m|\mathbf{z})p(\mathbf{z})d\mathbf{z}d\mathbf{x}^m \tag{2}$$

Notice that this integral is intractable. A variational distribution $q_\gamma(\mathbf{z}|\mathbf{x}^o)$ is added to be learnable in an importance sampling scheme. Adding the variational distribution allows to transform the contribution into

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

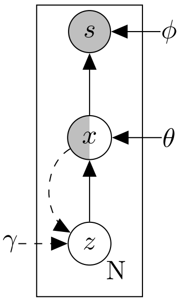


Figure 1: Graphical model of not-MIWAE.

The idea is to replace the expectation inside the logarithm by a Monte Carlo estimate of it [1]. This step will be directly present in the following section.

On the other hand, the missing process is basically a classification problem. The input in this case will be $\mathbf{x} = (\mathbf{x}^o, \hat{\mathbf{x}}^m)$ where $\hat{\mathbf{x}}^m$ is the output of the imputation model. The goal is to increase the ability of predicting the mask mechanism. Bernouilli distribution is considered for computing the probability of being masked or not,i.e., $p(\mathbf{s}|\mathbf{x}^o, \hat{\mathbf{x}}^m) = \text{Bern}(\mathbf{s}; \pi_\theta(\mathbf{x}))$. This probability is estimated with several approaches: *agnostic, self-masking* or *self-masking known.*

2.2 Double DLVM: not-2-MIWAE

In our approach, we only change the missing model. Instead of considering a Bernoulli distribution over the logits given by $\pi_\theta(\mathbf{x})$ we consider a DLVM for this part too. We assume a latent variable η such that $p(\eta|\mathbf{x}) = p(\eta) = \mathcal{N}(\eta; 0, I)$. The corresponding graphical model can be seen in Figure 2.

Including this latent variable the contribution of a data point to the loss can be written as

$$\log \int p_\phi(\mathbf{s}|\eta)p(\eta)p_\theta(\mathbf{x}^o|\mathbf{z})p_\theta(\mathbf{x}^m|\mathbf{z})p(\mathbf{z})d\mathbf{z}d\mathbf{x}^m d\eta \tag{4}$$

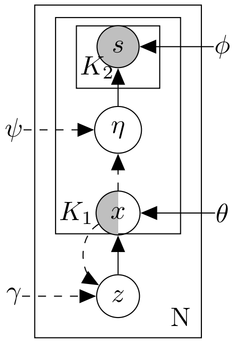


Figure 2: Graphical model of not-2-MIWAE.

Adding a new variational distribution $q_\psi(\eta|\mathbf{x})$ we have

$$\log p_{\theta,\phi}(\mathbf{x},\mathbf{s}) = \log \int \frac{p_\phi(\mathbf{s}|\eta)p(\eta)p_\theta(\mathbf{x}^o|\mathbf{z})p(\mathbf{z})}{q_\gamma(\mathbf{z}|\mathbf{x}^o)q_\psi(\eta|\mathbf{x})}\, q_\gamma(\mathbf{z}|\mathbf{x}^o)q_\psi(\eta|\mathbf{x})p_\theta(\mathbf{x}^m|\mathbf{z})\, d\mathbf{x}^m d\mathbf{z} d\eta \tag{5}$$

$$= \log E_{\substack{\eta\sim q_\psi(\eta|\mathbf{x})\\ \mathbf{z}\sim q_\gamma(\mathbf{z}|\mathbf{x}^o)\\ \mathbf{x}^m\sim p_\theta(\mathbf{x}^m|\mathbf{z})}} \left[\frac{p_\phi(\mathbf{s}|\eta)p(\eta)p_\theta(\mathbf{x}^o|\mathbf{z})p(\mathbf{z})}{q_\gamma(\mathbf{z}|\mathbf{x}^o)q_\psi(\eta|\mathbf{x})}\right]. \tag{6}$$

That can be translated, by a Monte Carlo sampling in both latent spaces [1], into the objective function

$$\mathcal{L}_{K_1,K_2}(\theta,\phi,\gamma,\psi) = \sum_{i=1}^{N} E\left[log \frac{1}{K_1K_2}\sum_{k_1=1}^{K_1}\sum_{k_2}^{K_2} w_{k_1,k_2,i}\right] \tag{7}$$

$$= \sum_{i=1}^{N} E\left[log \frac{1}{K_1K_2}\sum_{k_1=1}^{K_1}\sum_{k_2}^{K_2} \frac{p_\phi(\mathbf{s}_i|\eta_{k_1,k_2,i})p(\eta_{k_1,k_2,i})p_\theta(\mathbf{x}_i^o|\mathbf{z}_{k_1,i})p(\mathbf{z}_{k_1,i})}{q_\gamma(\mathbf{z}_{k_1,i}|\mathbf{x}_i^o)q_\psi(\eta_{k_1,k_2,i}|\mathbf{x}_{k_1,i}^m,\mathbf{x}_i^o)}\right] \tag{8}$$

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

Based on the unbiasedness of the Monte Carlo estimates, the objective is a lower-bound of the likelihood and it is proved that $(\mathcal{L}_K(\theta,\phi,\gamma))_{K\leq 1}$ converges monotonically to the likelihood [1] which can be extended to $(\mathcal{L}_{K_1,K_2}(\theta,\phi,\gamma))_{K_1.K_2\leq 1}$.

Notice that by adding a new latent variable we are integrating also with respect to this variable together with the other latent and part of the data (missing data). As in [3] we implement the reparametrization trick for $q_\gamma(\mathbf{z}|\mathbf{x}_i^o)$ and $p_\theta(\mathbf{x}_i^m|\mathbf{z})$ but we also do it for $q_\psi(\eta|\mathbf{x}_i^o,\hat{\mathbf{x}}_i^m)$. A mask $\mathbf{s}$ is generated for each η sample.

3 Material and Experimental set-up

Once the mathematical foundations of the models we use are defined, in this section we will provide more detailed information about the data used in the experiments, as well as the hyperparameters defined for each model and the metric selected to analyze the model's performance.

Due to the scope and variety of the experiments presented in [3], in this work we focus exclusively on the first of them. In this way, we aim for a better understanding of the procedure to follow, concentrating our efforts on a single objective to gain useful insights, while keeping in mind the academic scope of a final project of this magnitude.

3.1 Datasets

As said, following the first set of experiments reported in [3] we are using four datasets from the UCI database[1] presented in [2]. From this collection, we focus on the Banknote authentication, Wine quality, Yeast and Breast cancer. The details about each datasets are in Table 1. For each of the datasets a MNAR missing process is applied following the same frameworks as in [3]. For half of the features, the values are set to missing if they are greater than the feature's mean. Moreover, following the criteria that in real world scenarios it is not possible to assess the imputation error during training process, we use as validation set the same training set and do not apply any early stopping strategy.

3.2 Models setting

For all the experiments we handle the input observed values variability using the zero-imputation approach, a technique previously used in several works such as [5], [4] and [3] and several hyperparameters are fixed: `batch_size` is set to 16, `learning_rate` is equal to $1e-3$ and the number of epochs is fixed to 300. In [3] the number of epochs is considerably higher, we have decided to reduce it in order to make the experiments more manageable in terms of computationally cost even if doing this means giving up reproducing the results of [3]. As said, our goal is to gain insights about DLVMs and train our creativity to contribute to the community under our possibilities.

[1] link to the UCI database: UCI

	id	*num_examples*	*num_features*
Banknote	267	1372	4
Wine	186	4898	11
Yeast	110	1484	8
Breast	17	569	30

Table 1: Description of the datasets used in the experiments.

3.2.1 Baselines: not-MIWAE

As the main baseline for our proposed modification we use the not-MIWAE model proposed in [3]. Particularly, we focus on the no-PPCA variant following the argument given in [3] about the high inductive bias level in the data model for the PPCA like model. Thus, we compared against the more flexible, still efficient, not-MIWAE.

The encoder and decoder in the implemented version have 2 hidden layers of 128 units each and *tanh* activation function. Moreover, the latent space dimenssion is fixed as $p-1$ being p the number of features in the input data. Furthermore, for the importance sampling we used $K = 20$. The main information about this parameters values is in table 2.

For the Imputation Decoder, we use the three approaches reported in [3]:

- *agnostic*: a single fully connected Linear layer which outputs are going to be the logits for the Bernoulli distribution
- *self-masking:* logistic regression is applied to each feature
- *self-masking known:* the weights in the logistic regression are forced to be positive.

Following the idea from the experiments in [3], we also compare against mean and median imputation, missForest [8] and MICE [9].

3.2.2 not-2-MIWAE

The first DLVM for our proposal, corresponding to the data model, will have exactly the same architecture as the one in not-MIWAE, as you can notice in the Table 2. The main difference is that now, as explained before, instead of having a single Imputation Decoder, we introduce a new DLVM. For a first experimental approach its main architecture design is similar to the one for data generation. The exact parameters values can be seen in Table 2. Moreover, now, instead of having logits as output from the decoder, we will have the parameters of the s generative distribution.

-	*not-MIWAE*	*not-2-MIWAE*
K_1	20	20
K_2	-	20
h_layers	2	2
h_size	128	128
z_dim	$p-1$	$p-1$
η_dim	-	$p-1$

Table 2: Main hyperparameters for not-MIWAE and not-2-MIWAE

3.3 Evaluation Metric

Considering that in our case we are dealing with synthetically generated missing data, we are able to evaluate the imputation error. Particularly, we use the imputation RMSE following the same experimental setting as [3]. In all the cases, we use $1K$ importance samples.

4 Experiments

After detailing all the information of the datasets used, the baseline model not-MIWAE [3] and the new model not-2-MIWAE in the previous sections along with the set-up for each case. This section provides the outputs obtained in the experiment. Specifically, the RMSE in the imputation task are summarized in Table 3.

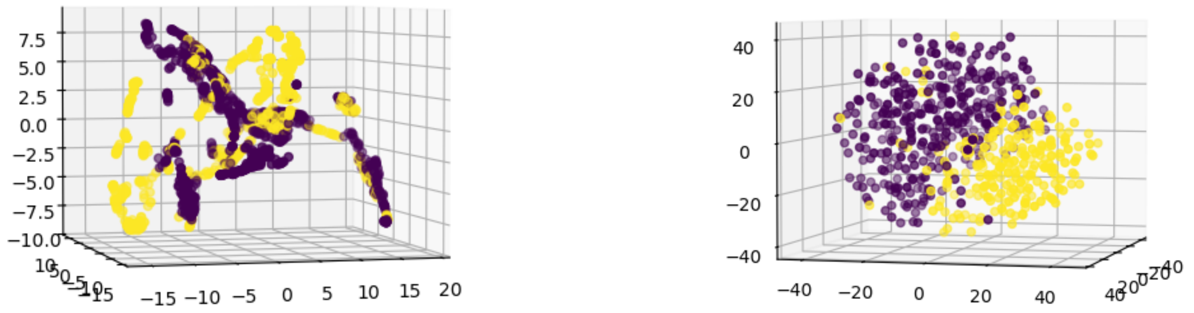
	Banknote	*Wine*	*Yeast*	*Breast*
not-MIWAE				
agnostic	4.93 ± 2.03	1.86 ± 0.09	1.53 ± 0.03	1.41 ± 0.05
self-masking	1.29 ± 0.23	1.97 ± 0.27	$\mathbf{1.50 \pm 0.04}$	1.44 ± 0.05
self-masking known	1.61 ± 0.04	1.81 ± 0.04	1.52 ± 0.01	1.95 ± 0.02
missForest	1.28 ± 0.00	1.68 ± 0.00	1.72 ± 0.00	1.49 ± 0.00
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 3: Imputation RMSE on UCI datasets affected by MNAR. Bolded numbers indicate the best performance for each dataset.

Based on the evaluation metrics obtained, it cannot be guaranteed that one of the models consistently outperforms the others in a general scenario. not-MIWAE provides the best results on the *Yeast* dataset, while not-2-MIWAE outperforms the baselines on the *Banknote* and *Wine Quality* datasets. As can be observed, each dataset has its own particularities fitting with a different imputation technique, which shows the difficulty of generalizing the imputation task to any given dataset. However, we want to highlight that not-2-MIWAE, being a more complex model, generally reports better results across all scenarios. Further experiments should be conducted to confirm this hypothesis, but the results presented in Figure 3 provide strong motivation to explore this direction further.

Following this idea, we were interested in understanding the type of information captured by the latent spaces, specifically in relation to the labels y of the models and the masking clusters (defined using K-Means on the mask matrix $\mathbf{S}$). More precisely, our goal is to determine whether the latent variables learn features that are useful for the masking model to recover the missing pattern distinguishing between masking clusters and for the imputation model to generate missing values consistent with the label to which each sample belongs.

Figure 3 illustrates the μ parameters learned in the latent space for the Banknote and Breast datasets through t-SNE. It can be observed that these values, particularly in the case of the Breast dataset, are related to the label to which the corresponding samples belong because both labels can be differenciated. Figure 6a further highlights that there is potential for improvement with more sophisticated experiments, although the values already appear to follow a discernible pattern.

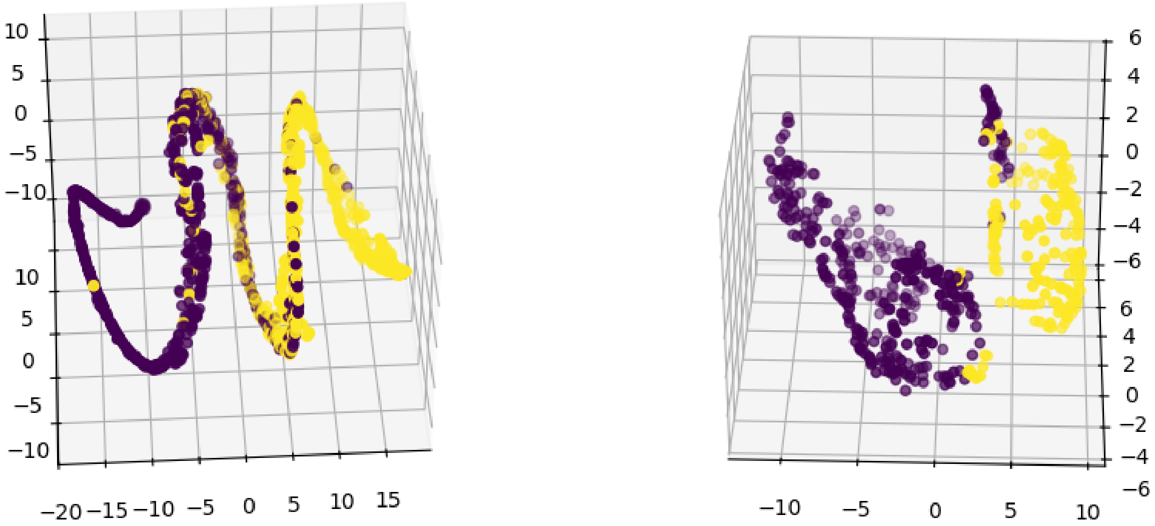


(a)

(b)

Figure 3: t-SNE in latent space z for Banknote (a) and Breast (b) datasets. Colored based on the labels y from the original dataset.

The same analysis using the latent variable η is presented in Figure 4. In this case, the colors indicate the cluster to which each mask vector ss belongs. Specifically, in Figure 5a for the Banknote dataset and in Figure 5b for the Wine Quality dataset, it can be observed that the μ values clearly depend on the cluster, successfully differentiating the clusters in most cases. This result demonstrates that the proposed DLVM for the masking model effectively learns patterns between features that are relevant and representative for this task. However, this insights cannot be infered to all datasets as seen in Figure 5 and Figure 6 and detailed further work should be done to increase the generality of the approach.



(a)

(b)

Figure 4: t-SNE in latent space η for Bankbone (a) and Breast (b) datasets. Colored based on k-means clustering done in masked vectors s.

5 Conclusion

It can be concluded that **each latent space sometimes appears to learn meaningful patterns that enhance the performance of the corresponding task**. These results highlight the potential of the proposed approach and should serve as motivation for further research aimed at gaining deeper insights and improving the methodology. More specific experiments could be conducted to achieve greater explainability and to confirm the hypotheses presented here. For instance, exploring alternative prior distributions for the latent variables could provide new perspectives on how the model captures underlying structures in the data. Additionally, testing the framework on real-world datasets, which inherently contain more noise and variability, would validate its robustness and generalizability.

A **particularly promising direction would involve addressing more complex missing data patterns**, where the Missing Not at Random (MNAR) mechanism does not solely rely on the feature values but may depend on unobserved factors or latent relationships within the data. These cases are more challenging but also more representative of real-world scenarios, making the contribution of this work even more significant.

Furthermore, improvements in the overall pipeline could yield better results and greater reliability. In particular, a more rigorous optimization of hyperparameters—such as the number of training epochs, the number of importance samples, the learning rate, and the depth or width of the neural network layers—should be pursued. These aspects were not the primary focus of this project and are therefore left as future work. Fine-tuning these parameters would likely lead to further improvements in model performance and a more efficient training process.

In summary, **this project lays the foundation for future studies to explore the interplay between latent space learning, missing data mechanisms, and model optimization**. Addressing these aspects systematically will not only enhance the explainability of the learned latent representations but also ensure that the proposed approach can tackle more realistic and complex scenarios where missing data mechanisms are inherently complicated.

6 Contributions

Both members of the team have worked or surpevised directly the full pipeline of the project. Javier has focused on transforming the code from TensorFlow to PyTorch and also has redone the chosen experiments of [3] while Carlos worked on the not-2-MIWAE approach mathematically and its code, which as said was discussed and fixed jointly by both members. The code of the project can be found on GitHub.

References

[1] Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. Importance weighted autoencoders. *arXiv preprint arXiv:1509.00519*, 2015.

[2] Dheeru Dua, Casey Graff, et al. Uci machine learning repository. 2017.

[3] 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.

[4] Pierre-Alexandre Mattei and Jes Frellsen. Miwae: Deep generative modelling and imputation of incomplete data sets. In *International conference on machine learning*, pages 4413–4423. PMLR, 2019.

[5] Alfredo Nazabal, Pablo M Olmos, Zoubin Ghahramani, and Isabel Valera. Handling incomplete heterogeneous data using vaes. *Pattern Recognition*, 107:107501, 2020.

[6] Donald B Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

[7] Yue Shi, Martha Larson, and Alan Hanjalic. Collaborative filtering beyond the user-item matrix: A survey of the state of the art and future challenges. *ACM Computing Surveys (CSUR)*, 47(1):1–45, 2014.

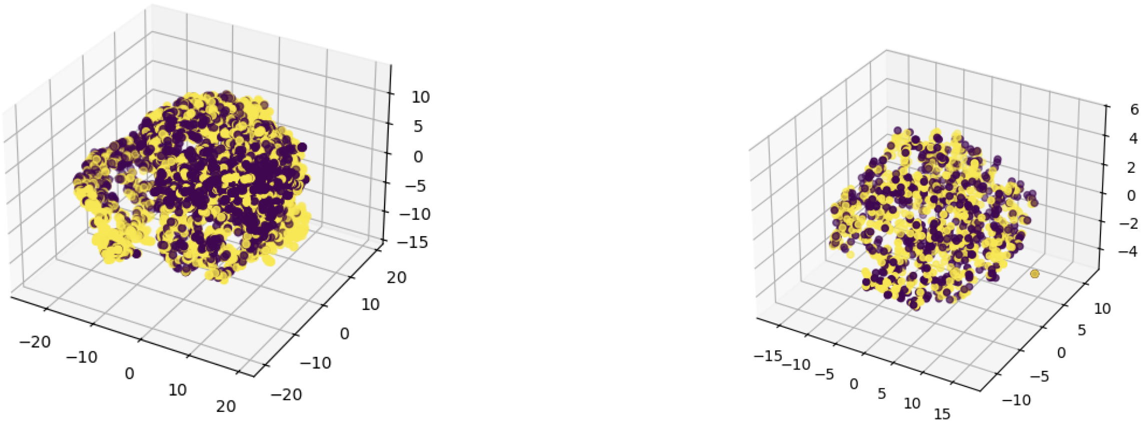
[8] Daniel J Stekhoven and Peter Bühlmann. Missforest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118, 2012.

[9] Stef Van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of statistical software*, 45:1–67, 2011.

A Latent spaces analysis for the rest of the datasets

Here we show the latent spaces in the other two datasets to show that the results are not good for all the datasets. We have to remark that in these cases the number of labels is higher than 2, which makes more difficult to differenciate them visually.

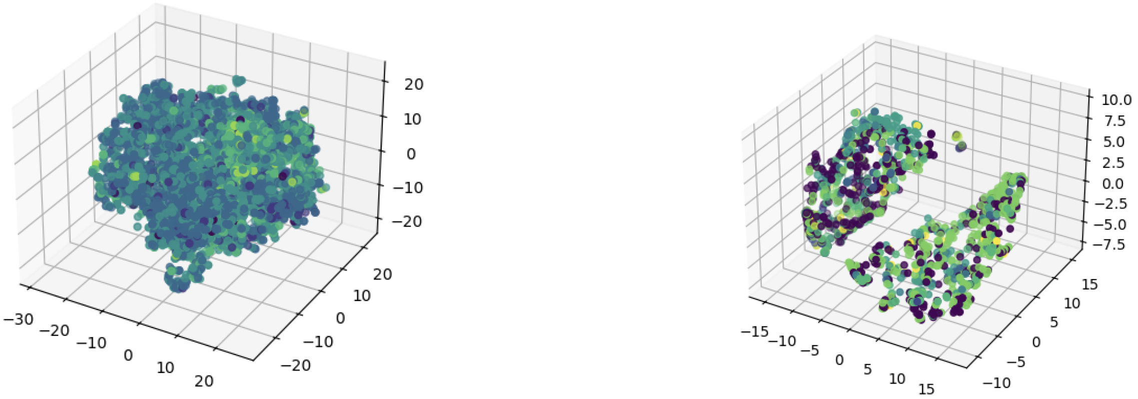
Figure 5 shows the latent space η where 2 clusters are considered as in the previous figures for the masking vectors. In this case the latent variable does not distinguish between the clusters. Figure 6 shows the same but 6b is special because two different groups can be differenciated but does not have relation between the labels but looking to Figure 7 it seems that the the latent variables are also affected by the missing pattern and it could be what the model discriminates in this case although the classes are not separable at all in each set one cluster dominates the other.



(a)

(b)

Figure 5: t-SNE in latent space η for Wine (a) and Yeast (b) datasets. Colored based on k-means clustering done in masked vectors s.



(a)

(b)

Figure 6: t-SNE in latent space z for Wine (a) and Yeast (b) datasets. Colored based on the labels y from the original dataset.

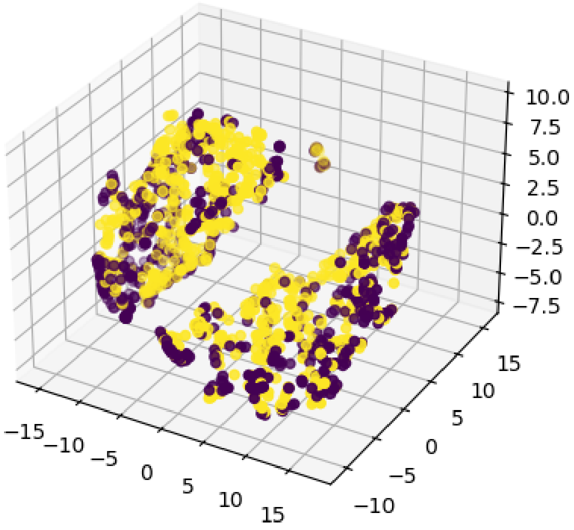


Figure 7: t-SNE in latent space z for Yeast dataset. Colored based on k-means clustering done in masked vectors s.