Project report: *not-MIWAE*: Deep Generative Modelling with Missing not at Random Data

Mohammad Ali JAUHAR
ENS Paris-Saclay
jauhar@ens-paris-saclay.fr

Antoine OLIVIER
Ecole des Ponts
antoine.olivier@eleves.enpc.fr

Joachim COLLIN
Ecole des Ponts
joachim.collin@eleves.enpc.fr

Abstract

Addressing missing data in datasets can prove crucial for robust predictive modeling. In "*not-MIWAE: Deep Generative modelling with missing not at random data*", Ipsen et al. introduces a method to building and training deep latent variable models (DLVMs) for scenarios where the missing process is related to the missing data themselves. In this report, we summarize *notMIWAE*, replicate the results, explore various missing patterns, and compare with other imputation methods on a number of datasets of various size to assess the strengths and weakenesses.

1 Introduction

In datasets, missing data are frequent occurences that analysts often need to adress. Missingness can stem from various sources; sensor sensitivity limits or participants opting not to answer specific survey questions are simple yet common cases. Adressing missing data can prove essential to ensure the quality of the predictive modeling. Different methods have been developed to handle these missing data, ranging from basic approaches like median or K-Nearest Neighbors to more sophisticated techniques such as multiple imputation methods Murray (2018). The accuracy of a model's predictions relies heavily on how the missing process was adressed and what the true nature of the missing mechanism is.

In 1976, Rubin Rubin (1976) categorizes different mechanisms behind missing data. If data are *missing completely at random* (MCAR), it implies that the missing pattern is entirely random and unrelated to either the missing or observed values. If data are *missing at random* (MAR), it indicates that missingness is connected solely to the observed data and is random conditionned on them. When the missing pattern relies both observed and missing data, it is known as *missing not at random* (MNAR) data. Addressing this dependency becomes crucial to avoid biased predictions. For instance, Marlin et al. **?** revealed that people tend to rate preferred items more frequently, resulting in missing patterns that prove highly detrimental to simple recommender systems assuming MAR or MCAR.

In the context of using deep latent variable models (DLVMs) to handle missing data scenarios, Ipsen et al. introduced the *not-missing-at-random importance-weighted autoencoder (not-MIWAE)* as a solution for MNAR cases. This model builds upon prior work that adapted the importance-weighted autoencoder (IWAE) Yuri Burda (2016) into the missing data importance-weighted autoencoder (MIVAE) Mattei and Frellsen (2019). *notMIWAE*, described in 1a, comprises two stochastic components: one mapping latent variables to observed and unobserved data, and the other predicting the missing pattern from the complete data. The model is then trained by maximizing a lower bound of its likelihood using a reparameterisation trick.

Preprint. Under review.

2 Related work and Background

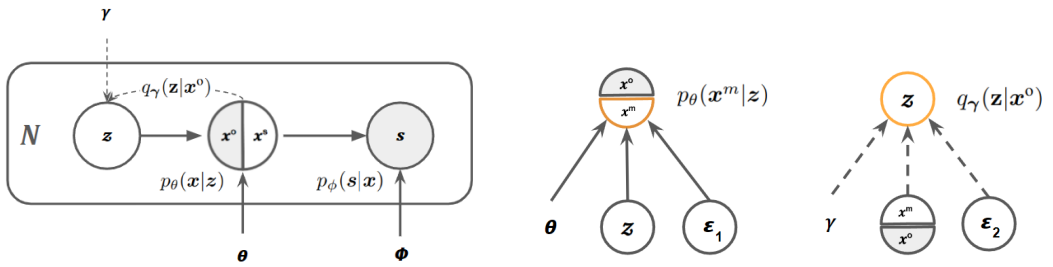
2.1 Related work

Rubin (1976) and Little and Rubin (2002) formulates the appropriateness of ignoring missing processes. Specifically, it shows that when data are MCAR or MAR, valid inferences can be obtained by ignoring the missing-data mechanism. Most imputation algorithms, therefore, have MCAR or MAR assumptions. Ma et al. (2019) and Mattei and Frellsen (2019) are some recent methods employing MCAR/MAR assumptions.

MNAR data lead to selection bias as the observed data are not the representative of the population. Hence, the missing data pattern must be taken into account. (Little and Rubin (2002)) provide an extensive framework for handling MNAR data. Among recent methods, Sportisse et al. (2020) propose matrix completion method based on low-rank assumptions for missing data imputation. It models the joint distribution of data and missing mask using an EM algorithm. This method encodes the latent representation of both data and the mask. Ghalebikesabi et al. (2021) propose a deep variational auto-encoder based model using a semi-supervised technique based on pattern-set mixture Little (1993) to tackle ignorable (MAR/MCAR) and non-ignorable (MNAR) data.

Deep latent variable models (DLVMs) have caught community attention as the generative part of the model can be used to sample the missing part of an observation. Methods on VAEs (A. Nazabal (2020), Mattei and Frellsen (2019))have been proposed to take into account he missing values. Partial-VAE by Ma et al. (2019) use the conditional independence of unobserved data on observed data to formulate the VAE lower bound based only on the observed data along in a permutation invariant encoder. Yoon et al. (2018) used GANs for missing data imputation.

2.2 Background



(a) **not-MIWAE: Graphical model** (b) **Reparameterisation trick**

Figure 1: Graphical Representation of (a) not-MIWAE and (b) the Reparameterization Trick. Observed data depicted in grey, unobserved data in white, and deterministic nodes highlighted in orange. In (b), the introduction of new random variables, denoted as ϵ_1 and ϵ_2, is used in the reparameterization trick to remove the randomness "inside" the model by treating it as inputs.

Let $(x_1, \ldots, x_n) \in \mathcal{X}^n$ represent n independent and identically distributed (iid) random variables representing data points, where $\mathcal{X}$ is a p-dimensional feature space. Following Ipsen et al. 's convention, x is used to denote x_i to enhance readability. As x comprises both observed (x^o) and missing (x^m) data, it can be represented as $x = (x^o, x^m)$. The missing pattern s of x, denoted as $s \in \{0, 1\}^p$, is a random variable where $s_j = 1$ when x_j is observed and $s_j = 0$ when x_j is missing.

In MNAR, the aim is to optimize the likelihood of both observing the observed and not observing the missing data. Considering this, Ipsen et al. introduced a factorizable parametric model, presented in 1a, such that $p_{\theta,\phi}(x, s) = p_\theta(x) \cdot p_\phi(s|x)$. In MCAR and MAR scenarios, $p_\phi(s|x)$ can be simplified as $p_\phi(s)$ and $p_\phi(s|x_o)$, disentangling the missing pattern from the latent variables. Under the assumtion that $p_\theta(s|x)$ is fully factorized, the log-likelihood of the model is:

$$l(\theta, \phi) = \sum_{i=1}^{n} \log p_{\theta,\phi}(x_i^0, s_i) = \sum_{i=1}^{n} \log \int p_\phi(s_i|x_i^o, x^m)\, p_\theta(x_i^o|z)\, p_\theta(x_i^m|z)\, p(z)\, dz\, dx^m \quad (1)$$

3 Method

The integral over the latent variable in 1 makes the likelihood intractable . Based on Yuri Burda (2016), Ipsen et al. introduce the *variational distribution* $q_\gamma(z|x^o)$, a conditional distribution that will be used as learnable proposal for the importance sampling technique.

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

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

The *variational distribution* parameters are computed by a neural network using x^o as input, where the length of x^o varies based on the presence of missing data. However, neural networks are not inherently designed to handle variable-length data. While R. Vedantam (2018) extended Gaussian distribution properties to non-linear models, this approach suffers from overfitting, high computational costs, and lacks parameter inference amortization. A. Nazabal (2020) addressed this by employing zero-imputation, leveraging the fact that zero inputs have a null impact on the encoder output and its derivative. However, differentiating real zero inputs from zero imputations can prove to be important to enhance uncertainty estimation during partial inference. Thus, C. Ma (2018) proposed a generalization of zero-imputation from a Point Net perspective, utilizing a permutation-invariant set function encoding followed by an amortized inference network. This methods allows to take into account only real zero values while maintaining computational efficiency and enabling amortization. In their code, Ipsen et al. implemented both zero imputations and a permutation-invariant network.

Using importance sampling allowed to get an estimate the expectation with a smaller variance than when using the Monte-Carlo method directly. To get the estimator of 3, we sample from $q_\gamma(z|x^o)$ and $p_\theta(x^m|z)$. Jensen's Inequality and un the unbiaisness of the Monte-Carlo method allow us to derive tracktable lower-bound of the likelihood $\mathcal{L}_K(\theta, \phi, \mu)$.

$$l(\theta, \phi) \geq \sum_{i=1}^{n} \mathbb{E}_{z\sim q_\gamma(z|x_i^o),\, x^m\sim p_\theta(x^m|z)} \left[\log \frac{1}{K} \sum_{k=1}^{K} w_{ki} \right] = \mathcal{L}_K(\theta, \phi, \mu) \tag{4}$$

$$\text{with } w_{ki} = \frac{p_\phi(s_{ki}|x_i^o, x_{ki}^m)\, p_\theta(x_i^o|z_{ki})\, p(z_{ki})}{q_\gamma(z_{ki}|x_i^o)} \tag{5}$$

Through Jensen's Inequality and the Law of Large Numbers, this lower bound demonstrates favorable properties, converging to the true likelihood as K increases with a rate of $\frac{1}{K}$:

$$\mathcal{L}_1(\theta, \phi, \mu) \leq \ldots \leq \mathcal{L}_K(\theta, \phi, \mu) \xrightarrow[K\to+\infty]{} l(\theta, \phi) \tag{6}$$

Maximizing the lower-bound involves computing its gradients. However, the REINFORCE estimate trains slowly, has a high variance and demandes numerous samples for accuracy. Instead, Ipsen et al. restructured the model by applying the reparameterization trick on both $q_\gamma(z|x^o)$ and $p_\theta(x^m|z)$ 1b. This allows backpropagation, which was impossible before due to random nodes, in order to update the model parameters. However, the reparameterzation trick requires both distributions to belong to a reparameterizable family, commonly including simpler distributions such as Gaussian or Student's distribution.

Information on the missing mechanism allow to better design the graphical model . The missing pattern $p(s|x^0, x^m)$ can be completly known, partially known (up to the distribution familly), or totally unknown. In a MNAR setting, using prior knowledge proves to be crucial, yet formulating robust model assumptions can pose challenges. As the missing pattern can be interpreted as a classifier on each feature of x, Ipsen et al. proposes a Bernoulli distribution to model the missing process: $p_\phi(s|x) = \prod_{j=1}^{n} \pi_{\phi,j}(x)^{s_j}(1 - \pi_{\phi,j}(x))^{1-s_j}$ with $\pi_{\phi,j}$ being the probability of x_j being observed knowing x. Different types of π_ϕ can be utilized to represent various situations, some of which will be explored in the upcoming sections.

4 Experiments

4.1 Datasets and Evaluation Metric

Following current research, we chose to run our experiments on few of the UCI datasets; the Red & White wine datasets, Boston Housing dataset, and Bank Note dataset. The datasets vary in size and therefore we think that they would serve as a good benchmark for evaluating the different algorithms for missing data imputation.

Boston Housing dataset Harrison and Rubinfeld (1978) consists of 506 housing data samples from the Boston area, each with 13 different features. The object is a regression task for predicting the median house value. 12 of the features are continuous and one of the features, on whether the house is one the banks of the Charles river, is categorical. Red & White wine datasets Aeberhard and Forina (1991) contain 4898 samples, each with 11 different features. The datasets are for the red and white variants of *vinho verde* wine from Portugal. For each wine color type, the features are all continuous. The BankNote dataset Lohweg (2013) contains 1372 samples, each with 4 continuous features for determining if a given banknote has been forged or not.

To be a good metric, the choice of error metric should conform with the expected probability distribution of the errors. For our current task, we could opt for either root mean squared error (RMSE) or mean average error (MAE). Following convention and Hodson (2022) which says that for Gaussian (normal) error distributions, RMSE is optimal, while for Laplacian errors, MAE is optimal, we opt for RMSE error as the metric. Mathematically, RMSE is,

where y is the ground truth, $\hat{y}$ is the model prediction, and N is the total number of samples.

4.2 Reproducing Ipsen et al. Experiments

We conducted experiments on UCI datasets as detailed in the paper for the agnostic case. We introduced the same MNAR missing process: for half of the features, when the feature value is higher than the feature mean it is set to missing. The utilized architecture mirrors the one presented in the paper, augmented by additional layers of Clipping and Softplus, as illustrated in 4 in the Annex. The training process remains consistent, configuring the latent space size to $(p-1)$ and incorporating $K = 20$ importance samples. The evaluation of imputation root mean square error (RMSE) involves $10,000$ importance samples, and the resulting mean and standard errors are computed over 5 iterations 1.

4.3 Imputation with *partial*-VAE

(Ma et al. (2019) propose a variational auto-encoder based method for performing imputation and inference on partial observations. It derives a variational lower bound that only depends on the observed data and forumlates a permutation-invariant encoder inspired from the *point-net* approach for point cloud classification Zaheer et al. (2017) and shows that *partial*-VAE is a generalization of the zero impute (ZI) VAE A. Nazabal (2020).

Computationally, the encoder is a neural network $h(.)$. It maps the input x and a learnable parameter, e, which is the embedding for the identity of the variable, to a latent space of K dimensions. Summation of the output of each node in $h(.)$ provides the characteristic permutation invariancy. We use the *point-net plus* (PNP) specification of the *partial*-VAE in our experiments. We perform a comparative study (Appendix A) of *partial*-VAE on Boston housing and Wine datasets with *notMIWAE*.

Imputation method	White	Red	BankNote	Concrete
MIWAE (Pytorch)	1.54 ± 0.013	1.63 ± 0.01	1.28 ± 0.02	1.68 ± 0.02499
not-MIWAE (Pytorch)	1.41 ± 0.012	1.47 ± 0.01	2.45 ± 0.34	2.66 ± 0.23
MIWAE (Tensorflow)	1.56 ± 0.005	1.65 ± 0.01	1.28 ± 0.024	1.73 ± 0.02
not-MIWAE (Tensorflow)	1.42 ± 0.009	1.36 ± 0.013	1.22 ± 0.27	2.82 ± 0.12
median	1.69 ± 0.0	1.84 ± 0.0	1.62 ± 0.0	1.86 ± 0.0
mean	1.74 ± 0.0	1.84 ± 0.0	1.73 ± 0.0	1.84 ± 0.0

Table 1: RSME Imputation for different methods for UCI Datasets.

The observed differences in values from the paper are likely attributed to variations in the standardization process. In the MNAR context, we typically achieve superior results for not-MIWAE compared to MIWAE, as anticipated. However, for not-MIWAE (PyTorch) on BankNote and both not-MIWAE (PyTorch) and not-MIWAE (TensorFlow) on Concrete, we encountered unexpectedly high RMSE, which can be attributed to the previously mentioned instability.

4.3.1 Simple Example

We reimplemented `not-MIWAE-demo.ipynb` from the notMIWAE Github repository in Pytorch. It allowed us to see that the model is able to learn the missing pattern 2.

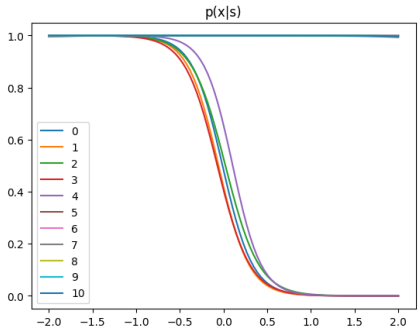


Figure 2: Representation of the missing pattern learned by the model

4.4 Designing Experiments

4.4.1 Missing Patterns

We designed various missing data patterns to assess the performance of the not-MIWAE model across distinct missing data scenarios. Columns containing missing data will be referred to as 'missing columns,' and the columns on which the missingness is based 'dependent columns'. In all our experiments, half of the dataset columns will be missing columns.

MCAR: Data are missing with a probability p.

MAR: The dependent columns are the same for the all the missing columns and no missing column is in the dependent columns. We explored *three* different patterns: **inter**: if at least one dependent column is above its mean, the probability of missingness is p; *sum*: the missing probability is equal to p times the number of dependent columns above their mean; **random_sum**: similar to *sum*, but each dependent column is associated to a discting missing probability inferior to p.

MNAR: for *inter* and *nn* the dependent columns are the same for all the missing columns but for *notsame* each missing columns has different dependent columns (mix of MAR and MNAR). Missing columns can be dependent columns: **inter**: same than *MAR inter*; **nn**: if the result of a Neural Network on the dependent columns is above a certain threshold, the probability of missingness is p; **notsame**: similar to *MAR sum* but each missing columns have different dependent columns.

4.4.2 Observations & Results

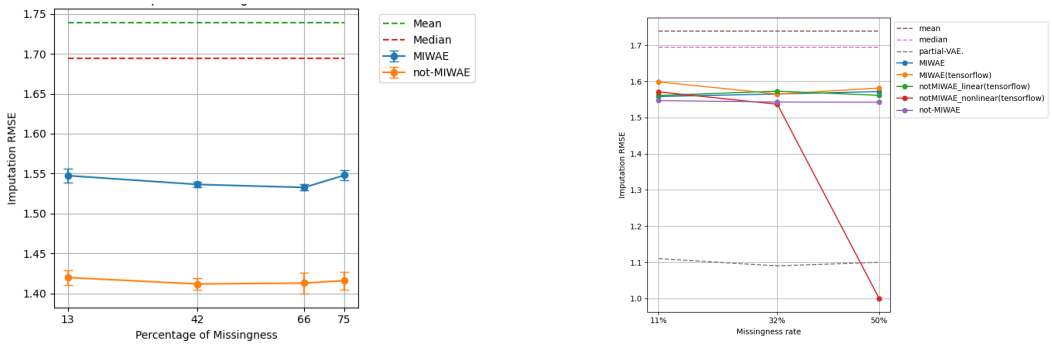
We explored two datasets to evalute the missing patterns described above. Our analysis focused first on the Boston housing dataset on which compared different imputation techniques: not-MIWAE, MIWAE, mean/median imputation, and partial-VAE. Results, presented in Table 2, were obtained with a missingness rate between 20% and 30% depending on the missing patterns. Our findings reveal that both MIWAE and not-MIWAE outperform traditional mean and median imputation methods, although they are not as effective as partial-VAE, even if it was designed for MAR cases. Notably, MIWAE consistently surpassed not-MIWAE in handling MNAR scenarios. We hypothesise that the relatively small size of the Boston Housing dataset might limit the efficacy of the MIWAE and not-MIWAE DVLMs. To test this theory, we replicated our experiments on the larger White Wine dataset from the UCI repository.

Further analyses yielded unexpected results. In the MAR-sum scenario 3a MIWAE outperformed not-MIWAE regardless of the missingness percentage. Until now, the not-MIWAE has relied on a linear model to learn the missing pattern. We're also exploring the application of a nonlinear model from here. In a broader comparison 3b, including MIWAE, not-MIWAE, mean/median imputation, and partial-VAE, partial-VAE generally led in performance, except under high missingness where non-linear not-MIWAE excelled. We hypothesized that our missing patterns might be a bit too challenging for the linear not-MIWAE model to work as well as described in Ipsen et al. paper. This suggests a need for further exploration into model adaptability under varying complexities of missing data.

4.5 Reproducibility

We implemented the not-MIWAE model using PyTorch and an updated version derived from the TensorFlow implementation. We came across frequent instability, indicated by an escalating loss potentially induced either by an exploding gradient or when converging to a local minimum that amplified the imputation error. Comparatively, the MIWAE model demonstrated greater stability, as might be expected.

In attempts to replicate the experiments outlined in the paper—utilizing the same architecture, parameters, and training process—we encountered instability. Notably, the only deviation was the undocumented standardization process. Additionally, despite thorough checks to ensure equivalent implementations, the TensorFlow version appeared more stable than its PyTorch counterpart under seemingly identical environment.



(a) **MAR-sum scenario** (b) **MNAR-notsame scenario**

Figure 3: Comparing various imputation techniques in both MAR and MNAR scenarios, examining RMSE with respect to the missingness in the missing columns.

5 Conclusions

This article, titled '*not-MIWAE: Deep Generative Modelling with Missing Not at Random Data*' by Ipsen et al. , addresses the challenge of handling missing data, particularly when the missing pattern depends on the data that are absent, using Deep Variational Latent Models (DVLMs). The method primarily focuses on minimizing a lower bound of the expected joint distribution of observed data and missing patterns through a double reparameterization trick.

While attempting to reproduce the results presented by Ipsen et al. on the UCI dataset, we encountered stability issues both with both PyTorch and TensorFlow implementations. To assess the method's robustness, we experimented with various missing patterns. Surprisingly, the performance of the not-MIWAE, except for a single instance with the nonlinear not-MIWAE, did not meet our expectations on either dataset. This observation led us to hypothesize that our missing patterns might have been too complex for the linear case. Consequently, we believe further investigation into the adaptability of models to different complexities of missing data is needed.

References

Z. Ghahramani I. Valera A. Nazabal, P. M. Olmos. 2020. Handling Incomplete Heterogeneous Data using VAEs. *arXiv preprint arXiv:1807.03653* (2020). https://arxiv.org/abs/1807.03653

Stefan Aeberhard and M. Forina. 1991. Wine. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5PC7J.

J. M. Hernández-Lobato1 N. Koenigstein S. Nowozin C. Zhang2 C. Ma, W. Gong. 2018. Partial VAE for Hybrid Recommender System. In *Bayesian Deep Learning*. http://bayesiandeeplearning.org/2018/papers/75.pdf Conference contribution, peer-reviewed.

Sahra Ghalebikesabi, Rob Cornish, Luke J. Kelly, and Chris Holmes. 2021. Deep Generative Pattern-Set Mixture Models for Nonignorable Missingness. arXiv:2103.03532 [stat.ML]

D. Harrison and D.L. Rubinfeld. 1978. Hedonic prices and the demand for clean air. J. Environ. Economics & Management.

T. O. Hodson. 2022. Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not. *Geoscientific Model Development* 15, 14 (2022), 5481–5487. https://doi.org/10.5194/gmd-15-5481-2022

Roderick J. A. Little. 1993. Pattern-Mixture Models for Multivariate Incomplete Data. *J. Amer. Statist. Assoc.* 88, 421 (1993), 125–134. http://www.jstor.org/stable/2290705

Roderick J. A. Little and Donald B. Rubin. 2002. *Statistical analysis with missing data*. John Wiley Sons, Inc.

Volker Lohweg. 2013. banknote authentication. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C55P57.

Chao Ma, Sebastian Tschiatschek, Konstantina Palla, José Miguel Hernández-Lobato, Sebastian Nowozin, and Cheng Zhang. 2019. EDDI: Efficient Dynamic Discovery of High-Value Information with Partial VAE. arXiv:1809.11142 [cs.LG]

Pierre-Alexandre Mattei and Jes Frellsen. 2019. MIWAE: Deep Generastive Modelling and Imputation of Incomplete Data. *arXiv.org* (Feb 2019). https://arxiv.org/pdf/1812.02633.pdf

Jared S. Murray. 2018. Multiple Imputation: A Review of Practical and Theoretical Findings. *Statist. Sci.* 33, 2 (2018), 142–159. https://dl.acm.org/doi/pdf/10.1145/1835804.1835895

J. Huang K. Murphy R. Vedantam, I. Fischer. 2018. Generative Models of Visually Grounded Imagination. *arXiv preprint arXiv:1705.10762* (2018). https://arxiv.org/abs/1705.10762

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

Aude Sportisse, Claire Boyer, and Julie Josse. 2020. Imputation and low-rank estimation with Missing Not At Random data. arXiv:1812.11409 [stat.ML]

Jinsung Yoon, James Jordon, and Mihaela van der Schaar. 2018. GAIN: Missing Data Imputation using Generative Adversarial Nets. In *Proceedings of the 35th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 80)*, Jennifer Dy and Andreas Krause (Eds.). PMLR, 5689–5698. https://proceedings.mlr.press/v80/yoon18a.html

Ruslan Salakhutdinov Yuri Burda, Roger Grosse. 2016. Importance Weighted Autoencoders. *arXiv.org* (Nov 2016). https://arxiv.org/pdf/1509.00519.pdf

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R Salakhutdinov, and Alexander J Smola. 2017. Deep Sets. In *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/f22e4747da1aa27e363d86d40ff442fe-Paper.pdf

Appendices

A Comparative Results

	Model	RMSE
MCAR	MIWAE	0.85549 ± 0.02362
	not-MIWAE	0.96068 ± 0.07481
	partial-VAE	0.7791
	Mean	1.00668 ± 0.00000
	Median	1.06276 ± 0.00000
MAR - inter	MIWAE	0.95149 ± 0.02171
	not-MIWAE	1.03265 ± 0.05264
	partial-VAE	0.8394
	Mean	1.10156 ± 0.00000
	Median	1.16560 ± 0.00000
MAR - sum	MIWAE	0.94962 ± 0.01779
	not-MIWAE	0.95360 ± 0.02854
	partial-VAE	0.7187
	Mean	1.03896 ± 0.00000
	Median	1.09641 ± 0.00000
MAR - random	MIWAE	0.87280 ± 0.00821
	not-MIWAE	0.91266 ± 0.08426
	partial-VAE	0.6878
	Mean	0.97863 ± 0.00000
	Median	1.02634 ± 0.00000
MNAR - inter	MIWAE	0.83364 ± 0.04012
	not-MIWAE	0.83783 ± 0.02115
	partial-VAE	0.6670
	Mean	0.94240 ± 0.00000
	Median	0.97278 ± 0.00000
MNAR - nn	MIWAE	0.82772 ± 0.02211
	not-MIWAE	0.84026 ± 0.02143
	partial-VAE	0.6834
	Mean	0.96712 ± 0.00000
	Median	1.00043 ± 0.00000
MNAR - notsame	MIWAE	1.09671 ± 0.02569
	not-MIWAE	1.20506 ± 0.10965
	partial-VAE	0.8520
	Mean	1.17536 ± 0.00000
	Median	1.24651 ± 0.00000

Table 2: RMSE for various missing patterns on the Boston Housing dataset.

B Network Architecture

(a) Encoder

Layer
Input x (p)
Linear (128)
Tanh
Linear (128)
Tanh
μ: Linear (p-1)
$\log \sigma^2$: Linear (p-1)
Clip(-10,10)

(b) Decoder

Layer
Input z (p - 1)
Linear (128)
Tanh
Linear (128)
Tanh
μ:
Linear (p-1)
σ:
Linear (p-1)
Softplus

(c) Missing model

Layer
Input x (p)
Linear (p)
Sigmoid

Figure 4: notMIWAE Architecture