# Deep Generative Missingness Pattern-Set Mixture Models

Sahra Ghalebikesabi, Rob Cornish, Chris Holmes, Luke J Kelly

# Deep Generative Missingness Pattern-Set Mixture Models

**Sahra Ghalebikesabi**
University of Oxford

**Rob Cornish**
University of Oxford

**Chris Holmes**[1]
University of Oxford

**Luke J. Kelly**
CEREMADE, CNRS
Université Paris-Dauphine

## Abstract

We propose a variational autoencoder architecture to model both ignorable and nonignorable missing data using pattern-set mixtures as proposed by Little (1993). Our model explicitly learns to cluster the missing data into missingness pattern sets based on the observed data and missingness masks. Underpinning our approach is the assumption that the data distribution under missingness is probabilistically semi-supervised by samples from the observed data distribution. Our setup trades off the characteristics of ignorable and nonignorable missingness and can thus be applied to data of both types. We evaluate our method on a wide range of data sets with different types of missingness and achieve state-of-the-art imputation performance. Our model outperforms many common imputation algorithms, especially when the amount of missing data is high and the missingness mechanism is nonignorable.

## 1 INTRODUCTION

Missing data is a ubiquitous problem in real-world applications. In imaging, for instance, inpainting algorithms are used to restore damaged images or to remove selected objects (Bertalmio et al., 2000). In medical applications, some patient records never get collected because of missed appointments or because information acquisition was too cost intensive. In other cases, some of the data may have simply been lost. Since the 1970s, a rapidly growing body of literature has developed imputation algorithms that appropriately fill in the missing data. Missing data is commonly classified into three categories: missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR) (Rubin, 1976). When data are MCAR, the missingness is independent of both the observed and missing variables. Data are MAR when the missingness depends on the observed variables only. And they are said to be MNAR when the missingness mechanism depends not only on the observed variables but also on the missing values. Imputing MNAR data bears the highest difficulties as it requires modeling the missing data mechanism. It is thus not sufficiently analyzed in the current machine learning literature.

In this paper, we exemplary explain some existing deep generative imputation methods in the framework of nonignorable missingness models (Rubin, 1976). We propose a VAE-based imputation algorithm derived from one of these nonignorable missingness models, the pattern-set mixture model, which was introduced by Little (1993). Our model encodes the nonignorable missing data mechanism in a latent variable. To prevent underidentification, we impose probabilistic semi-supervision. This approach prevents overfitting to the observed data. An additional benefit of our method is that it explicitly provides meaningful clusters of the data that can be exploited to cluster the population based on the similarity of their missing data mechanism. We show that the model achieves state-of-the-art performance on a variety of data sets. Furthermore, we provide an implementation of our algorithm and all corresponding experiments online.[2]

## 2 RELATED WORK

Common imputation approaches include the EM algorithm (Dempster et al., 1977), MICE (Buuren and Groothuis-Oudshoorn, 2010), matrix completion (Mazumder et al., 2010) and MissForest (Stekhoven and Bühlmann, 2012). With the recent development

[1] denotes senior author

Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130. 

[2] `https://github.com/sghalebikesabi/PSMVAE`

of deep generative models, there has been an increase in generative imputation algorithms. Deep generative models such as generative adversarial nets (GANs) (Goodfellow et al., 2014) or variational autoencoders (VAEs) (Kingma and Welling, 2013; Rezende et al., 2014) have been used for missingness imputation since their introduction.

One of the most popular examples is GAIN (Yoon et al., 2018). It uses a generative adversarial network in which the generator takes the observed data vector as an input and outputs an imputed data vector. The discriminator determines which variables were actually observed based on partial information of the true missingness mask. Another GAN-based imputation algorithm, MisGAN (Li et al., 2019), learns one generator and one discriminator solely for the missingness mask, and simultaneously learns a separate generator and discriminator for the data and the missingness mask.

Nazabal et al. (2020) introduce an objective function in order to directly train VAEs with a Gaussian Mixture prior on data with missingness. MIWAE (Mattei and Frellsen, 2019) is an importance-weighted autoencoder that achieves state-of-the-art imputation performance by maximizing a tight lower bound of the log-likelihood of the observed data.

All these imputation algorithms have in common that they do not incorporate assumptions on the missingness mechanisms and thus cannot be applied when the data is MNAR. To tackle this problem, Ipsen et al. (2021) introduced an extension of MIWAE, the not-MIWAE, that handles MNAR data by explicitly modeling the conditional distribution of the missingness mask given the data. Such an approach is, however, not robust to the misspecification of the missingness mechanism. Collier et al. (2020) develop a latent variable model that considers the observed data to be the result of a corruption process which depends on a binary missingness mask and can be applied on data with ignorable and nonignorable missingness. Nonetheless, the model architecture has to be chosen dependent on the underlying missingness mechanism. Recently, Sportisse et al. (2020) have shown the identifiability of the parameters of probabilistic principal components analysis for a class of MNAR mechanisms. Missingness not at random has also been tackled using discriminative approaches such as matrix completion in the recommender system literature (Wang et al., 2019).

## 3 PROBLEM FORMULATION

Let $\mathbf{x} = (x_1, ..., x_d)$ be a random variable taking values in the $d$-dimensional space $\mathcal{X} = \mathcal{X}_1 \times ... \times \mathcal{X}_d$. We further assume that the missingness mask $\mathbf{m}$ is a random variable defined on $\{0,1\}^d$. The missingness mask is defined such that $x_j$ is observed for $m_j = 1$, and missing otherwise. We denote the joint distribution of $\mathbf{x}$ and $\mathbf{m}$ by $P_\theta(\mathbf{x}, \mathbf{m})$. Following Yoon et al. (2018), we also introduce a random variable $\mathbf{x_{obs}} = (x_{obs,1}, ..., x_{obs,d})$ which takes values in $\mathcal{X}_{obs} = (\mathcal{X}_1 \cup \{*\}) \times \cdots \times (\mathcal{X}_d \cup \{*\})$ where $*$ is a point not in $\mathcal{X}_1 \cup .... \cup \mathcal{X}_d$ and represents unobserved data points. The random variable $\mathbf{x_{obs}}$ is then defined by

$$x_{obs,j} = \begin{cases} x_j, & \text{if } m_j = 1 \\ *, & \text{otherwise.} \end{cases}$$

Building upon this, we introduce another random variable $\mathbf{x_{mis}}$ with $x_{mis,j} = x_j$ if $m_j = 0$ and $x_{mis,j} = *$ otherwise. We can now retrieve $\mathbf{x}$ as

$$\mathbf{x} = \mathbf{m} \odot \mathbf{x_{obs}} + (\mathbf{1} - \mathbf{m}) \odot \mathbf{x_{mis}}, \tag{1}$$

where $\odot$ denotes the Hadamard product. Note that both $\mathbf{x_{mis}}$ and $\mathbf{x_{obs}}$ are defined to be $d$-dimensional random variables. We can interpret such an approach by imagining self-masked missingness, which means that the missingness probability of each covariate only depends on the covariate itself such that $P_\theta(m_j|\mathbf{x}) = P_\theta(m_j|x_j)$. Then $x_{mis,j}$ denotes the outcome of that covariate under the treatment $m_j = 0$ and $x_{obs,j}$ the outcome under the treatment $m_j = 1$. When we want to refer to only the observed or missing observations, we instead write $\mathbf{x'_{obs}} = \{x_j \mid m_j = 1 \text{ for } j \in \{1, ..., d\}\}$ and $\mathbf{x'_{mis}} = \{x_j \mid m_j = 0 \text{ for } j \in \{1, ..., d\}\}$.

In the imputation setting, we assume that we are given $n$ i.i.d. copies $\mathbf{x_{obs}^1}, ..., \mathbf{x_{obs}^n}$ of $\mathbf{x_{obs}}$. From these observations, we can infer the missingness masks $\mathbf{m^1}, ...., \mathbf{m^n}$ as realizations of $\mathbf{m}$. We write the observed data as $\mathcal{D} = \{\mathbf{x_{obs}^1}, ..., \mathbf{x_{obs}^n}, \mathbf{m^1}, ...., \mathbf{m^n}\}$. When we impute the missing data, we aim to recover $P_\theta(\mathbf{x_{mis}}|\mathbf{x_{obs}}, \mathbf{m})$.

## 4 NONIGNORABLE MISSINGNESS MODELS

When data are MCAR (i.e. $P_\theta(\mathbf{x}, \mathbf{m}) = P_\theta(\mathbf{x})P_\theta(\mathbf{m})$) or MAR (i.e. $P_\theta(\mathbf{x}, \mathbf{m}) = P_\theta(\mathbf{x})P_\theta(\mathbf{m}|\mathbf{x'_{obs}})$), we can maximize the data likelihood without modelling the missing-data mechanism since

$$\begin{aligned} P_\theta(\mathbf{x'_{obs}}, \mathbf{m}) &= \int P_\theta(\mathbf{x'_{obs}}, \mathbf{x'_{mis}}) P_\theta(\mathbf{m}|\mathbf{x'_{obs}}, \mathbf{x'_{mis}}) d\mathbf{x'_{mis}} \\ &= P_\theta(\mathbf{x'_{obs}}) P_\theta(\mathbf{m}|\mathbf{x'_{obs}}). \end{aligned}$$

When data are MNAR, the missing-data mechanism is nonignorable and has to be modeled within a maximum likelihood framework. Little and Rubin (2019)

differentiate three ways of modelling the joint distribution of $\mathbf{x}$ and $\mathbf{m}$ in this case.

*Selection models* factorize the joint distribution as $P_\theta(\mathbf{x}, \mathbf{m}) = P_\theta(\mathbf{x})P_\theta(\mathbf{m}|\mathbf{x})$. This factorization goes along with intuitive missing-data mechanisms: In the MNAR case, the complete data $x$ can be seen as the cause why some variables are missing. Not-MIWAE (Ipsen et al., 2021) can be categorized as a selection model that uses MIWAE to model $P_\theta(\mathbf{x})$ and an additional layer (through a neural network layer or a logistic regression) on top of $P_\theta(\mathbf{x})$ to model $P_\theta(\mathbf{m}|\mathbf{x})$. GAIN (Yoon et al., 2018) makes an MCAR assumption in the setting of a selection model. As such it imputes the missing values by modeling the data distribution $P_\theta(\mathbf{x})$ using a neural network with one hidden layer, and then estimates $P_\theta(\mathbf{m}|\mathbf{x}, \mathbf{h})$ where $\mathbf{h}$ is a hint variable that indicates what $\mathbf{m}$ looks like.

*Pattern mixture models* factorize the joint distribution as $P_\theta(\mathbf{x}, \mathbf{m}) = P_\theta(\mathbf{x}|\mathbf{m})P_\theta(\mathbf{m})$, where $P_\theta(\mathbf{m})$ is a categorical distribution and $P_\theta(\mathbf{x}, \mathbf{m})$ is as a result a mixture of distributions. The drawback of such a parameterization is that $P_\theta(\mathbf{x}|\mathbf{m})$ is conditional on a high-dimensional categorical variable whose categories are often not completely observed which leads to the distribution of the missing data being underidentified without any additional assumptions (Little, 1993). Despite this, pattern mixture models are applied when there is an interest in $P_\theta(\mathbf{x}|\mathbf{m})$. For instance, a drug company might be interested in predicting the lab values of patients that dropped out of a clinical trial. Collier et al. (2020) propose a latent variable model that optimizes the lower bound of $P_\theta(\mathbf{x_{obs}}|\mathbf{m})$ and as such constitutes a special type of pattern mixture model.

In order to combine the benefits of both factorizations, Little (1993) introduced *pattern-set mixture models*. These models have an additional latent variable $\mathbf{r}$ with realizations in $\{1, ..., k\}$ that clusters the missingness patterns into $k$ missingness pattern-sets. In each missingness pattern-set, the missing data mechanism is modeled using a selection model. The joint distribution can then be written as

$$P_\theta(\mathbf{x}, \mathbf{m}, \mathbf{r}) = P_\theta(\mathbf{r})P_\theta(\mathbf{x}|\mathbf{r})P_\theta(\mathbf{m}|\mathbf{x}, \mathbf{r}). \tag{2}$$

For $k = 1$ this model reduces to a selection model. Compared to pattern mixture models, it requires fewer parameters (because we are borrowing strengths across the clusters), is less prone to underidentification and has thus more statistical power. Regardless of this, it still allows us to cluster the population into interesting categories. In a clinical trial, there might be some patients that dropped out for reasons unrelated to the study, while other patients dropped out because of adverse reactions to the treatment. The objective is that the missingness pattern-set $\mathbf{r}$ corresponds to these different missingness mechanisms. It summarizes similar missingness patterns and thus simplifies downstream tasks. The HIVAE model (Nazabal et al., 2020), a Gaussian Mixture VAE, can be seen as a pattern-set mixture model for data that are MAR and where $P_\theta(\mathbf{m}|\mathbf{x}, \mathbf{r})$ is ignored in the maximum likelihood inference.

Wu and Carroll (1988) propose *shared-parameter models* which build upon the assumption that there exists a latent random variable $\mathbf{z} \in \mathbb{R}^b$ for some $b < n$ conditional on which the missing model and the data model are independent: $P_\theta(\mathbf{x}, \mathbf{m}|\mathbf{z}) = P_\theta(\mathbf{x}|\mathbf{z})P_\theta(\mathbf{m}|\mathbf{z})$.

These specifications are equivalent only when the data is MCAR (Little and Rubin, 2019). The choice of the model should be made based on the underlying data problem. We argue that in the case of high-dimensional data sets, such as images, where machine learning algorithms proved to be more useful than classical statistical approaches, pattern mixture models are not feasible considering the high-dimensional space of $\mathbf{m}$. As selection models fall within the class of pattern-set mixture models, we use a combination of pattern-set mixture models and shared-parameter models for building a deep generative pattern-set mixture model.

# 5 DEEP GENERATIVE PATTERN-SET MIXTURE MODEL

We now introduce an imputation approach that combines ideas of variational autoencoders and pattern-set mixture models. In contrast to other machine learning imputation methods such as HIVAE (Nazabal et al., 2020) or MIWAE (Mattei and Frellsen, 2019), we thus aim to model the joint distribution $P_\theta(\mathbf{x}, \mathbf{m})$ instead of the marginal $P_\theta(\mathbf{x})$.

## 5.1 Generative Model

We will now define a generative model for $P_\theta(\mathbf{x}, \mathbf{m})$. More specifically, we will model $P_\theta(\mathbf{x_{obs}}, \mathbf{m}, \mathbf{x_{mis}})$ where we assume for now that $\mathbf{x_{mis}}$ is a latent variable. Then, $P_\theta(\mathbf{x}, \mathbf{m})$ follows from Equation 1. Assuming the pattern-set mixture model holds, we introduce an additional latent categorical variable $\mathbf{r}$ which groups the missingness patterns into sets. Following Equation 2 and assuming that $P_\theta(\mathbf{m}|\mathbf{x_{obs}}, \mathbf{x_{mis}}, \mathbf{r}) = P_\theta(\mathbf{m}|\mathbf{x}, \mathbf{r})$, we can now write the joint distribution as

$$\begin{aligned} &P_\theta(\mathbf{x_{obs}}, \mathbf{m}, \mathbf{x_{mis}}, \mathbf{r}) \\ &\quad = P_\theta(\mathbf{r})P_\theta(\mathbf{x_{obs}}, \mathbf{x_{mis}}|\mathbf{r})P_\theta(\mathbf{m}|\mathbf{x}, \mathbf{r}). \end{aligned}$$

Since $\mathbf{r}$ is a categorical variable that only captures the pattern-set of an observation, we introduce an addi-

tional continuous latent variable $\mathbf{z}$ that models the latent interaction of $\mathbf{x_{mis}}$ and $\mathbf{x_{obs}}$. Given $\mathbf{z}$ and $\mathbf{r}$, we then assume the joint of $\mathbf{x_{mis}}$ and $\mathbf{x_{obs}}$ to fully factorize implying

$$P_\theta(\mathbf{x_{mis}}, \mathbf{x_{obs}}) = \prod_{j=1}^{d} P_\theta(x_{mis,j}|\mathbf{z}, \mathbf{r}) P_\theta(x_{obs,j}|\mathbf{z}, \mathbf{r}).$$

If additional information on the missing data mechanism $P_\theta(\mathbf{m}|\mathbf{x}, \mathbf{r})$ is available, we can write the generative model as

$$\begin{aligned} &P_\theta(\mathbf{x_{obs}}, \mathbf{m}, \mathbf{x_{mis}}, \mathbf{z}, \mathbf{r}) \qquad (3)\\ &= P_\theta(\mathbf{m}|\mathbf{x}, \mathbf{r}) P_\theta(\mathbf{x_{obs}}|\mathbf{r}, \mathbf{z}) P_\theta(\mathbf{x_{mis}}|\mathbf{r}, \mathbf{z}) P_\theta(\mathbf{z}|\mathbf{r}) P_\theta(\mathbf{r}), \end{aligned}$$

as is visualized in Figure 1a. While Ipsen et al. (2021) only show how one uniform missing model can be parameterized for the whole population, our approach allows to account for different missingness models. This can be interesting when part of the data is assumed to be MCAR while some observations can be MNAR.

We parameterize the generative model of the VAE using a neural network for improved data fit. Allowing the missing model $P_\theta(\mathbf{m}|\mathbf{x}, \mathbf{r})$ to be parameterized by a neural network has the disadvantage that the fit of $P_\theta(\mathbf{x_{obs}}, \mathbf{x_{mis}}|\mathbf{r})$ suffers from the flexibility of the missing model. This statement is strengthened by the empirical results of Ipsen et al. (2021). Since for any MNAR model there is an MAR model with equal fit to the observed data, it is not possible to test for MNAR without any further assumptions on the missing data mechanism (Molenberghs et al., 2008). For this reason, it is important to choose an imputation algorithm that finds a trade off between flexibility of the missingness model and the distortion it induces into the data model when the data are MCAR. We find that this dilemma can be solved by the assumption made when using shared parameter models that $\mathbf{x_{obs}}, \mathbf{x_{mis}}$ and $\mathbf{m}$ are independent conditional on the pattern-set indicator $\mathbf{r}$ and the continuous latent representation $\mathbf{z}$. Such a model has been proven to be more robust to model specification (Harel and Schafer, 2009). When no additional information on $P_\theta(\mathbf{m}|\mathbf{x})$ is available, we thus formalize the generative model as

$$\begin{aligned} &P_\theta(\mathbf{x_{obs}}, \mathbf{m}, \mathbf{x_{mis}}, \mathbf{z}, \mathbf{r}) \qquad (4)\\ &= P_\theta(\mathbf{r}) P_\theta(\mathbf{z}|\mathbf{r}) P_\theta(\mathbf{x_{mis}}|\mathbf{r}, \mathbf{z}) P_\theta(\mathbf{x_{obs}}|\mathbf{r}, \mathbf{z}) P_\theta(\mathbf{m}|\mathbf{r}, \mathbf{z}), \end{aligned}$$

as illustrated in Figure 1b. The missing values are imputed by sampling from $P_\theta(\mathbf{x_{mis}}|\mathbf{x_{obs}}, \mathbf{m}) = P_\theta(\mathbf{x_{obs}}|\mathbf{r}, \mathbf{z})$ which follows from the conditional independence of $\mathbf{x_{obs}}, \mathbf{m}$ and $\mathbf{x_{mis}}$ given $\mathbf{r}$ and $\mathbf{z}$.

### 5.2 Probabilistic Semi-Supervision

We use likelihood-based variational inference to learn the joint distribution $P_\theta(\mathbf{x}, \mathbf{m})$. As we do not observe the missing data, we can however only optimize for the parameters of $P_\theta(\mathbf{x_{obs}}, \mathbf{m})$. A sensible choice is to treat $\mathbf{x_{mis}}$ as a latent factor learned from $\mathbf{x_{obs}}$ and $\mathbf{m}$ (Nazabal et al., 2020). Without any further assumptions, the distribution of $\mathbf{x_{mis}}$ would be underidentified given the observed data as, among other reasons, any orthogonal transformation of $\mathbf{x_{mis}}$ would not affect the observed data likelihood $P_\theta(\mathbf{x_{obs}}, \mathbf{m})$ (Khemakhem et al., 2020).

We can regularize the problem and introduce smoothness through a novel process involving information sharing and semi-supervision (Kingma et al., 2014; Joy et al., 2020). For any covariate of any sample $x^i_{obs,j}$ with $j \in \{1, ..., d\}$ and $i \in \{1, ..., n\}$, we assume $x^i_{obs,j}$ is not only an observation of $x_{obs,j}$, but also of $x_{mis,j}$ with probability $1 - \pi_{i,j}$ if $m_{i,j} = 1$. Put simply, we sample each $x^i_{mis,j}$ from

$$y_{i,j} P_\theta(\widetilde{x}_{mis,j}) + (1 - y_{i,j}) \mathbb{1}(x^i_{obs,j}),$$

where $\mathbb{1}$ is the indicator function, $\widetilde{x}_{mis,j}$ is a latent auxiliary variable that describes the unobserved dynamics of the missing data, and $y_j$ is an independent Bernoulli random variable with known success probability $\pi'$ if $m_j = 1$, and with probability 1 otherwise. We then define the augmented data set as

$$\mathcal{D}_{\boldsymbol{\pi}}(\mathbf{y^1}, ..., \mathbf{y^n}) := \mathcal{D} \cup \{x^i_{mis,j}; y^i_j = 0\}_{i,j}. \qquad (5)$$

For simplicity, let us assume for now that we observe a single univariate data point $x^0_{obs,0}$ with $m^0_0 = 0$ such that $\mathcal{D} = \{x^0_{obs,0}, m^0_0\}$. With probability $1 - \pi'$, it holds that $y^0_0 = 0$. We then assume that $x^0_{mis,0}$ is also observed with value $x^0_{obs,0}$ and the augmented data set is thus $\mathcal{D}_\pi(y^0_0 = 0) = \{x^0_{obs,0}, m^0_0, x^0_{mis,0}\}$. In this case, we maximize the likelihood $P_\theta(x_{obs,0}, m_0, x_{mis,0}|\mathcal{D}_\pi(y^0_0 = 0))$. With probability $\pi'$, however, it holds that $y^0_0 = 1$ and $x^0_{mis,0}$ is assumed to be unobserved. We then have $\mathcal{D}_\pi(y^0_0 = 1) = \{x^0_{obs,0}, m^0_0\} = \mathcal{D}$. We now maximize the likelihood $P_\theta(x_{obs,0}, m_0|\mathcal{D}_\pi(y^0_0 = 1))$. Since we know the true distribution of $y^0_0$, we can also marginalize out $y_0$ and maximize the weighted likelihood

$$\begin{aligned} &\pi' P_\theta(x_{obs,0}, m_0|\mathcal{D}_\pi(y^0_0 = 1))\\ &\quad + (1 - \pi') P_\theta(x_{obs,0}, m_0, x_{mis,0}|\mathcal{D}_\pi(y^0_0 = 0)). \end{aligned}$$

In a more general setting, we can write the expected likelihood given the augmented data set as

$$\begin{aligned} &\mathbb{E}_{\mathbf{y}}[P_\theta(\mathbf{x_{obs}}, \mathbf{x_{mis,1-y}}, \mathbf{m}|\mathcal{D}_{\boldsymbol{\pi}}(\mathbf{y}))]\\ &\quad = \boldsymbol{\pi} P_\theta(\mathbf{x_{obs}}, \mathbf{m}|\mathcal{D}_{\boldsymbol{\pi}}(\mathbf{1}))\\ &\qquad + (\mathbf{1} - \boldsymbol{\pi}) P_\theta(\mathbf{x_{obs}}, \mathbf{m}, \mathbf{x_{mis}}|\mathcal{D}_{\boldsymbol{\pi}}(\mathbf{0})), \end{aligned}$$

where $\mathbf{x}_{\mathbf{mis},\mathbf{1}-\mathbf{y}} := \{x_{mis,j} | y_j = 0 \text{ for } j \in \{1, ..., d\}\}$, and $\mathbf{1}$ and $\mathbf{0}$ are $d$-vectors of ones and zeros respectively. We only assume semi-supervision for the covariates $x_{mis,j} | (m_j = 1) \in \{*\}$ which drop out in the generation process of $x_j$ (1). The parameter $\pi$ can thus be interpreted as confidence on the ignorability of the missing model: the greater $\pi$ is, the less likely $x_{mis,j}$ stems from an observed distribution. Note that this approach is equivalent to a biased data augmentation approach and that we do not modify the generative model here. As proven in the supplements, it holds:

**Proposition 1.** *Assume that we observe data $\mathcal{D}$ sampled from one of the generative models defined according to* (3) *or* (4)*. For $\pi' < 1$, it holds that the distributional parameter $\mu$ of $\mathbf{x}_{\mathbf{mis}} | \mathbb{E}_{\mathbf{y}}(\mathcal{D}_\pi(\mathbf{y}))) \sim \mathcal{N}(\mu, c)$ for some fixed constant $c$ and $\mathcal{D}_\pi(\mathbf{y})$ as defined in* (5) *is identifiable under the maximum likelihood estimation method.*

Without any additional model assumptions such as the specification of the missing model, the distribution of $\widetilde{\mathbf{x}}_{\mathbf{mis}}$ will be underidentified and the estimator of $\mathbf{x}_{\mathbf{mis}}$ will be biased. In that case, the question arises why we should not just assume $P_\theta(\mathbf{x}_{\mathbf{mis}}) = P_\theta(\mathbf{x}_{\mathbf{obs}})$ from the beginning. We argue that our unconventional approach to model both $\mathbf{x}_{\mathbf{obs}}$ and $\mathbf{x}_{\mathbf{mis}}$ as $d$-dimensional vectors and include such a semi-supervised approach introduces smoothing and prevents overfitting to the observed data distribution which is especially beneficial when missingness is high. A similar methodology has been proposed by Szegedy et al. (2016) under the term label-smoothing regularization. Contrary to our approach, Nazabal et al. (2020) assume the union of $\mathbf{x}_{\mathbf{obs}}$ and $\mathbf{x}_{\mathbf{mis}}$ to be $d$-dimensional. In other words, they assume that only the missing components of $\mathbf{x}$ are latent. As a result the missing data imputation of the HIVAE model can be seen as a special case of our model for $\pi' = 0$.

### 5.3 Recognition Model

As already noted, the generative model is learned by maximum likelihood inference. The marginal likelihood of the observed variables, $P_\theta(\mathbf{x}_{\mathbf{obs}}, \mathbf{m}, \mathbf{x}_{\mathbf{mis},\mathbf{1}-\mathbf{y}})$, is intractable but we can follow Kingma and Welling (2013) and define a recognition model $Q_\phi$ for modelling the unobserved latent variables $\mathbf{x}_{\mathbf{mis},\mathbf{y}}, \mathbf{z}$ and $\mathbf{r}$ given the observed observations $\mathbf{x}_{\mathbf{obs}}, \mathbf{m}$ and $\mathbf{x}_{\mathbf{mis},\mathbf{1}-\mathbf{y}}$ assuming a simple parametric form:

$$\begin{aligned} Q' :=& Q_\phi(\mathbf{x}_{\mathbf{mis},\mathbf{y}}, \mathbf{z}, \mathbf{r} | \mathbf{x}_{\mathbf{obs}}, \mathbf{m}, \mathbf{x}_{\mathbf{mis},\mathbf{1}-\mathbf{y}}) \\ =& Q_\phi(\mathbf{r} | \mathbf{x}_{\mathbf{obs}}, \mathbf{m}) Q_\phi(\mathbf{z} | \mathbf{r}, \mathbf{x}_{\mathbf{obs}}, \mathbf{m}, \mathbf{x}_{\mathbf{mis},\mathbf{1}-\mathbf{y}}) \\ & \times Q_\phi(\mathbf{x}_{\mathbf{mis},\mathbf{y}} | \mathbf{z}, \mathbf{r}, \mathbf{x}_{\mathbf{mis},\mathbf{1}-\mathbf{y}}, \mathbf{x}_{\mathbf{obs}}, \mathbf{m}). \end{aligned}$$

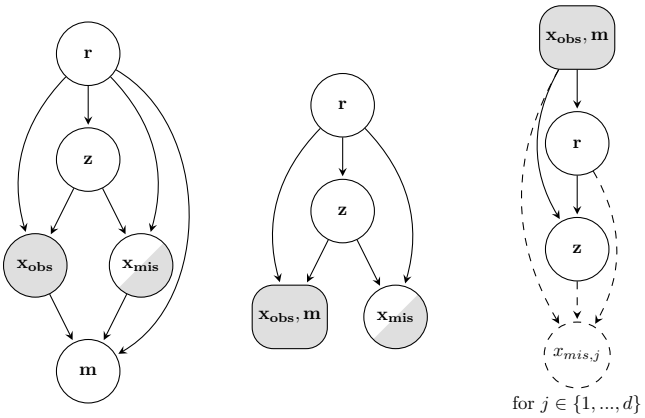
Using Jensen's equality it follows that

$$\begin{aligned} &\log P_\theta(\mathbf{x}_{\mathbf{obs}}, \mathbf{m}, \mathbf{x}_{\mathbf{mis},\mathbf{1}-\mathbf{y}}) \\ &\quad \geq \mathbb{E}_{Q'}[-\log Q' + \log P_\theta(\mathbf{x}_{\mathbf{obs}}, \mathbf{m}, \mathbf{x}_{\mathbf{mis}}, \mathbf{z}, \mathbf{r})]. \end{aligned}$$

We can thus fit the model by maximizing this evidence lower bound (ELBO).

Since the data sampled for $\mathbf{x}_{\mathbf{mis},\mathbf{1}-\mathbf{y}}$ is not more informative than $\mathbf{x}_{\mathbf{obs}}$ and $\mathbf{y}$, and $\mathbf{y}$ is defined independent of the missing data mechanism, we will assume in the following that $Q_\phi(\cdot | \mathbf{x}_{\mathbf{obs}}, \mathbf{m}, \mathbf{x}_{\mathbf{mis},\mathbf{1}-\mathbf{y}}) = Q_\phi(\cdot | \mathbf{x}_{\mathbf{obs}}, \mathbf{m})$. The recognition model can be seen in Figure 1c. Please refer to Table 1 for the ELBO that results from integrating out $\mathbf{y}$. This result is proved in the supplementary material.

Since we parameterize the recognition model $Q_\phi(\mathbf{x}_{\mathbf{mis},\mathbf{y}}, \mathbf{z}, \mathbf{r}, | \mathbf{x}_{\mathbf{obs}}, \mathbf{m})$ as a neural network, the input has to be of a fixed size. We thus implement an input-dropout layer (Nazabal et al., 2020) which takes all data (missing or observed) as input and drops out all the missing observations. The mechanics of this approach are the same as mean imputing the standardized input before applying the VAE model.




(a) Generative model with $P_\theta(\mathbf{m} | \mathbf{x}, \mathbf{r})$

(b) Generative model with $P_\theta(\mathbf{m} | \mathbf{z}, \mathbf{r})$

(c) Recognition model.

Figure 1: The architecture of the proposed models. Dashed lines and nodes exist with probability $\pi_j = 1$ if $m_j = 0$ and probability $\pi'$ if $m_j = 1$

### 5.4 Imputation

The VAE model defined above allows us to learn the predictive distribution of $P_\theta(\mathbf{x}_{\mathbf{mis}} | \mathbf{x}_{\mathbf{obs}}, \mathbf{m})$. When a single value is used for imputing the missing data, we speak of single imputation. For continuous data and when the $l_2$ norm is a relevant error metric, we follow Burda et al. (2015) and Mattei and Frellsen (2019) and impute the missing data by estimating $\mathbb{E}(\mathbf{x}_{\mathbf{mis}} | \mathbf{x}_{\mathbf{obs}}, \mathbf{m})$ using importance sampling. This

Table 1: ELBO of the proposed model.

$$\begin{aligned}\mathcal{L}(\mathbf{x_{obs}}, \mathbf{m}) = \mathbb{E}_{Q_\phi(\mathbf{r},\mathbf{z}|\mathbf{x_{obs}},\mathbf{m})}\Bigg[\log P_\theta(\mathbf{x_{obs}} \mid \mathbf{r}, \mathbf{z}) + \log \frac{P_\theta(\mathbf{z} \mid \mathbf{r})}{Q_\phi(\mathbf{z} \mid \mathbf{r}, \mathbf{x_{obs}}, \mathbf{m})} + \log \frac{P_\theta(\mathbf{r})}{Q_\phi(\mathbf{r} \mid \mathbf{x_{obs}}, \mathbf{m})}\Bigg] \\ + \pi \cdot \mathbb{E}_{Q_\phi(\mathbf{r},\mathbf{z},\mathbf{x_{mis}}|\mathbf{x_{obs}},\mathbf{m})}\Bigg[\log P_\theta(\mathbf{m} \mid \mathbf{r}, \mathbf{z}, \mathbf{x_{obs}}, \mathbf{x_{mis}}) + \log \frac{P_\theta(\mathbf{x_{mis}} \mid \mathbf{r}, \mathbf{z})}{Q_\phi(\mathbf{x_{mis}} \mid \mathbf{x_{obs}}, \mathbf{m}, \mathbf{z}, \mathbf{r})}\Bigg] \\ + (1-\pi) \cdot \mathbb{E}_{Q_\phi(\mathbf{r},\mathbf{z}|\mathbf{x_{obs}},\mathbf{m},\mathbf{x_{mis}})}\Bigg[\log P_\theta(\mathbf{m} \mid \mathbf{r}, \mathbf{z}, \mathbf{x_{obs}}, \mathbf{x_{mis}}) + \log P_\theta(\mathbf{x_{mis}} \mid \mathbf{r}, \mathbf{z})\Bigg]\end{aligned}$$

procedure will reduce the noise in our estimation compared to sampling a single value from $P_\theta(\mathbf{x_{mis}}|\mathbf{z}, \mathbf{r})$. The above expectation can also be written as

$$\begin{aligned}&\mathbb{E}(h(\mathbf{x_{mis}})|\mathbf{x_{obs}}, \mathbf{m}) \\ &= \int h(\mathbf{x_{mis}}) P_\theta(\mathbf{x_{mis}}|\mathbf{x_{obs}}, \mathbf{m}) d\mathbf{x_{mis}} \\ &= \int h(\mathbf{x_{mis}}) P_\theta(\mathbf{x_{mis}}|\mathbf{z}) P_\theta(\mathbf{z}|\mathbf{x_{obs}}, \mathbf{m})\, d\mathbf{z}\; d\mathbf{x_{mis}},\end{aligned}$$

where $h$ is an absolutely integrable function of $\mathbf{x_{mis}}$ and equal to the identity function in the case of single imputation. We use a self-normalized importance weighted estimator with the importance distribution $P_\theta(\mathbf{x_{mis}}|\mathbf{z})Q_\phi(\mathbf{z}|\mathbf{x_{obs}}, \mathbf{m})$ and the resulting estimator

$$\mathbb{E}(h(\mathbf{x_{mis}})|\mathbf{x_{obs}}, \mathbf{m}) \approx \frac{1}{\sum_{l=1}^{L} w^{(l)}} \sum_{l=1}^{L} w^{(l)} h(\mathbf{x_{mis}^{(l)}}),$$

where $(\mathbf{x_{mis}}^{(l)}, \mathbf{z}^{(l)})_{l=1}^{L}$ are samples from $P_\theta(\mathbf{x_{mis}}|\mathbf{z})Q_\phi(\mathbf{z}|\mathbf{x_{obs}})$ obtained by ancestral sampling. More specifically, we get samples from $Q_\phi(\mathbf{z}|\mathbf{x_{obs}})$ by marginalizing out $\mathbf{r}$ in $Q_\phi(\mathbf{z}|\mathbf{r}, \mathbf{x_{obs}}, \mathbf{m})$ and $Q_\phi(\mathbf{x_{mis}}|\mathbf{z}, \mathbf{r}, \mathbf{x_{obs}}, \mathbf{m})$. The importance weights are then defined by

$$w^{(l)} = \sum_{\mathbf{r}=1}^{k} \frac{P_\theta(\mathbf{x_{mis}}^{(l)}|\mathbf{z}^{(l)}, \mathbf{r}) P_\theta(\mathbf{z}^{(l)}|\mathbf{r}) P_\theta(\mathbf{r})}{Q_\phi(\mathbf{z}^{(l)}|\mathbf{r}, \mathbf{x_{obs}}) Q_\phi(\mathbf{r}|\mathbf{x_{obs}}, \mathbf{m})}.$$

Note that, in contrast to Mattei and Frellsen (2019), we do not rely on such an IWAE-based approach in the training stage. This stems from the fact that we propose a complex variational distribution and thus do not need the additional complexity induced by an importance-weighted approach (Cremer et al., 2017; Mattei and Frellsen, 2019). This hypothesis is further supported by our results in the Experiments section.

While single imputation returns a single estimator for the missing data, it is usually more interesting to sample multiple values from the predictive distribution for uncertainty quantification and statistically robust inference in downstream tasks (Rubin, 1996). Multiple imputations can be obtained by sampling multiple times from the generative model. Again we follow Mattei and Frellsen (2019) and use sampling importance resampling. For this, we generate a set of imputations and weight them using the weights defined above. The multiple imputations are then sampled from this weighted set.

# 6 EXPERIMENTS

In this section, we quantitatively evaluate the imputation performance of our model (from now on PSMVAE) and several state-of-the-art imputation approaches on UCI data sets (Lichman et al., 2013) commonly used in the machine learning for imputation literature (Yoon et al., 2018; Nazabal et al., 2020). Each experiment is repeated five times with different seeds. We report the root mean squared error (RMSE) of the estimators of the missing data along with its standard deviations across the experiments. Unless otherwise specified, we choose $\pi' = 0.5$. We further use the same hyperparameter constellation for each data set. Please see the supplementary material for details, the complete results and additional experiments such as the multiple imputation setting. The complete code for all experiments can be found online and figures are created using Weights & Biases (Biewald, 2020).

We synthetically introduce missingness into our data sets. The MNAR data are generated by self-masking one of the features. Every time this randomly selected variable is larger than its median we set it to zero with probability equal to the missingness rate, similar to Ipsen et al. (2021). The MCAR characteristic is introduced by randomly setting a feature to be missing with probability equal to the missingness rate. We present an overview of some results in Table 2 and Table 4. Note that the best score of each group is highlighted in bold, but that some of the approximate confidence intervals overlap.

## 6.1 Properties Of The Model

First, we analyze how the performance of our model changes due to incremental modifications in its architecture. For this purpose we compare a basic VAE,

Table 2: RMSE (Average±Std of RMSE) for single imputation on data with a missingness rate of 20%.

| Model class | Adult | | Letter | | Wine | |
|---|---|---|---|---|---|---|
| Missingness | MCAR | MNAR | MCAR | MNAR | MCAR | MNAR |
| PSMVAE(a) | .2494±.0021 | .4943±.3187 | .0964±.0013 | **.0835±.0153** | .0958±.0054 | **.1034±.0026** |
| PSMVAE(b) | .2426±.0019 | .5249±.2387 | .0936±.0008 | .0864±.0152 | .0890±.0029 | .1158±.0095 |
| ↳ K=10.000 | **.2306±.0019** | .4981±.2176 | **.0879±.0006** | .0854±.0162 | **.0832±.0022** | .1069±.0100 |
| ↳ w/o M | .2450±0.031 | **.4815±.2778** | .0941±.0009 | .0908±.0185 | .0885±.0024 | .1167±.0096 |
| DLGM | .2467±.0033 | .5468±.2328 | .0947±.0172 | .0947±.0172 | .0923±.0036 | .1234±.0104 |
| HIVAE | .2693±.0338 | .4907±.2072 | .1023±.0008 | .0947±.0179 | .0940±.0032 | .1246±.0156 |
| VAE | .2562±.0027 | .5012±.2313 | .1119±.0008 | .1061±.0194 | .1067±.0035 | .1255±.0105 |
| MIWAE | .2845±.0121 | .6081±.2423 | .1183±.0018 | **.1024±.0191** | .1129±.0034 | .1253±.0259 |
| ↳ K=10.000 | **.2373±.0015** | .5872±.3065 | **.1149±.0004** | .1242±.0063 | **.0915±.0017** | .0803±.0117 |
| not-MIWAE | .2374±.0011 | **.5201±.2640** | .1153±.0005 | .1192±.0317 | .0928±.0022 | **.0756±.0089** |
| GAIN | .2570±.0084 | .5940±.3744 | .1518±.0074 | .1316±.0061 | .1749±.0042 | .1151±.0175 |
| MICE | .2383±.0013 | .5879±.3079 | .1167±.0008 | .1235±.0069 | .0881±.0030 | .0782±.0117 |
| ↳ sample | .3166±.0015 | .6100±.2375 | .1575±.0007 | .1664±.0128 | .1264±.0025 | .1073±.0135 |
| MissForest | **.2246±.0026** | **.4513±.1774** | **.0650±.0018** | **.0534±.0113** | **.0738±.0023** | **.0698±.0035** |
| mean | .2510±.0012 | .6676±.3350 | .1560±.0005 | .1938±.0341 | .1765±.0041 | .1591±.0292 |

a simplified HIVAE model (HIVAE without different losses for different categories), and a deep latent Gaussian Variable model (DLGM, which is a VAE with one categorical and two Gaussian latent variables) with our model. We train our algorithm once with one importance sample and no missingness mask as input and output (PSMVAE w/o M), once with one importance sample and generative model as specified in Figure 1b (PSMVAE(b)), once as PSMVAE(b) with 10,000 importance samples, and finally with generative model as specified in Figure 1a (PSMVAE(a)) and 10,000 importance samples.

We note in our experiments that adding a categorical variable to a basic VAE always leads to a reduction in the RMSE loss of up to 1.4 percentage points. Moreover, we find that the PSMVAE w/o M is performing better than the DLGM. The main difference between the models is the assumed probabilistic semi-supervision of $\mathbf{x_{mis}}$ in the PSMVAE(b) while $\pi'$ is equal to 0 in the DLGM.[3] The average improvement of approximately 4.78% compared to the DLGM loss across all data sets speaks for the semi-supervised approach. Our model is also consistently performing better than the HIVAE which is a special case of our model when the union of $\mathbf{x_{mis}}$ and $\mathbf{x_{obs}}$ are assumed to be $d$-dimensional and $\pi' = 0$. In four of the displayed twelve experiments, from which three were run on data with MNAR, including the missingness mask led to a higher loss. We hypothesize that this stems from the class imbalance: While only one of the variables is set to be MNAR, the other variables are always observed.

[3]Please refer to the supplementary material for a detailed description of the DLGM.

In Table 3, we see that our models learn to cluster the data into meaningful missingness pattern-sets. For this purpose we ran our models and a HIVAE, whose input and output was a concatenation of the observed data and the corresponding missingness masks, on a variety of data sets. We did not only model MNAR, but also created a version of the data sets where all variables were MCAR and one variable was additionally MNAR. We assume that the model learns to cluster the data if the categories of the (here two-dimensional) latent variable $\mathbf{r}$ correspond to the cases when the variable that is MNAR is higher resp. lower than its median, which is the threshold for the assignment to the different missingness mechanisms. The pattern-set accuracy was then determined by computing the accuracy of a permutation $\widetilde{p}$ of the learned categories $\mathbf{r}|\mathbf{x_{obs}}, \mathbf{m}$ where the permutation was chosen such that $\widetilde{p}(\mathbf{r}|\mathbf{x_{obs}}, \mathbf{m}) = 0$ contained the highest fraction of observations of the MNAR pattern-set. Even though we only induce 20% missingness we see that all models learn to cluster the data.

### 6.2 Comparison With Other Imputation Approaches

We compare our proposed method with common imputation methods from the deep learning and statistical literature. As the losses of MIWAE and PSMVAE(b) indicate, importance sampling is essential in the imputation step of VAEs to reduce the noise of the imputations. We further note that traditional methods, such as MissForest and MICE, often outperform neural network based methods. We claim that this roots from the fact that single imputations in linear models, such as MICE learned with Bayesian Ridge,

Table 3: Pattern-set accuracy on data with a missingness rate of 20%.

| | Breast | | Wine | | Spam | |
|---|---|---|---|---|---|---|
| Algorithm | MNAR | MNAR+MCAR | MNAR | MNAR+MCAR | MNAR | MNAR+MCAR |
| PSMVAE (a) | **.7819±.0937** | **.7116±.1110** | .6113±.0564 | .6435±.0659 | .7349±.0638 | .8137±.0965 |
| PSMVAE (b) | .7709±.0824 | .6975±.1104 | .6068±.0535 | .6476±.0843 | **.7375±.0622** | .8123±.1019 |
| HIVAE w.M | .7050±.1161 | .6887±.1079 | **.6544±.0813** | **.7267±.0725** | .7244±.0833 | **.8145±.0949** |

Table 4: RMSE (Average±Std of RMSE) for single imputation on data with a missingness rate of 80%.

| | Breast | | Credit | | Spam | |
|---|---|---|---|---|---|---|
| Algorithm | MCAR | MNAR | MCAR | MNAR | MCAR | MNAR |
| PSMVAE(a) | **.1003±.0035** | .1114±.0417 | .1715±.0023 | **.0948±.061** | .0603±.0022 | .0801±.0232 |
| PSMVAE(b) | .1004±.0013 | .1085±.0258 | .1712±.0010 | .1095±.0533 | .0591±.0023 | .0889±.0427 |
| ↳ K=10,000 | .1125±.0034 | **.0729±.0184** | **.1701±.0011** | .0973±.0437 | **.0552±.0035** | **.0800±.0377** |
| ↳ w/o M | .1058±.0025 | .1025±.0276 | .1730±.0033 | .1202±.0555 | .0594±.0023 | .0879±.0433 |
| DLGM | .1179±.0011 | .1303±.0363 | .1742±.0026 | .1458±.0507 | .0624±.0027 | .0943±.0415 |
| HIVAE | .1207±.0023 | .1406±.0537 | .1743±.11 | .1301±.0542 | .0621±.0020 | .0924±.0426 |
| VAE | .1214±.0013 | .1214±.0475 | .1748±.0011 | .1267±.0576 | .0614±.0023 | .0929±.0425 |
| MIWAE | .1926±.0256 | .1452±.0316 | .1807±.0009 | .1716±.1262 | .0690±.0025 | .1133±.0709 |
| ↳ K=10,000 | **.0953±.0020** | **.1229±.0381** | **.1582±.0004** | .1076±.0391 | **.0579±.0021** | **.0858±.0364** |
| not-MIWAE | .2795±.0496 | .1393±.0425 | .2744±.0045 | **.1027±.0293** | .1112±.0035 | .0875±.0456 |
| GAIN | .6115±.1014 | .9170±.7307 | .1654±.0024 | .1334±.0649 | .0603±.0020 | .0886±.0423 |
| MICE | .1343±.0044 | .1445±.0424 | .1671±.0036 | **.1113±.0356** | .0619±.0023 | **.0886±.0372** |
| ↳ sample | .1419±.0021 | .1526±.0392 | .1961±.0010 | .1664±.0128 | .0804±.0027 | .1000±.0418 |
| MissForest | **.1155±.0020** | **.1260±.0269** | .1759±.0028 | .1145±.0384 | .0732±.0026 | .0909±.0397 |
| Mean | .1496±.0010 | .1545±.0297 | **.1640±.0005** | .1666±.0976 | **.0600±.0020** | .0926±.046 |

rely on consistent estimators of the expectation over the missing data distributions. When we sample a single value from the predictive distribution of MICE instead, we note that the single sample imputation of our method outperforms MICE. In Table 4 we see that the deep generative models perform better than MICE and MissForest on multiple data sets when the missingness rate is high. Another drawback of MICE and MissForest towards deep generative methods is that they are typically not scalable to high-dimensional data sets while our proposed models are.

We further note that our model achieves state-of-the-art imputation performance on both MCAR and MNAR data while MIWAE and Not-MIWAE typically only outperform other methods on one of the missingness types and are thus less robust to the misspecification of the missingness mechanism. In Figure 2 we see that the PSMVAE model with $\pi'(m_j) = \frac{P_\theta(m_j=1)}{1+P_\theta(m_j=1)}$ (blue line) is the only method with a decreasing imputation loss the longer the method learns. This choice ensures that the distributions of $x_{obs,j}$ and $x_{mis,j}$ are closer whenever $x_j$ was more likely to be missing and that $x_{mis,j}$ is always observed with a probability of at least 50%. Choosing a small constant $\pi'$ leads to high volatility in the results. Note that we choose a misspecified missing model (learned by self-masked logistic regression) for training PSMVAE(a) and not-MIWAE. While the loss of both models starts to increase after 500 steps, the loss of PSMVAE(a) increases less steeply than the the loss of not-MIWAE.

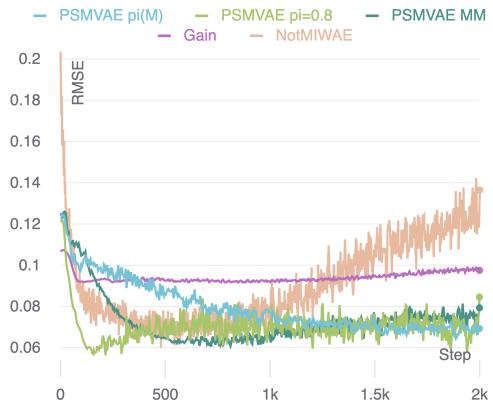



Figure 2: RMSE for single imputation (K=10,000) of various models as a function of iteration steps trained on the Breast data set with 80% induced MNAR missingness.

# 7 CONCLUSION AND FUTURE WORK

We propose a generative model combining ideas from VAEs and pattern-set mixture models for missing data imputation. This architecture specifies a variational autoencoder that prevents overfitting to data by introducing the assumption of probabilistic semi-supervision of the missing data. We also allow for the specification of a missing model. We see in various experiments with real-world data sets that our model achieves state-oft-the-art imputation performance for different missingness types and explicitly models meaningful clusters which add to the interpretability of the model when no information on the missingness mechanism is available. We recommend our model for use when the underlying data mechanisms are unknown, the missingness is high and/or the data is high-dimensional. Future research will focus on incorporating assumptions on the differences between the missing models of distinct missingness pattern-sets.

### Acknowledgments

We thank the reviewers for their constructive feedback. SG is a student of the EPSRC CDT in Modern Statistics and Statistical Machine Learning (EP/S023151/1) and receives funding from the Oxford Radcliffe Scholarship and Novartis. RC is supported by the Engineering and Physical Sciences Research Council (EPSRC) through the Bayes4Health programme Grant EP/R018561/1. LJK is supported by the French government under management of Agence Nationale de la Recherche as part of the "ABSint" programme, reference ANR-18-CE40-0034. CH is supported by The Alan Turing Institute, Health Data Research UK, the Medical Research Council UK, the EPSRC through the Bayes4Health programme Grant EP/R018561/1, and AI for Science and Government UK Research and Innovation (UKRI).

### References

Bertalmio, M., Sapiro, G., Caselles, V., and Ballester, C. (2000). Image inpainting. In *SIGGRAPH '00: Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques*, pages 417–424.

Biewald, L. (2020). Experiment tracking with weights and biases. Software available from wandb.com.

Burda, Y., Grosse, R., and Salakhutdinov, R. (2015). Importance weighted autoencoders. *arXiv preprint arXiv:1509.00519*.

Buuren, S. v. and Groothuis-Oudshoorn, K. (2010). MICE: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, pages 1–68.

Collier, M., Nazabal, A., and Williams, C. K. (2020). VAEs in the presence of missing data. *arXiv preprint arXiv:2006.05301*.

Cremer, C., Morris, Q., and Duvenaud, D. (2017). Reinterpreting importance-weighted autoencoders. *arXiv preprint arXiv:1704.02916*.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680.

Harel, O. and Schafer, J. L. (2009). Partial and latent ignorability in missing-data problems. *Biometrika*, 96(1):37–50.

Ipsen, N. B., Mattei, P.-A., and Frellsen, J. (2021). not-MIWAE: Deep Generative Modelling with Missing not at Random Data. In *International Conference on Learning Representations*.

Joy, T., Schmon, S. M., Torr, P. H., Siddharth, N., and Rainforth, T. (2020). Rethinking Semi-Supervised Learning in VAEs. *arXiv preprint arXiv:2006.10102*.

Khemakhem, I., Kingma, D., Monti, R., and Hyvarinen, A. (2020). Variational Autoencoders and Nonlinear ICA: A Unifying Framework. In *International Conference on Artificial Intelligence and Statistics*, pages 2207–2217.

Kingma, D. P., Mohamed, S., Rezende, D. J., and Welling, M. (2014). Semi-supervised Learning with Deep Generative Models. In *Advances in Neural Information Processing Systems*, pages 3581–3589.

Kingma, D. P. and Welling, M. (2013). Auto-Encoding Variational Bayes. *arXiv preprint arXiv:1312.6114*.

Li, S. C.-X., Jiang, B., and Marlin, B. (2019). Learning from Incomplete Data with Generative Adversarial Networks. In *International Conference on Learning Representations*.

Lichman, M. et al. (2013). UCI machine learning repository.

Little, R. J. (1993). Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, 88(421):125–134.

Little, R. J. and Rubin, D. B. (2019). *Statistical analysis with missing data*, volume 793. John Wiley & Sons.

Mattei, P.-A. and Frellsen, J. (2019). MIWAE: deep generative modelling and imputation of incomplete data sets. In *International Conference on Machine Learning*, pages 4413–4423.

Mazumder, R., Hastie, T., and Tibshirani, R. (2010). Spectral regularization algorithms for learning large incomplete matrices. *The Journal of Machine Learning Research*, 11:2287–2322.

Molenberghs, G., Beunckens, C., Sotto, C., and Kenward, M. G. (2008). Every missingness not at random model has a missingness at random counterpart with equal fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(2):371–388.

Nazabal, A., Olmos, P. M., Ghahramani, Z., and Valera, I. (2020). Handling Incomplete Heterogeneous Data using VAEs. *Pattern Recognition*, page 107501.

Rezende, D. J., Mohamed, S., and Wierstra, D. (2014). Stochastic Backpropagation and Approximate Inference in Deep Generative Models. In *International Conference on Machine Learning*.

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

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91(434):473–489.

Sportisse, A., Boyer, C., and Josse, J. (2020). Estimation and imputation in probabilistic principal component analysis with missing not at random data. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 7067–7077. Curran Associates, Inc.

Stekhoven, D. J. and Bühlmann, P. (2012). MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826.

Wang, X., Zhang, R., Sun, Y., and Qi, J. (2019). Doubly robust joint learning for recommendation on data missing not at random. In *International Conference on Machine Learning*, pages 6638–6647.

Wu, M. C. and Carroll, R. J. (1988). Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics*, pages 175–188.

Yoon, J., Jordon, J., and Van Der Schaar, M. (2018). GAIN: Missing Data Imputation using Generative Adversarial Nets. In Dy, J. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 5689–5698.