Deep Generative Pattern-Set Mixture Models for Nonignorable Missingness

Sahra Ghalebikesabi (sahra.ghalebikesabi@stats.ox.ac.uk), Rob Cornish, Luke J. Kelly, Chris Holmes

UNIVERSITY OF OXFORD

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 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 [12], 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 $\mathbf{x'_{mis}} = \{x_j \mid m_j = 0 \text{ for } j \in \{1, ..., d\}\}$.

(1) When data are MCAR, i.e.

$$P_\theta(\mathbf{x}, \mathbf{m}) = P_\theta(\mathbf{x})P_\theta(\mathbf{m})$$

(2) 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.

(3) When data are MNAR, the missing-data mechanism is nonignorable and has to be modeled within a maximum likelihood framework. [7] differentiate three ways of modelling the joint distribution of $\mathbf{x}$ and $\mathbf{m}$ in this case.

2 Nonignorable Missingness Models

(1) *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 [3] and GAIN [12] can be categorized as such models.

(2) *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[6]. [1] propose a latent variable model that optimizes the lower bound of $P_\theta(\mathbf{x_{obs}}|\mathbf{m})$.

(3) In order to combine the benefits of both factorizations, [6] 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

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

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. The HIVAE model [10], a Gaussian Mixture VAE, can be seen as a pattern-set mixture model.

(4) [11] 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:

$$P_\theta(\mathbf{x}, \mathbf{m}|\mathbf{z}) = P_\theta(\mathbf{x}|\mathbf{z})P_\theta(\mathbf{m}|\mathbf{z}).$$

3 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[10] or MIWAE[8], we thus aim to model the joint distribution $P_\theta(\mathbf{x}, \mathbf{m})$ instead of the marginal $P_\theta(\mathbf{x})$.

3.1 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

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

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

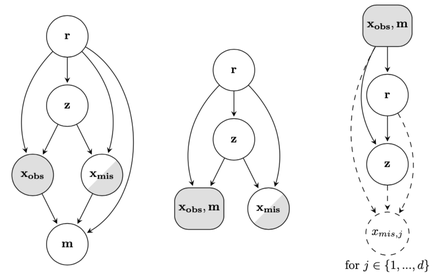
$$P_\theta(\mathbf{x_{mis}}, \mathbf{x_{obs}}) = \prod_{j=1}^{d} P_\theta(x_{mis,j}|\mathbf{z}, \mathbf{r})P_\theta(x_{obs,j}|\mathbf{z}, \mathbf{r}).$$

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

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

While [3] 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 [3]. 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[9]. 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[2]. When no additional information on $P_\theta(\mathbf{m}|\mathbf{x})$ is available, we thus formalize the generative model as

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



(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.

3.2 Probabilistic Semi-Supervision

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 $\mathbf{m}$[10]. Without any further assumptions, the distribution of $\mathbf{x_{mis}}$ would be underidentified given the observed data.
We can regularize the problem and introduce smoothness through a novel process involving information sharing and semi-supervision[5, 4]. 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 π' if $m_j = 1$, and with probability 1 otherwise. We then define the augmented data set as

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

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 π', 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

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

where $\mathbf{x_{mis,1-y}} := \{x_{mis,j}|y_j = 0 \text{ for } j \in \{1, ..., d\}\}$, 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 π can thus be interpreted as confidence on the ignorability of the missing model: the greater π 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.

References

[1] Mark Collier, Alfredo Nazabal, and Christopher KI Williams. VAEs in the Presence of Missing Data. *arXiv preprint arXiv:2006.05301* (2020).

[2] Ofer Harel and Joseph L Schafer. Partial and latent ignorability in missing-data problems. *Biometrika* 96.1 (2009), pp. 37–50.

[3] Niels Bruun Ipsen, Pierre-Alexandre Mattei, and Jes Frellsen. "not-MIWAE: Deep Generative Modelling with Missing not at Random Data". *International Conference on Learning Representations*. 2021.

[4] Tom Joy et al. Rethinking Semi-Supervised Learning in VAEs. *arXiv preprint arXiv:2006.10102* (2020).

[5] Durk P Kingma et al. "Semi-supervised Learning with Deep Generative Models". *Advances in Neural Information Processing Systems*. 2014, pp. 3581–3589.

[6] Roderick JA Little. Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association* 88.421 (1993), pp. 125–134.

[7] Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*. Vol. 793. John Wiley & Sons, 2019.

[8] Pierre-Alexandre Mattei and Jes Frellsen. "MIWAE: Deep generative modelling and imputation of incomplete data sets". *International Conference on Machine Learning*. 2019, pp. 4413–4423.

[9] Geert Molenberghs et al. 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 (2008), pp. 371–388.

[10] Alfredo Nazabal et al. Handling Incomplete Heterogeneous Data using VAEs. *Pattern Recognition* (2020), p. 107501.

[11] Margaret C Wu and Raymond J Carroll. Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics* (1988), pp. 175–188.

[12] Jinsung Yoon, James Jordon, and Mihaela Van Der Schaar. "GAIN: Missing Data Imputation using Generative Adversarial Nets". *Proceedings of the 35th International Conference on Machine Learning*. Ed. by Jennifer Dy and Andreas Krause. Vol. 80. 2018, pp. 5689–5698.