MIWAE: Deep Generative Modelling and Imputation of Incomplete Data Sets

Pierre-Alexandre Mattei

IT University of Copenhagen

`http://pamattei.github.io/`

@pamattei

ICML 2019

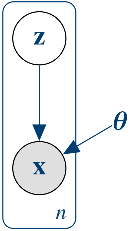
Joint work with **Jes Frellsen** (ITU Copenhagen)

IT UNIVERSITY OF CPH

How to handle missing data with deep generative models?

Let $(\mathbf{x}_i, \mathbf{z}_i)_{i \leq n}$ be i.i.d. random variables driven by a deep generative model:

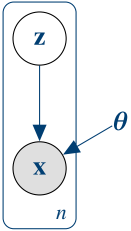
$$\begin{cases} \mathbf{z} \sim p(\mathbf{z}) & \text{(prior)} \\ \mathbf{x} \sim p_{\boldsymbol{\theta}}(\mathbf{x} \mid \mathbf{z}) & \text{(observation model)} \end{cases}$$



How to handle missing data with deep generative models?

Let $(\mathbf{x}_i, \mathbf{z}_i)_{i \leq n}$ be i.i.d. random variables driven by a deep generative model:

$$\begin{cases} \mathbf{z} \sim p(\mathbf{z}) & \text{(prior)} \\ \mathbf{x} \sim p_{\boldsymbol{\theta}}(\mathbf{x} \mid \mathbf{z}) & \text{(observation model)} \end{cases}$$



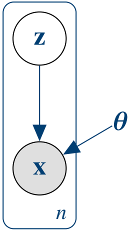
Assume that some of the training data are **missing-at-random** (MAR).
We can then split each sample $i \in \{1, \ldots, n\}$ into

- the **observed features** $\mathbf{x}_i^{\mathrm{o}}$ and
- the **missing features** $\mathbf{x}_i^{\mathrm{m}}$.

How to handle missing data with deep generative models?

Let $(\mathbf{x}_i, \mathbf{z}_i)_{i \leq n}$ be i.i.d. random variables driven by a deep generative model:

$$\begin{cases} \mathbf{z} \sim p(\mathbf{z}) & \text{(prior)} \\ \mathbf{x} \sim p_{\boldsymbol{\theta}}(\mathbf{x} \mid \mathbf{z}) & \text{(observation model)} \end{cases}$$



Assume that some of the training data are **missing-at-random** (MAR).
We can then split each sample $i \in \{1, \ldots, n\}$ into

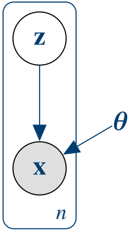
- the **observed features** $\mathbf{x}_i^{\mathbf{o}}$ and
- the **missing features** $\mathbf{x}_i^{\mathbf{m}}$.

1. Can we train $p_{\boldsymbol{\theta}}$ in a VAE fashion in spite of the missingness?

How to handle missing data with deep generative models?

Let $(\mathbf{x}_i, \mathbf{z}_i)_{i \leq n}$ be i.i.d. random variables driven by a deep generative model:

$$\begin{cases} \mathbf{z} \sim p(\mathbf{z}) & \text{(prior)} \\ \mathbf{x} \sim p_{\boldsymbol{\theta}}(\mathbf{x} \mid \mathbf{z}) & \text{(observation model)} \end{cases}$$



Assume that some of the training data are **missing-at-random** (MAR). We can then split each sample $i \in \{1, \dots, n\}$ into

- the **observed features** $\mathbf{x}_i^{\mathbf{o}}$ and
- the **missing features** $\mathbf{x}_i^{\mathbf{m}}$.

1. Can we train $p_{\boldsymbol{\theta}}$ in a VAE fashion in spite of the missingness?

2. Can we impute the missing values?

1. Can we train p_θ in a VAE fashion in spite of the missingness?

Under the MAR assumption, the relevant quantity to maximise is the **likelihood of the observed data** equal to

$$\ell^{\mathrm{o}}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log p_{\boldsymbol{\theta}}(\mathbf{x}_i^{\mathrm{o}}) = \sum_{i=1}^{n} \log \int p_{\boldsymbol{\theta}}(\mathbf{x}_i^{\mathrm{o}} \mid \mathbf{z}) p(\mathbf{z}) d\mathbf{z}.$$

1. Can we train p_θ in a VAE fashion in spite of the missingness?

Under the MAR assumption, the relevant quantity to maximise is the **likelihood of the observed data** equal to

$$\ell^{\mathrm{o}}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log p_{\boldsymbol{\theta}}(\mathbf{x}_i^{\mathrm{o}}) = \sum_{i=1}^{n} \log \int p_{\boldsymbol{\theta}}(\mathbf{x}_i^{\mathrm{o}} \mid \mathbf{z}) p(\mathbf{z}) d\mathbf{z}.$$

Building on the **importance weighted autoencoder (IWAE)** of Burda et al. (2016), we derive an approachable stochastic lower bound of $\ell^{\mathrm{o}}(\boldsymbol{\theta})$, the **missing IWAE (MIWAE)** bound:

$$\mathcal{L}_K(\boldsymbol{\theta}, \boldsymbol{\gamma}) = \sum_{i=1}^{n} \mathbb{E}_{\mathbf{z}_{i1},\ldots,\mathbf{z}_{iK} \sim q_{\boldsymbol{\gamma}}(\mathbf{z}|\mathbf{x}_i^{\mathrm{o}})} \left[\log \frac{1}{K} \sum_{k=1}^{K} \frac{p_{\boldsymbol{\theta}}(\mathbf{x}_i^{\mathrm{o}}|\mathbf{z}_{ik}) p(\mathbf{z}_{ik})}{q_{\boldsymbol{\gamma}}(\mathbf{z}_{ik}|\mathbf{x}_i^{\mathrm{o}})} \right] \leq \ell^{\mathrm{o}}(\boldsymbol{\theta}).$$

Like for the IWAE, the MIWAE bound gets tighter when the number of importance weights K grows.

2. Can we impute the missing values?

For the **single imputation problem** we use **self-normalised importance sampling** to approximate $\mathbb{E}[\mathbf{x}^{\mathrm{m}}|\mathbf{x}^{\mathrm{o}}]$:

$$\mathbb{E}[\mathbf{x}^{\mathrm{m}}|\mathbf{x}^{\mathrm{o}}] \approx \sum_{l=1}^{L} w_l \, \mathbf{x}^{\mathrm{m}}_{(l)},$$

where $(\mathbf{x}^{\mathrm{m}}_{(1)}, \mathbf{z}_{(1)}), \ldots, (\mathbf{x}^{\mathrm{m}}_{(L)}, \mathbf{z}_{(L)})$ are i.i.d. samples from $p_{\boldsymbol{\theta}}(\mathbf{x}^{\mathrm{m}}|\mathbf{x}^{\mathrm{o}}, \mathbf{z})q_{\boldsymbol{\gamma}}(\mathbf{z}|\mathbf{x}^{\mathrm{o}})$ and

$$w_l = \frac{r_l}{r_1 + \ldots + r_L}, \text{ with } r_l = \frac{p_{\boldsymbol{\theta}}(\mathbf{x}^{\mathrm{o}}|\mathbf{z}_{(l)})p(\mathbf{z}_{(l)})}{q_{\boldsymbol{\gamma}}(\mathbf{z}_{(l)}|\mathbf{x}^{\mathrm{o}})}.$$

2. Can we impute the missing values?

For the **single imputation problem** we use **self-normalised importance sampling** to approximate $\mathbb{E}[\mathbf{x}^{\mathrm{m}}|\mathbf{x}^{\mathrm{o}}]$:

$$\mathbb{E}[\mathbf{x}^{\mathrm{m}}|\mathbf{x}^{\mathrm{o}}] \approx \sum_{l=1}^{L} w_l \, \mathbf{x}^{\mathrm{m}}_{(l)},$$

where $(\mathbf{x}^{\mathrm{m}}_{(1)}, \mathbf{z}_{(1)}), \ldots, (\mathbf{x}^{\mathrm{m}}_{(L)}, \mathbf{z}_{(L)})$ are i.i.d. samples from $p_{\boldsymbol{\theta}}(\mathbf{x}^{\mathrm{m}}|\mathbf{x}^{\mathrm{o}}, \mathbf{z}) q_{\boldsymbol{\gamma}}(\mathbf{z}|\mathbf{x}^{\mathrm{o}})$ and

$$w_l = \frac{r_l}{r_1 + \ldots + r_L}, \text{ with } r_l = \frac{p_{\boldsymbol{\theta}}(\mathbf{x}^{\mathrm{o}}|\mathbf{z}_{(l)}) p(\mathbf{z}_{(l)})}{q_{\boldsymbol{\gamma}}(\mathbf{z}_{(l)}|\mathbf{x}^{\mathrm{o}})}.$$

Multiple imputation, i.e. sampling from $p_{\boldsymbol{\theta}}(\mathbf{x}^{\mathrm{m}}|\mathbf{x}^{\mathrm{o}})$, can be done using **sampling importance resampling** according to the weights w_l for large L.

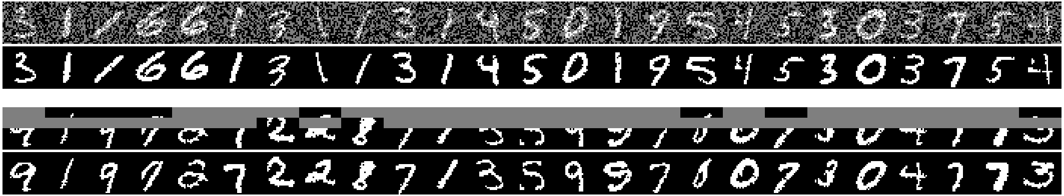
Single imputation of UCI data sets (50% MCAR)

	Banknote	*Breast*	*Concrete*	*Red*	*White*	*Yeast*
MIWAE	**0.446 (0.038)**	**0.280 (0.021)**	**0.501 (0.040)**	**0.643 (0.026)**	**0.735 (0.033)**	**0.964(0.057)**
MVAE	0.593 (0.059)	0.318 (0.018)	0.587(0.026)	0.686 (0.120)	0.782 (0.018)	0.997 (0.064)
missForest	0.676 (0.040)	0.291 (0.026)	0.510 (0.11)	0.697 (0.050)	0.798 (0.019)	1.41 (0.02)
PCA	0.682 (0.016)	0.729 (0.068)	0.938 (0.033)	0.890 (0.033)	0.865 (0.024)	1.05(0.061)
kNN	0.744 (0.033)	0.831 (0.029)	0.962(0.034)	0.981 (0.037)	0.929 (0.025)	1.17 (0.048)
Mean	1.02 (0.032)	1.00 (0.04)	1.01 (0.035)	1.00 (0.03)	1.00 (0.02)	1.06 (0.052)

Mean-squared error for single imputation for various continuous UCI data sets.

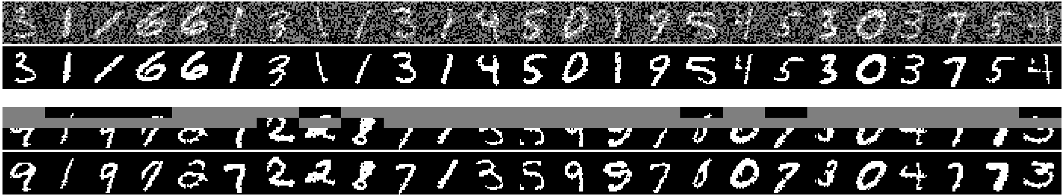
Imputation incomplete versions of binary MNIST

Single imputations:

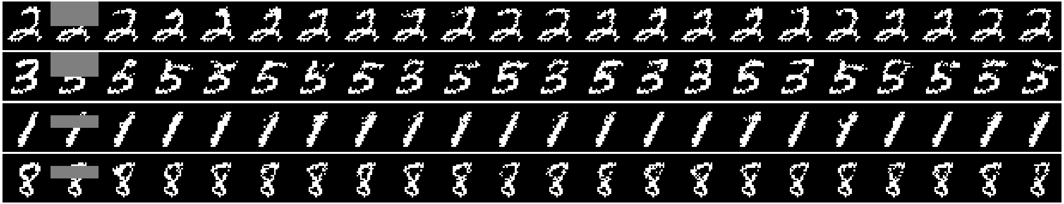


Imputation incomplete versions of binary MNIST

Single imputations:



Multiple imputations :



Classification of binary MNIST (50% MCAR pixels)

	Test accuracy	*Test cross-entropy*
Zero imputation	0.9739 (0.0018)	0.1003 (0.0092)
missForest imputation	0.9805 (0.0018)	0.0645 (0.0066)
MIWAE single imputation	0.9847 (0.0009)	0.0510 (0.0035)
MIWAE multiple imputation	**0.9870 (0.0003)**	**0.0396 (0.0003)**
Complete data	**0.9866 (0.0007)**	**0.0464 (0.0026)**

Learn more about MIWAE at poster 9 in the Pacific ballroom at 6.30!

Thanks for your attention :)