Enhanced Variational Autoencoder Estimation from Incomplete Data using Mixture Variational Families

Vaidotas Šimkus Michael U. Gutmann

School of Informatics, The University of Edinburgh

Summary

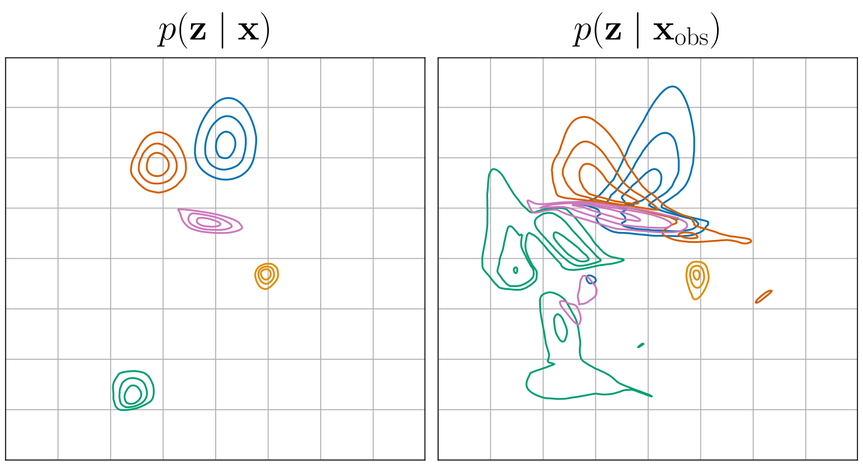
- We show why estimating VAEs from incomplete data is harder compared to the fully-observed case.
- As a result, fitting VAEs from incomplete data requires more flexible variational families, compared to the fully-observed case. But, more flexible families may lack the useful inductive biases!
- We propose two approaches based on variational mixture distributions that can improve VAE estimation from incomplete data while reusing the families from the fully-observed case.

Why is VAE estimation from *incomplete* data hard(er)?

- A VAE model is typically trained by maximising the ELBO:

$$\log p_{\boldsymbol{\theta}}(\boldsymbol{y}) \geqslant \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{y})}\left[\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{y} \mid \boldsymbol{z}) p_{\boldsymbol{\theta}}(\boldsymbol{z})}{q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{y})}\right] \tag{1}$$
$$= \log p_{\boldsymbol{\theta}}(\boldsymbol{y}) - D_{\mathrm{KL}}(q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{y}) \,||\, p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{y}))$$

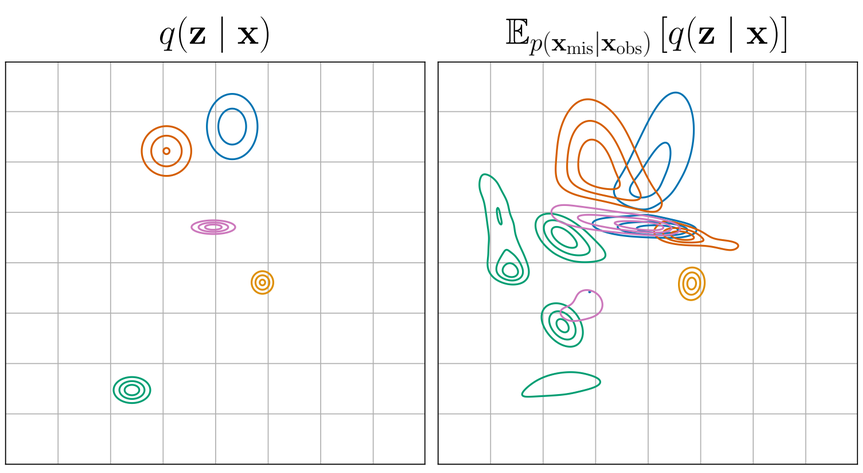
- When data is fully-observed $\boldsymbol{y} \stackrel{!}{=} \boldsymbol{x}$ and when incomplete $\boldsymbol{y} \stackrel{!}{=} \boldsymbol{x}_{\text{obs}}$, where $\boldsymbol{x}_{\text{obs}} = \boldsymbol{x}[\boldsymbol{m}]$ for a missingness pattern $\boldsymbol{m} \in \{0,1\}^D$.
- The objective for fully-observed and incomplete data is similar, so why is it typically harder to train a VAE from incomplete data?
- **Effective training needs to well-minimise the KL term:**
  - When data is fully-observed: $D_{\mathrm{KL}}(q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x}) \,||\, p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{x}))$,
  - When data is incomplete: $D_{\mathrm{KL}}(q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}) \,||\, p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}))$.



- Need more flexible variational distributions q with incomplete data!
- Potential loss of inductive biases when going from fully-observed to incomplete case, due to modifying the family of q.

Can we re-use variational family from complete case?

- Re-use complete-data approximation $q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x})$ for incomplete-data: $p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}) = \mathbb{E}_{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{mis}}|\boldsymbol{x}_{\text{obs}})}[p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{x})] \approx \mathbb{E}_{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{mis}}|\boldsymbol{x}_{\text{obs}})}[q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x})]$.



- Enables re-using the variational family from the fully-observed case.
- As a result we can save some of the useful inductive biases.
- But, the missing-variable distribution $p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$ in $\mathbb{E}_{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{mis}}|\boldsymbol{x}_{\text{obs}})}[q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x})]$ is intractable!

Opt. 1: Use finite-mixture variational distribution

- Make the approach more tractable by using finite-mixtures.
- Approximate $\mathbb{E}_{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{mis}}|\boldsymbol{x}_{\text{obs}})}[q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x})]$ with a finite-mixture:

$$q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}) = \sum_{k=1}^{K} q_{\boldsymbol{\phi}}(k \mid \boldsymbol{x}_{\text{obs}}) q_{\boldsymbol{\phi}}^k(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}).$$

- Where the component distributions $q_{\boldsymbol{\phi}}^k(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})$ re-use the complete-data variational family, and hence they may roughly take the role of $q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}})$ for some $\boldsymbol{x}_{\text{mis}}$.
- Optimise the model using eq. (1) or importance-weighted ELBO.
- Several options available that trade-off variance versus computational cost (see paper).

Opt. 2: Use imputation-mixture variational distribution

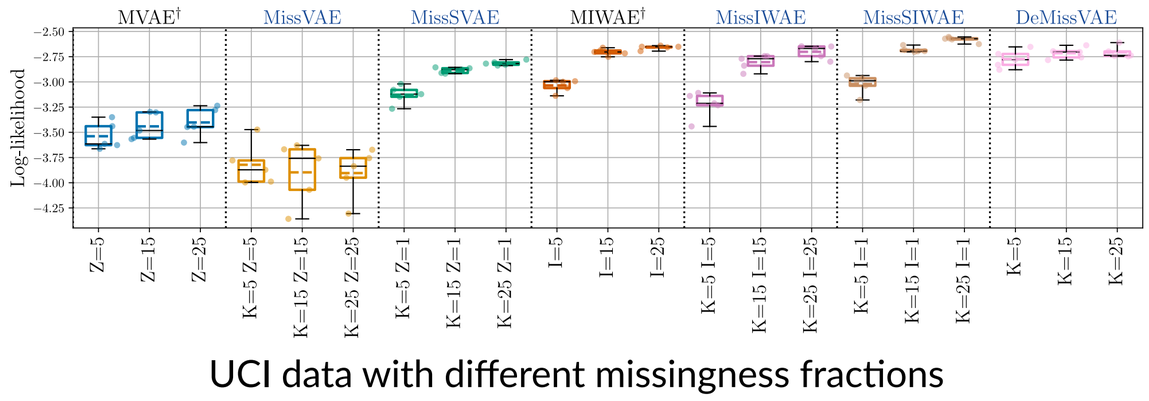
- **Decompose the model estimation task into two: data imputation and model estimation (similar to Monte Carlo EM).**
- Approximate $\mathbb{E}_{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{mis}}|\boldsymbol{x}_{\text{obs}})}[q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x})]$ with an imputation-mixture:

$$q_{\boldsymbol{\phi},f^t}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}) = \mathbb{E}_{f^t(\boldsymbol{x}_{\text{mis}}|\boldsymbol{x}_{\text{obs}})}\left[q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}})\right]$$

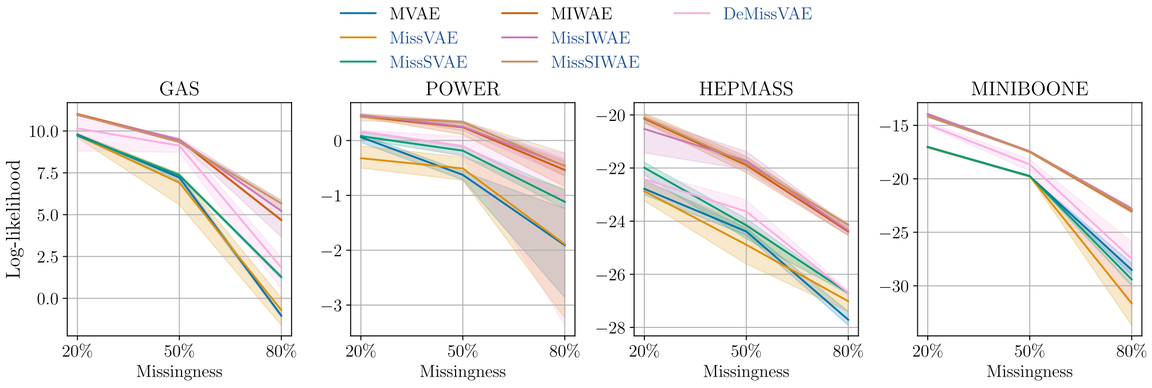
- The imputation distribution $f^t(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$ is represented using *approximate* samples from $p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$, using cheap sampling methods such as pseudo-Gibbs, Metropolis-within-Gibbs, or others.
- And, $q_{\boldsymbol{\phi}}$ is fitted such that $q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}}) \approx p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}})$.
- We thus have a flexible approximation $q_{\boldsymbol{\phi},f^t}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}}) \approx p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{x}_{\text{obs}})$, while retaining the variational family from the complete-data case.
- Method uses separate objectives for $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ to minimise the potential bias due to approximation errors of the imputation distribution $f^t(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$.

Results (proposed methods highlighted in blue)

Toy MoG data with varying number of latent samples per iteration



UCI data with different missingness fractions



MNIST and Omniglot missing 2/4 quadrants at random