# How to deal with missing data in supervised deep learning?

Niels Bruun Ipsen, Pierre-Alexandre Mattei, Jes Frellsen

July 15, 2020

How to approach $p(\boldsymbol{y}|\boldsymbol{x}^{\text{obs}}, \boldsymbol{x}^{\text{miss}})$ , assuming MAR?

$$\begin{aligned} &\text{0-imputation:} && \iota_0(\boldsymbol{x}^{\text{obs}}) = \boldsymbol{x} \odot \boldsymbol{s} + \boldsymbol{0} \odot (1 - \boldsymbol{s}) && (1) \\ &\text{learnable imputation:} && \iota_\lambda(\boldsymbol{x}^{\text{obs}}) = \boldsymbol{x} \odot \boldsymbol{s} + \boldsymbol{\lambda} \odot (1 - \boldsymbol{s}) && (2) \\ &\text{concatenation in separate channels:} && \iota_0(\boldsymbol{x}^{\text{obs}}),\ \boldsymbol{\lambda},\ \text{and } \boldsymbol{s} && (3) \end{aligned}$$

where $\mathbf{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^\intercal \in \mathcal{X}^n$ contain $n$ i.i.d. copies of the random variable $\boldsymbol{x} \in \mathcal{X}$ and the positions of observed entries in the data matrix $\mathbf{X}$ are contained in a mask matrix $\mathbf{S} = (\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n)^\intercal \in \{0, 1\}^{n \times p}$ and $\boldsymbol{x} = (\boldsymbol{x}^{\text{obs}}, \boldsymbol{x}^{\text{miss}})$.

$$\text{Deep Latent Variable Model, DLVM:} \quad p(\boldsymbol{z})p(\boldsymbol{x}|\boldsymbol{z}) \tag{4}$$

$$\text{Joint DLVM and discriminative model :} \quad p(\boldsymbol{z})p(\boldsymbol{x}^{\text{obs}}|\boldsymbol{z})p(\boldsymbol{x}^{\text{miss}}|\boldsymbol{z})p(\boldsymbol{y}|\boldsymbol{x}^{\text{obs}},\boldsymbol{x}^{\text{miss}}) \tag{5}$$

## Lower bound for training

$$\mathcal{L}_K = \mathbb{E}\left[\log\left(\frac{1}{K}\sum_{k=1}^{K}\frac{p(\boldsymbol{z}_k)p(\boldsymbol{x}^{\text{obs}}|\boldsymbol{z}_k)p(\boldsymbol{y}|\boldsymbol{x}^{\text{obs}},\boldsymbol{x}_k^{\text{miss}})}{q(\boldsymbol{z}_k|\boldsymbol{x}^{\text{obs}},\boldsymbol{s})}\right)\right] \leq \log p(\boldsymbol{y},\boldsymbol{x}^{\text{obs}}) \tag{6}$$

where $q(\boldsymbol{z}_k|\boldsymbol{x}^{\text{obs}},\boldsymbol{s})$ is the *variational distribution* (learnable proposal) and $(\boldsymbol{z}_k,\boldsymbol{x}_k^{\text{miss}})_{k\in\{1,\dots,K\}}$ are i.i.d. samples from $p(\boldsymbol{x}^{\text{miss}}|\boldsymbol{z})q(\boldsymbol{z}|\boldsymbol{x}^{\text{obs}},\boldsymbol{s})$.

## Prediction: self-normalized importance sampling

$$p(\boldsymbol{y}|\boldsymbol{x}^{\text{obs}}) \approx \sum_{i=1}^{K} w_k p(\boldsymbol{y}|\boldsymbol{x}^{\text{obs}}, \boldsymbol{x}_k^{\text{miss}}), \tag{7}$$

where

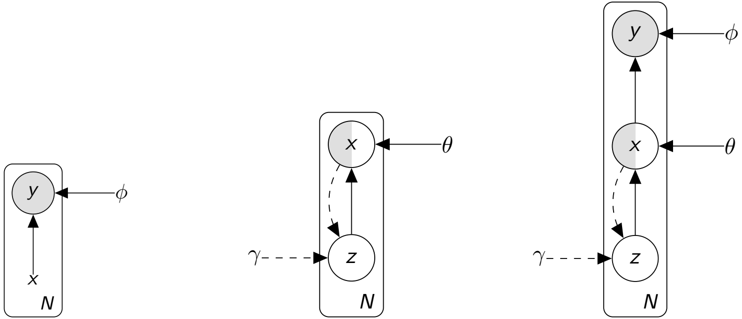
$$w_k = \frac{r_k}{r_1 + \ldots + r_K}, \quad \text{and} \quad r_k = \frac{p(\boldsymbol{z}_k)p(\boldsymbol{x}^{\text{obs}}|\boldsymbol{z}_k)}{q(\boldsymbol{z}_k|\boldsymbol{x}^{\text{obs}}, \boldsymbol{s})}, \tag{8}$$

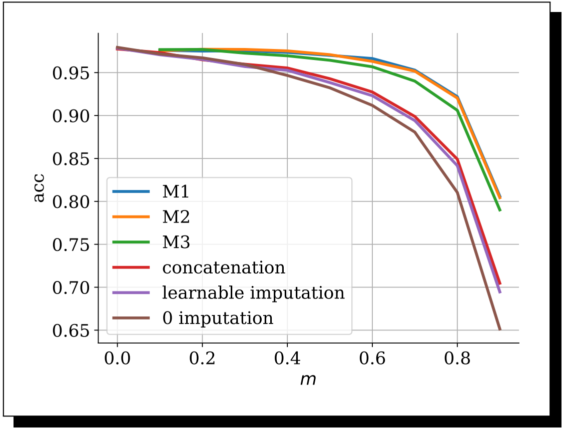
and $(z_k, \boldsymbol{x}_k^{\text{miss}})_{k\in\{1,\ldots,K\}}$ are i.i.d. samples from $p(\boldsymbol{x}^{\text{miss}}|\boldsymbol{z})q(\boldsymbol{z}|\boldsymbol{x}^{\text{obs}}, \boldsymbol{s})$.

Discriminative model: $$p_\phi(\boldsymbol{y}|\boldsymbol{x}) \quad (9)$$

Deep Latent Variable Model, DLVM: $$p(\boldsymbol{z})p_\theta(\boldsymbol{x}|\boldsymbol{z}) \quad (10)$$

Joint DLVM and discriminative model : $$p(\boldsymbol{z})p_\theta(\boldsymbol{x}^{\text{obs}}|\boldsymbol{z})p_\theta(\boldsymbol{x}^{\text{miss}}|\boldsymbol{z})p_\phi(\boldsymbol{y}|\boldsymbol{x}^{\text{obs}},\boldsymbol{x}^{\text{miss}}) \quad (11)$$






$M1$, joint model, trained jointly
$M2$, joint model, DLVM and discriminative model trained separately
$M3$, DLVM used for imputing. Discriminative model trained on imputed dataset.

# Thank You