Variational Gibbs Inference for Statistical Model Estimation from Incomplete Data

Vaidotas Šimkus Benjamin Rhodes Michael U. Gutmann

School of Informatics, The University of Edinburgh

About the paper

- General method for estimating models from incomplete data.
- Needs only D Gibbs conditionals $q_{\boldsymbol{\phi}_j}(x_j \mid \boldsymbol{x}_{\setminus j})$ instead of 2^D variational distributions (one for each pattern of missingness).
- Achieves better or competitive performance on normalising flow and VAE estimation from incomplete data.
- Published at Journal of Machine Learning Research, 2023.

Key challenges with deep statistical models

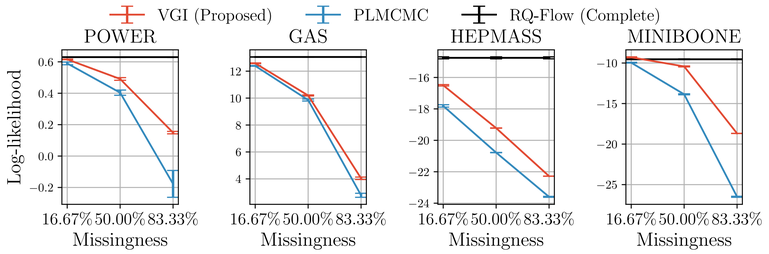
- The models $p_{\boldsymbol{\theta}}(\boldsymbol{x})$ are specified for fully-observed data $\boldsymbol{x} \in \mathcal{D}$.
- **Marginalising the missing variables $\int p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}})\, \mathrm{d}\boldsymbol{x}_{\text{mis}}$ is generally not tractable.** (With $\boldsymbol{x}_{\text{obs}} \cup \boldsymbol{x}_{\text{mis}} = \boldsymbol{x}$ and $\boldsymbol{x}_{\text{obs}} \cap \boldsymbol{x}_{\text{mis}} = \varnothing$.)
- Conditional sampling of $p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$ is inefficient.

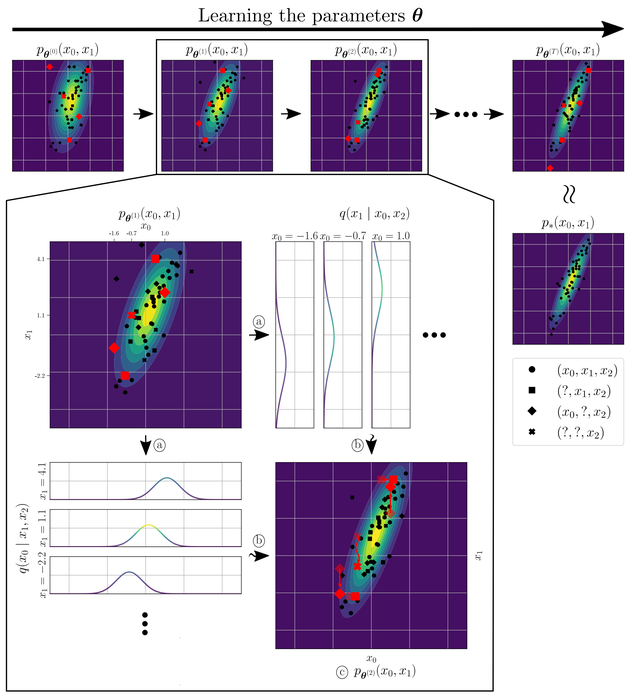
Why not use expectation maximisation?

$$\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{obs}}) \geqslant \mathbb{E}_{f(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})}\left[\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}})}{f(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})}\right], \qquad \text{“ELBO”}$$

- E-step: Maximise the ELBO w.r.t. $f(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}^i_{\text{obs}})$ for $\forall \boldsymbol{x}^i_{\text{obs}} \in \mathcal{D}$: $f(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}^i_{\text{obs}}) = p_{\boldsymbol{\theta}^t}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}^i_{\text{obs}})$.
- M-step: Maximise the ELBO w.r.t. $\boldsymbol{\theta}$: $\boldsymbol{\theta}^{t+1} = \arg\max_{\boldsymbol{\theta}} \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_{p_{\boldsymbol{\theta}^t}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}^i_{\text{obs}})}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}^i_{\text{obs}}, \boldsymbol{x}_{\text{mis}})\right]$
- Requires sampling $p_{\boldsymbol{\theta}^t}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}^i_{\text{obs}})$ for $\forall \boldsymbol{x}^i_{\text{obs}} \in \mathcal{D}$ → inefficient!

Results: Normalising flows





Issues with (amortised) variational inference

- $\forall \boldsymbol{x}^i_{\text{obs}} \in \mathcal{D}$ specify a $f_{\boldsymbol{\phi}}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}^i_{\text{obs}}) \in \mathcal{Q}(\boldsymbol{\phi})$.
- E-step: Maximise the ELBO w.r.t. $\boldsymbol{\phi}$. (Inference → optimisation.)
- Is inefficient if $|\mathcal{D}|$ is large.
- **Amortised VI:** Parametrise all $f_{\boldsymbol{\phi}}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}^i_{\text{obs}})$ with a *single* neural network $\text{NN}_{\boldsymbol{\phi}}(\boldsymbol{x}^i_{\text{obs}})$ shared for $\forall \boldsymbol{x}^i_{\text{obs}} \in \mathcal{D}$?
- Makes VI efficient for latent-variable models with large $|\mathcal{D}|$.
- But can we use amortised VI for estimation with missing data?

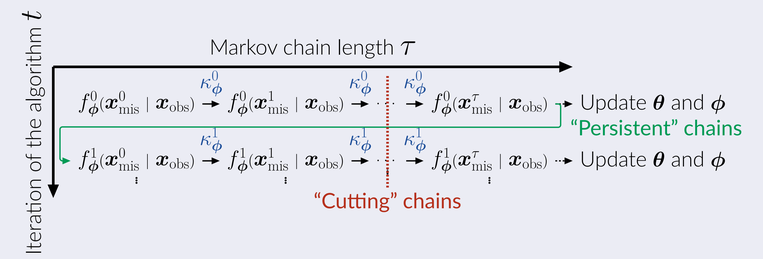
	d_1	d_2	d_3	d_4
$\boldsymbol{x}^1$	x^1_1	?	x^1_3	x^1_4
$\boldsymbol{x}^2$	?	x^2_2	x^2_3	?
$\boldsymbol{x}^3$	?	?	?	x^3_4
⋮			⋮	

$f_{\boldsymbol{\phi}}(\boldsymbol{x}^i_{\text{mis}} \mid \boldsymbol{x}^i_{\text{obs}})$
$f_{\boldsymbol{\phi}}(x^1_2 \mid x^1_1, x^1_3, x^1_4)$
$f_{\boldsymbol{\phi}}(x^2_1, x^2_4 \mid x^2_2, x^2_3)$
$f_{\boldsymbol{\phi}}(x^3_1, x^3_2, x^3_3 \mid x^3_4)$
⋮

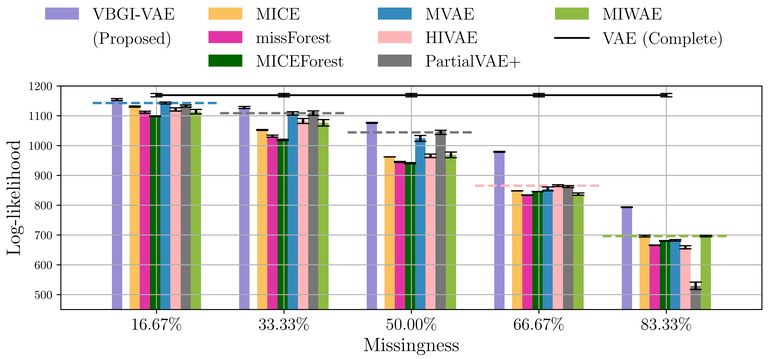
- Would require 2^D amortised variational distributions, one for each pattern of missingness → inefficient!

Variational Gibbs Inference

1. To make $f^t(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$ flexible: Specify it to be the marginal of a Markov chain with a *learnable* kernel $\kappa_{\boldsymbol{\phi}}(\boldsymbol{x}^{\tau+1}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{x}^{\tau}_{\text{mis}})$.
2. To tackle the 2^D scaling problem: Specify the kernel to be Gibbs (updates one dimension of $\boldsymbol{x}_{\text{mis}}$ at a time):
$$\kappa_{\boldsymbol{\phi}}(\boldsymbol{x}^{\tau+1}_{\text{mis}} \mid \boldsymbol{x}^{\tau}_{\text{mis}}, \boldsymbol{x}_{\text{obs}}) = \mathbb{E}_{\boldsymbol{\pi}(j)}\left[q_{\boldsymbol{\phi}_j}(x_j \mid \boldsymbol{x}^{\tau}_{\text{mis}\setminus j}, \boldsymbol{x}_{\text{obs}}) \delta_{\boldsymbol{x}^{\tau}_{\text{mis}\setminus j}}(\boldsymbol{x}^{\tau+1}_{\text{mis}\setminus j})\right]$$
3. To make inference efficient use “persistent” chains: initialise the chains at the last state of the previous iteration.
4. To make learning $\kappa_{\boldsymbol{\phi}}$ efficient avoid computing the marginal of the Markov chain $f^t(\boldsymbol{x}^{\tau}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$: “cut” the chains just before the last transition and optimise over the last step of the chain.



Results: Variational autoencoder



NEURAL INFORMATION PROCESSING SYSTEMS