Variational Gibbs Inference for Statistical Model Estimation from Incomplete Data

Vaidotas Šimkus Ben Rhodes Michael Gutmann

School of Informatics
The University of Edinburgh

December 2023

THE UNIVERSITY *of* EDINBURGH informatics

Statistical models and missing data problem

THE UNIVERSITY *of* EDINBURGH informatics

- Statistical models $p_{\boldsymbol{\theta}}(\boldsymbol{x})$ are typically specified for fully-observed data $\boldsymbol{x} \in \mathcal{D}$,
- And are often fitted via maximum-likelihood estimation (MLE).
- What can we do if part of the data is missing?

1. **Marginalising the missing variables** $\int p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}})\, \mathrm{d}\boldsymbol{x}_{\text{mis}}$ is generally **intractable**.
2. **Expectation-maximisation (EM)** requires sampling of $p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}}) \to$ **intractable**.

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

3. **Variational EM** requires fitting of $f_{\boldsymbol{\phi}}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$ for each $\boldsymbol{x}_{\text{obs}} \in \mathcal{D} \to$ **inefficient**.
4. **Amortised variational inference** requires 2^D variational distributions, one for each pattern of missingness $\to$ **inefficient**!

	d_1	d_2	d_3	d_4	$f_{\boldsymbol{\phi}}(\boldsymbol{x}^i_{\text{mis}} \mid \boldsymbol{x}^i_{\text{obs}})$
$\boldsymbol{x}^1$	x^1_1	?	x^1_3	x^1_4	$f_{\boldsymbol{\phi}}(x^1_2 \mid x^1_1, x^1_3, x^1_4)$
$\boldsymbol{x}^2$	?	x^2_2	x^2_3	?	$f_{\boldsymbol{\phi}}(x^2_1, x^2_4 \mid x^2_2, x^2_3)$
$\boldsymbol{x}^3$	?	?	?	x^3_4	$f_{\boldsymbol{\phi}}(x^3_1, x^3_2, x^3_3 \mid x^3_4)$
$\vdots$		$\vdots$			$\vdots$

A wish-list

THE UNIVERSITY of EDINBURGH informatics

- A general-purpose method for any statistical model $p_{\boldsymbol{\theta}}(\boldsymbol{x})$ via (approximate) MLE.
  - Do not make unnecessary simplifying assumptions to accommodate data missingness.
- Efficiently represent and sample the 2^D conditional distributions for large datasets.

Variational Gibbs Inference: Core idea

1. Core idea: Turn the 2^D conditional distribution problem into D conditional distributions.
2. To make $f^t_{\boldsymbol{\phi}}(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}})$ flexible:
   - Specify it to be the marginal of a Markov chain with a *learnable* kernel $\kappa_{\boldsymbol{\phi}}(\boldsymbol{x}^{\tau+1}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{x}^{\tau}_{\text{mis}})$.
3. To address the 2^D pattern problem:
   - We specify the kernel to be Gibbs (updates one dimension of $\boldsymbol{x}_{\text{mis}}$ at a time):

$$\kappa_{\boldsymbol{\phi}}(\boldsymbol{x}^{\tau+1}_{\text{mis}} \mid \boldsymbol{x}^{\tau}_{\text{mis}}, \boldsymbol{x}_{\text{obs}}) = \mathbb{E}_{\boldsymbol{\pi}(j \mid \mathrm{idx}(\boldsymbol{m}))}\left[q_{\boldsymbol{\phi}_j}(x_j \mid \boldsymbol{x}^{\tau}_{\text{mis}\smallsetminus j}, \boldsymbol{x}_{\text{obs}})\delta(\boldsymbol{x}^{\tau+1}_{\text{mis}\smallsetminus j} - \boldsymbol{x}^{\tau}_{\text{mis}\smallsetminus j})\right],$$

     where $\boldsymbol{\pi}(j \mid \mathrm{idx}(\boldsymbol{m}))$ is the selection probability for the j-th dimension of a Gibbs sampler.
   - Hence we have to learn only D variational Gibbs conditional $q_{\boldsymbol{\phi}_j}(x_j \mid \boldsymbol{x}_{\text{mis}\smallsetminus j}, \boldsymbol{x}_{\text{obs}})$.

See our JMLR paper for

- Full method: how to efficiently sample and optimise the transition kernel.
- Details on the variational model of the Gibbs conditionals.
- Applications to variational autoencoders and normalising flows.

References

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data Via the EM Algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22. (Cited on slide 2)

Gershman, S. J. and Goodman, N. D. (2014). Amortized Inference in Probabilistic Reasoning. In *Annual Meeting of the Cognitive Science Society*, volume 36. (Cited on slide 4)

Simkus, V., Rhodes, B., and Gutmann, M. U. (2023). Variational Gibbs Inference for Statistical Model Estimation from Incomplete Data. *Journal of Machine Learning Research*, 24(196):1–72. (Cited on slide 4)