

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 $p_{\theta}(\mathbf{x})$ are typically specified for fully-observed data $\mathbf{x} \in \mathcal{D}$,
- 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** $\int p_{\theta}(\mathbf{x}_{\text{obs}}, \mathbf{x}_{\text{mis}}) d\mathbf{x}_{\text{mis}}$ is generally **intractable**.
2. **Expectation-maximisation (EM)** requires sampling of $p_{\theta}(\mathbf{x}_{\text{mis}} | \mathbf{x}_{\text{obs}}) \rightarrow$ **intractable**.

$$\log p_{\theta}(\mathbf{x}_{\text{obs}}) \geq \mathbb{E}_{f(\mathbf{x}_{\text{mis}} | \mathbf{x}_{\text{obs}})} \left[\log \frac{p_{\theta}(\mathbf{x}_{\text{obs}}, \mathbf{x}_{\text{mis}})}{f(\mathbf{x}_{\text{mis}} | \mathbf{x}_{\text{obs}})} \right], \quad \text{“ELBO”}$$

3. **Variational EM** requires fitting of $f_{\phi}(\mathbf{x}_{\text{mis}} | \mathbf{x}_{\text{obs}})$ for each $\mathbf{x}_{\text{obs}} \in \mathcal{D} \rightarrow$ **inefficient**.

4. **Amortised variational inference**

requires 2^D variational distributions, one for each pattern of missingness $\rightarrow$ **inefficient!**

	d_1	d_2	d_3	d_4	$f_{\phi}(\mathbf{x}_{\text{mis}}^i \mathbf{x}_{\text{obs}}^i)$
$\mathbf{x}^1$	x_1^1	?	x_3^1	x_4^1	$f_{\phi}(x_2^1 x_1^1, x_3^1, x_4^1)$
$\mathbf{x}^2$	?	x_2^2	x_3^2	?	$f_{\phi}(x_1^2, x_4^2 x_2^2, x_3^2)$
$\mathbf{x}^3$	?	?	?	x_4^3	$f_{\phi}(x_1^3, x_2^3, x_3^3 x_4^3)$
$\vdots$	$\vdots$				$\vdots$

- A general-purpose method for any statistical model $p_{\theta}(x)$ 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.

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

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

where $\pi(j \mid \text{idx}(\mathbf{m}))$ is the selection probability for the j -th dimension of a Gibbs sampler.

- Hence we have to learn only D variational Gibbs conditional $q_{\phi_j}(x_j \mid \mathbf{x}_{\text{mis}_{\setminus j}}, \mathbf{x}_{\text{obs}})$.

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.





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)