Variational Gibbs Inference for Statistical Model Estimation from Incomplete Data

Vaidotas Šimkus Ben Rhodes Michael Gutmann

School of Informatics
The University of Edinburgh

November 2023

THE UNIVERSITY *of* EDINBURGH
informatics

Topic of the talk

THE UNIVERSITY of EDINBURGH informatics

Variational Gibbs Inference for Statistical Model Estimation from Incomplete Data

- General-purpose method for estimating statistical models from incomplete data.
- Journal of Machine Learning Research, 2023: jmlr.org/papers/v24/21-1373.html.
- Code: github.com/vsimkus/variational-gibbs-inference.
- Demo: nbviewer.org/github/vsimkus/variational-gibbs-inference/blob/main/notebooks/VGI_demo.ipynb.

Overview

1. Statistical models and the missing data issue
2. Some problems with direct estimation from incomplete data
3. Variational Gibbs Inference

Statistical models and the missing data issue

THE UNIVERSITY of EDINBURGH informatics

- Statistical models $p_{\boldsymbol{\theta}}(\boldsymbol{x})$ of complex phenomena are a key component of many important tasks in machine learning.

Statistical models and the missing data issue

THE UNIVERSITY *of* EDINBURGH **informatics**

- Statistical models $p_{\boldsymbol{\theta}}(\boldsymbol{x})$ of complex phenomena are a key component of many important tasks in machine learning.
- The models $p_{\boldsymbol{\theta}}(\boldsymbol{x})$ are specified for fully-observed data $\boldsymbol{x}$,

Statistical models and the missing data issue

- Statistical models $p_{\boldsymbol{\theta}}(\boldsymbol{x})$ of complex phenomena are a key component of many important tasks in machine learning.
- The models $p_{\boldsymbol{\theta}}(\boldsymbol{x})$ are specified for fully-observed data $\boldsymbol{x}$,
- And are typically fitted via maximum-likelihood estimation (MLE)

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \frac{1}{N} \sum_{i=1}^{N} \log p_{\boldsymbol{\theta}}(\boldsymbol{x}^i), \qquad \text{where} \qquad \boldsymbol{x}^i \in \mathcal{D}.$$

Statistical models and the missing data issue

- Statistical models $p_{\boldsymbol{\theta}}(\boldsymbol{x})$ of complex phenomena are a key component of many important tasks in machine learning.
- The models $p_{\boldsymbol{\theta}}(\boldsymbol{x})$ are specified for fully-observed data $\boldsymbol{x}$,
- And are typically fitted via maximum-likelihood estimation (MLE)

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \frac{1}{N} \sum_{i=1}^{N} \log p_{\boldsymbol{\theta}}(\boldsymbol{x}^i), \quad \text{where} \quad \boldsymbol{x}^i \in \mathcal{D}.$$

- Real-world data is often incomplete due to: non-response, sensor failure, occlusion, etc.

Statistical models and the missing data issue

- Statistical models $p_{\boldsymbol{\theta}}(\boldsymbol{x})$ of complex phenomena are a key component of many important tasks in machine learning.
- The models $p_{\boldsymbol{\theta}}(\boldsymbol{x})$ are specified for fully-observed data $\boldsymbol{x}$,
- And are typically fitted via maximum-likelihood estimation (MLE)

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \frac{1}{N} \sum_{i=1}^{N} \log p_{\boldsymbol{\theta}}(\boldsymbol{x}^i), \quad \text{where} \quad \boldsymbol{x}^i \in \mathcal{D}.$$

- Real-world data is often incomplete due to: non-response, sensor failure, occlusion, etc.
- What can we do?

Statistical models and the missing data issue

- Statistical models $p_{\boldsymbol{\theta}}(\boldsymbol{x})$ of complex phenomena are a key component of many important tasks in machine learning.
- The models $p_{\boldsymbol{\theta}}(\boldsymbol{x})$ are specified for fully-observed data $\boldsymbol{x}$,
- And are typically fitted via maximum-likelihood estimation (MLE)

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \frac{1}{N} \sum_{i=1}^{N} \log p_{\boldsymbol{\theta}}(\boldsymbol{x}^i), \quad \text{where} \quad \boldsymbol{x}^i \in \mathcal{D}.$$

- Real-world data is often incomplete due to: non-response, sensor failure, occlusion, etc.
- What can we do?
- Denote $\boldsymbol{x}_{\mathrm{o}}$ and $\boldsymbol{x}_{\mathrm{m}}$ as the observed and missing elements of $\boldsymbol{x} = \boldsymbol{x}_{\mathrm{o}} \cup \boldsymbol{x}_{\mathrm{m}}$.

Statistical models and the missing data issue

THE UNIVERSITY of EDINBURGH informatics

- Statistical models $p_{\boldsymbol{\theta}}(\boldsymbol{x})$ of complex phenomena are a key component of many important tasks in machine learning.
- The models $p_{\boldsymbol{\theta}}(\boldsymbol{x})$ are specified for fully-observed data $\boldsymbol{x}$,
- And are typically fitted via maximum-likelihood estimation (MLE)

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \frac{1}{N} \sum_{i=1}^{N} \log p_{\boldsymbol{\theta}}(\boldsymbol{x}^i), \quad \text{where} \quad \boldsymbol{x}^i \in \mathcal{D}.$$

- Real-world data is often incomplete due to: non-response, sensor failure, occlusion, etc.
- What can we do?
- Denote $\boldsymbol{x}_\mathrm{o}$ and $\boldsymbol{x}_\mathrm{m}$ as the observed and missing elements of $\boldsymbol{x} = \boldsymbol{x}_\mathrm{o} \cup \boldsymbol{x}_\mathrm{m}$.

Options:

1. Discard data-points with missing values

Statistical models and the missing data issue

- Statistical models $p_{\boldsymbol{\theta}}(\boldsymbol{x})$ of complex phenomena are a key component of many important tasks in machine learning.
- The models $p_{\boldsymbol{\theta}}(\boldsymbol{x})$ are specified for fully-observed data $\boldsymbol{x}$,
- And are typically fitted via maximum-likelihood estimation (MLE)

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \frac{1}{N} \sum_{i=1}^{N} \log p_{\boldsymbol{\theta}}(\boldsymbol{x}^i), \quad \text{where} \quad \boldsymbol{x}^i \in \mathcal{D}.$$

- Real-world data is often incomplete due to: non-response, sensor failure, occlusion, etc.
- What can we do?
- Denote $\boldsymbol{x}_{\mathrm{o}}$ and $\boldsymbol{x}_{\mathrm{m}}$ as the observed and missing elements of $\boldsymbol{x} = \boldsymbol{x}_{\mathrm{o}} \cup \boldsymbol{x}_{\mathrm{m}}$.

Options:

1. Discard data-points with missing values → loss of information, not sustainable, bias ✗

Statistical models and the missing data issue

- Statistical models $p_{\boldsymbol{\theta}}(\boldsymbol{x})$ of complex phenomena are a key component of many important tasks in machine learning.
- The models $p_{\boldsymbol{\theta}}(\boldsymbol{x})$ are specified for fully-observed data $\boldsymbol{x}$,
- And are typically fitted via maximum-likelihood estimation (MLE)

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \frac{1}{N} \sum_{i=1}^{N} \log p_{\boldsymbol{\theta}}(\boldsymbol{x}^i), \qquad \text{where} \qquad \boldsymbol{x}^i \in \mathcal{D}.$$

- Real-world data is often incomplete due to: non-response, sensor failure, occlusion, etc.
- What can we do?
- Denote $\boldsymbol{x}_\text{o}$ and $\boldsymbol{x}_\text{m}$ as the observed and missing elements of $\boldsymbol{x} = \boldsymbol{x}_\text{o} \cup \boldsymbol{x}_\text{m}$.

Options:

1. Discard data-points with missing values → loss of information, not sustainable, bias ✗
2. Impute-then-fit

Statistical models and the missing data issue

- Statistical models $p_{\boldsymbol{\theta}}(\boldsymbol{x})$ of complex phenomena are a key component of many important tasks in machine learning.
- The models $p_{\boldsymbol{\theta}}(\boldsymbol{x})$ are specified for fully-observed data $\boldsymbol{x}$,
- And are typically fitted via maximum-likelihood estimation (MLE)

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \frac{1}{N} \sum_{i=1}^{N} \log p_{\boldsymbol{\theta}}(\boldsymbol{x}^i), \qquad \text{where} \qquad \boldsymbol{x}^i \in \mathcal{D}.$$

- Real-world data is often incomplete due to: non-response, sensor failure, occlusion, etc.
- What can we do?
- Denote $\boldsymbol{x}_{\text{o}}$ and $\boldsymbol{x}_{\text{m}}$ as the observed and missing elements of $\boldsymbol{x} = \boldsymbol{x}_{\text{o}} \cup \boldsymbol{x}_{\text{m}}$.

Options:

1. Discard data-points with missing values → loss of information, not sustainable, bias ✗
2. Impute-then-fit → selecting appropriate imputation method, imputation incongeniality ✗

Statistical models and the missing data issue

THE UNIVERSITY *of* EDINBURGH informatics

- Statistical models $p_{\boldsymbol{\theta}}(\boldsymbol{x})$ of complex phenomena are a key component of many important tasks in machine learning.
- The models $p_{\boldsymbol{\theta}}(\boldsymbol{x})$ are specified for fully-observed data $\boldsymbol{x}$,
- And are typically fitted via maximum-likelihood estimation (MLE)

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \frac{1}{N} \sum_{i=1}^{N} \log p_{\boldsymbol{\theta}}(\boldsymbol{x}^i), \quad \text{where} \quad \boldsymbol{x}^i \in \mathcal{D}.$$

- Real-world data is often incomplete due to: non-response, sensor failure, occlusion, etc.
- What can we do?
- Denote $\boldsymbol{x}_{\mathrm{o}}$ and $\boldsymbol{x}_{\mathrm{m}}$ as the observed and missing elements of $\boldsymbol{x} = \boldsymbol{x}_{\mathrm{o}} \cup \boldsymbol{x}_{\mathrm{m}}$.

Options:

1. Discard data-points with missing values → loss of information, not sustainable, bias ✗
2. Impute-then-fit → selecting appropriate imputation method, imputation incongeniality ✗
3. Direct fitting by marginalising the missing variables $\boldsymbol{x}_{\mathrm{m}}$

Statistical models and the missing data issue

- Statistical models $p_{\boldsymbol{\theta}}(\boldsymbol{x})$ of complex phenomena are a key component of many important tasks in machine learning.
- The models $p_{\boldsymbol{\theta}}(\boldsymbol{x})$ are specified for fully-observed data $\boldsymbol{x}$,
- And are typically fitted via maximum-likelihood estimation (MLE)

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \frac{1}{N} \sum_{i=1}^{N} \log p_{\boldsymbol{\theta}}(\boldsymbol{x}^i), \quad \text{where} \quad \boldsymbol{x}^i \in \mathcal{D}.$$

- Real-world data is often incomplete due to: non-response, sensor failure, occlusion, etc.
- What can we do?
- Denote $\boldsymbol{x}_{\mathrm{o}}$ and $\boldsymbol{x}_{\mathrm{m}}$ as the observed and missing elements of $\boldsymbol{x} = \boldsymbol{x}_{\mathrm{o}} \cup \boldsymbol{x}_{\mathrm{m}}$.

Options:

1. Discard data-points with missing values → loss of information, not sustainable, bias ✗
2. Impute-then-fit → selecting appropriate imputation method, imputation incongeniality ✗
3. Direct fitting by marginalising the missing variables $\boldsymbol{x}_{\mathrm{m}}$?

Overview

1. Statistical models and the missing data issue
2. Some problems with direct estimation from incomplete data
3. Variational Gibbs Inference

Expectation Maximisation (EM)

- Marginalising the missing variables $\int p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\mathsf{o}}, \boldsymbol{x}_{\mathsf{m}}) \, \mathrm{d}\boldsymbol{x}_{\mathsf{m}}$ is generally not tractable.

Expectation Maximisation (EM)

- Marginalising the missing variables $\int p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\mathsf{o}}, \boldsymbol{x}_{\mathsf{m}})\,\mathrm{d}\boldsymbol{x}_{\mathsf{m}}$ is generally not tractable.
- What can we do if simplifying assumptions cannot be inserted?

Expectation Maximisation (EM)

- Marginalising the missing variables $\int p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\mathsf{o}}, \boldsymbol{x}_{\mathsf{m}})\,\mathrm{d}\boldsymbol{x}_{\mathsf{m}}$ is generally not tractable.
- What can we do if simplifying assumptions cannot be inserted?
- Expectation-maximisation (EM) (assuming ignorable missingness)

$$\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\mathsf{o}}) = \log \int f(\boldsymbol{x}_{\mathsf{m}} \mid \boldsymbol{x}_{\mathsf{o}}) \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\mathsf{o}}, \boldsymbol{x}_{\mathsf{m}})}{f(\boldsymbol{x}_{\mathsf{m}} \mid \boldsymbol{x}_{\mathsf{o}})}\,\mathrm{d}\boldsymbol{x}_{\mathsf{m}} \geqslant \mathbb{E}_{f(\boldsymbol{x}_{\mathsf{m}} \mid \boldsymbol{x}_{\mathsf{o}})}\left[\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\mathsf{o}}, \boldsymbol{x}_{\mathsf{m}})}{f(\boldsymbol{x}_{\mathsf{m}} \mid \boldsymbol{x}_{\mathsf{o}})}\right], \quad \text{“ELBO”}$$

Expectation Maximisation (EM)

- Marginalising the missing variables $\int p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\mathsf{o}}, \boldsymbol{x}_{\mathsf{m}})\, \mathrm{d}\boldsymbol{x}_{\mathsf{m}}$ is generally not tractable.
- What can we do if simplifying assumptions cannot be inserted?
- Expectation-maximisation (EM) (assuming ignorable missingness)

$$\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\mathsf{o}}) = \log \int f(\boldsymbol{x}_{\mathsf{m}} \mid \boldsymbol{x}_{\mathsf{o}}) \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\mathsf{o}}, \boldsymbol{x}_{\mathsf{m}})}{f(\boldsymbol{x}_{\mathsf{m}} \mid \boldsymbol{x}_{\mathsf{o}})}\, \mathrm{d}\boldsymbol{x}_{\mathsf{m}} \geqslant \mathbb{E}_{f(\boldsymbol{x}_{\mathsf{m}} \mid \boldsymbol{x}_{\mathsf{o}})}\left[\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\mathsf{o}}, \boldsymbol{x}_{\mathsf{m}})}{f(\boldsymbol{x}_{\mathsf{m}} \mid \boldsymbol{x}_{\mathsf{o}})}\right], \quad \text{“ELBO”}$$

  - **E-step:** Maximise w.r.t. $f(\boldsymbol{x}_{\mathsf{m}} \mid \boldsymbol{x}_{\mathsf{o}}^i)$ for $\forall \boldsymbol{x}_{\mathsf{o}}^i \in \mathcal{D}$: $f(\boldsymbol{x}_{\mathsf{m}} \mid \boldsymbol{x}_{\mathsf{o}}^i) = p_{\boldsymbol{\theta}^t}(\boldsymbol{x}_{\mathsf{m}} \mid \boldsymbol{x}_{\mathsf{o}}^i)$.

Expectation Maximisation (EM)

- Marginalising the missing variables $\int p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\mathsf{o}}, \boldsymbol{x}_{\mathsf{m}})\, \mathrm{d}\boldsymbol{x}_{\mathsf{m}}$ is generally not tractable.
- What can we do if simplifying assumptions cannot be inserted?
- Expectation-maximisation (EM) (assuming ignorable missingness)

$$\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\mathsf{o}}) = \log \int f(\boldsymbol{x}_{\mathsf{m}} \mid \boldsymbol{x}_{\mathsf{o}}) \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\mathsf{o}}, \boldsymbol{x}_{\mathsf{m}})}{f(\boldsymbol{x}_{\mathsf{m}} \mid \boldsymbol{x}_{\mathsf{o}})}\, \mathrm{d}\boldsymbol{x}_{\mathsf{m}} \geqslant \mathbb{E}_{f(\boldsymbol{x}_{\mathsf{m}} \mid \boldsymbol{x}_{\mathsf{o}})}\left[\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\mathsf{o}}, \boldsymbol{x}_{\mathsf{m}})}{f(\boldsymbol{x}_{\mathsf{m}} \mid \boldsymbol{x}_{\mathsf{o}})}\right], \quad \text{"ELBO"}$$

  - **E-step:** Maximise w.r.t. $f(\boldsymbol{x}_{\mathsf{m}} \mid \boldsymbol{x}_{\mathsf{o}}^i)$ for $\forall \boldsymbol{x}_{\mathsf{o}}^i \in \mathcal{D}$: $f(\boldsymbol{x}_{\mathsf{m}} \mid \boldsymbol{x}_{\mathsf{o}}^i) = p_{\boldsymbol{\theta}^t}(\boldsymbol{x}_{\mathsf{m}} \mid \boldsymbol{x}_{\mathsf{o}}^i)$.
  - **M-step:** Maximise w.r.t. $\boldsymbol{\theta}$: $\boldsymbol{\theta}^{t+1} = \arg\max_{\boldsymbol{\theta}} \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_{p_{\boldsymbol{\theta}^t}(\boldsymbol{x}_{\mathsf{m}} \mid \boldsymbol{x}_{\mathsf{o}}^i)}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\mathsf{o}}^i, \boldsymbol{x}_{\mathsf{m}})\right]$

Expectation Maximisation (EM)

- Marginalising the missing variables $\int p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\mathsf{o}}, \boldsymbol{x}_{\mathsf{m}})\, \mathrm{d}\boldsymbol{x}_{\mathsf{m}}$ is generally not tractable.
- What can we do if simplifying assumptions cannot be inserted?
- Expectation-maximisation (EM) (assuming ignorable missingness)

$$\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\mathsf{o}}) = \log \int f(\boldsymbol{x}_{\mathsf{m}} \mid \boldsymbol{x}_{\mathsf{o}}) \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\mathsf{o}}, \boldsymbol{x}_{\mathsf{m}})}{f(\boldsymbol{x}_{\mathsf{m}} \mid \boldsymbol{x}_{\mathsf{o}})}\, \mathrm{d}\boldsymbol{x}_{\mathsf{m}} \geqslant \mathbb{E}_{f(\boldsymbol{x}_{\mathsf{m}} \mid \boldsymbol{x}_{\mathsf{o}})} \left[\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\mathsf{o}}, \boldsymbol{x}_{\mathsf{m}})}{f(\boldsymbol{x}_{\mathsf{m}} \mid \boldsymbol{x}_{\mathsf{o}})} \right], \quad \text{“ELBO”}$$

  - **E-step:** Maximise w.r.t. $f(\boldsymbol{x}_{\mathsf{m}} \mid \boldsymbol{x}_{\mathsf{o}}^i)$ for $\forall \boldsymbol{x}_{\mathsf{o}}^i \in \mathcal{D}$: $f(\boldsymbol{x}_{\mathsf{m}} \mid \boldsymbol{x}_{\mathsf{o}}^i) = p_{\boldsymbol{\theta}^t}(\boldsymbol{x}_{\mathsf{m}} \mid \boldsymbol{x}_{\mathsf{o}}^i)$.
  - **M-step:** Maximise w.r.t. $\boldsymbol{\theta}$: $\boldsymbol{\theta}^{t+1} = \arg\max_{\boldsymbol{\theta}} \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_{p_{\boldsymbol{\theta}^t}(\boldsymbol{x}_{\mathsf{m}} \mid \boldsymbol{x}_{\mathsf{o}}^i)} \left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\mathsf{o}}^i, \boldsymbol{x}_{\mathsf{m}}) \right]$
- Monte Carlo EM: Approximate the expectation using Monte Carlo average.

Expectation Maximisation (EM)

- Marginalising the missing variables $\int p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\mathsf{o}}, \boldsymbol{x}_{\mathsf{m}})\,\mathrm{d}\boldsymbol{x}_{\mathsf{m}}$ is generally not tractable.
- What can we do if simplifying assumptions cannot be inserted?
- Expectation-maximisation (EM) (assuming ignorable missingness)

$$\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\mathsf{o}}) = \log \int f(\boldsymbol{x}_{\mathsf{m}} \mid \boldsymbol{x}_{\mathsf{o}}) \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\mathsf{o}}, \boldsymbol{x}_{\mathsf{m}})}{f(\boldsymbol{x}_{\mathsf{m}} \mid \boldsymbol{x}_{\mathsf{o}})}\,\mathrm{d}\boldsymbol{x}_{\mathsf{m}} \geqslant \mathbb{E}_{f(\boldsymbol{x}_{\mathsf{m}}|\boldsymbol{x}_{\mathsf{o}})}\left[\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\mathsf{o}}, \boldsymbol{x}_{\mathsf{m}})}{f(\boldsymbol{x}_{\mathsf{m}} \mid \boldsymbol{x}_{\mathsf{o}})}\right], \quad \text{“ELBO”}$$

  - **E-step:** Maximise w.r.t. $f(\boldsymbol{x}_{\mathsf{m}} \mid \boldsymbol{x}_{\mathsf{o}}^i)$ for $\forall \boldsymbol{x}_{\mathsf{o}}^i \in \mathcal{D}$: $f(\boldsymbol{x}_{\mathsf{m}} \mid \boldsymbol{x}_{\mathsf{o}}^i) = p_{\boldsymbol{\theta}^t}(\boldsymbol{x}_{\mathsf{m}} \mid \boldsymbol{x}_{\mathsf{o}}^i)$.
  - **M-step:** Maximise w.r.t. $\boldsymbol{\theta}$: $\boldsymbol{\theta}^{t+1} = \arg\max_{\boldsymbol{\theta}} \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{p_{\boldsymbol{\theta}^t}(\boldsymbol{x}_{\mathsf{m}}|\boldsymbol{x}_{\mathsf{o}}^i)}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\mathsf{o}}^i, \boldsymbol{x}_{\mathsf{m}})\right]$
- Monte Carlo EM: Approximate the expectation using Monte Carlo average.
- Then, M-step corresponds to fitting $p_{\boldsymbol{\theta}}(\boldsymbol{x})$ with completed data.

Expectation Maximisation (EM)

- Marginalising the missing variables $\int p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\mathsf{o}}, \boldsymbol{x}_{\mathsf{m}})\, \mathrm{d}\boldsymbol{x}_{\mathsf{m}}$ is generally not tractable.
- What can we do if simplifying assumptions cannot be inserted?
- Expectation-maximisation (EM) (assuming ignorable missingness)

$$\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\mathsf{o}}) = \log \int f(\boldsymbol{x}_{\mathsf{m}} \mid \boldsymbol{x}_{\mathsf{o}}) \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\mathsf{o}}, \boldsymbol{x}_{\mathsf{m}})}{f(\boldsymbol{x}_{\mathsf{m}} \mid \boldsymbol{x}_{\mathsf{o}})}\, \mathrm{d}\boldsymbol{x}_{\mathsf{m}} \geqslant \mathbb{E}_{f(\boldsymbol{x}_{\mathsf{m}} \mid \boldsymbol{x}_{\mathsf{o}})}\left[\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\mathsf{o}}, \boldsymbol{x}_{\mathsf{m}})}{f(\boldsymbol{x}_{\mathsf{m}} \mid \boldsymbol{x}_{\mathsf{o}})}\right], \quad \text{“ELBO”}$$

  - **E-step:** Maximise w.r.t. $f(\boldsymbol{x}_{\mathsf{m}} \mid \boldsymbol{x}_{\mathsf{o}}^i)$ for $\forall \boldsymbol{x}_{\mathsf{o}}^i \in \mathcal{D}$: $f(\boldsymbol{x}_{\mathsf{m}} \mid \boldsymbol{x}_{\mathsf{o}}^i) = p_{\boldsymbol{\theta}^t}(\boldsymbol{x}_{\mathsf{m}} \mid \boldsymbol{x}_{\mathsf{o}}^i)$.
  - **M-step:** Maximise w.r.t. $\boldsymbol{\theta}$: $\boldsymbol{\theta}^{t+1} = \arg\max_{\boldsymbol{\theta}} \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_{p_{\boldsymbol{\theta}^t}(\boldsymbol{x}_{\mathsf{m}} \mid \boldsymbol{x}_{\mathsf{o}}^i)} \left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\mathsf{o}}^i, \boldsymbol{x}_{\mathsf{m}})\right]$
- Monte Carlo EM: Approximate the expectation using Monte Carlo average.
- Then, M-step corresponds to fitting $p_{\boldsymbol{\theta}}(\boldsymbol{x})$ with completed data.

Issue with Monte Carlo EM

- Conditional sampling of $p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\mathsf{m}} \mid \boldsymbol{x}_{\mathsf{o}})$ is generally intractable or inefficient.

Variational approximation to $p_{\boldsymbol{\theta}^t}(\boldsymbol{x}_\mathsf{m} \mid \boldsymbol{x}_\mathsf{o})$

THE UNIVERSITY of EDINBURGH informatics

Variational inference (VI)

- $\forall \boldsymbol{x}_\mathsf{o} \in \mathcal{D}$ specify a $f_{\boldsymbol{\phi}}(\boldsymbol{x}_\mathsf{m} \mid \boldsymbol{x}_\mathsf{o}) \in \mathcal{Q}(\boldsymbol{\phi})$.
- E-step: Maximise the ELBO w.r.t. ϕ.
- M-step: Sample $f_{\boldsymbol{\phi}}(\boldsymbol{x}_\mathsf{m} \mid \boldsymbol{x}_\mathsf{o})$ to approximate the expectation.

Variational approximation to $p_{\boldsymbol{\theta}^t}(\boldsymbol{x}_\mathrm{m} \mid \boldsymbol{x}_\mathrm{o})$

Variational inference (VI)

- $\forall \boldsymbol{x}_\mathrm{o} \in \mathcal{D}$ specify a $f_{\boldsymbol{\phi}}(\boldsymbol{x}_\mathrm{m} \mid \boldsymbol{x}_\mathrm{o}) \in \mathcal{Q}(\boldsymbol{\phi})$.
- E-step: Maximise the ELBO w.r.t. $\boldsymbol{\phi}$.
- M-step: Sample $f_{\boldsymbol{\phi}}(\boldsymbol{x}_\mathrm{m} \mid \boldsymbol{x}_\mathrm{o})$ to approximate the expectation.

Advantages of VI

- Choice of $\mathcal{Q}(\boldsymbol{\phi})$ is in our control.
- Turns inference to optimisation.
- Can fit using SGD.
- Efficient if $|\mathcal{D}|$ is small.

Variational approximation to $p_{\boldsymbol{\theta}^t}(\boldsymbol{x}_\mathrm{m} \mid \boldsymbol{x}_\mathrm{o})$

THE UNIVERSITY of EDINBURGH informatics

Variational inference (VI)

- $\forall \boldsymbol{x}_\mathrm{o} \in \mathcal{D}$ specify a $f_\phi(\boldsymbol{x}_\mathrm{m} \mid \boldsymbol{x}_\mathrm{o}) \in \mathcal{Q}(\phi)$.
- E-step: Maximise the ELBO w.r.t. ϕ.
- M-step: Sample $f_\phi(\boldsymbol{x}_\mathrm{m} \mid \boldsymbol{x}_\mathrm{o})$ to approximate the expectation.

Advantages of VI

- Choice of $\mathcal{Q}(\phi)$ is in our control.
- Turns inference to optimisation.
- Can fit using SGD.
- Efficient if $|\mathcal{D}|$ is small.

Disadvantages of VI

- Is inefficient if $|\mathcal{D}|$ is large.

Variational approximation to $p_{\boldsymbol{\theta}^t}(\boldsymbol{x}_\mathrm{m} \mid \boldsymbol{x}_\mathrm{o})$

Variational inference (VI)

- $\forall \boldsymbol{x}_\mathrm{o} \in \mathcal{D}$ specify a $f_\phi(\boldsymbol{x}_\mathrm{m} \mid \boldsymbol{x}_\mathrm{o}) \in \mathcal{Q}(\phi)$.
- E-step: Maximise the ELBO w.r.t. ϕ.
- M-step: Sample $f_\phi(\boldsymbol{x}_\mathrm{m} \mid \boldsymbol{x}_\mathrm{o})$ to approximate the expectation.

Amortised VI

- Parametrise $f_\phi(\boldsymbol{x}_\mathrm{m} \mid \boldsymbol{x}_\mathrm{o})$ with a *single* neural network $\mathrm{NN}_\phi(\boldsymbol{x}_\mathrm{o})$ for $\forall \boldsymbol{x}_\mathrm{o} \in \mathcal{D}$.

Advantages of VI

- Choice of $\mathcal{Q}(\phi)$ is in our control.
- Turns inference to optimisation.
- Can fit using SGD.
- Efficient if $|\mathcal{D}|$ is small.

Disadvantages of VI

- Is inefficient if $|\mathcal{D}|$ is large.

Variational approximation to $p_{\boldsymbol{\theta}^t}(\boldsymbol{x}_\mathrm{m} \mid \boldsymbol{x}_\mathrm{o})$

THE UNIVERSITY of EDINBURGH informatics

Variational inference (VI)

- $\forall \boldsymbol{x}_\mathrm{o} \in \mathcal{D}$ specify a $f_\phi(\boldsymbol{x}_\mathrm{m} \mid \boldsymbol{x}_\mathrm{o}) \in \mathcal{Q}(\phi)$.
- E-step: Maximise the ELBO w.r.t. ϕ.
- M-step: Sample $f_\phi(\boldsymbol{x}_\mathrm{m} \mid \boldsymbol{x}_\mathrm{o})$ to approximate the expectation.

Amortised VI

- Parametrise $f_\phi(\boldsymbol{x}_\mathrm{m} \mid \boldsymbol{x}_\mathrm{o})$ with a *single* neural network $\mathrm{NN}_\phi(\boldsymbol{x}_\mathrm{o})$ for $\forall \boldsymbol{x}_\mathrm{o} \in \mathcal{D}$.

Advantages of VI

- Choice of $\mathcal{Q}(\phi)$ is in our control.
- Turns inference to optimisation.
- Can fit using SGD.
- Efficient if $|\mathcal{D}|$ is small.

Disadvantages of VI

- Is inefficient if $|\mathcal{D}|$ is large.

Advantages of amortised VI

- Efficient for large $|\mathcal{D}|$.

Variational approximation to $p_{\boldsymbol{\theta}^t}(\boldsymbol{x}_\mathrm{m} \mid \boldsymbol{x}_\mathrm{o})$

Variational inference (VI)

- $\forall \boldsymbol{x}_\mathrm{o} \in \mathcal{D}$ specify a $f_{\boldsymbol{\phi}}(\boldsymbol{x}_\mathrm{m} \mid \boldsymbol{x}_\mathrm{o}) \in \mathcal{Q}(\boldsymbol{\phi})$.
- E-step: Maximise the ELBO w.r.t. $\boldsymbol{\phi}$.
- M-step: Sample $f_{\boldsymbol{\phi}}(\boldsymbol{x}_\mathrm{m} \mid \boldsymbol{x}_\mathrm{o})$ to approximate the expectation.

Amortised VI

- Parametrise $f_{\boldsymbol{\phi}}(\boldsymbol{x}_\mathrm{m} \mid \boldsymbol{x}_\mathrm{o})$ with a *single* neural network $\mathrm{NN}_{\boldsymbol{\phi}}(\boldsymbol{x}_\mathrm{o})$ for $\forall \boldsymbol{x}_\mathrm{o} \in \mathcal{D}$.

	d_1	d_2	d_3	d_4	$f_{\boldsymbol{\phi}}(\boldsymbol{x}^i_\mathrm{m} \mid \boldsymbol{x}^i_\mathrm{o})$
$\boldsymbol{x}^1$	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	$f_{\boldsymbol{\phi}}(x^3_1, x^3_2, x^3_3 \mid x^3_4)$
$\vdots$		$\vdots$			$\vdots$

Advantages of VI

- Choice of $\mathcal{Q}(\boldsymbol{\phi})$ is in our control.
- Turns inference to optimisation.
- Can fit using SGD.
- Efficient if $|\mathcal{D}|$ is small.

Disadvantages of VI

- Is inefficient if $|\mathcal{D}|$ is large.

Advantages of amortised VI

- Efficient for large $|\mathcal{D}|$.

Disadvantages of amortised VI

- Need one $f_{\boldsymbol{\phi}}(\boldsymbol{x}_\mathrm{m} \mid \boldsymbol{x}_\mathrm{o})$ for each pattern of missingness (2^M in total).

Overview

1. Statistical models and the missing data issue
2. Some problems with direct estimation from incomplete data
3. Variational Gibbs Inference

Variational Gibbs Inference: Core idea

Variational Gibbs Inference for Statistical Model Estimation from Incomplete Data, JMLR, 2023

- General-purpose method for estimating $p_{\boldsymbol{\theta}}(\boldsymbol{x})$ from incomplete data.
- Efficient for large $|\mathcal{D}|$ and mitigates the need for 2^M conditional distributions.

Variational Gibbs Inference: Core idea

THE UNIVERSITY *of* EDINBURGH informatics

Variational Gibbs Inference for Statistical Model Estimation from Incomplete Data, JMLR, 2023

- General-purpose method for estimating $p_{\boldsymbol{\theta}}(\boldsymbol{x})$ from incomplete data.
- Efficient for large $|\mathcal{D}|$ and mitigates the need for 2^M conditional distributions.

1. Core idea: Turn the 2^M conditional distribution problem into M conditional distributions.

Variational Gibbs Inference: Core idea

THE UNIVERSITY *of* EDINBURGH informatics

Variational Gibbs Inference for Statistical Model Estimation from Incomplete Data, JMLR, 2023

- General-purpose method for estimating $p_{\boldsymbol{\theta}}(\boldsymbol{x})$ from incomplete data.
- Efficient for large $|\mathcal{D}|$ and mitigates the need for 2^M conditional distributions.

1. Core idea: Turn the 2^M conditional distribution problem into M conditional distributions.
2. To make $f^t_{\boldsymbol{\phi}}(\boldsymbol{x}_{\mathsf{m}} \mid \boldsymbol{x}_{\mathsf{o}})$ flexible:

Variational Gibbs Inference: Core idea

THE UNIVERSITY *of* EDINBURGH informatics

Variational Gibbs Inference for Statistical Model Estimation from Incomplete Data, JMLR, 2023

- General-purpose method for estimating $p_{\boldsymbol{\theta}}(\boldsymbol{x})$ from incomplete data.
- Efficient for large $|\mathcal{D}|$ and mitigates the need for 2^M conditional distributions.

1. Core idea: Turn the 2^M conditional distribution problem into M conditional distributions.
2. To make $f^t_{\boldsymbol{\phi}}(\boldsymbol{x}_\mathrm{m} \mid \boldsymbol{x}_\mathrm{o})$ flexible:
   - Specify it to be the marginal of a Markov chain with a *learnable* kernel $\kappa_{\boldsymbol{\phi}}(\boldsymbol{x}^{\tau+1}_\mathrm{m} \mid \boldsymbol{x}_\mathrm{o}, \boldsymbol{x}^{\tau}_\mathrm{m})$.

Variational Gibbs Inference: Core idea

Variational Gibbs Inference for Statistical Model Estimation from Incomplete Data, JMLR, 2023

- General-purpose method for estimating $p_{\boldsymbol{\theta}}(\boldsymbol{x})$ from incomplete data.
- Efficient for large $|\mathcal{D}|$ and mitigates the need for 2^M conditional distributions.

1. Core idea: Turn the 2^M conditional distribution problem into M conditional distributions.
2. To make $f^t_{\boldsymbol{\phi}}(\boldsymbol{x}_\mathsf{m} \mid \boldsymbol{x}_\mathsf{o})$ flexible:
    - Specify it to be the marginal of a Markov chain with a *learnable* kernel $\kappa_{\boldsymbol{\phi}}(\boldsymbol{x}^{\tau+1}_\mathsf{m} \mid \boldsymbol{x}_\mathsf{o}, \boldsymbol{x}^{\tau}_\mathsf{m})$.
3. To address the 2^M pattern problem:

Variational Gibbs Inference: Core idea

THE UNIVERSITY *of* EDINBURGH informatics

Variational Gibbs Inference for Statistical Model Estimation from Incomplete Data, JMLR, 2023

- General-purpose method for estimating $p_{\boldsymbol{\theta}}(\boldsymbol{x})$ from incomplete data.
- Efficient for large $|\mathcal{D}|$ and mitigates the need for 2^M conditional distributions.

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

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

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

Variational Gibbs Inference: Core idea

THE UNIVERSITY of EDINBURGH informatics

Variational Gibbs Inference for Statistical Model Estimation from Incomplete Data, JMLR, 2023

- General-purpose method for estimating $p_{\boldsymbol{\theta}}(\boldsymbol{x})$ from incomplete data.
- Efficient for large $|\mathcal{D}|$ and mitigates the need for 2^M conditional distributions.

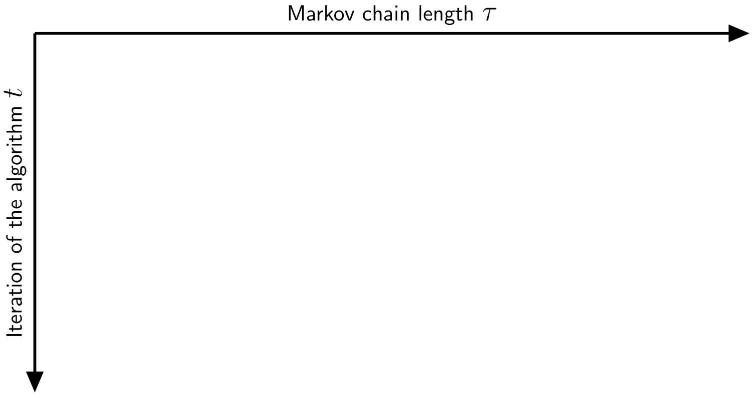
1. Core idea: Turn the 2^M conditional distribution problem into M conditional distributions.
2. To make $f^t_{\boldsymbol{\phi}}(\boldsymbol{x}_\mathrm{m} \mid \boldsymbol{x}_\mathrm{o})$ flexible:
   - Specify it to be the marginal of a Markov chain with a *learnable* kernel $\kappa_{\boldsymbol{\phi}}(\boldsymbol{x}^{\tau+1}_\mathrm{m}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

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

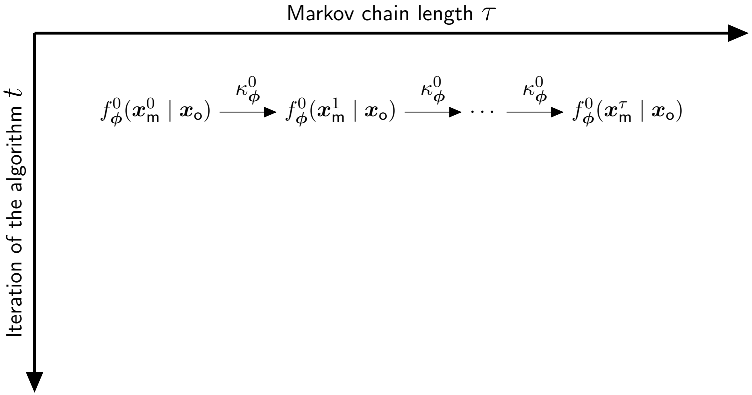
     where $\boldsymbol{\pi}(j \mid \mathrm{idx}(\boldsymbol{m}))$ is the selection probability for the j-th dimension of a Gibbs sampler.
   - Hence we have to learn only M variational Gibbs conditional $q_{\boldsymbol{\phi}_j}(x_j \mid \boldsymbol{x}_{\mathrm{m}\smallsetminus j}, \boldsymbol{x}_\mathrm{o})$.

Variational Gibbs Inference: Persistent chains

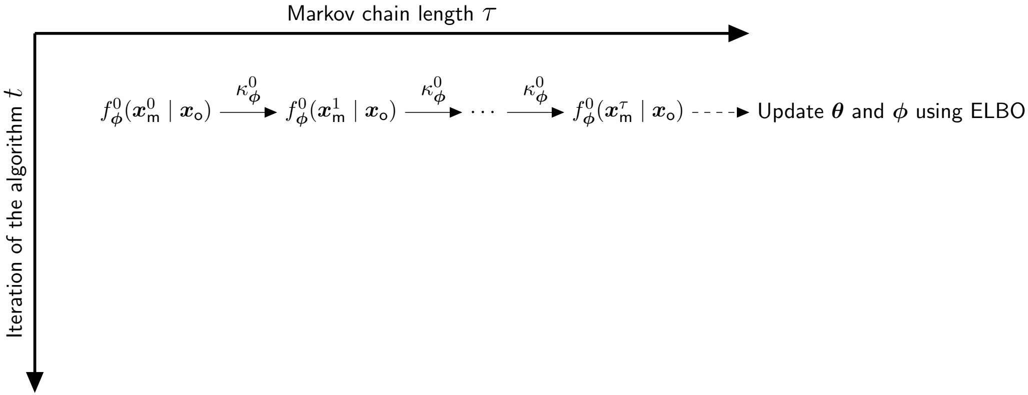
THE UNIVERSITY *of* EDINBURGH informatics



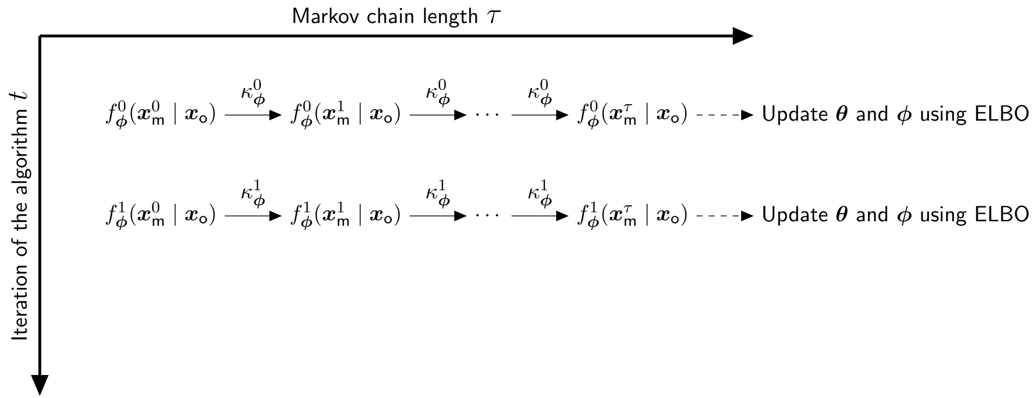
Variational Gibbs Inference: Persistent chains



Variational Gibbs Inference: Persistent chains



Variational Gibbs Inference: Persistent chains



Variational Gibbs Inference: Persistent chains

Markov chain length τ

Iteration of the algorithm t

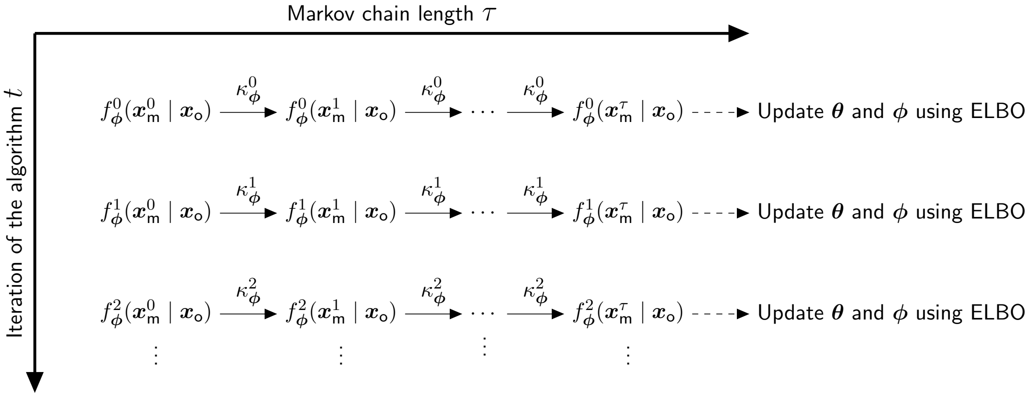
$$f^0_\phi(\boldsymbol{x}^0_\mathsf{m} \mid \boldsymbol{x}_\mathsf{o}) \xrightarrow{\kappa^0_\phi} f^0_\phi(\boldsymbol{x}^1_\mathsf{m} \mid \boldsymbol{x}_\mathsf{o}) \xrightarrow{\kappa^0_\phi} \cdots \xrightarrow{\kappa^0_\phi} f^0_\phi(\boldsymbol{x}^\tau_\mathsf{m} \mid \boldsymbol{x}_\mathsf{o}) \dashrightarrow \text{Update } \boldsymbol{\theta} \text{ and } \phi \text{ using ELBO}$$

$$f^1_\phi(\boldsymbol{x}^0_\mathsf{m} \mid \boldsymbol{x}_\mathsf{o}) \xrightarrow{\kappa^1_\phi} f^1_\phi(\boldsymbol{x}^1_\mathsf{m} \mid \boldsymbol{x}_\mathsf{o}) \xrightarrow{\kappa^1_\phi} \cdots \xrightarrow{\kappa^1_\phi} f^1_\phi(\boldsymbol{x}^\tau_\mathsf{m} \mid \boldsymbol{x}_\mathsf{o}) \dashrightarrow \text{Update } \boldsymbol{\theta} \text{ and } \phi \text{ using ELBO}$$

$$f^2_\phi(\boldsymbol{x}^0_\mathsf{m} \mid \boldsymbol{x}_\mathsf{o}) \xrightarrow{\kappa^2_\phi} f^2_\phi(\boldsymbol{x}^1_\mathsf{m} \mid \boldsymbol{x}_\mathsf{o}) \xrightarrow{\kappa^2_\phi} \cdots \xrightarrow{\kappa^2_\phi} f^2_\phi(\boldsymbol{x}^\tau_\mathsf{m} \mid \boldsymbol{x}_\mathsf{o}) \dashrightarrow \text{Update } \boldsymbol{\theta} \text{ and } \phi \text{ using ELBO}$$

$$\vdots \qquad \vdots \qquad \vdots \qquad \vdots$$

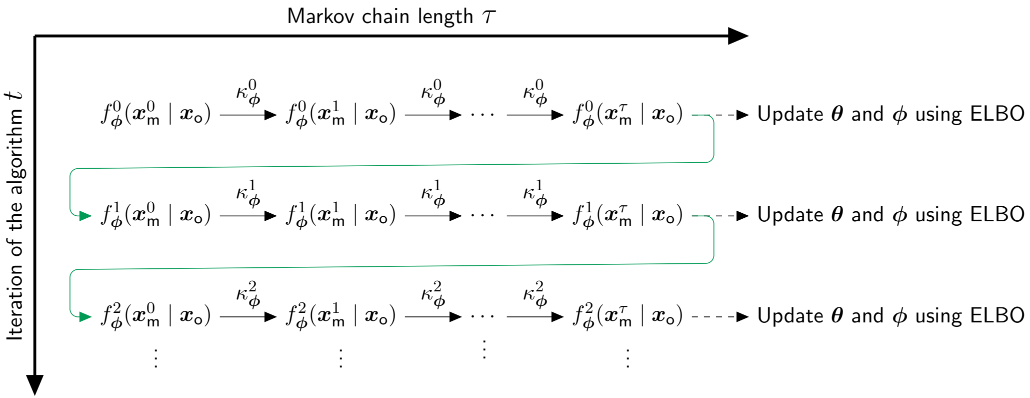
- Sampling long Markov chains at each iteration t of the algorithm is costly.



Variational Gibbs Inference: Persistent chains

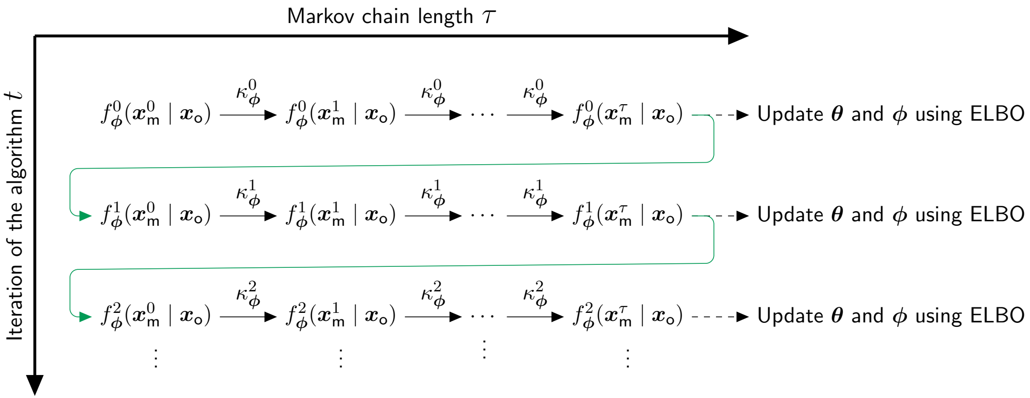
THE UNIVERSITY of EDINBURGH informatics

- Sampling long Markov chains at each iteration t of the algorithm is costly.
- Use "persistent" chains: initialise the chains at the last state of the previous iteration.



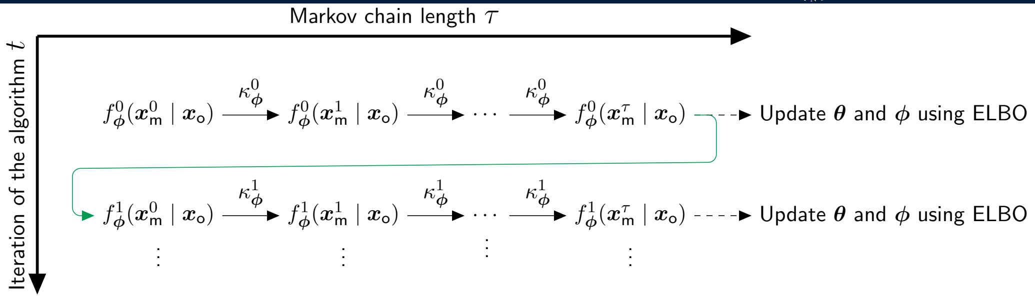
Variational Gibbs Inference: Persistent chains

- Sampling long Markov chains at each iteration t of the algorithm is costly.
- Use "persistent" chains: initialise the chains at the last state of the previous iteration.
- Can now use short chains, that is using small τ, at every iteration t.

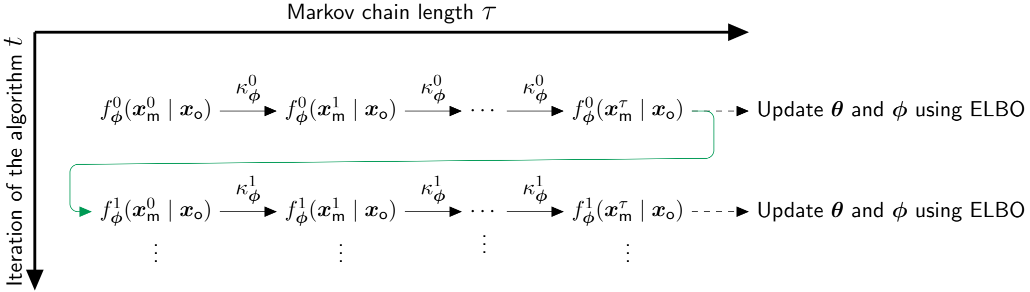


Variational Gibbs Inference: "Cutting" chains

THE UNIVERSITY of EDINBURGH informatics



Variational Gibbs Inference: "Cutting" chains

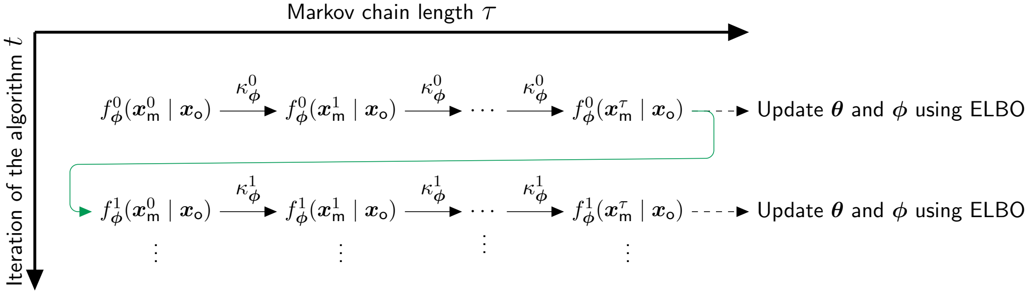


- Computing the marginal density $f^t_{\boldsymbol{\phi}}(\boldsymbol{x}^\tau_{\mathsf{m}} \mid \boldsymbol{x}_{\mathsf{o}})$ of a Markov chain remains intractable:

$$f^t_{\boldsymbol{\phi}}(\boldsymbol{x}^\tau_{\mathsf{m}} \mid \boldsymbol{x}_{\mathsf{o}}) = \int f^t_{\boldsymbol{\phi}}(\boldsymbol{x}^0_{\mathsf{m}} \mid \boldsymbol{x}_{\mathsf{o}}) \prod_{h=0}^{\tau-1} \kappa_{\boldsymbol{\phi}}(\boldsymbol{x}^{h+1}_{\mathsf{m}} \mid \boldsymbol{x}_{\mathsf{o}}, \boldsymbol{x}^h_{\mathsf{m}}) \, \mathrm{d}\boldsymbol{x}^0_{\mathsf{m}} \ldots \mathrm{d}\boldsymbol{x}^{\tau-1}_{\mathsf{m}}.$$

Variational Gibbs Inference: “Cutting” chains

THE UNIVERSITY of EDINBURGH informatics

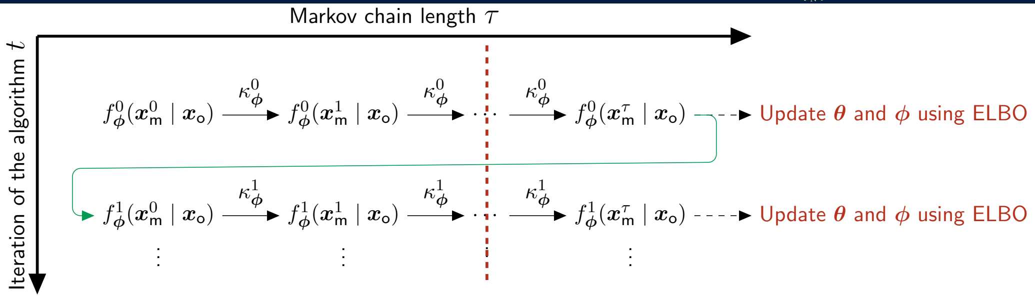


- Computing the marginal density $f^t_\phi(\boldsymbol{x}^\mathcal{T}_\mathrm{m} \mid \boldsymbol{x}_\mathrm{o})$ of a Markov chain remains intractable:

$$f^t_\phi(\boldsymbol{x}^\mathcal{T}_\mathrm{m} \mid \boldsymbol{x}_\mathrm{o}) = \int f^t_\phi(\boldsymbol{x}^0_\mathrm{m} \mid \boldsymbol{x}_\mathrm{o}) \prod_{h=0}^{\mathcal{T}-1} \kappa_\phi(\boldsymbol{x}^{h+1}_\mathrm{m} \mid \boldsymbol{x}_\mathrm{o}, \boldsymbol{x}^h_\mathrm{m})\, \mathrm{d}\boldsymbol{x}^0_\mathrm{m} \ldots \mathrm{d}\boldsymbol{x}^{\mathcal{T}-1}_\mathrm{m}.$$

- So how can we optimise the parameters ϕ of the kernel κ_ϕ?

Variational Gibbs Inference: "Cutting" chains



- Computing the marginal density $f^t_\phi(\boldsymbol{x}^\tau_\mathrm{m} \mid \boldsymbol{x}_\mathrm{o})$ of a Markov chain remains intractable:

$$f^t_\phi(\boldsymbol{x}^\tau_\mathrm{m} \mid \boldsymbol{x}_\mathrm{o}) = \int f^t_\phi(\boldsymbol{x}^0_\mathrm{m} \mid \boldsymbol{x}_\mathrm{o}) \prod_{h=0}^{\tau-1} \kappa_\phi(\boldsymbol{x}^{h+1}_\mathrm{m} \mid \boldsymbol{x}_\mathrm{o}, \boldsymbol{x}^h_\mathrm{m})\, \mathrm{d}\boldsymbol{x}^0_\mathrm{m} \ldots \mathrm{d}\boldsymbol{x}^{\tau-1}_\mathrm{m}.$$

- So how can we optimise the parameters ϕ of the kernel κ_ϕ?
- Instead of optimising ϕ over the full length of the Markov chains, we "cut" the chains just before the last transition and optimise over the last step of the chain.

Variational Gibbs Inference: Learning objective

THE UNIVERSITY *of* EDINBURGH informatics

- Objective for learning $\boldsymbol{\theta}$ and ϕ:

$$\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\mathsf{o}}) \geqslant \mathbb{E}_{\boldsymbol{\pi}(j|\mathrm{idx}(\boldsymbol{m}))f^{t-1}(\boldsymbol{x}_{\mathsf{m}\setminus j}|\boldsymbol{x}_{\mathsf{o}})q_{\boldsymbol{\phi}_j}(x_j|\boldsymbol{x}_{\mathsf{m}\setminus j},\boldsymbol{x}_{\mathsf{o}})}\left[\log \frac{p_{\boldsymbol{\theta}}(x_j, \boldsymbol{x}_{\mathsf{m}\setminus j}, \boldsymbol{x}_{\mathsf{o}})}{q_{\boldsymbol{\phi}_j}(x_j \mid \boldsymbol{x}_{\mathsf{m}\setminus j}, \boldsymbol{x}_{\mathsf{o}})}\right] + \text{Const.}$$

- Objective for learning $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$:

$$\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\mathsf{o}}) \geqslant \mathbb{E}_{\boldsymbol{\pi}(j|\mathrm{idx}(\boldsymbol{m}))f^{t-1}(\boldsymbol{x}_{\mathsf{m}\setminus j}|\boldsymbol{x}_{\mathsf{o}})q_{\boldsymbol{\phi}_j}(x_j|\boldsymbol{x}_{\mathsf{m}\setminus j},\boldsymbol{x}_{\mathsf{o}})}\left[\log \frac{p_{\boldsymbol{\theta}}(x_j, \boldsymbol{x}_{\mathsf{m}\setminus j}, \boldsymbol{x}_{\mathsf{o}})}{q_{\boldsymbol{\phi}_j}(x_j \mid \boldsymbol{x}_{\mathsf{m}\setminus j}, \boldsymbol{x}_{\mathsf{o}})}\right] + \text{Const.}$$

- We only need samples from penultimate step of the Markov chain f^{t-1}.

Variational Gibbs Inference: Learning objective

- Objective for learning $\boldsymbol{\theta}$ and ϕ:

$$\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\mathsf{o}}) \geqslant \mathbb{E}_{\boldsymbol{\pi}(j|\mathrm{idx}(\boldsymbol{m}))f^{t-1}(\boldsymbol{x}_{\mathsf{m}\setminus j}|\boldsymbol{x}_{\mathsf{o}})q_{\boldsymbol{\phi}_j}(x_j|\boldsymbol{x}_{\mathsf{m}\setminus j},\boldsymbol{x}_{\mathsf{o}})}\left[\log \frac{p_{\boldsymbol{\theta}}(x_j, \boldsymbol{x}_{\mathsf{m}\setminus j}, \boldsymbol{x}_{\mathsf{o}})}{q_{\boldsymbol{\phi}_j}(x_j \mid \boldsymbol{x}_{\mathsf{m}\setminus j}, \boldsymbol{x}_{\mathsf{o}})}\right] + \text{Const.}$$

- We only need samples from penultimate step of the Markov chain f^{t-1}.
- Can optimise w.r.t. $\boldsymbol{\theta}$ and ϕ using stochastic gradient ascent.

Variational Gibbs Inference: Learning objective

THE UNIVERSITY *of* EDINBURGH informatics

- Objective for learning $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$:

$$\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\mathsf{o}}) \geqslant \mathbb{E}_{\boldsymbol{\pi}(j|\mathrm{idx}(\boldsymbol{m}))f^{t-1}(\boldsymbol{x}_{\mathsf{m}\setminus j}|\boldsymbol{x}_{\mathsf{o}})q_{\boldsymbol{\phi}_j}(x_j|\boldsymbol{x}_{\mathsf{m}\setminus j},\boldsymbol{x}_{\mathsf{o}})}\left[\log \frac{p_{\boldsymbol{\theta}}(x_j, \boldsymbol{x}_{\mathsf{m}\setminus j}, \boldsymbol{x}_{\mathsf{o}})}{q_{\boldsymbol{\phi}_j}(x_j \mid \boldsymbol{x}_{\mathsf{m}\setminus j}, \boldsymbol{x}_{\mathsf{o}})}\right] + \text{Const.}$$

- We only need samples from penultimate step of the Markov chain f^{t-1}.
- Can optimise w.r.t. $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ using stochastic gradient ascent.
- Maximising the above w.r.t. $\boldsymbol{\phi}$ corresponds to minimising the KL divergence:

$$\mathbb{E}_{\boldsymbol{\pi}(j|\mathrm{idx}(\boldsymbol{m}))f^{t-1}(\boldsymbol{x}_{\mathsf{m}\setminus j}|\boldsymbol{x}_{\mathsf{o}})}\left[D_{\mathsf{KL}}(q_{\boldsymbol{\phi}_j}(x_j \mid \boldsymbol{x}_{\mathsf{m}\setminus j}, \boldsymbol{x}_{\mathsf{o}}) \mid\mid p_{\boldsymbol{\theta}}(x_j \mid \boldsymbol{x}_{\mathsf{m}\setminus j}, \boldsymbol{x}_{\mathsf{o}}))\right]$$

Variational Gibbs Inference: Learning objective

- Objective for learning $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$:

$$\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\mathsf{o}}) \geqslant \mathbb{E}_{\boldsymbol{\pi}(j|\mathrm{idx}(\boldsymbol{m}))f^{t-1}(\boldsymbol{x}_{\mathsf{m}\setminus j}|\boldsymbol{x}_{\mathsf{o}})q_{\boldsymbol{\phi}_j}(x_j|\boldsymbol{x}_{\mathsf{m}\setminus j},\boldsymbol{x}_{\mathsf{o}})}\left[\log \frac{p_{\boldsymbol{\theta}}(x_j, \boldsymbol{x}_{\mathsf{m}\setminus j}, \boldsymbol{x}_{\mathsf{o}})}{q_{\boldsymbol{\phi}_j}(x_j \mid \boldsymbol{x}_{\mathsf{m}\setminus j}, \boldsymbol{x}_{\mathsf{o}})}\right] + \text{Const.}$$

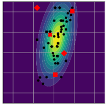
- We only need samples from penultimate step of the Markov chain f^{t-1}.
- Can optimise w.r.t. $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ using stochastic gradient ascent.
- Maximising the above w.r.t. $\boldsymbol{\phi}$ corresponds to minimising the KL divergence:

$$\mathbb{E}_{\boldsymbol{\pi}(j|\mathrm{idx}(\boldsymbol{m}))f^{t-1}(\boldsymbol{x}_{\mathsf{m}\setminus j}|\boldsymbol{x}_{\mathsf{o}})}\left[D_{\mathsf{KL}}(q_{\boldsymbol{\phi}_j}(x_j \mid \boldsymbol{x}_{\mathsf{m}\setminus j}, \boldsymbol{x}_{\mathsf{o}}) \,||\, p_{\boldsymbol{\theta}}(x_j \mid \boldsymbol{x}_{\mathsf{m}\setminus j}, \boldsymbol{x}_{\mathsf{o}}))\right]$$

- The fitted $\kappa_{\boldsymbol{\phi}}$ approximates the Gibbs kernel with the stationary distribution $p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\mathsf{m}} \mid \boldsymbol{x}_{\mathsf{o}})$.

Algorithm 1 Variational Gibbs inference

1: **Create** K-times imputed data $\mathcal{D}_K$ using f_0



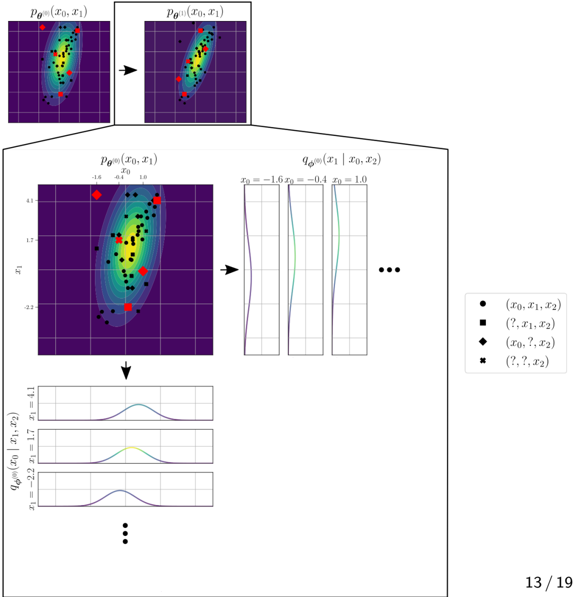
Algorithm 1 Variational Gibbs inference

1: **Create** K-times imputed data $\mathcal{D}_K$ using f_0

2: **for** t in $[1, \mathit{max\_epochs}]$ **do**

3: **Sample** mini-batch $\mathcal{B}_K$ from $\mathcal{D}_K$

7: **end for**

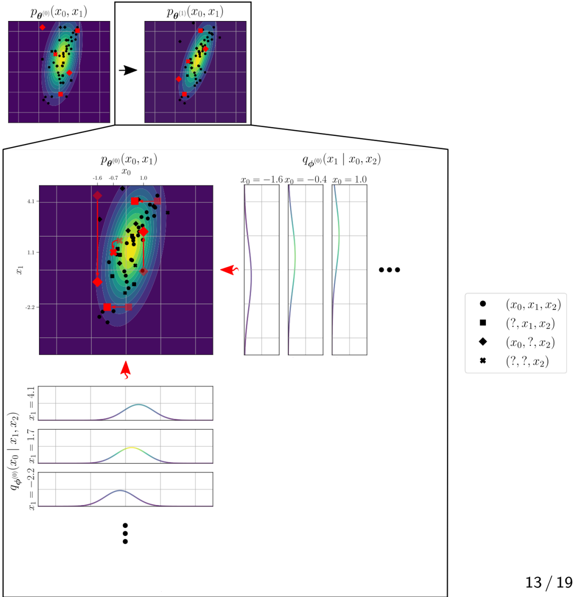


Variational Gibbs Inference: Algorithm

Algorithm 1 Variational Gibbs inference

Create K-times imputed data $\mathcal{D}_K$ using f_0
for t in $[1, \textit{max\_epochs}]$ **do**
  Sample mini-batch $\mathcal{B}_K$ from $\mathcal{D}_K$
  Update the imputations in $\mathcal{B}_K$:
  $\bar{\boldsymbol{x}}_{\mathrm{m}}^{(i,k)} \sim \mathrm{Gibbs}_{\tau}(\boldsymbol{x}_{\mathrm{o}}^{i}, \kappa_{\boldsymbol{\phi}}; \boldsymbol{x}_{\mathrm{m}}^{(i,k)}), \forall \boldsymbol{x}_{\mathrm{m}}^{(i,k)} \in \mathcal{B}_K$
  Persist the imputations in $\mathcal{B}_K$ to $\mathcal{D}_K$

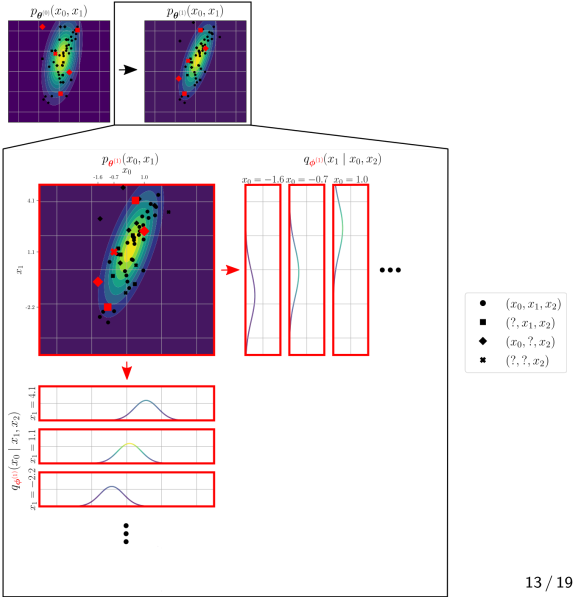
end for



Variational Gibbs Inference: Algorithm

Algorithm 1 Variational Gibbs inference

Create K-times imputed data $\mathcal{D}_K$ using f_0
for t in $[1, \textit{max\_epochs}]$ **do**
  Sample mini-batch $\mathcal{B}_K$ from $\mathcal{D}_K$
  Update the imputations in $\mathcal{B}_K$:
  $\bar{\boldsymbol{x}}_{\mathrm{m}}^{(i,k)} \sim \mathrm{Gibbs}_{\tau}(\boldsymbol{x}_{\mathrm{o}}^{i}, \kappa_{\phi}; \boldsymbol{x}_{\mathrm{m}}^{(i,k)}), \forall \boldsymbol{x}_{\mathrm{m}}^{(i,k)} \in \mathcal{B}_K$
  Persist the imputations in $\mathcal{B}_K$ to $\mathcal{D}_K$
  Update $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ with SGA.
end for

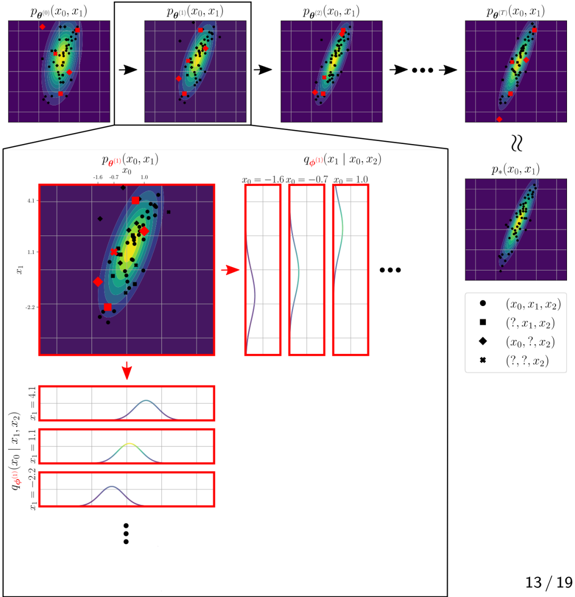


Variational Gibbs Inference: Algorithm

THE UNIVERSITY of EDINBURGH informatics

Algorithm 1 Variational Gibbs inference

Create K-times imputed data $\mathcal{D}_K$ using f_0
for t in $[1, \textit{max\_epochs}]$ **do**
  Sample mini-batch $\mathcal{B}_K$ from $\mathcal{D}_K$
  Update the imputations in $\mathcal{B}_K$:
  $\bar{\boldsymbol{x}}_{\mathrm{m}}^{(i,k)} \sim \mathrm{Gibbs}_{\tau}(\boldsymbol{x}_{\mathrm{o}}^{i}, \kappa_{\boldsymbol{\phi}}; \boldsymbol{x}_{\mathrm{m}}^{(i,k)}), \forall \boldsymbol{x}_{\mathrm{m}}^{(i,k)} \in \mathcal{B}_K$
  Persist the imputations in $\mathcal{B}_K$ to $\mathcal{D}_K$
  Update $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ with SGA.
end for



Variational Gibbs Inference: Summary

THE UNIVERSITY *of* EDINBURGH informatics

- Direct fitting by marginalising the missing variables x_m ?

Variational Gibbs Inference: Summary

- Direct fitting by (approximately) marginalising the missing variables x_{m} ✓

Variational Gibbs Inference: Summary

THE UNIVERSITY *of* EDINBURGH informatics

- Direct fitting by (approximately) marginalising the missing variables $\boldsymbol{x}_{\mathrm{m}}$ ✓

- General-purpose method for estimating $p_{\boldsymbol{\theta}}(\boldsymbol{x})$ from incomplete data.

Variational Gibbs Inference: Summary

- Direct fitting by (approximately) marginalising the missing variables $\boldsymbol{x}_{\mathrm{m}}$ ✓

- General-purpose method for estimating $p_{\boldsymbol{\theta}}(\boldsymbol{x})$ from incomplete data.
- Mitigated the need for 2^M conditional distributions to just M by representing the variational distribution via a learnable Gibbs kernel.

Variational Gibbs Inference: Summary

THE UNIVERSITY of EDINBURGH informatics

- Direct fitting by (approximately) marginalising the missing variables $\boldsymbol{x}_{\mathrm{m}}$ ✓

- General-purpose method for estimating $p_{\boldsymbol{\theta}}(\boldsymbol{x})$ from incomplete data.
- Mitigated the need for 2^M conditional distributions to just M by representing the variational distribution via a learnable Gibbs kernel.
- Used “persistent” chains to efficiently sample imputations using the learnt Gibbs kernel.

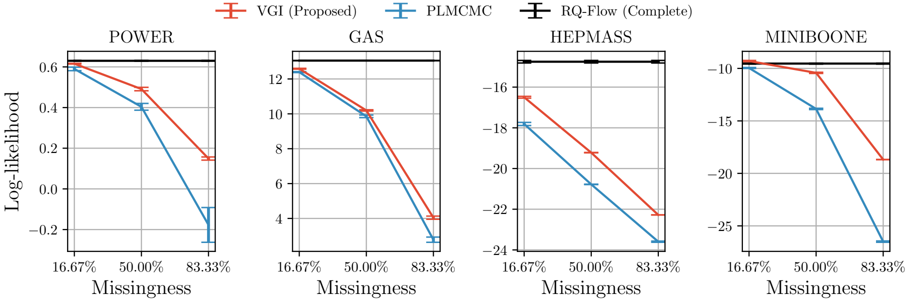
Variational Gibbs Inference: Summary

- Direct fitting by (approximately) marginalising the missing variables $\boldsymbol{x}_\text{m}$ ✓

- General-purpose method for estimating $p_{\boldsymbol{\theta}}(\boldsymbol{x})$ from incomplete data.
- Mitigated the need for 2^M conditional distributions to just M by representing the variational distribution via a learnable Gibbs kernel.
- Used "persistent" chains to efficiently sample imputations using the learnt Gibbs kernel.
- "Cut" the Markov chains to make optimisation of ϕ efficient.

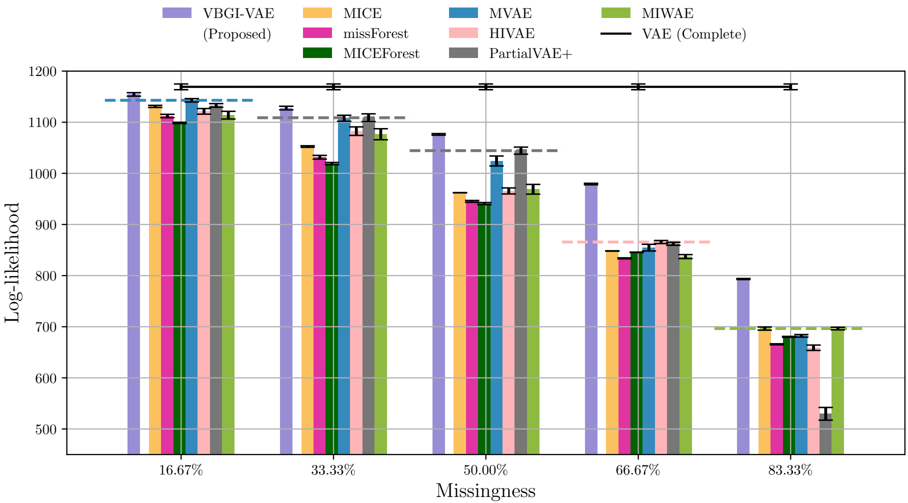
Variational Gibbs Inference: Results (Flows)

THE UNIVERSITY of EDINBURGH
informatics



Variational Gibbs Inference: Results (VAE)

THE UNIVERSITY of EDINBURGH informatics



Summary

- Statistical models and the missing data issue.

Summary

- Statistical models and the missing data issue.
  - Modern models, such as normalising flows and VAEs, are very flexible.

Summary

- Statistical models and the missing data issue.
  - Modern models, such as normalising flows and VAEs, are very flexible.
  - But, they are formulated for complete data.

Summary

THE UNIVERSITY *of* EDINBURGH informatics

- Statistical models and the missing data issue.
  - Modern models, such as normalising flows and VAEs, are very flexible.
  - But, they are formulated for complete data.
- Some problems with direct estimation from incomplete data.

Summary

- Statistical models and the missing data issue.
  - Modern models, such as normalising flows and VAEs, are very flexible.
  - But, they are formulated for complete data.
- Some problems with direct estimation from incomplete data.
  - Marginalisation $\int p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\mathsf{o}}, \boldsymbol{x}_{\mathsf{m}}) \, \mathrm{d}\boldsymbol{x}_{\mathsf{m}}$ is generally intractable.

Summary

- Statistical models and the missing data issue.
  - Modern models, such as normalising flows and VAEs, are very flexible.
  - But, they are formulated for complete data.
- Some problems with direct estimation from incomplete data.
  - Marginalisation $\int p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\mathsf{o}}, \boldsymbol{x}_{\mathsf{m}}) \, \mathrm{d}\boldsymbol{x}_{\mathsf{m}}$ is generally intractable.
  - EM algorithm requires sampling conditionals $p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\mathsf{m}} \mid \boldsymbol{x}_{\mathsf{o}})$ for $\forall \boldsymbol{x}_{\mathsf{o}} \in \mathcal{D}$, which is expensive.

Summary

- Statistical models and the missing data issue.
  - Modern models, such as normalising flows and VAEs, are very flexible.
  - But, they are formulated for complete data.
- Some problems with direct estimation from incomplete data.
  - Marginalisation $\int p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\mathsf{o}}, \boldsymbol{x}_{\mathsf{m}}) \,\mathrm{d}\boldsymbol{x}_{\mathsf{m}}$ is generally intractable.
  - EM algorithm requires sampling conditionals $p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\mathsf{m}} \mid \boldsymbol{x}_{\mathsf{o}})$ for $\forall \boldsymbol{x}_{\mathsf{o}} \in \mathcal{D}$, which is expensive.
  - Standard amortised VI requires 2^M variational distributions, which is inefficient.

Summary

- Statistical models and the missing data issue.
  - Modern models, such as normalising flows and VAEs, are very flexible.
  - But, they are formulated for complete data.
- Some problems with direct estimation from incomplete data.
  - Marginalisation $\int p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\mathsf{o}}, \boldsymbol{x}_{\mathsf{m}})\, \mathrm{d}\boldsymbol{x}_{\mathsf{m}}$ is generally intractable.
  - EM algorithm requires sampling conditionals $p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\mathsf{m}} \mid \boldsymbol{x}_{\mathsf{o}})$ for $\forall \boldsymbol{x}_{\mathsf{o}} \in \mathcal{D}$, which is expensive.
  - Standard amortised VI requires 2^M variational distributions, which is inefficient.
- Variational Gibbs Inference.

Summary

- Statistical models and the missing data issue.
  - Modern models, such as normalising flows and VAEs, are very flexible.
  - But, they are formulated for complete data.
- Some problems with direct estimation from incomplete data.
  - Marginalisation $\int p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\mathrm{o}}, \boldsymbol{x}_{\mathrm{m}})\, \mathrm{d}\boldsymbol{x}_{\mathrm{m}}$ is generally intractable.
  - EM algorithm requires sampling conditionals $p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\mathrm{m}} \mid \boldsymbol{x}_{\mathrm{o}})$ for $\forall \boldsymbol{x}_{\mathrm{o}} \in \mathcal{D}$, which is expensive.
  - Standard amortised VI requires 2^M variational distributions, which is inefficient.
- Variational Gibbs Inference.
  - General purpose method for model estimation from incomplete data.

Summary

- Statistical models and the missing data issue.
  - Modern models, such as normalising flows and VAEs, are very flexible.
  - But, they are formulated for complete data.
- Some problems with direct estimation from incomplete data.
  - Marginalisation $\int p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\mathrm{o}}, \boldsymbol{x}_{\mathrm{m}}) \, \mathrm{d}\boldsymbol{x}_{\mathrm{m}}$ is generally intractable.
  - EM algorithm requires sampling conditionals $p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\mathrm{m}} \mid \boldsymbol{x}_{\mathrm{o}})$ for $\forall \boldsymbol{x}_{\mathrm{o}} \in \mathcal{D}$, which is expensive.
  - Standard amortised VI requires 2^M variational distributions, which is inefficient.
- Variational Gibbs Inference.
  - General purpose method for model estimation from incomplete data.
  - Achieves good performance on normalising flow and VAE estimation, compared to other methods.

Thank you for listening.
Questions?

References I

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 6)

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 2, 9)

Tieleman, T. (2008). Training restricted Boltzmann machines using approximations to the likelihood gradient. In *International Conference on Machine Learning (ICML)*, pages 1064–1071. (Cited on slide 10)

Wei, G. C. G. and Tanner, M. A. (1990). A Monte Carlo Implementation of the EM Algorithm and the Poor Man's Data Augmentation Algorithms. *Journal of the American Statistical Association*, 85(411):699–704. (Cited on slide 6)