Variational Inference
A Review for Statisticians[1]

Eric Duong
eduong@arizona.edu

University of Arizona, Tucson

September 14, 2022

Outline

Introduction

Bayes' Rule

$$p(z \mid x, \theta) = \frac{p(x, z \mid \theta)}{p(x \mid \theta)}$$

- z is latent variables
- x is observed data
- θ is model parameters

Goal

- We want to compute posterior $p(z \mid x, \theta)$ and parameters θ that are maximally likely given observations x.
- However, computing exact evidence $p(x \mid \theta)$ by integrating $p(x, z \mid \theta)$ is intractable.
- With variational inference, we can approximate evidence $p(x \mid \theta)$ and posterior $p(z \mid x, \theta)$.

Evidence Lower Bound (ELBO)

$$
\begin{aligned}
\log p(x \mid \theta) &= \log p(x \mid \theta) \int_z q(z) \\
&= \int_z q(z) \log p(x \mid \theta) \\
&= \mathbb{E}_q[\log p(x \mid \theta)] \\
&= \mathbb{E}_q \left[\log p(x \mid \theta) \frac{q(z)}{q(z)} \right] \\
&= \mathbb{E}_q \left[\log \frac{p(x, z \mid \theta) q(z)}{p(z \mid x, \theta) q(z)} \right] \\
&= \mathbb{E}_q \left[\log \frac{q(z)}{p(z \mid x, \theta)} + \log \frac{p(x, z \mid \theta)}{q(z)} \right] \\
&= \mathbb{E}_q \left[\log \frac{q(z)}{p(z \mid x, \theta)} \right] + \mathbb{E}_q \left[\log \frac{p(x, z \mid \theta)}{q(z)} \right] \\
\log p(x \mid \theta) &= \mathrm{KL}(q(z) \| p(z \mid x, \theta)) + \mathcal{L}(q(z), \theta)
\end{aligned}
$$

Evidence Lower Bound (ELBO)

$$\log p(x \mid \theta) = \mathrm{KL}(q(z) \| p(z \mid x, \theta)) + \mathcal{L}(q(z), \theta)$$

- $\log p(x \mid \theta)$ is the bound we want to approximate.
- $\mathrm{KL}(q(z) \| p(z \mid x, \theta))$ is bound gap we want to minimize.
- $\mathcal{L}(q(z), \theta)$ is evidence lower bound (ELBO).

Maximizing $\mathcal{L}(q(z), \theta)$ gives us both $\log p(x \mid \theta)$ and $p(z \mid x, \theta)$

$$\log p(x \mid \theta) \geq \max_{q(z) \in \mathcal{Q}} \mathcal{L}(q(z), \theta)$$

$$p(z \mid x, \theta) \approx q^*(z) = \arg\max_{q(z) \in \mathcal{Q}} \mathcal{L}(q(z), \theta)$$

Parameter learning

Find θ that is maximizes the likelihood $p(x \mid \theta)$

$$\log p(x \mid \theta) = \mathrm{KL}(q(z) \| p(z \mid x, \theta)) + \mathcal{L}(q(z), \theta)$$

$$\theta^* = \arg\max_{\theta} \log p(x \mid \theta)$$

Parameter learning

Find θ that is maximizes the likelihood $p(x \mid \theta)$

$$\log p(x \mid \theta) = \mathrm{KL}(q(z) \| p(z \mid x, \theta)) + \mathcal{L}(q(z), \theta)$$

$$\theta^* = \arg\max_\theta \log p(x \mid \theta)$$

Expectation Maximization

- If we can compute $p(z \mid x, \theta)$, then set $q(z) = p(z \mid x, \theta)$ and then find θ that maximizes ELBO $\mathcal{L}(q(z), \theta)$

$$\theta^* = \arg\max_\theta \mathcal{L}(q(z), \theta)$$

Parameter learning

Find θ that is maximizes the likelihood $p(x \mid \theta)$

$$\log p(x \mid \theta) = \mathrm{KL}(q(z) \| p(z \mid x, \theta)) + \mathcal{L}(q(z), \theta)$$

$$\theta^* = \arg\max_\theta \log p(x \mid \theta)$$

Expectation Maximization

- If we can compute $p(z \mid x, \theta)$, then set $q(z) = p(z \mid x, \theta)$ and then find θ that maximizes ELBO $\mathcal{L}(q(z), \theta)$

$$\theta^* = \arg\max_\theta \mathcal{L}(q(z), \theta)$$

Variational Inference

- If we cannot compute $p(z \mid x, \theta)$, then we optimize $q(z)$ to best approximate $p(z \mid x, \theta)$.

Variational Inference

Finding $q(z)$ that approximates $p(z \mid x, \theta)$

$$q^*(z) = \arg\min_{q(z) \in \mathcal{Q}} \mathrm{KL}(q(z) \| p(z \mid x, \theta))$$

- This is called "variational" because we are optimizing over function $q(z)$, not a scalar or vector.
- However, we cannot minimize the KL divergence because we cannot compute $p(z \mid x, \theta)$. Therefore, we optimize $q(z)$ to maximize the ELBO $\mathcal{L}(q(z), \theta)$ instead.

$$q^*(z) = \arg\max_{q(z) \in \mathcal{Q}} \mathcal{L}(q(z), \theta)$$

- This objective approximates both $\log p(x \mid \theta)$ and $p(z \mid x, \theta)$.
- In practice, we choose a family $\mathcal{Q}$ (e.g., mean field) and then optimize the parameters of $q(z)$ to maximize ELBO.

Mean Field

(Naive) Mean Field

Assume that each latent variable z_j is independent of other latent variables, $z_j \perp z_i, j \neq i$. We construct a probability factor $q_j(z_j)$ for each z_j

$$q(\mathbf{z}) = \prod_{j=1}^{M} q_j(z_j)$$

Structured Mean Field

We add dependencies between latent variables.

Mixture-Based Mean Field

We add more latent variables.

Coordinate Ascent Variational Inference (CAVI)

Let $\mathcal{Q}$ be a mean field variational family. Our goal is to optimize $q(\mathbf{z})$ to maximize ELBO $\mathcal{L}(q(\mathbf{z}), \theta)$ (note that θ is parameters of model $p(\mathbf{z}, \mathbf{x} \mid \theta)$, not parameter of factors $q(\mathbf{z})$).

$$q^*(\mathbf{z}) = \arg\max_{q(\mathbf{z}) \in \mathcal{Q}} \mathcal{L}(q(\mathbf{z}), \theta)$$

Instead of jointly optimizing all $q(\mathbf{z})$, we optimize each factor $q_j(z_j)$ while keeping other factors $q_\ell(z_\ell)$, $\ell \neq j$, fixed.

$$q_j^*(z_j) = \arg\max_{q_j(z_j),\, j \neq \ell} \mathcal{L}(q(\mathbf{z}), \theta)$$

This can be solved by setting $\frac{\partial \mathcal{L}}{\partial q_j} = 0$ and solving for $q_j(z_j)$. We obtain an update rule for $q_j(z_j)$

$$\begin{aligned} q_j^*(z_j) &\propto \exp\{\mathbb{E}_{-j}[\log p(z_j \mid \mathbf{z}_{-j}, \mathbf{x})]\} \\ &\propto \exp\{\mathbb{E}_{-j}[\log p(z_j, \mathbf{z}_{-j}, \mathbf{x})]\} \end{aligned}$$

Coordinate Ascent Variational Inference (CAVI)

Typically we write $q_j^*(z_j)$ in the exponential family form to derive the update rules for parameters of $q_j^*(z_j)$. The exponential family form of $p(z_j \mid \mathbf{z}_{-j}, \mathbf{x})$ is

$$p(z_j \mid \mathbf{z}_{-j}, \mathbf{x}) = h(z_j) \exp\left\{\eta_j(\mathbf{z}_{-j}, \mathbf{x})^\top \phi(z_j) - A(\eta_j(\mathbf{z}_{-j}, \mathbf{x}))\right\}$$

The exponential family form of $q_j^*(z_j)$ is

$$\begin{aligned} q_j^*(z_j) &\propto \exp\left\{\mathbb{E}_{-j}\left[\log p(z_j \mid \mathbf{z}_{-j}, \mathbf{x})\right]\right\} \\ &\propto h(z_j) \exp\left\{\mathbb{E}_{-j}\left[\eta_j(\mathbf{z}_{-j}, \mathbf{x})\right]^\top \phi(z_j)\right\} \end{aligned}$$

We see that the natural parameter of $q_j^*(z_j)$ is

$$\upsilon_j = \mathbb{E}_{-j}\left[\eta_j(\mathbf{z}_{-j}, \mathbf{x})\right]$$

Coordinate Ascent Variational Inference (CAVI)

Data: Model $p(\mathbf{x}, \mathbf{z})$ and data $\mathbf{x}$
Result: Variational density $q(\mathbf{z}) = \prod_{j=1}^{m} q_j(z_j)$

```
Initialize factors q_j(z_j)
while ELBO has not converged do
    for j = 1, ..., m do
        /* Compute parameters of q_j(z_j) here          */
        Set q_j(z_j) ∝ exp{E_{-j}[log p(z_j | z_{-j}, x)]}
    end
    Compute ELBO L(q(z), θ)
end
return q(z)
```

Algorithm 1: CAVI algorithm

Note that $\mathbb{E}_{-j}$ is an expectation over joint factors excluding target variable z_j, i.e., $\mathbb{E}[\mathbb{E}[...\mathbb{E}[\prod_{\ell \neq j} q_\ell(z_\ell)]...]]$, but will collapse into individual expectation of only each factor.

Coordinate Ascent Variational Inference (CAVI)

Conditionally Conjugate Models

There is a special case of exponential family models called *conditionally conjugate models* that has a global latent variable vector β and local latent variables $\mathbf{z}$ whose i-th element only governs the i-th data. The joint density is

$$p(\beta, \mathbf{z}, \mathbf{x}) = p(\beta) \prod_{i=1}^{n} p(z_i, x_i \mid \beta)$$

Goal

Given mean field factors $q(\beta; \lambda)$ and $q_j(z_j; \varphi_j)$, compute λ and $\boldsymbol{\varphi}$ that maximizes ELBO $\mathcal{L}(q(\beta, \mathbf{z}; \lambda, \boldsymbol{\varphi}))$

$$\lambda^*, \boldsymbol{\varphi}^* = \arg\max_{\lambda, \boldsymbol{\varphi}} \mathcal{L}(q(\beta, \mathbf{z}; \lambda, \boldsymbol{\varphi}))$$

Coordinate Ascent Variational Inference (CAVI)

Derive exponential family form of $p(\beta \mid \mathbf{z}, \mathbf{x})$

The exponential family form of the likelihood $p(z_i, x_i \mid \beta)$ is

$$p(z_i, x_i \mid \beta, \theta) = h(z_i, x_i) \exp\left\{\beta^\top \phi(z_i, x_i) - A(\beta)\right\}$$

and exponential family form of the conjugate prior $p(\beta)$ is

$$p(\beta) = h(\beta) \exp\left\{\alpha^\top \begin{bmatrix} \beta \\ -A(\beta) \end{bmatrix} - B(\beta)\right\}$$

Then the posterior $p(\beta \mid \mathbf{z}, \mathbf{x})$ has the exponential family form

$$p(\beta \mid \mathbf{z}, \mathbf{x}) \propto \exp\left\{\begin{bmatrix} \alpha_1 + \sum_{i=1}^n \phi(z_i, x_i) \\ \alpha_2 + n \end{bmatrix}^\top \begin{bmatrix} \beta \\ -A(\beta) \end{bmatrix}\right\}$$

Coordinate Ascent Variational Inference (CAVI)

Update parameter of mean field factor $q^*(\beta; \lambda)$

The updated natural parameter for $p(\beta \mid \mathbf{z}, \mathbf{x})$ is

$$\hat{\alpha} = \alpha + \left[\frac{\sum_{i=1}^{n} \phi(z_i, x_i)}{n} \right]$$

We can update CAVI parameter λ of factor $q^*(\beta; \lambda)$ as

$$\lambda = \mathbb{E}_{q(\mathbf{z};\boldsymbol{\varphi})}\left[\hat{\alpha}\right] = \alpha + \left[\frac{\sum_{i=1}^{n} \mathbb{E}_{q(z_i;\varphi_i)}\left[\phi(z_i, x_i)\right]}{n} \right]$$

Coordinate Ascent Variational Inference (CAVI)

Derive exponential family form of $p(z_j \mid \mathbf{z}_{-j}, \mathbf{x}, \beta)$

z_i is independent of other variables given x_j and β. Therefore

$$p(z_j \mid \mathbf{z}_{-j}, \mathbf{x}, \beta) = p(z_j \mid x_j, \beta)$$

The exponential family form of $p(z_j \mid x_j, \beta)$ is

$$p(z_j \mid x_j, \beta) = h(z_j) \exp\left\{\eta_j(x_j, \beta)^\top \phi(z_j) - A(\eta_j(x_j, \beta))\right\}$$

Update parameter of mean field factor $q^*(z_j; \varphi_j)$

We can update CAVI parameter φ_j of factor $q^*(z_j; \varphi_j)$ as

$$\varphi_i = \mathbb{E}\left[\eta_j(x_j, \beta)\right]$$

Coordinate Ascent Variational Inference (CAVI)

Data: Model $p(\beta, \mathbf{z}, \mathbf{x})$ and data $\mathbf{x}$
Result: Variational density $q(\beta, \mathbf{z}) = p(\beta; \lambda) \prod_{j=1}^{m} q_j(z_j; \varphi_j)$
Initialize factors $q(\beta, \mathbf{z})$
while *ELBO has not converged* **do**
    $\lambda \leftarrow \alpha + \left[\dfrac{\sum_{i=1}^{n} \phi(z_i, x_i)}{n}\right]$
    for $j = 1, \ldots, m$ **do**
        $\varphi_j \leftarrow \mathbb{E}\left[\eta_j(x_j, \beta)\right]$
    end
    Compute ELBO $\mathcal{L}(q(\beta, \mathbf{z}; \lambda, \boldsymbol{\varphi}))$
end
return $q(\beta, \mathbf{z})$
Algorithm 2: CAVI algorithm for conditionally conjugate models

Coordinate Ascent Variational Inference (CAVI)

Results of CAVI

CAVI optimizes mean field factors $q(\mathbf{z})$ to maximize ELBO $\mathcal{L}(q(\mathbf{z}), \theta)$, which gives us both $\log p(\mathbf{x} \mid \theta)$ and $p(\mathbf{z} \mid \mathbf{x}, \theta)$

$$\log p(\mathbf{x} \mid \theta) \geq \max_{q(\mathbf{z}) \in \mathcal{Q}} \mathcal{L}(q(\mathbf{z}), \theta)$$

$$p(\mathbf{z} \mid \mathbf{x}, \theta) \approx q^*(\mathbf{z}) = \arg\max_{q(\mathbf{z}) \in \mathcal{Q}} \mathcal{L}(q(\mathbf{z}), \theta)$$

Properties of (Naive) Mean Field and CAVI

- Parameters are typically updated using exponential family.
- ELBO monotonically increases, which guarantees to converge.
- Since ELBO is non-convex, CAVI only guarantees convergence at local optimum.
- Because latent variables are assumed to be independent, CAVI cannot capture the correlation between variables.

Mean Field cannot capture the correlation between latent variables

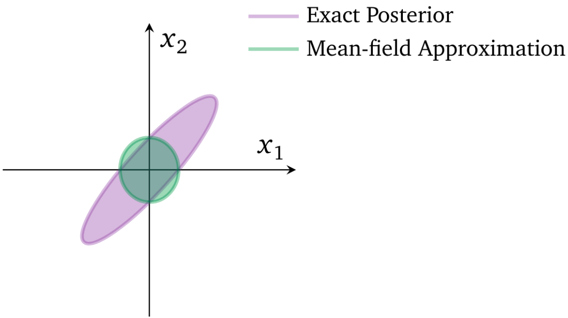

Figure: Mean-field approximation of 2-dimensional Gaussian posterior.

By minimizing $\mathrm{KL}(q(z) \| p(z \mid x, \theta))$, CAVI penalizes any data point outside the variance of the exact posterior, hence the approximated mean field stays inside the variance of the exact posterior.

Example: Bayesian Mixture of Gaussians

Bayesian Mixture of Gaussians Model

$$\begin{aligned} \mu_k &\sim \mathcal{N}(0, \sigma^2), & k &= 1, \ldots, K \\ c_i &\sim \text{Categorical}\left(\frac{1}{K}, \ldots, \frac{1}{K}\right), & i &= 1, \ldots, n \\ x_i \mid c_i, \boldsymbol{\mu} &\sim \mathcal{N}(c_i^\top \boldsymbol{\mu}, 1), & i &= 1, \ldots, n \end{aligned}$$

$$p(\boldsymbol{\mu}, \mathbf{c}, \mathbf{x}; \sigma^2, K) = p(\boldsymbol{\mu}; \sigma^2) \prod_{i=1}^{n} p(c_i; K) p(x_i \mid c_i, \boldsymbol{\mu})$$

Goal

- Given observed data $\mathbf{x}$, infer means $\mathbf{m}$ and variance $\mathbf{s}^2$ of $q(\boldsymbol{\mu}; \mathbf{m}, \mathbf{s}^2)$ of K components and component assignment probability φ_i of each data x_i.
- Assume component variances σ^2 as a known hyperparameter.

Mean Field of Bayesian Mixture of Gaussians

Find $\mathbf{m}$, $\mathbf{s}^2$, and $\boldsymbol{\varphi}$ that maximizes likelihood $p(\mathbf{x} \mid \sigma^2, K)$

$$\mathbf{m}^*, \mathbf{s}^{2*}, \boldsymbol{\varphi}^* = \underset{\mathbf{m},\mathbf{s}^2,\boldsymbol{\varphi}}{\arg\max} \log p(\mathbf{x} \mid \sigma^2, K)$$

$$= \underset{\mathbf{m},\mathbf{s}^2,\boldsymbol{\varphi}}{\arg\max} \mathcal{L}(q(\boldsymbol{\mu}, \mathbf{c}; \mathbf{m}, \mathbf{s}^2, \boldsymbol{\varphi}), \sigma^2, K)$$

Note that $\mathbf{m}$, $\mathbf{s}^2$, and $\boldsymbol{\varphi}$ are used by ELBO to compute expectation with respect to factors $q(\boldsymbol{\mu}, \mathbf{c}; \mathbf{m}, \mathbf{s}^2, \boldsymbol{\varphi})$.

Mean Field of Bayesian Mixture of Gaussians

$$q(\boldsymbol{\mu}, \mathbf{c}; \mathbf{m}, \mathbf{s}^2, \boldsymbol{\varphi}) = \prod_{k=1}^{K} q(\mu_k; m_k, s_k^2) \prod_{i=1}^{n} q(c_i; \varphi_i)$$

$$q(\mu_k; m_k, s_k^2) \sim \mathcal{N}(m_k, s_k^2)$$

$$q(c_i; \varphi_i) \sim \text{Categorical}(\varphi_i)$$

CAVI for Bayesian Mixture of Gaussians

$$p(\boldsymbol{\mu}, \mathbf{c}, \mathbf{x}; \sigma^2, K) = p(\boldsymbol{\mu}; \sigma^2) \prod_{i=1}^{n} p(c_i; K) p(x_i \mid c_i, \boldsymbol{\mu})$$

$$q(\boldsymbol{\mu}, \mathbf{c}; \mathbf{m}, \mathbf{s}^2, \boldsymbol{\varphi}) = \prod_{k=1}^{K} q(\mu_k; m_k, s_k^2) \prod_{i=1}^{n} q(c_i; \varphi_i)$$

ELBO for Bayesian Mixture of Gaussians

$$\begin{aligned}
&\mathcal{L}(q(\boldsymbol{\mu}, \mathbf{c}; \mathbf{m}, \mathbf{s}^2, \boldsymbol{\varphi}), \sigma^2, K) \\
&= \sum_{k=1}^{K} \mathbb{E}\left[\log p(\mu_k; \sigma^2); m_k, s_k^2\right] \\
&+ \sum_{i=1}^{n} \left(\mathbb{E}\left[\log p(c_i; K); \varphi_i\right] + \mathbb{E}\left[\log p(x_i \mid c_i, \boldsymbol{\mu}); \varphi_i, \mathbf{m}, \mathbf{s}^2\right]\right) \\
&- \sum_{i=1}^{n} \mathbb{E}\left[\log q(c_i; \varphi_i)\right] - \sum_{k=1}^{K} \mathbb{E}\left[\log q(\mu_k; m_k, s_k^2)\right]
\end{aligned}$$

CAVI for Bayesian Mixture of Gaussians

$$
\begin{aligned}
q^*(c_i; \varphi_i) &\propto \exp\{\mathbb{E}\left[\log p(\mathbf{c}, \mathbf{x}, \boldsymbol{\mu}; \sigma^2, K); \mathbf{m}, \mathbf{s}^2\right]\} \\
&\propto \exp\{\log p(c_i; K) + \mathbb{E}\left[\log p(x_i \mid c_i, \boldsymbol{\mu}); \mathbf{m}, \mathbf{s}^2\right]\}
\end{aligned}
$$

$q(c_i; \varphi_i)$ is a categorical distribution. We can compute φ_i by formulating the categorical distribution in the form of an exponential family and read natural parameters. Recall that the exponential family form of categorical distribution is

$$
\text{Categorical}(c_i; \varphi_i) \propto \exp\left\{\sum_{k=1}^{K} c_{ik}\varphi_{ik}\right\}
$$

Note that the equation above omits the base measure and log-normalizer. It only shows natural parameters and sufficient statistics.

CAVI for Bayesian Mixture of Gaussians

$$
\begin{aligned}
q^*(c_i;\varphi_i) &\propto \exp\{\mathbb{E}\left[\log p(\mathbf{c},\mathbf{x},\boldsymbol{\mu};\sigma^2,K);\mathbf{m},\mathbf{s}^2\right]\} \\
&\propto \exp\{\log p(c_i;K) + \mathbb{E}\left[\log p(x_i \mid c_i,\boldsymbol{\mu});\mathbf{m},\mathbf{s}^2\right]\} \\
&\propto \exp\{\mathbb{E}\left[\log p(x_i \mid c_i,\boldsymbol{\mu});\mathbf{m},\mathbf{s}^2\right]\} \\
&= \exp\{\mathbb{E}\left[\log \prod_{k=1}^{K} p(x_i \mid \mu_k)^{c_{ik}};\mathbf{m},\mathbf{s}^2\right]\} \\
&\propto \exp\{\sum_{k=1}^{K} c_{ik}\mathbb{E}\left[-(x_i-\mu_k)^2/2;m_k,s_k^2\right]\} \\
&\propto \exp\left\{\sum_{k=1}^{K} c_{ik}(\mathbb{E}[\mu_k;m_k,s_k^2]x_i - \mathbb{E}[\mu_k;m_k^2,s_k^2]/2)\right\}
\end{aligned}
$$

CAVI for Bayesian Mixture of Gaussians

$$q^*(c_i; \varphi_i) \propto \exp\left\{\sum_{k=1}^{K} c_{ik}(\mathbb{E}[\mu_k; m_k, s_k^2]x_i - \mathbb{E}[\mu_k; m_k^2, s_k^2]/2)\right\}$$

We have sufficient statistics $\{c_i\}$ and natural parameters $\{\mathbb{E}[\mu_k; m_k, s_k^2]x_i - \mathbb{E}[\mu_k; m_k^2, s_k^2]/2\}$ of categorical distribution. Recall that the exponential family form of categorical distribution (with only natural parameters and sufficient statistics)

$$\text{Categorical}(c_i; \varphi_i) \propto \exp\left\{\sum_{k=1}^{K} c_{ik}\varphi_{ik}\right\}$$

The update rule for φ_{ik} is

$$\varphi_{ik} \propto \exp\{\mathbb{E}[\mu_k; m_k, s_k^2]x_i - \mathbb{E}[\mu_k; m_k^2, s_k^2]/2\}$$

CAVI for Bayesian Mixture of Gaussians

$$
\begin{aligned}
q^*(\mu_k; m_k, s_k^2) &\propto \exp\{\mathbb{E}\left[\log p(\boldsymbol{\mu}, \mathbf{c}, \mathbf{x}; \sigma^2, K); \boldsymbol{\varphi}, \mathbf{m}_{-k}, \mathbf{s}^2_{-k}\right]\} \\
&\propto \exp\{\log p(\mu_k; \sigma^2) \\
&+ \sum_{i=1}^{n} \mathbb{E}\left[\log p(x_i \mid c_i, \boldsymbol{\mu}); \varphi_i, \mathbf{m}_{-k}, \mathbf{s}^2_{-k}\right]\}
\end{aligned}
$$

Because $q(\mu_k; m_k, s_k^2)$ is a Gaussian distribution, we compute m_k and s_k^2 by formulating the Gaussian distribution in the form of an exponential family and read natural parameters. Recall that the exponential family form of Gaussian is

$$
\mathcal{N}(\mu_k; m_k, s_k^2) \propto \exp\left\{\frac{m_k}{s_k^2}\mu_k - \frac{1}{2s_k^2}\mu_k^2\right\}
$$

Note that the equation above omits the base measure and log-normalizer. It only shows natural parameters and sufficient statistics.

CAVI for Bayesian Mixture of Gaussians

$$
\begin{aligned}
&\log q^*(\mu_k; m_k, s_k^2) \\
&\propto \log p(\mu_k; \sigma^2) + \sum_{i=1}^{n} \mathbb{E}\left[\log p(x_i \mid c_i, \boldsymbol{\mu}); \varphi_i, \mathbf{m}_{-k}, \mathbf{s}^2_{-k}\right] \\
&\propto \log p(\mu_k; \sigma^2) + \sum_{i=1}^{n} \mathbb{E}\left[c_{ik} \log p(x_i \mid \mu_k); \varphi_i\right] \\
&\propto \log p(\mu_k; \sigma^2) + \sum_{i=1}^{n} \mathbb{E}\left[c_{ik} \log p(x_i \mid \mu_k); \varphi_i\right] \\
&\propto -\frac{\mu_k^2}{2\sigma^2} + \sum_{i=1}^{n} \mathbb{E}[c_{ik}; \varphi_i] \log p(x_i \mid \mu_k) \\
&\propto -\frac{\mu_k^2}{2\sigma^2} + \sum_{i=1}^{n} \varphi_{ik} \left(\frac{-(x_i - \mu_k)^2}{2}\right) \\
&\propto -\frac{\mu_k^2}{2\sigma^2} + \sum_{i=1}^{n} \varphi_{ik} x_i \mu_k - \frac{\varphi_{ik} \mu_k^2}{2}
\end{aligned}
$$

CAVI for Bayesian Mixture of Gaussians

$$\log q^*(\mu_k; m_k, s_k^2) \propto \left(\sum_{i=1}^n \varphi_{ik} x_i \right) \mu_k - \left(\frac{1}{2\sigma^2} + \sum_{i=1}^n \frac{\varphi_{ik}}{2} \right) \mu_k^2$$

We have sufficient statistics $\{\mu_k, \mu_k^2\}$ and natural parameters $\{\sum_{i=1}^n \varphi_{ik} x_i, -\frac{1}{2\sigma^2} - \sum_{i=1}^n \frac{\varphi_{ik}}{2}\}$ of Gaussian. Recall that the exponential family form of Gaussian (with only natural parameters and sufficient statistics)

$$\log \mathcal{N}(\mu_k; m_k, s_k^2) \propto \frac{m_k}{s_k^2} \mu_k - \frac{1}{2s_k^2} \mu_k^2$$

The update rule for m_k and s_k^2 is

$$m_k = \frac{\sum_{i=1}^n \varphi_{ik} x_i}{\frac{1}{\sigma^2} + \sum_{i=1}^n \varphi_{ik}}, \qquad s_k^2 = \frac{1}{\frac{1}{\sigma^2} + \sum_{i=1}^n \varphi_{ik}}$$

CAVI algorithm for Bayesian Mixture of Gaussians

Data: Data $x_{1:n}$, number of components K, prior variance of component means σ^2

Result: Variational density $q(\mu_k; m_k, s_k^2)$ and $q(c_i, \varphi_i)$

Initialize parameters $\mathbf{m} = m_{1:K}$, $\mathbf{s^2} = s^2_{1:K}$, and $\boldsymbol{\varphi} = \varphi_{1:n}$

while *ELBO has not converged* **do**

- **for** $i = 1, \ldots, m$ **do**
  - Compute and normalize $\varphi_{ik} \propto \exp\{\mathbb{E}[\mu_k; m_k, s_k^2] x_i - \mathbb{E}[\mu_k; m_k^2, s_k^2]/2\}$
- **end**
- **for** $k = 1, \ldots, K$ **do**
  - Compute $m_k \leftarrow \frac{\sum_{i=1}^n \varphi_{ik} x_i}{\frac{1}{\sigma^2} + \sum_{i=1}^n \varphi_{ik}}$
  - Compute $s_k^2 \leftarrow \frac{1}{\frac{1}{\sigma^2} + \sum_{i=1}^n \varphi_{ik}}$
- **end**
- Compute ELBO $\mathcal{L}(q(\boldsymbol{\mu}, \mathbf{c}), \mathbf{m}, \mathbf{s}^2, \boldsymbol{\varphi})$

end

return $q(\boldsymbol{\mu}, \mathbf{c}; \mathbf{m}, \mathbf{s}^2, \boldsymbol{\varphi})$

Results of CAVI for Bayesian Mixture of Gaussians

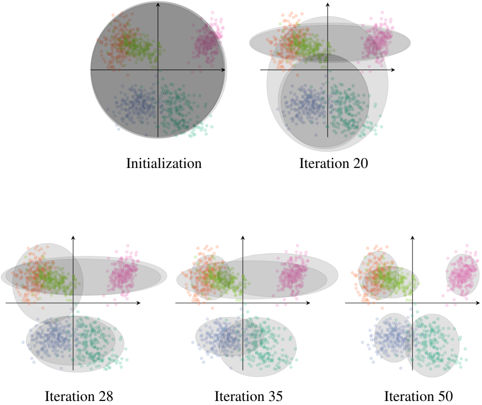

Figure: CAVI result of 2D Bayesian Mixture of Gaussians

Because latent variables are independent of other latent variables, covariance is not inferred; hence the ellipses are not diagonal.

Results of CAVI for Bayesian Mixture of Gaussians

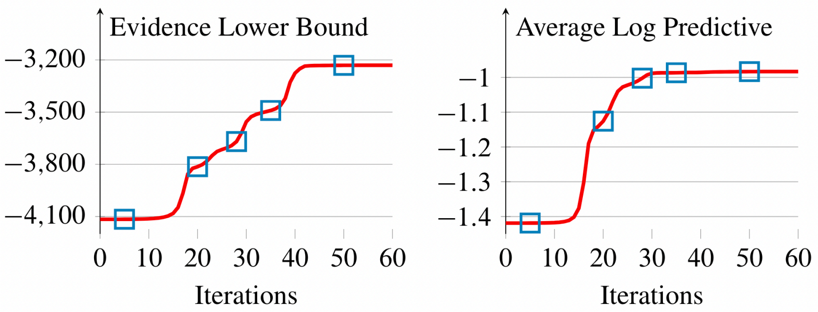

Figure: CAVI result of 2D Bayesian Mixture of Gaussians

Notice that ELBO monotonically increases and converges.

Questions

- How is CAVI's similar to and different from Gibbs sampling?
- Can variational inference be used in real-time inference?
- What are the potential applications of mean-field and CAVI?
- What are other variational inference methods that can capture variances better than CAVI?

References I

[1] David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe.
Variational Inference: A Review for Statisticians.
112(518):859–877.