Fitting Gaussian Mixture Models with Variational Inference

Michael Thomas Wojnowicz

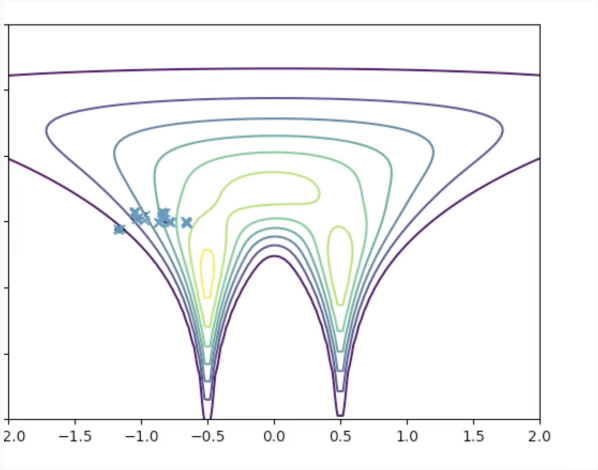
November 9, 2020

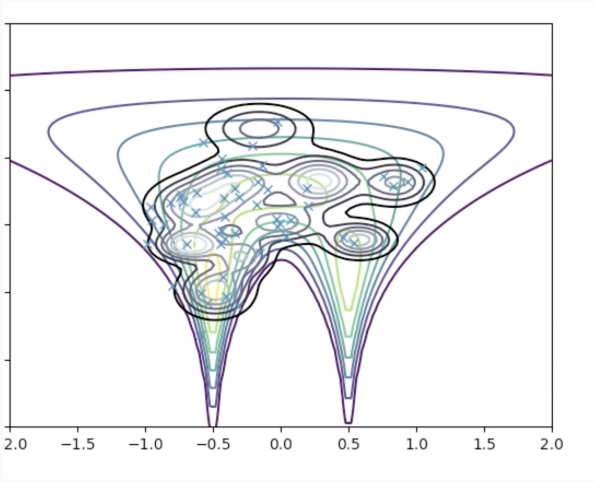
Table of contents

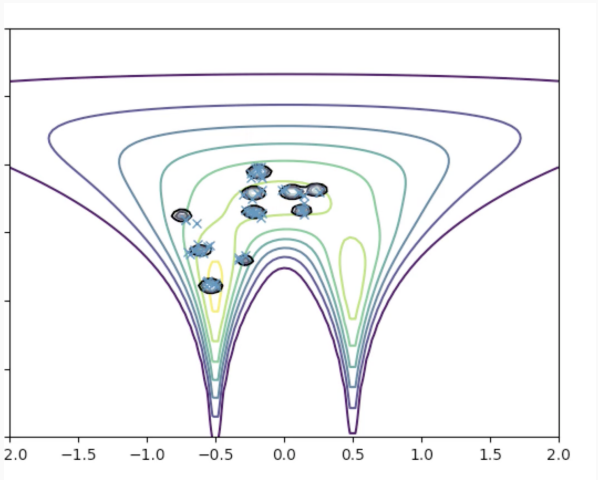
Variational Inference: Overview

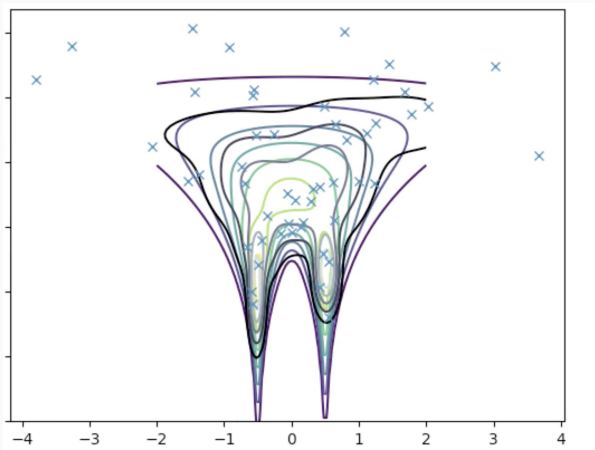
Illustration

Here we approximate an probability density by finding the best approximation from tractable family $\mathcal{Q} = \{\text{10-component Gaussian mixture models}\}$









Statistical inference in Bayesian models

By Bayes' rule, we cannot get the posterior without the **evidence**

$$p(\boldsymbol{x}) = \int p(\boldsymbol{x}, \boldsymbol{u}) \, d\, \boldsymbol{u} \tag{1.1}$$

where

- $\boldsymbol{x}$: observed data
- $\boldsymbol{u}$: unobserved random variables

How variational inference works

We construct a lower bound on the evidence.

Evidence Lower Bound (ELBO)

Let q be any probability density over $\boldsymbol{u}$. Then:

$$\begin{aligned}
\ln p(\boldsymbol{x}) &= \ln \int p(\boldsymbol{u}, \boldsymbol{x}) \, d\,\boldsymbol{u} \\
&= \ln \int q(\boldsymbol{u}) \, \frac{p(\boldsymbol{u}, \boldsymbol{x})}{q(\boldsymbol{u})} \, d\,\boldsymbol{u} \\
&\overset{\textit{Jensen's}}{\geq} \int q(\boldsymbol{u}) \, \ln \left(\frac{p(\boldsymbol{u}, \boldsymbol{x})}{q(\boldsymbol{u})} \right) d\,\boldsymbol{u} \\
&:= \texttt{ELBO}\,(q)
\end{aligned}$$

We want to find $q \in \mathcal{Q}$ to maximize the $\texttt{ELBO}$.

What maximizes the ELBO also minimizes the KL divergence to the posterior

By definition, the KL divergence from the target posterior to the variational density is given by

$$\mathtt{KL}(q(\boldsymbol{u}) \,||\, p(\boldsymbol{u} \mid \boldsymbol{x})) = \mathbb{E}_q\left[\log \frac{q(\boldsymbol{u})}{p(\boldsymbol{u} \mid \boldsymbol{x})}\right]$$

By the chain rule, we get

$$\begin{aligned} \mathtt{KL}(q(\boldsymbol{u}) \,||\, p(\boldsymbol{u} \mid \boldsymbol{x})) &= \mathbb{E}_q[\log q(\boldsymbol{u})] - \mathbb{E}_q[\log p(\boldsymbol{x}, \boldsymbol{u})] && + \log p(\boldsymbol{x}) \\ &= \quad - \mathtt{ELBO}\,(q) && + \text{constant} \end{aligned}$$

Mean Field Coordinate Ascent Variational Inference (MF-CAVI)

Mean field variational families

A variational family $\mathcal{Q}$ is mean field if it factorizes

$$q(u_1, ..., u_K) = \prod_{k=1}^{K} q_k(u_k) \tag{1.2}$$

Mean field coordinate ascent variational inference (MF-CAVI) iteratively optimizes each variational density, while holding the others fixed.

Note: Since the `ELBO` maps probability densities to a real number, it is a functional, so such optimization can be handled with variational calculus.

Update equations for MF-CAVI

Given the mean field assumption (1.2), the optimal updates to $\{q_k\}_k$ for coordinate ascent on the ELBO satisfy

$$q_k(u_k) \propto \exp\left\{\mathbb{E}_{q_{-k}}\left[\log p(\boldsymbol{u}, \boldsymbol{x})\right]\right\} \tag{1.3}$$

Example: Bayesian Gaussian Mixture Model

Example: Bayesian Gaussian Mixture Model

To see the mean field CAVI algorithm (1.3) in a concrete context, consider a version of the Bayesian Gaussian Mixture Model.

$$\begin{aligned} \mu_k &\sim \text{Normal}(0, \sigma^2) && k = 1, ..., K \\ c_i &\sim \text{Categorical}(\pi_1, ..., \pi_K) && i = 1, ..., n \\ x_i \mid c_i, \mu &\sim \text{Normal}(c_i^T \mu, 1) && i = 1, ..., n \end{aligned}$$

(The model is simple in that it assumes that each mixture component has unit variance.)

The joint density, by chain rule, is

$$p(x, c, \mu) = p(\mu) \prod_{i=1}^{n} p(c_i) p(x_i \mid c_i, \mu)$$

And a mean-field variational family is given by

$$q(c, \mu) = \prod_{k=1}^{K} q(\mu_k) \prod_{i=1}^{n} q(c_i)$$

Example: Bayesian Gaussian Mixture Model

We apply (1.3) to determine the coordinate updates for q_{c_i}, the variational factors governing cluster assignments. We take the log of the joint and discard terms that do not depend upon c_i to obtain

$$\begin{aligned} q(c_{ik}) &\propto \exp\left\{\mathbb{E}_{q_{\mu_k}}\left[\log p(c_i = k) + \log p(x_i \mid c_i = k, \mu)\right]\right\} \\ &\propto \exp\left\{\mathbb{E}_{q_{\mu_k}}\left[\log \pi_k + x_i\mu_k - \frac{1}{2}\mu_k^2\right]\right\} \\ &\propto \pi_k \exp\left\{x_i\mathbb{E}_{q_{\mu_k}}[\mu_k] - \frac{1}{2}\mathbb{E}_{q_{\mu_k}}[\mu_k^2]\right\} \end{aligned}$$

The coordinate updates for q_{μ_k} are derived similarly. They reveal that q_{μ_k} are Gaussian, and hence the above expectations are easy to compute.

Appendix

Bayesian Gaussian Mixture Model: Updates to mixture component means

Using the same strategy as when updating cluster assignments c_i, we obtain

$$
\begin{aligned}
q(\mu_k) &\propto \exp\left\{ \mathbb{E}_{-q_{\mu_k}} \left[\log p(\mu_k) + \sum_{i=1}^{n} \log p(x_i \mid c_i = k, \mu) \right] \right\} \\
&\propto \exp\left\{ -\frac{1}{2\sigma^2}\mu_k^2 + \sum_{i=1}^{n} \mathbb{E}_{q_c} \left[1_{c_i=k} \left(x_i \mu_k - \frac{1}{2}\mu_k^2 \right) \right] \right\} \\
&\propto \exp\left\{ -\frac{1}{2}\left(\frac{1}{\sigma^2} + \sum_{i=1}^{n} q(c_{ik}) \right) \mu_k^2 \; + \; \left(\sum_{i=1}^{n} q(c_{ik}) x_i \right) \mu_k \right\}
\end{aligned}
$$

which is an exponential family distribution with sufficient statistics (μ_k, μ_k^2), and hence Gaussian.