Variational inference

Consider a simple Bayesian Gaussian mixture model:

$$\pi \sim \mathrm{Dir}(\alpha)$$

$$\mu_j \sim \mathcal{N}(m_0, I\lambda_0^{-1})$$

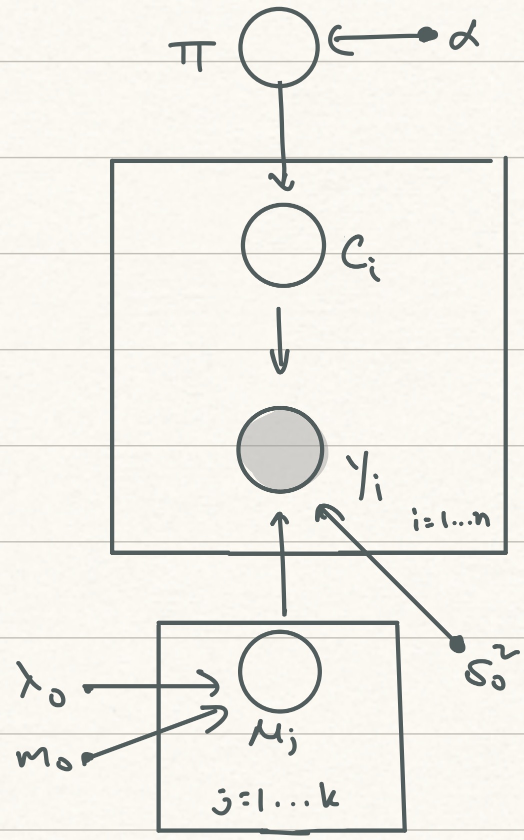
$$c_i \sim \mathrm{Cat}(\pi)$$

$$y_i \sim \mathcal{N}(\mu_{c_i}, \sigma_0^2)$$

$$\theta = \{\mu_{1:K}, \pi\}$$

$$\eta_0 = \{\alpha, m_0, \lambda_0, \sigma_0^2\}$$

Here's the graphical model:



In lecture 6 we used the EM algorithm to do MAP inference in this model. Today we will be interested in full Bayesian posterior inference over all variables.

$$p(c, \theta \mid y, \eta_0)$$

The full posterior is intractable (as before). One thing we could do is run a Gibbs sampler to approximate the posterior with samples. This would be easy to implement since all the complete conditionals are closed-form. However MCMC has some downsides, such as being a stochastic algorithm, and potentially taking a long time to collect many samples.

Variational inference is an alternative. It turns posterior inference into **optimization**.

We begin by setting up a **variational distribution over all variables (i.e., over cluster assignments and parameters)**. We will want this distribution to approximate the posterior.

$$q(c, \theta) \approx p(c, \theta \mid y, \eta_0)$$

The variational distribution will have its own **variational parameters**

$$q(c, \theta) \equiv q(c, \theta; \nu)$$

We will want to fit these variational parameters so that q(...) approximates p(...).

To ease notation, let's lump together the cluster assignments and parameters:

$$z \equiv \{C, \pi, \mu\}$$

In the past, we used Z to denote the cluster assignments and referred to them as **latent variables** as the latent variables, while the cluster means and mixture weights were the **parameters.** As a rule-of-thumb the difference between latent variables and parameters is what you want to do with them: for latent variables, we want a posterior distribution, whereas for parameters we want a point estimate.

In this case, we want a posterior distribution over all variables, so we will now refer to all of them as latent variables. (Note that this is a norm which is inconsistently applied, both in this class, and in the broader literature.)

To recap, using simpler notation, we want the following:

$$q(z) \approx p(z \mid y)$$

Consider the optimization problem:

$$q^*(z) = \underset{q \in Q}{\operatorname{argmin}} \; KL(q(z) \,\|\, p(z \mid y))$$

For some family of densities Q, this will find the member that is closest in KL divergence to the exact posterior p(Z | Y). This is the objective function for (the most commonly used form) of variational inference often known as **variational Bayes.**

How do you minimize the KL divergence to the intractable posterior? Let's perform a manipulation of the KL divergence (which should look familiar):

$$\begin{aligned} KL(q \,\|\, p) &= \mathbb{E}_q\left[\log \frac{q(z)}{p(z \mid x)}\right] \\ &= \underbrace{\mathbb{E}_q\left[\log \frac{q(z)}{p(z, x)}\right]}_{= -ELBO(q)} + \underbrace{\mathbb{E}_q\left[\log p(x)\right]}_{= \log p(x)} \end{aligned}$$

evidence lower bound (ELBO)

log evidence

$$\propto_q -ELBO(q)$$

So we can minimize the KL divergence to an intractable density by maximizing the ELBO:

$$\min_q KL(q \,\|\, p) = \max_q ELBO(q)$$

We saw this same fact when deriving EM. One difference here is that the ELBO is now just a functional of the q-distribution (rather than a functional of q and the parameters).

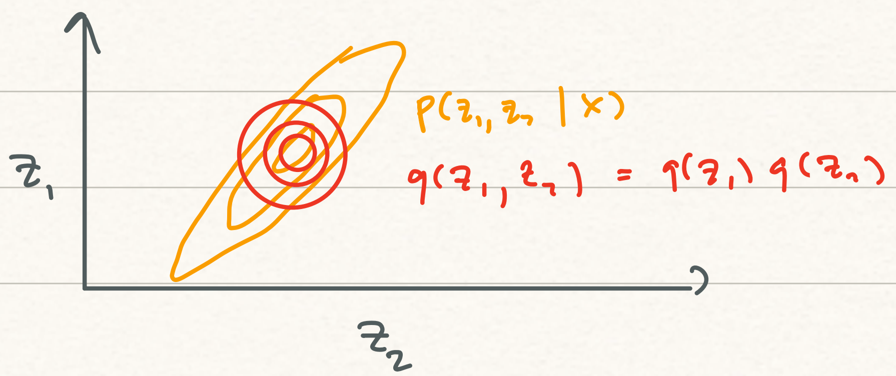
We now have a tractable objective function, but not necessarily a tractable way to optimize it. The next move that variational Bayes makes is to assume that q(...) comes from a simple parametric family of densities that is easy to search over. Specifically, assume that q(...) is a **factorized family**:

$$q(z_{1:D}) = \prod_d q(z_d)$$

In other words, all latent variables are marginally independent under the q-distribution. For example in the Bayesian mixture model:

$$\text{e.g.} \quad q(c, \mu, \pi) = q(\pi)\left[\prod_j q(\mu_j)\right]\left[\prod_i q(c_i)\right]$$

This is called the **mean-field approximation**. Consider the simple picture below:



How does this assumption help us? It facilitates the following **coordinate ascent algorithm**

<u>coordinate ascent variational inference (CAVI)</u>

until convergence:

for $d = 1 \ldots D$:

$$q^*(z_d) \leftarrow \underset{q(z_d)}{\arg\max}\ \mathrm{ELBO}(q)$$

Here is a fact from Bishop (2006): the maximizer of the ELBO takes the following form:

$$q^*(z_d) \propto \exp\left(\mathbb{E}_{q_{\setminus z_d}}[\log p(z_d \mid z_{\setminus d}, x)]\right)$$

Let's confirm it. Consider the KL divergence from any other setting of the density q(zd) to the optimal setting:

$$\begin{aligned} KL(q(z_d) \,\|\, q^*(z_d)) &= \mathbb{E}_{q(z_d)}[\log q(z_d) - \log q^*(z_d)] \\ &\propto \mathbb{E}_{q(z_d)}[\log q(z_d)] - \mathbb{E}_{q(z_d)}\left[\mathbb{E}_{q(z_{\setminus d})}[\log p(z_d \mid z_{\setminus d}, x)]\right] \\ &\propto -H(q(z_d)) - \mathbb{E}_q[\log(z, x)] \\ &\propto -\mathrm{ELBO}(q) \end{aligned}$$

Therefore maximizing the ELBO wrt q(zd) would be minimizing the KL to q*(zd), which is only achieved if q(zd)=q*(zd).

As we saw with EM, if the <u>complete</u> conditional p(zd | -) is exponential family, then the optimal setting q*(zd) will be

$$p(z_d \mid -) = f(z_d; \eta(\cdots)) \propto h(z_d)\exp(t(z_d)^\top \eta(\cdots))$$

where its posterior natural parameter will be a function of other latent variables and data.

$$q^*(z_d) = f(z_d; \mathbb{E}_{q_{\setminus z_d}}[\eta(\cdots)])$$

We will see this in practice next time.