Approximating The Posterior

Expectation Maximization and Friends

DS6040 Summer 2024
Teague R. Henry

University of Virginia | School of Data Science

Outline

- Review
- Why Approximate?
- Unsupervised Clustering as Mixture Models
- The Expectation-Maximization Algorithm
- Variational Inference Writ Large

Review

We want to obtain the posterior $\mathrm{P}(\theta|X)$:

- If we have conjugacy, we already *know* the posterior distribution.
    - It'll be the same distribution as the prior
    - It's parameters are a function of the data and the priors' hyperparameters
- If we don't have conjugacy, we can sample:
    - Samplers crawl over the posterior space until they find the posterior
    - Gibbs Samplers are for setups with proper conditionals
    - Metropolis/Metropolis Hastings Samplers are for any model setup
    - Hamiltonian MCMC is a modern way of sampling from high dimensional space.

But what are some problems with samplers?

Problems with Samplers

In order to sample from the posterior, a sampler has to find the posterior first.

- If your start values are really far away from the posterior, this might take some time.
- If you've got a really odd likelihood surface, you might not be able to traverse it with a sampler
- Categorical parameters (like class membership) are very finicky to sample from.
  - And the issues get much larger very quickly with increasing amounts of data.

To help with these issues, we can try mode approximation.

Mode Approximation

Instead of trying to sample from the posterior, we can try to find the mode of the posterior and use that to guide our sampler.

- Useful for determining starting values for a sampler

We also might have so much data, we don't care about our estimates of uncertainty. In that case, we can use the mode as a point estimate of the parameters.

- In this case, the mode of a posterior can be viewed as a penalized maximum likelihood estimate, with the penalty being the log of the prior distribution…

Optimizers!

When the model is simple, you can use your standard optimizers to find the mode of the posterior

$$\mathrm{P}(\theta|X) \propto P(X|\theta)P(\theta)$$

Newton-Raphson:

Use gradient information of the log of the posterior to climb to the mode.

$$L(\theta|X) = \log\big(P(\theta|X)\big)$$
$$\theta^t = \theta^{t-1} - [L''(\theta^{t-1})]^{-1}L'(\theta^{t-1})$$

You need to know what the first and second derivative of the log posterior is…

- Not always easy to know!
- But you can use numerical derivatives to get it.

Mixture Models and Latent Variables

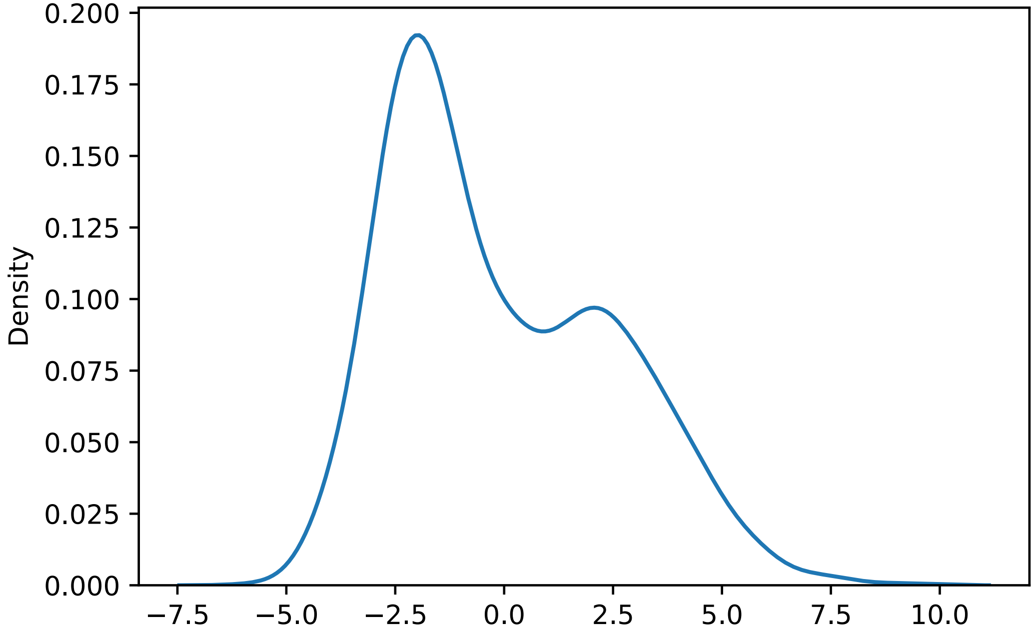
Data is rarely "nicely" distributed…

- Assumption – Once external factors are controlled for, the data is nicely distributed…

We rarely (never) have all possible important features.

- Sometimes this is ignorable…

How can we model data with features we need but don't have?

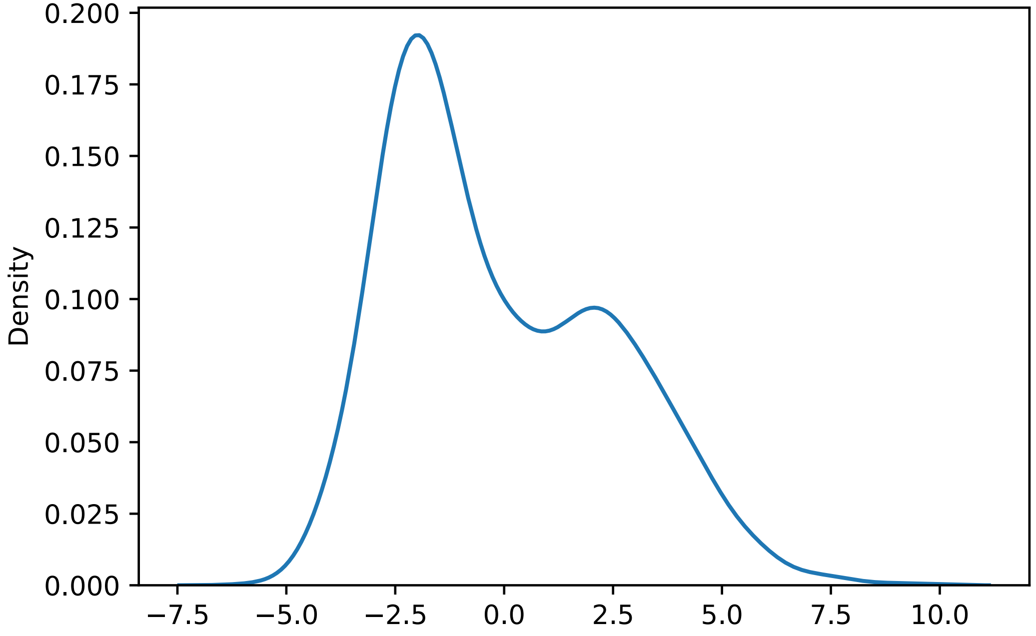


Cases to Approximate the Mode

Consider a mixture model (a parameterized cluster model)

- The class membership of individual observations are all parameters…
- The parameters that govern the clusters depend on the observations in the clusters.
- The membership of observations depends on the membership of other observations!

Very tricky to directly sample from!

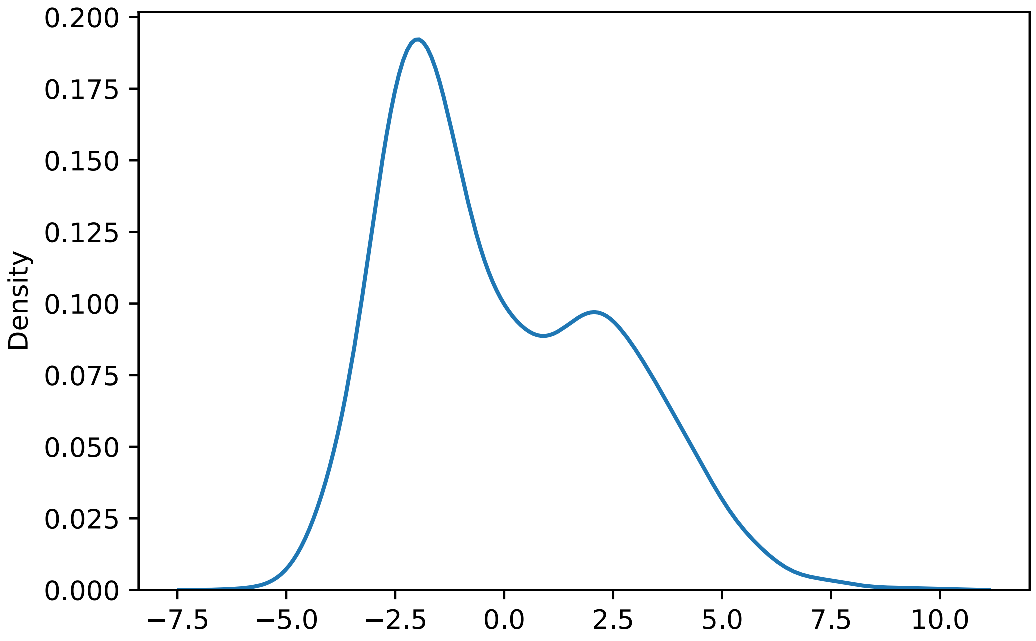


Mixture Models and Latent Variables

<u>Data Augmentation</u> – Propose unobserved (latent) variables to explain your data…

Latent variables are defined by the relations you impose between them observed data.

<u>Example</u> – Mixture membership…



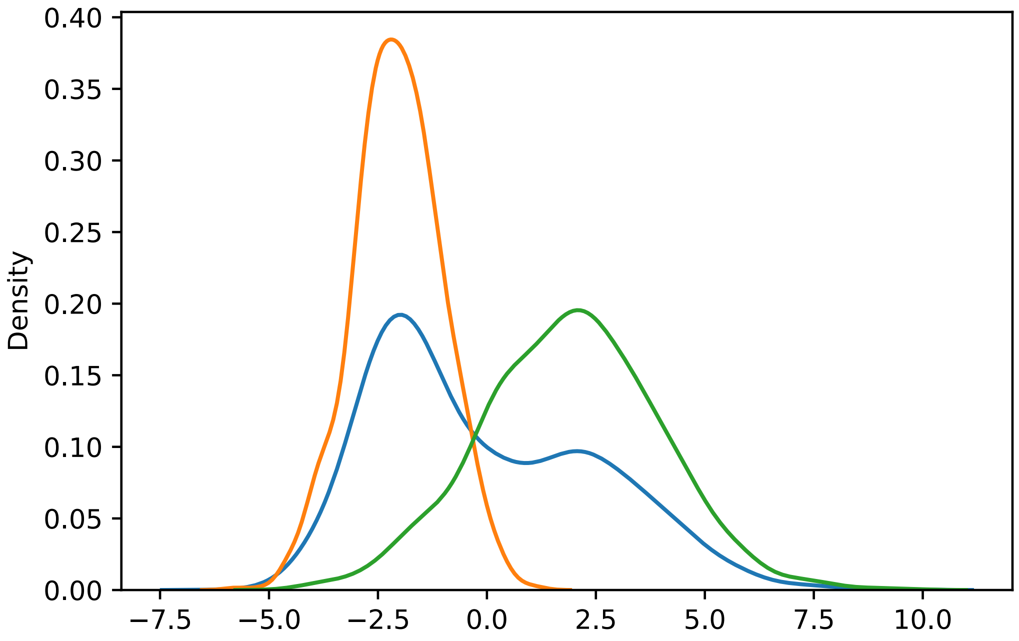
Simple Gaussian Mixture Model

Observed data - X_i

Model - $X_i \sim N\left(\mu_{z_i}, \sigma_{z_i}^2\right)$

What is z_i?

- Not a parameter…
- Not observed data…



z_i is a latent variable we made up.

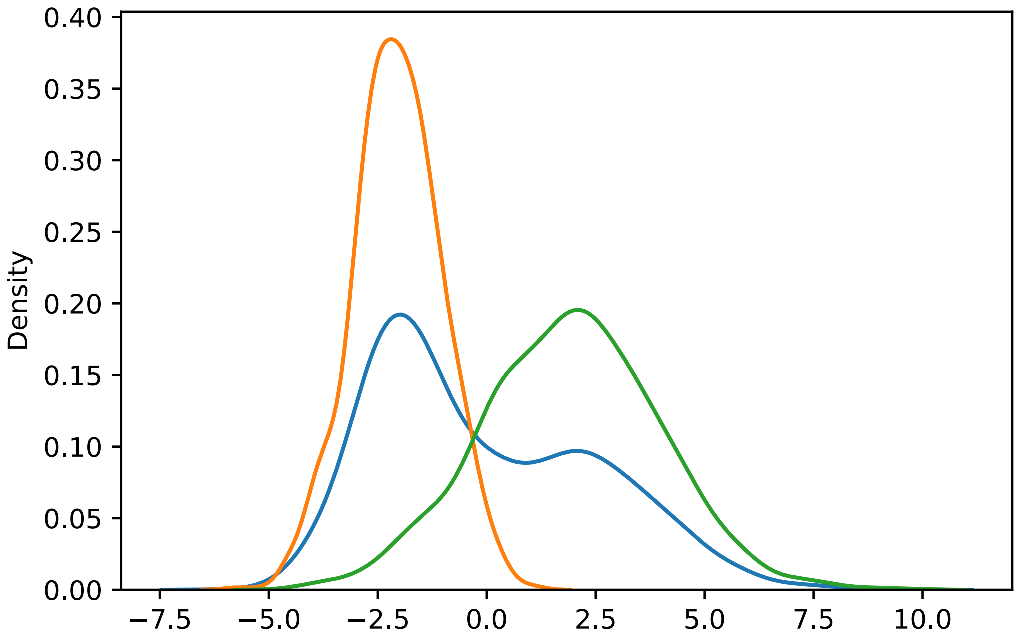
- We are making a strong claim here, that X_i is normal, conditional on z_i.

A Word of Warning

Beware clustering, mixture models and unsupervised learning methods...

<u>Reification Fallacy -</u>

Just because you find a cluster / component, doesn't mean it's a real thing...



Sometimes a weird distribution is just that, a weird distribution...

The EM Algorithm

To determine the parameters of the mixture distributions, we need to know z_i, but to know z_i we need to determine the parameters of the mixture…

- What does this sound like? (Hint: Starts with "G" and ends with "ibbs Sampler")

The Expectation Maximization Algorithm (Dempster, Laird, & Rubin (1976), ***63963 cites!***)

- Let $\mathbf{X}$ be observed data, $\mathbf{Z}$ be latent/missing data, and $\boldsymbol{\theta}$ be parameters.
- Likelihood: $p(\mathbf{X}|\boldsymbol{\theta}) = \int p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) d\mathbf{Z} = \int p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}) p(\mathbf{X}|\boldsymbol{\theta}) d\mathbf{Z}$

- <u>Expectation Step:</u> Using your current estimate of θ_t, calculate $E[\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}_t] = \mathbf{Z}_t$
- <u>Maximization Step:</u> Using your current estimate of $\mathbf{Z}_t$ as observed, estimate MLE solution for $\boldsymbol{\theta}_{t+1}$

The EM Algorithm

The Expectation Maximization Algorithm(Dempster, Laird, & Rubin (1976), *63963 cites!*)

- Let $\mathbf{X}$ be observed data, $\mathbf{Z}$ be latent/missing data, and $\boldsymbol{\theta}$ be parameters.
- Likelihood: $p(\mathbf{X}|\boldsymbol{\theta}) = \int p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) d\mathbf{Z} = \int p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}) p(\mathbf{X}|\boldsymbol{\theta}) d\mathbf{Z}$

- **Expectation Step:** Using your current estimate of $\boldsymbol{\theta}_t$, calculate $E[\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}_t] = \mathbf{Z}_t$
- **Maximization Step:** Using your current estimate of $\mathbf{Z}_t$ as observed, estimate MLE solution for $\boldsymbol{\theta}_{t+1}$

Advantages –

- By adding in “fake” data, you can make an intractable optimization problem tractable.
- EM algorithm is guaranteed to converge to a (local) minima. It just works, and works well.

Disadvantages –

- Uncertainty is underestimated, as you are feeding in $\mathbf{z}_t$ as known in each step.

Variational Inference

As we know, any "interesting" Bayesian model has an improper posterior.

<u>Samplers</u> – Construct a Markov Chain that samples from the posterior…

- Still approximate, but we are sampling from the full posterior.

<u>Variational Inference</u>

- Instead of trying to sample from an improper posterior, why don't we just… not?
- Let's instead approximate the posterior using a much simpler set of proper distributions
- That way, if we get a good approximation, we can calculate our EAP, MAP, and credible intervals directly.

Variational Inference – Technical Deets

Let $p(\theta|X)$ be our standard posterior distribution.

- Assume hard to sample from, and we don't have the analytic expression for it.

Goal: Find a distribution $q(\theta)$ that approximates $p(\theta|X)$

- Specifically $q^*(\theta) = \text{argmin}_{q(\theta)\in\mathcal{Q}} KL(q(\theta)||p(\theta|X))$
- We are looking for the distribution $q(\theta)$ that minimizes the KL divergence between it and the posterior distribution.
- But… (All expectations taken over $q(\theta)$)

$$KL(q(\theta)||p(\theta|X)) = E[\log q(\theta)] - E[\log p(\theta \cap X)] + \boxed{\log p(X)}$$

The marginal strikes again!

Variational Inference – Technical Deets

$$KL(q(\theta)||p(\theta|X)) = E[\log q(\theta)] - E[\log p(\theta \cap X)] + \log(p(X))$$

Getting closer, but we still have that pesky marginal hanging around…

Rearranging

$$\log(p(X)) = KL(q(\theta)\big|\big|p(\theta|X)\big) + E[\log p(\theta \cap X)] - E[\log q(\theta)]$$

KL is strictly positive, so the <u>**Evidence Lower Bound**</u> is defined as:

$$ELBO(q) = E[\log p(\theta \cap X)] - E[\log q(\theta)]$$

Using some probability rules:

$$ELBO(q) = \underbrace{\boxed{E[\log p(X|\theta)]}}_{\text{Just uses the likelihood!}} + \underbrace{\boxed{E[\log p(\theta)]}}_{\text{Just uses the priors!}} - \underbrace{\boxed{E[\log q(\theta)]}}_{\text{Uses our variational distribution}}$$

Variational Inference – Technical Deets

$$ELBO(q) = E[\log p(X|\theta)] + E[\log p(\theta)] - E[\log q(\theta)]$$

Maximizing the ELBO is equivalent to minimizing the KL divergence between $q(\theta)$ and $p(\theta|X)$.

<u>**To Recap:**</u>

- Approximate $p(\theta|X)$ using $q(\theta)$, the variational distribution.
- Choose $q(\theta)$ so that $ELBO(q)$ is maximized.
- Once you have your ELBO maximizing $q(\theta)$, you can treat $q(\theta)$ as your best approximation to $p(\theta|X)$
- How do you chose what $q(\theta)$ looks like?

Mean Field Approximation

How do you chose what $q(\theta)$ looks like?

- First, note that $q(\theta)$ refers to a distribution of θ, which has its own governing parameters η.
- The goal is to determine both the family of distributions and the values of η that maximize the ELBO…

Consider the situation when $\boldsymbol{\theta}$ is multidimensional (i.e. there are more than one parameter we need a posterior for.)

Mean Field Approximation

Consider the situation when θ is multidimensional (i.e. there are more than one parameter we need a posterior for.)

- $p(\boldsymbol{\theta}|\mathbf{X})$ is a complex distribution with dependencies between each marginal distribution.

Mean Field Approximation –

$$q(\boldsymbol{\theta}|\boldsymbol{\eta}) = \prod_{i=1}^{p} q(\theta_i|\eta_i)$$

The variational distributions of all parameters are independent from one another.

Mean Field Approximation

<u>Mean Field Approximation</u> –

$$q(\boldsymbol{\theta}|\boldsymbol{\eta}) = \prod_{i=1}^{p} q(\theta_i|\eta_i)$$

Advantages –

- Independent distributions are simpler to optimize η_i

Disadvantages –

- Cannot capture dependency between posterior dimensions

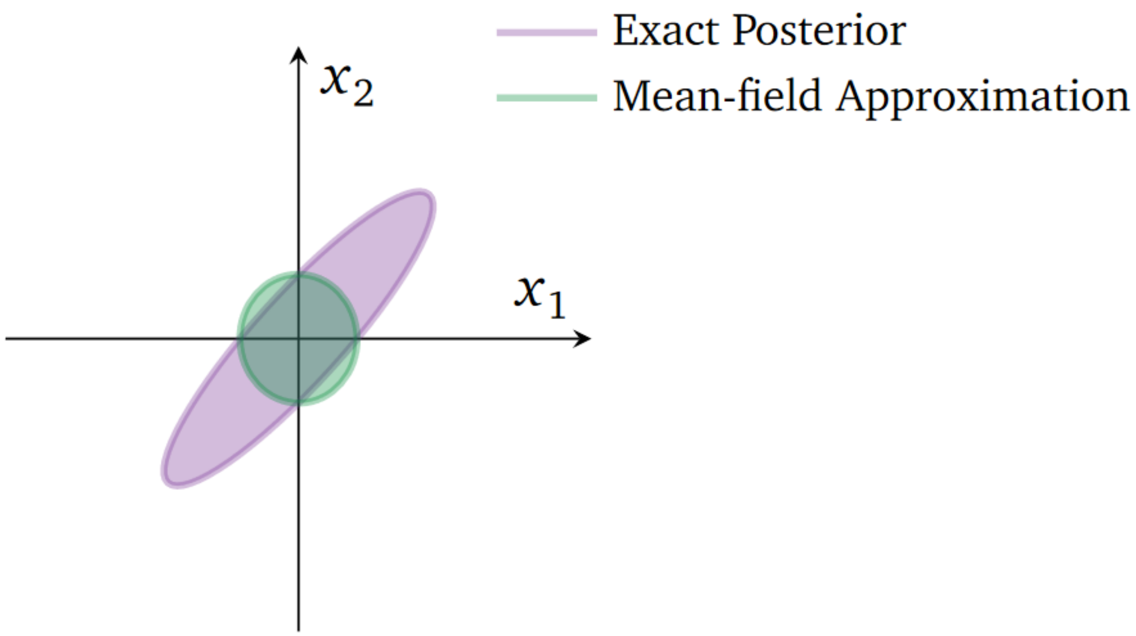


Figure from: https://arxiv.org/pdf/1601.00670.pdf
An excellent overview of variational inference ^

Determining $q^*(\theta_i)$

$$q^*(\theta_i) \propto \exp\{E_{-i}[\log p(\theta_i | X, \boldsymbol{\theta}_{-i})]\}$$

$$q^*(\theta_i) \propto \exp\{E_{-i}[\underbrace{\boxed{\log p(X | \theta_i, \boldsymbol{\theta}_{-i})}}_{\text{Just your likelihood!}} + \underbrace{\boxed{\log p(\theta_i | \boldsymbol{\theta}_{-i})}}_{\text{Given by your prior setup}} + \underbrace{\boxed{\log p(\boldsymbol{\theta}_{-i})}}_{\text{Constant with respect to } \theta_i}]\}$$

$$\log q^*(\theta_i) \propto E_{-i}[\log p(X | \theta_i, \boldsymbol{\theta}_{-i}) + \underbrace{\boxed{\log p(\theta_i | \boldsymbol{\theta}_{-i})}}_{\text{This is inside an expectation over } \boldsymbol{\theta}_{-i}}]$$

$$\log q^*(\theta_i) \propto E_{-i}[\log p(X | \theta_i, \boldsymbol{\theta}_{-i})] + \underbrace{\boxed{\log p(\theta_i)}}_{\text{Just the prior on } \theta_i}$$

The optimal $q^*(\theta_i)$ (*<u>using mean field approximation</u>*) is proportional to a function of the data (which is fixed) and **<u>fixed</u>** values of $\boldsymbol{\theta}_{-i}$

Optimizing ELBO using CAVI

$$\log q^*(\theta_i) \propto E_{-i}[\log p(X|\theta_i, \boldsymbol{\theta}_{-i})] + \log p(\theta_i)$$

This equation tells you the family of your variational distribution, and how the parameters η_i are a function of the data, priors and $\boldsymbol{\theta}_{-i}$

<u>Coordinate Ascent Variational Inference (CAVI) –</u>

- For $i \in [1, \dots, p]$:
  - Find the optimal η_i using $E_{-i}[\log p(X|\theta_i, \boldsymbol{\theta}_{-i})] + \log p(\theta_i)$, holding $\boldsymbol{\theta}_{-i}$ fixed
- Repeat until the ELBO has converged.
- Similar in theory to a Gibbs Sampler, and identical to EM.

Summary

<u>Variational Inference</u> –

- Approximate $p(\boldsymbol{\theta}|X)$ using $q(\boldsymbol{\theta})$
- Assume that $q(\boldsymbol{\theta}) = \prod_{i=1}^{p} q(\theta_i)$ (Mean Field Approximation)
- Do math to determine optimal family and variational parameters η_i
    - $\log q^*(\theta_i) \propto E_{-i}[\log p(X|\theta_i, \boldsymbol{\theta}_{-i})] + \log p(\theta_i)$
- Update the optimal η_i using X and $\boldsymbol{\theta}_{-i}$ as fixed values.
- Continue to update until ELBO converges.

At the end of this you will have $q^*(\boldsymbol{\theta}|\boldsymbol{\eta}^*)$ as your approximation to $p(\boldsymbol{\theta}|\mathbf{X})$

- You can then directly calculate EAP, MAP, and credible intervals because $q^*(\boldsymbol{\theta}|\boldsymbol{\eta}^*)$ is, by design, a proper set of distributions (which are nicely independent of one another)

Next Week

<u>Introduction to Stan!</u>

- Fundamentals of probabilistic programming
- Basic syntax
- Running Stan code in R

<u>Then, more Stan!</u>

- Complex models in Stan
- Diagnostics
- Results interpretation