Approximate Inference

Maggie Makar

January 25, 2017

This presentation combines material from lectures by David MacKay, David Blei and Michael Jordan

Overview

1. A quick introduction
2. The Generative Process
3. Inference and Intractability
4. Variational Inference
5. Monte Carlo methods (Sampling)

A quick introduction

Overview of today

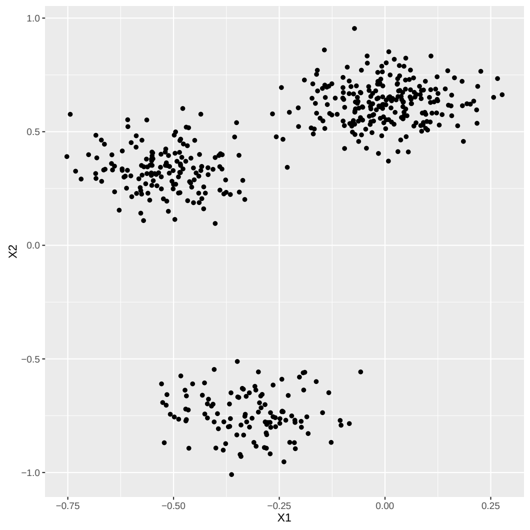
1. Why is (approximate) inference important?
2. The paradigm of probabilistic modeling
3. The importance of the posterior
4. The posterior is sometimes hard to compute!
5. The solution $\Rightarrow$ Approximate inference (MFVB, Gibbs Sampling)
6. The main example: Gaussian Mixture models

Why is (approximate) inference important?

1. A lot of really cool models require approximate inference
2. Even if not needed, approximations can make things faster
3. It is an area of very active research
4. Inference is the best section of a paper to hide a dead body

The Generative Process

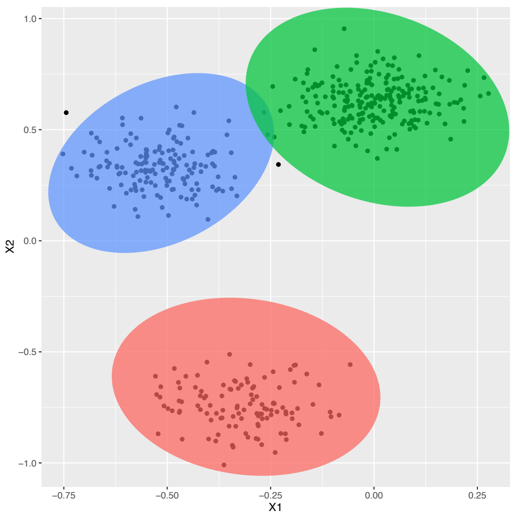
Raw data



Basics of the generative process

Asking the question: How did nature generate this data?

A hypothesis



GMM: Generative process

1. Draw the cluster location $\mu_{1:K} \sim \mathcal{N}(0, \tau^2)$
2. For $i = 1...N$:
    2.1 Draw the cluster assignment $z_i \sim Mult(\pi)$
    2.2 Draw the data point $x_i \sim \mathcal{N}(\mu_{z_i}, \sigma^2)$

GMM: Plate diagram

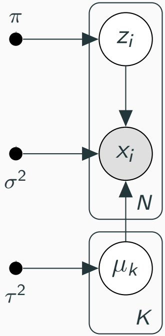
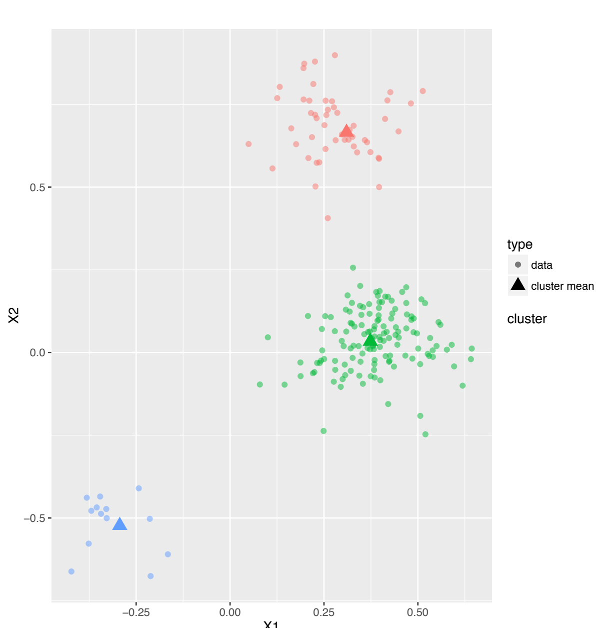


Figure 1: Plate diagram for GMM

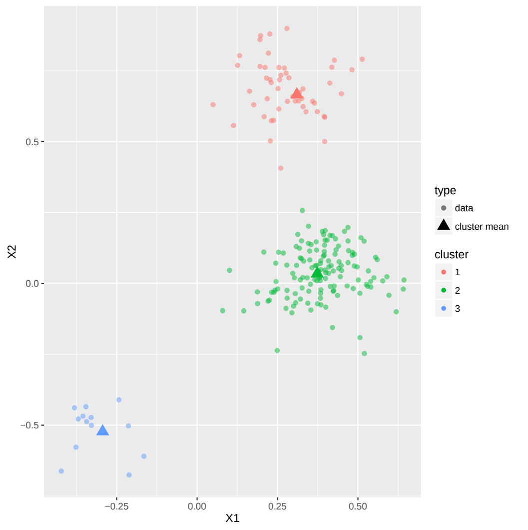
GMM: Small τ and σ

$$N = 200$$
$$\pi = (0.2, 0.7, 0.1)$$
$$\tau = 0.5$$
$$\sigma = 0.1$$

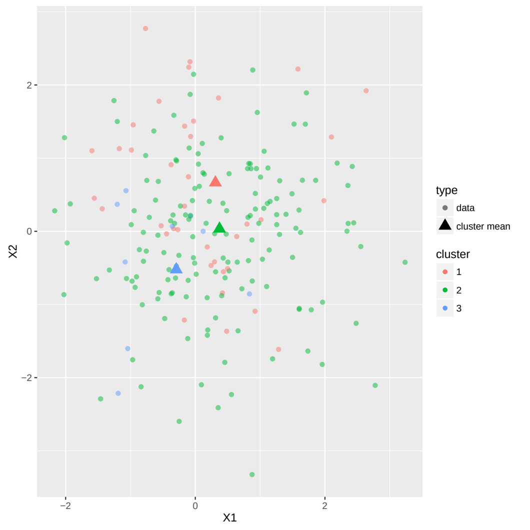
GMM: Small τ and σ



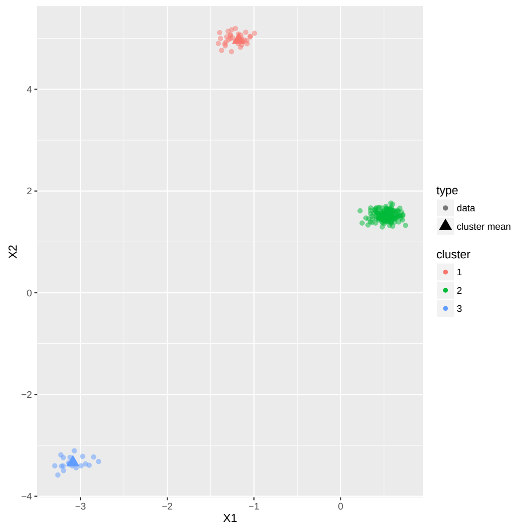
GMM: Small τ and σ



GMM: Small τ, large σ

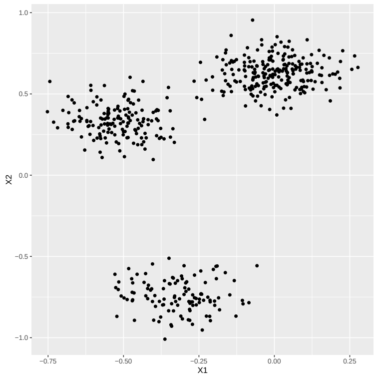


GMM: Large τ, small σ



Problem...

We do *not* have an oracle



Inference and Intractability

What's the point of inference, anyway?

1. After dreaming up the model that generates the data, we need to learn all these parameters
2. Specifically, we need the posterior distribution

The posterior distribution

$$
\begin{aligned}
\text{Posterior} &= \frac{\text{Joint}}{\text{Evidence}} \\
&= \frac{\text{Joint}}{\text{Marginalizing over all possible configurations}}
\end{aligned}
$$

The posterior distribution

$$p(\mu_{1:K}, z_{1:N}|X) = \frac{p(X|\mu_{1:K}, z_{1:N})p(\mu_{1:K}, z_{1:N})}{p(X)} \tag{1}$$

The posterior distribution

$$
\begin{aligned}
&p(\mu_{1:K}, z_{1:N} | X) \qquad (2)\\
&= \frac{\prod_{k=1}^{K} p(\mu_k) \prod_{i=1}^{N} p(z_i) p(x_i | z_i, \mu_{1:K})}{\int_{\mu_{1:K}} \sum_{z_{1:N}} \prod_{k=1}^{K} p(\mu_k) \prod_{i=1}^{N} p(z_i) p(x_i | z_i, \mu_{1:K})}
\end{aligned}
$$

The posterior distribution (not as easy on the eyes)

$$
\begin{aligned}
& p(\mu_{1:K}, z_{1:N} | X, \pi, \tau, \sigma) \\
&= \frac{\prod_{k=1}^{K} p(\mu_k|\tau_k) \prod_{i=1}^{N} p(z_i|\pi) p(x_i|z_i, \mu_{1:K})}{\int_{\mu_{1:K}} \sum_{z_{1:N}} \prod_{k=1}^{K} p(\mu_k|\tau_k) \prod_{i=1}^{N} p(z_i|\pi) p(x_i|z_i, \mu_{1:K})}
\end{aligned} \tag{3}
$$

Computing the posterior

The numerator is easy. What about the denominator?

$$
\begin{aligned}
p(X) &= \int_{\mu_1}\int_{\mu_2}\int_{\mu_3} p(\mu_1)p(\mu_2)p(\mu_3)\prod_i^N p(x_i|\mu_1,\mu_2,\mu_3) \\
&= \int_{\mu_1}\int_{\mu_2}\int_{\mu_3} p(\mu_1)p(\mu_2)p(\mu_3)\prod_i^N \sum_{k=1}^K \pi_k p(x_i|\mu_K)
\end{aligned}
$$

The denominator

Looks scary. But is it intractable?

$$\int_{\mu_1}\int_{\mu_2}\int_{\mu_3} p(\mu_1)p(\mu_2)p(\mu_3)\prod_{i}^{N}\sum_{k=1}^{K}\pi_k p(x_i|\mu_k) \tag{4}$$

Now what?

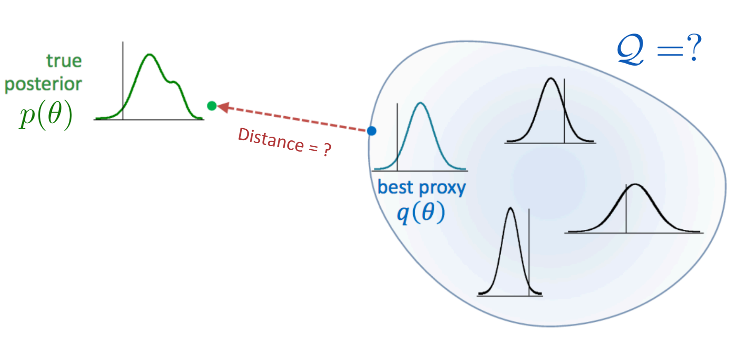
The solution? Approximate inference...

1. Laplacian approximations
2. Variational Inference
3. Monte Carlo methods

Variational Inference

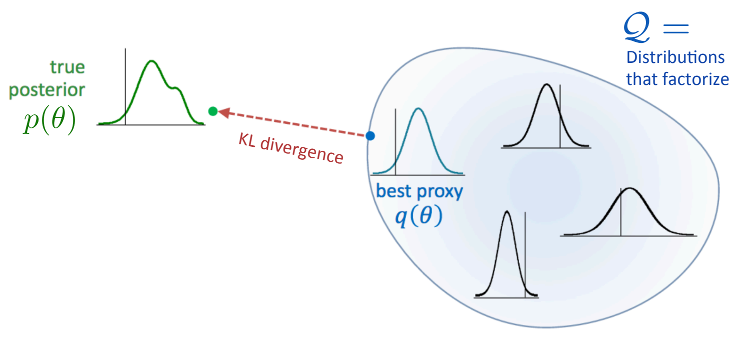
Intuition

Since this distribution is too hard to compute, let me find a "nice" distribution that is "closest" to my intractable distribution



Intuition

Close = small KL divergence, Nice = Distributions that factorize



KL divergence

$$KL(q(\theta)||p(\theta|\mathcal{D})) = \int_{-\infty}^{\infty} q(\theta) \log \frac{q(\theta)}{p(\theta|\mathcal{D})} d\theta \tag{5}$$

$$= \mathbb{E}_q \left[\log \frac{q(\theta)}{p(\theta|\mathcal{D})} \right] \tag{6}$$

KL divergence

Problem: Minimize distance to what?

KL divergence

$$
\begin{aligned}
KL(q(\theta)||p(\theta|\mathcal{D})) &= \mathbb{E}_q\left[\log \frac{q(\theta)}{p(\theta|\mathcal{D})}\right] && (7)\\
&= \mathbb{E}_q[\log q(\theta)] - \mathbb{E}_q[\log p(\theta|\mathcal{D})] && (8)\\
&= \mathbb{E}_q[\log q(\theta)] - \mathbb{E}_q[\log p(\theta,\mathcal{D})] + \mathbb{E}_q[\log p(\mathcal{D})] && (9)\\
&= \mathbb{E}_q[\log q(\theta)] - \mathbb{E}_q[\log p(\theta,\mathcal{D})] + \log p(\mathcal{D}) && (10)
\end{aligned}
$$

Evidence Lower Bound for GMM

$$\log p(X) = \log \int_{\mu_{1:K}} \sum_{z_{1:N}} p(\mu_{1:K}, z_{1:N}, X) \tag{11}$$

$$= \log \int_{\mu_{1:K}} \sum_{z_{1:N}} p(\mu_{1:K}, z_{1:N}, X) \frac{q(\mu_{1:K}, z_{1:N})}{q(\mu_{1:K}, z_{1:N})} \tag{12}$$

$$\geq \mathbb{E}_q[\log p(\mu_{1:K}, z_{1:N}, X)] - \mathbb{E}_q[\log q(\mu_{1:K}, z_{1:N})] \tag{13}$$

$$\triangleq \mathcal{L}(q) \tag{14}$$

It turns out that the KL divergence can be decomposed to:

$$KL(q(\theta)||p(\theta|\mathcal{D})) = -\mathcal{L}(q) + \log p(X) \tag{15}$$

A nice distribution for GMM

Mean Field Variational Bayes: When we choose a $\mathcal{Q}$ that factorizes

$$q(\mu_{1:K}, z_{1:N}) = \prod_{k=1}^{K} q(\mu_k | m_k, s_k^2) \prod_{i=1}^{N} q(z_i | \phi_i) \tag{16}$$

We call $m_{1:K}, s^2_{1:K}, \phi_{1:N}$ the variational parameters

Coordinate Ascent Algorithm

1. Treat all except for one variational distribution (say $q(z_5|\phi_5)$) as "fixed"
2. Find the value of $q(z_5|\phi_5)$ that maximizes the ELBO
3. Iterate through each of the variational distributions

Practically

Derive the ELBO

- Step 1: Write the joint probability distribution
- Step 2: Subtract the entropy of the variational distribution
- Step 3: Fully expand each term
- Step 4: Every time you come across a latent variable, replace it with the expectation under q

Compute the updates

- Step 5: Pick one of the variational distributions that you have (e.g., $q(z_5|\phi_5)$)
- Step 6: Collect all the terms in the ELBO that depend on $q(z_5|\phi_5)$
- Step 7: Set the derivative $\frac{\partial \mathcal{L}}{\partial q(z_5)} = 0$, solve for ϕ_5 to get the value of ϕ_5 that maximizes the EBLO
- Step 8: Repeat steps from 5-7 for all your variational distributions till convergence.

MFVB for GMM

Algorithm 1 MFVB for GMM

Input: data X, number of components K

Initialize Variational parameter: $m_{1:K}, s^2_{1:K}, \phi_{1:N}$, $\mathit{Converged} = \mathit{FALSE}$

repeat

  for $i \in \{1, .., N\}$ **do**

    Set $\phi_{i,k} \propto \exp\{\mathbb{E}_q[\mu_k]x_i - \frac{\mathbb{E}_q[\mu_k^2]}{2}\}$

  end for

  for $K \in \{1, .., K\}$ **do**

    Set $m_k \leftarrow \frac{\sum_i \phi_{i,k} x_i}{1/\sigma^2 + \sum_i \phi_{i,k}}$

    Set $s_k^2 \leftarrow \frac{1}{1/\sigma^2 + \sum_i \phi_{i,k}}$

  end for

  Compute ELBO

  Test for convergence

until *Converged* is *TRUE*

Monte Carlo methods (Sampling)

Sampling overview

1. Approximates the intractable distribution by sampling
2. Law of large numbers $\Rightarrow$ unbiased estimators

$$\frac{1}{N}\sum_{i}^{N} f(x_i) = \mathbb{E}[f(x)] \tag{17}$$

Monte Carlo methods: a background

Sampling methods that we can use when

1. We want to evaluate $p(\theta) = \frac{p^*(\theta)}{Z}$
2. Can evaluate $p^*(\theta)$
3. We don't know Z

Monte Carlo methods: a background

- **Step 1** Find “good candidates” of θ
- **Step 2** Estimate the some expression at the value of θ

E.g: Importance and rejection sampling

Monte Carlo methods: a background

- **Step 1** $\theta^{(r)} \sim \mathcal{S}(\theta)$
- **Step 2** Estimate $p^*(\theta^{(r)})$

Monte Carlo Markov Chain (MCMC)

- **Step 1** At time step t: $\theta^{(r)} \sim \mathcal{S}(\theta|\theta^{(t-1)})$
- **Step 2** Estimate $p^*(\theta^{(r)})$

Gibbs sampling

- Assume that I can sample from *conditionals*
- Initialize at random
- Pick one variable at a time and sample it's value conditional on all the others and the data

Conditionals for GMM

Cluster assignment conditionals

$$p(z_i|\mu_{1:K}, z_{-i}, X) = p(z_i|\mu_{1:K}, x_i) \tag{18}$$

Conditionals for GMM

Cluster location conditionals

$$p(\mu_k|\mu_{-k}, z_{1:N}, X) = p(\mu_k|z_{1:N}, X) \tag{19}$$

Gibbs for GMM

Algorithm 2 Gibbs for GMM

Input: data X, number of components K

Initialize mixture locations $\mu_{1:K}$

repeat

  for $i \in \{1, .., N\}$ **do**

    Sample $z_i | \mu_{1:K}, z_{-i}, X$

  end for

  for $K \in \{1, .., K\}$ **do**

    Sample $\mu_k | \mu_{-k}, z_{1:N}, X$

  end for

until *Converged* is *TRUE*

But which method of inference is better?

Neither!

VI vs Sampling

Optimization vs. Monte Carlo principle (random numbers)

VI vs Sampling

Biased vs. Slow

But...

1. There is a lot of work on making sampling faster (e.g., Neal. MCMC using Hamiltonian dynamics)
2. And there is a lot of work on quantifying the bias in VI (e.g, Giordano, Broderick, and Jordan. *Robust Inference with Variational Bayes*, NIPS 2015.)

Some references

- Tutorial on VI: http://digitalassets.lib.berkeley.edu/techreports/ucb/text/CSD-98-980.pdf
- A Review of recent work on VI: Section 5 in https://arxiv.org/pdf/1602.05221v2.pdf
- Tutorial on Sampling methods http://www.cs.ubc.ca/~arnaud/andrieu_defreitas_doucet_jordan_intromontecarlomachinelearning.pdf
- A review (and really cool demos) of recent work on sampling http://chifeng.scripts.mit.edu/stuff/mcmc-demo/