Bayesian Mixture Models and the Gibbs Sampler

David M. Blei
Columbia University

October 19, 2015

We have discussed probabilistic modeling, and have seen how the posterior distribution is the critical quantity for understanding data through a model.

The goal of probabilistic modeling is use domain and data-knowledge to build structured joint distributions, and then to reason about the domain (and exploit our knowledge) through the posterior and posterior predictive distributions.

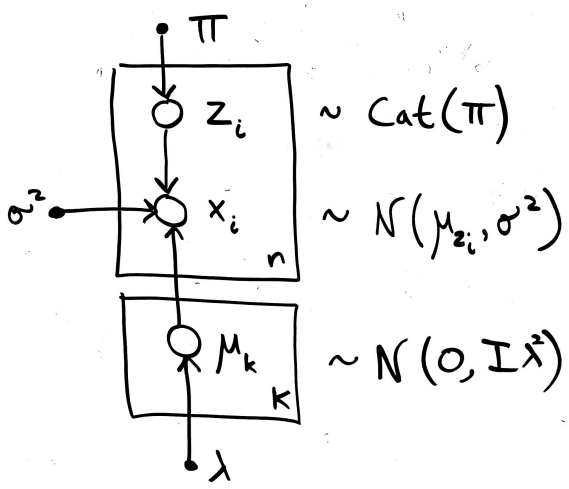
We have discussed tree propagation, a method for computing posterior marginals of any variables in a tree-shaped graphical model. In theory, if our graphical model was a tree, we could shade the observations and do useful inferences about the posterior.

For many interesting models, however, the posterior is not tractable to compute. Either the model is not a tree or the messages are not tractable to compute (because of the form of the potentials). Most modern applications of probabilistic modeling rely on *approximate posterior inference* algorithms.

We now discuss an important method for approximate posterior inference. In parallel, we begin to discuss the building blocks of complex models.

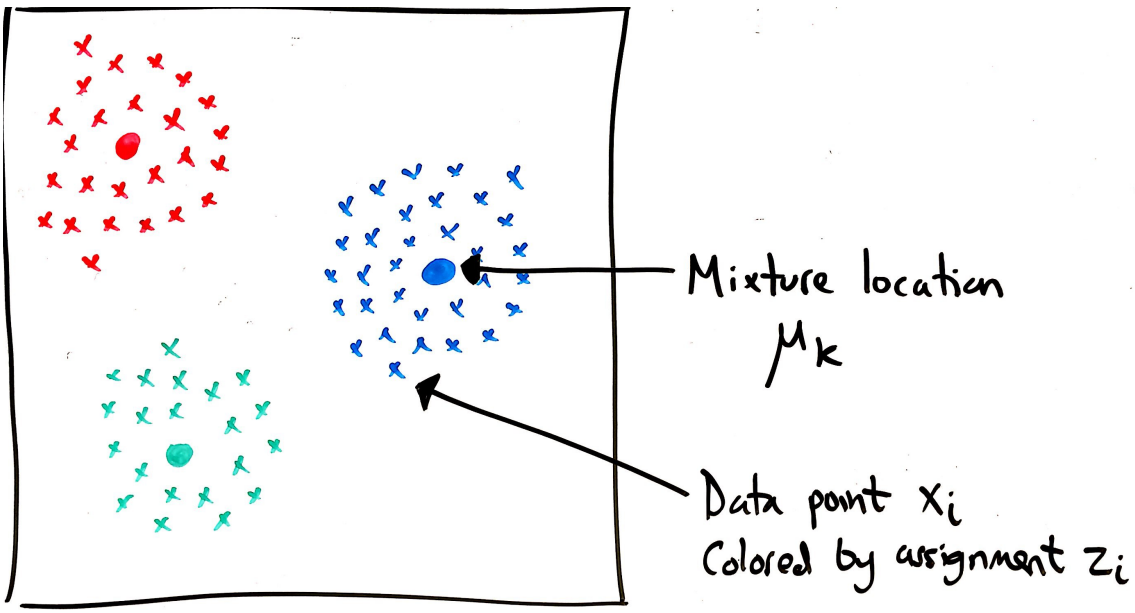
Bayesian mixture of Gaussians

To lock ideas, and to give you a flavor of the simplest interesting probabilistic model, we will first discuss Bayesian mixture models. Here is the Bayesian mixture of Gaussians,



(We haven't yet discussed the multivariate Gaussian. But we don't need to yet. Here each mixture component is $\mu_k = [\mu_x, \mu_y]$ and we generate the x and y coordinates independently from their respective distributions.)

To get a feel for what this model is about, we generate data from it.

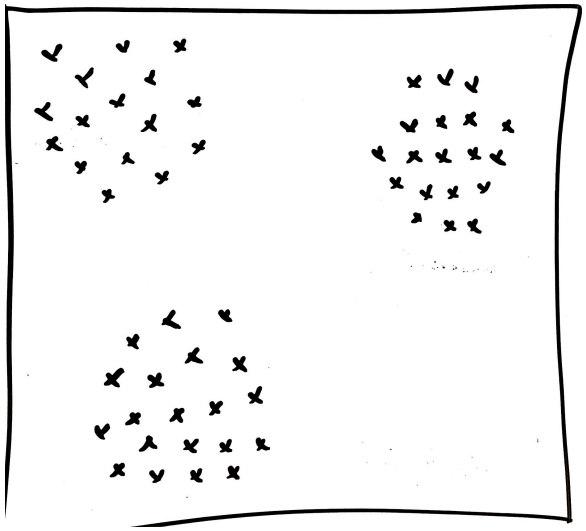


The posterior and posterior predictive distributions

¶ The posterior distribution is a distribution over the latent variables, the cluster locations and the cluster assignments. Let $\mathbf{x} = x_{1:n}$, $\mathbf{z} = z_{1:n}$, and $\boldsymbol{\mu} = \mu_{1:K}$. The posterior is

$$p(\mathbf{z}, \boldsymbol{\mu} \mid \mathbf{x}). \tag{1}$$

This gives an understanding of the data (at least, a grouping into K groups).



What is this posterior? The mixture model assumes that each data point came from one of K distributions. However, it is unknown what those distributions are and how the data were assigned to them. The posterior is a conditional distribution over these quantities.

¶ As usual, the posterior also gives a posterior predictive distribution,

$$\begin{aligned} p(x_{n+1} \mid \mathbf{x}) &= \int p(x_{n+1}, \boldsymbol{\mu} \mid \mathbf{x}) \mathrm{d}\boldsymbol{\mu} && (2)\\ &= \int p(x_{n+1} \mid \boldsymbol{\mu}, \mathbf{x}) p(\boldsymbol{\mu} \mid \mathbf{x}) \mathrm{d}\boldsymbol{\mu} && (3)\\ &= \int p(x_{n+1} \mid \boldsymbol{\mu}) p(\boldsymbol{\mu} \mid \mathbf{x}) \mathrm{d}\boldsymbol{\mu} && (4)\\ &= \mathbb{E}\left[p(x_{n+1} \mid \boldsymbol{\mu}) \mid \mathbf{x}\right]. && (5) \end{aligned}$$

Note we removed $\mathbf{x}$ from the RHS of the conditional distribution. This is thanks to conditional independence. From Bayes ball we have that

$$x_{n+1} \perp\!\!\!\perp \mathbf{x} \mid \boldsymbol{\mu} \tag{6}$$

In more detail, the posterior predictive distribution is

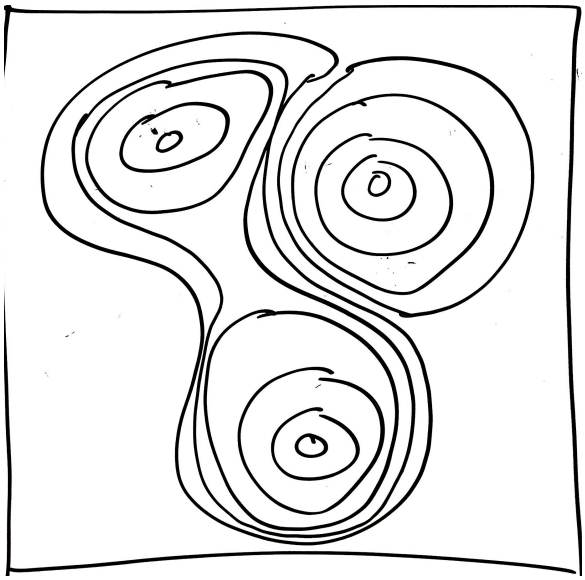
$$p(x_{n+1} \mid \mathbf{x}) = \int_{\boldsymbol{\mu}} p(x_{n+1} \mid \boldsymbol{\mu}) p(\boldsymbol{\mu} \mid \mathbf{x}) d\boldsymbol{\mu} \tag{7}$$

$$= \int_{\boldsymbol{\mu}} \left(\sum_{k=1}^{K} p(z_{n+1} = k) p(x_{n+1} \mid \mu_k) \right) d\boldsymbol{\mu} \tag{8}$$

$$= \sum_{k=1}^{K} p(z_{n+1} = k) \left(\int_{\mu_k} p(x_{n+1} \mid \mu_k) p(\mu_k \mid \mathbf{x}) d\mu_k \right) \tag{9}$$

¶ What is this? We consider x_{n+1} as coming from each of the possible mixture locations (one through K) and then take a weighted average of its posterior density at each.

This is a multi-modal distribution over the next data point. Here is a picture:



This predictive distribution involves the posterior through $p(\mu_k \mid \mathbf{x})$, the posterior distribution of the kth component given the data.

Contrast this with the predictive distribution we might obtain if we used a single Gaussian to model the data. In that case, the mean will be at a place where there is little data.

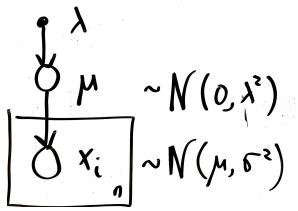
¶ Through the posterior, a mixture model tells us about a grouping of our data, and captures complex predictive distributions of future data.

The posterior is intractable to compute

¶ We cannot compute the posterior exactly. Let's see why.

¶ First an aside: the Gaussian is conjugate to the Gaussian.

¶ Consider a simple model, where we draw a Gaussian mean μ from a Gaussian prior $\mathcal{N}(0, \lambda)$ and then generate n data points from a Gaussian $\mathcal{N}(\mu, \sigma^2)$. (We fix the variance σ^2.) Here is the graphical model



¶ You have seen the beta-Bernoulli; this is another example of a conjugate pair. Given $\mathbf{x}$ the posterior distribution of μ is $\mathcal{N}(\hat{\mu}, \hat{\lambda})$, where

$$\hat{\mu} = \left(\frac{n/\sigma^2}{n/\sigma^2 + 1/\lambda^2} \right) \bar{x} \tag{10}$$

$$\hat{\lambda} = \left(n/\sigma^2 + 1/\lambda^2 \right)^{-1}, \tag{11}$$

where $\bar{x}$ is the sample mean.

As for the beta-Bernoulli, as n increases the posterior mean approaches the sample mean and the posterior variance approaches zero. (Note: this is the posterior mean and variance of the unknown *mean*. The data variance σ^2 is held fixed in this analysis.)

¶ But now suppose we are working with a mixture of Gaussians. In that case, $p(\boldsymbol{\mu} \mid \mathbf{x})$ is not easy. Suppose the prior proportions π are fixed and $K = 3$,

$$p(\mu'_1, \mu'_2, \mu'_3 \mid \mathbf{x}) = \frac{p(\mu'_1, \mu'_2, \mu'_3, \mathbf{x})}{\int_{\mu_1} \int_{\mu_2} \int_{\mu_3} p(\mu_1, \mu_2, \mu_3, \mathbf{x})}. \tag{12}$$

¶ The numerator is easy,

$$\text{numerator} = p(\mu'_1) p(\mu'_2) p(\mu'_3) \prod_{i=1}^{n} p(x_i \mid \mu'_1, \mu'_2, \mu'_3) \tag{13}$$

where each likelihood term marginalizes out the z_i variable,

$$p(x_i \mid \mu'_1, \mu'_2, \mu'_3) = \sum_{k=1}^{K} \pi_k \, p(x_i \mid \mu'_k). \tag{14}$$

¶ But consider the denominator, which is the marginal probability of the data,

$$p(\mathbf{x}) = \int_{\mu_1} \int_{\mu_2} \int_{\mu_3} p(\mu_1) p(\mu_2) p(\mu_3) \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k \, p(x_i \mid \mu_k). \tag{15}$$

¶ One way to see this is to simply believe me. Another way is to bring the summation to the outside of the integral

$$p(\mathbf{x}) = \sum_{\mathbf{z}} \int p(\mu_1)p(\mu_2)p(\mu_3) \prod_{i=1}^{n} p(x_i \mid \mu_{z_i}). \tag{16}$$

This can be decomposed by partitioning the data according to $\mathbf{z}$,

$$p(\mathbf{x}) = \sum_{\mathbf{z}} \prod_{k=1}^{3} \left(\int_{\mu_k} p(\mu_k) \prod_{\{i:z_i=k\}} p(x_i \mid \mu_k) \right). \tag{17}$$

Each term in the product is an integral under the conjugate prior, which is an expression we can compute. (We will see that later on.) But there are 3^n different assignments of the data to consider.

¶ To work with Bayesian mixtures of Gaussians (and many other models), we need approximate inference.

¶ Show a mixture model fit to real data, e.g., the image mixture model.

The Gibbs sampler

¶ The main idea behind Gibbs sampling (and all of MCMC) is to approximate a distribution with a set of samples. For example, in the mixture model,

$$p(\mu, z \mid x) \approx \frac{1}{B} \sum_{b=1}^{B} \delta_{(\mu^{(b)}, z^{(b)})}(\mu, z), \tag{18}$$

where we shorthand $\mu = \boldsymbol{\mu}$ and $z = \mathbf{z}$.

¶ Let's first discuss Gibbs sampling for mixtures of Gaussians. Then we will see how it generalizes and why it works.

¶ In the Gibbs sampler, we maintain a value for each latent variable. In each iteration, sample from each latent variable conditional on the other latent variables and the observations. I like to call this distribution a *complete conditional*.

¶ See Figure 1 for Gibbs sampling for Gaussian mixtures.

¶ Note that within an iteration, when we sample one variable its value changes in what we subsequently condition on. E.g., when we sample μ_1, this changes what μ_1 is for the subsequent samples.

¶ The theory around Gibbs sampling says that if we do this many times, the resulting sample will be a sample from the true posterior.

Input: Data $\mathbf{x}$ and a number of components K
Initialize: Mixture locations $\boldsymbol{\mu}$.

Maintain mixture locations $\boldsymbol{\mu}$ and mixture assignments $\mathbf{z}$.
Repeat:

1. For each $i \in \{1, \ldots, n\}$:
   Sample $z_i \mid \{\boldsymbol{\mu}, z_{-i}, \mathbf{x}\}$ from Equation (21)
2. For each $k \in \{1, \ldots, K\}$:
   Sample $\mu_k \mid \{\mu_{-k}, \mathbf{z}, \mathbf{x}\}$ from Equation (23)

Figure 1: The Gibbs sampler for mixture of Gaussians.

.

The reason is that we have defined a Markov chain whose state space are the latent variables and whose *stationary distribution* is the posterior we care about. After a long time, a sample of $\boldsymbol{\mu}$ and $\mathbf{z}$ is a sample from the posterior. Doing this multiple times, we can obtain B samples from the posterior.

Details about the complete conditionals

¶ Let's work out each step of the algorithm, beginning with the complete conditional of z_i. We first look at the graphical model and observe a conditional independence,

$$p(z_i \mid \boldsymbol{\mu}, z_{-i}, \mathbf{x}) = p(z_i \mid \boldsymbol{\mu}, x_i). \tag{19}$$

Now we calculate the distribution,

$$p(z_i \mid \boldsymbol{\mu}, x_i) \propto p(z_i)p(x_i \mid \mu_{z_i}) \tag{20}$$

$$= \pi_{z_i}\phi(x_i; \mu_{z_i}, \sigma^2). \tag{21}$$

What is this? To keep things simple, assume $\pi_k = 1/K$. Then this is a categorical distribution where the probability of the the kth is proportional to the likelihood of the ith data point under the kth cluster.

Notes:

- Categorical distributions are easy to sample from.
- This distribution requires that we know $\boldsymbol{\mu}$.

¶ Now let's derive the complete conditional of μ_k. Again, we observe a conditional independence from the graphical model,

$$p(\mu_k \mid \mu_{-k}, \mathbf{z}, \mathbf{x}) = p(\mu_k \mid \mathbf{z}, \mathbf{x}) \tag{22}$$

Here let's calculate the distribution intuitively. If we know the cluster assignments, what is the conditional distribution of μ_k? It is simple a posterior Gaussian, conditional on the data that were assigned to the kth cluster.

Technically: Let z_i be an indicator vector, a K-vector with a single one. Then,

$$\mu_k \mid \mathbf{z}, \mathbf{x} \sim \mathcal{N}(\hat{\mu}_k, \hat{\lambda}_k) \tag{23}$$

where

$$\hat{\mu}_k = \left(\frac{n_k/\sigma^2}{n_k/\sigma^2 + 1/\lambda^2} \right) \bar{x}_k \tag{24}$$

$$\hat{\lambda} = \left(n_k/\sigma^2 + 1/\lambda^2 \right)^{-1}, \tag{25}$$

and

$$n_k = \sum_{i=1}^{n} z_i^k \tag{26}$$

$$\bar{x}_k = \frac{\sum_{i=1}^{n} z_i^k x_i}{n_k}. \tag{27}$$

Important: Conjugacy helps us, even when we cannot compute the posterior.

¶ This is an approximate inference algorithm for mixtures of Gaussians. At each iteration, we first sample each mixture assignment from Equation (21) and then sample each mixture location from Equation (23).

¶ Discussion:

- This sampler results in one sample from the posterior. To get B samples, we run several times. In practice, we begin from an *initial state* and run for a fixed number of *burn in* iterations. We then continue to run the algorithm, collecting samples a specified *lag*.

  Initialization, burn-in, and lag are important practical issues. There are no good principled solutions, but many ad-hoc ones work well.

- Notice that conditional independencies in the complete conditionals give us opportunities to parallelize. What can be parallelized here?

- Gibbs sampling closely relates to the expectation-maximization (EM) algorithm.

  In the EM algorithm we iterate between the E-step and M-step. In the E-step we compute the conditional distribution of each assignment given the locations. This is precisely Equation (21).

  In the M-step we update the locations at maximum likelihood estimates under expected sufficient statistics. As we know, the MLE relates to the Bayesian posterior in Equation (24).

- The theory implies that we need infinite lag time and infinite burn-in. Practical decisions around Gibbs sampling can be difficult to make. (But, happily, in practice it's easy to come up with sensible unjustified choices.) One quantity to monitor is $\log p(\boldsymbol{\mu}^{(t)}, \mathbf{z}^{(t)}, \mathbf{x})$, i.e., the log joint of the assignments of the latent variables and observations.

  This further relates to EM, which optimizes the conditional expectation (over the mixture assignments z) of this quantity.

The collapsed Gibbs sampler

¶ Sometimes we can integrate out hidden random variables from a complete conditional. This is called *collapsing*.

¶ It is a good idea because it usually leads to faster convergence of the Markov chain to the stationary distribution. But it is also usually more costly per iteration.

¶ In the mixture of Gaussians, consider collapsing the mixture locations,

$$p(z_i = k \mid z_{-i}, \mathbf{x}) \propto p(z_i = k)\, p(x_i \mid z_{-i}, x_{-i}, z_i = k). \tag{28}$$

¶ The second term is simply a posterior predictive distribution

$$p(x_i \mid z_{-i}, x_{-i}, z_i) = \int_{\mu_k} p(x_i \mid \mu_k)\, p(\mu_k \mid z_{-i}, x_{-i}). \tag{29}$$

¶ Figure 2 shows collapsed Gibbs sampling for Gaussian mixtures.

¶ Collapsed Gibbs sampling can be more expensive at each iteration, but converges faster. Typically, if you can collapse then it is worth it.

Gibbs sampling in general

¶ The ease of implementing a Gibbs sampler depends on how easy it is to compute and sample from the various complete conditionals.

¶ In a graphical model, the complete conditional depends on the *Markov blanket* of the node.

¶ Suppose the nodes are $x_1, \ldots, x_k$ (observed and unobserved).

In an undirected graphical model, the complete conditional only depends on a node's neighbors,

$$p(x_i \mid x_{-i}) = p(x_i \mid x_{\mathcal{N}(i)}). \tag{32}$$

Input: Data $\mathbf{x}$ and a number of components K
Initialize: Mixture assignments $\mathbf{z}$.

Maintain mixture assignments $\mathbf{z}$ and two derived quantities:

$$n_k \triangleq \sum_{i=1}^{n} z_i^k \qquad \text{(number of items per cluster)} \tag{30}$$

$$s_k \triangleq \sum_{i=1}^{n} z_i^k x_i \qquad \text{(cluster sum)} \tag{31}$$

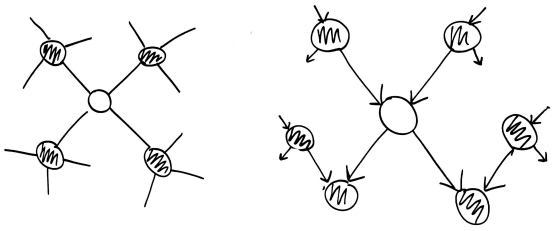
Repeatedly cycle through each data point $i \in \{1, \ldots, n\}$:

1. "Knock out" x_i from its currently assigned cluster z_i. Update n_k and s_k for its assigned cluster.
2. Sample z_i from Equation (28). The posterior Gaussian $p(\mu_k \mid z_{-i}, x_{-i})$ can be computed from the n_k and s_k.

Figure 2: Collapsed Gibbs sampling for Gaussian mixtures.

In a directed model, the complete conditional depends on a node's parents, children, and other parents of its children,

¶ We can see these facts from the graphical model and separation / d-separation



¶ Theme: Difficult global computation is made easier in terms of many local computations. Notice how Gibbs sampling is a form of "message passing".