Bayesian Modeling
Mixture models

Yixin Wang

11 Unsupervised learning and nonparametric Bayes

Unsupervised learning is a term that originates from machine learning, but it basically refers to a class of learning problems and techniques that involves latent variable models.

The most basic instance of unsupervised learning is the problem of clustering. The problem of clustering is often vaguely formulated as follows: given n data points $X_1, \ldots, X_n$ residing in some space, say $\mathbb{R}^d$, how do one subdivide these data into a number of clusters of points, in a way so that the data points belong to the same cluster are more similar than those from different clusters.

A popular method is called the k-means algorithm, which is a simple and fast procedure for obtaining k clusters for a given $k < \infty$. There is only limited theoretical basis for such an algorithm.

To provide a firm foundation for clustering, a powerful approach is to introduce additional probabilistic structures for the data. Such modeling is important to provide guarantee that we are doing the right thing under certain assumptions, but more importantly it opens up new venues for developing more sophisticated clustering algorithms as additional information about the data set or requirement about the inference become available.

The most common statistical modeling tool is mixture models. A mixture distribution admits the following density:

$$p(x|\boldsymbol{p}, \boldsymbol{\phi}) = \sum_{j=1}^{k} p_j f(x|\phi_j)$$

where f is a known density kernel, k is the number of mixing components. p_j and ϕ_j are the mixing probability and parameter associated with component j. When k is finite, this is the pdf of a *finite mixture model*.

Given n-iid sample $\boldsymbol{X} := (x_1, \ldots, x_n)$ from this mixture density, it is possible to obtain the parameters ϕ_j via maximum likelihood estimation, which can be achieved by the Expectation-Maximization (EM) algorithm. In fact, the EM algorithm can be viewed to be a generalization of the popular k-means algorithm mentioned above. [14]

By taking a Bayesian approach to the learning of mixture model, we will see that a Gibbs sampler for posterior inference with a suitable choice of conjugate priors is a probabilistic version of the EM algorithm (and k-means algorithm). Thus, the Bayesian approach can produce comparable estimate as that of EM, but with the advantage of uncertainty quantification.

[14] You may ignore any references to the k-means and the EM algorithm in this set of notes if you have not seen these algorithms before.

The question of model selection, i.e., how to select k the number of mixture components, requires the development of a new framework known as *Bayesian nonparametrics* (BNP): The number of relevant parameters will be unknown, random, and potentially unbounded. Thus the totality of all potential parameters will be infinite. This requires new ideas for the prior construction and computational methods. The outcome is an elegant solution to the model selection in that the number of the parameters will be shown to be increasing *a posteriori* as the data sample size increases.

In this chapter we use the unsupervised learning as a setting to introduce BNP, but this is a new, general and powerful way to apply Bayesian analysis to any class of statistical problems.[15]

In the Bayesian nonparametric framework for unsupervised learning, the corresponding model for the clustering problem will be called *infinite mixture models* that are endowed with suitable nonparametric Bayesian priors.

[15]Good references for Bayesian nonparametrics include Hjort et al. [2010], Ghosh and Ramamoorthi [2002], Ghosal and van der Vaart [2017].

11.1 Finite mixture models

Consider a finite mixture of normal distribution on the real line:

$$p(x|\boldsymbol{p}, \boldsymbol{\phi}) = \sum_{j=1}^{k} p_j \, \mathrm{N}(x|\phi_j, \sigma^2),$$

where the parameters are $\boldsymbol{p} = (p_1, \ldots, p_k)$ and the mean parameters $\phi_1, \ldots, \phi_k \in \mathbb{R}$. σ^2 is assumed known for simplicity.

For prior specification, we take

$$\phi_j \overset{indep}{\sim} N(\mu, \tau^2)$$

for $j = 1, \ldots, k$, for some hyperparameters μ and τ. The mixing probability vector $\boldsymbol{p} = (p_1, \ldots, p_k) \in \Delta^{k-1}$ will be endowed with a Dirichlet prior,

$$\boldsymbol{p} \sim \mathcal{D}_k(\boldsymbol{\alpha}).$$

Recall that the Dirichlet distribution on Δ^{k-1} requires positive valued hyperparameters $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_k)$.

11.1.1 Auxiliary variables

Now we introduce a very common and powerful technique in Bayesian inference: instead of working directly with the original (mixture) model, we shall introduce additional auxiliary latent variables in a joint model. When the auxiliary variables are integrated out, we get back the original model.

The main advantage of this technique is in the posterior computation. The joint posterior distribution (with the auxiliary variables included) tend to be easier too work with via Gibbs sampling or other MCMC updates, because the full conditional distributions are easy to compute: in the presence of the auxiliary variables, the prior that was not semiconjugate with respect to the original model becomes semiconjugate with respect to the joint model.

For our current mixture model, we need on auxiliary variable for each sample x_n: $\boldsymbol{Z} := (Z_1, \ldots, Z_n)$, where each $Z_i \in \{1, \ldots, k\}$. Z_i is interpreted as the (unknown and random) label of the mixture component from which the data X_i is generated.

The joint model $p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{p}, \boldsymbol{\phi})$ with the auxiliary $\boldsymbol{Z}$ included is defined as follows:

$$\begin{aligned} Z_i &\overset{iid}{\sim} \text{Cat}(\boldsymbol{p}) \quad i = 1, \ldots, n; \\ X_i|\boldsymbol{\phi}, Z_i = j &\sim \text{N}(\cdot|\phi_j, \sigma^2), \quad i = 1, \ldots, n; \quad j = 1, \ldots, k. \end{aligned}$$

The priors for $\boldsymbol{p}$ and $\boldsymbol{\phi}$ are given as before.

Now we proceed to compute the posterior distribution for the quantities of interest $p(\boldsymbol{Z}, \boldsymbol{p}, \boldsymbol{\phi}|\boldsymbol{X})$ via Gibbs sampling. The full conditional distributions are easy to derive.

- For $\boldsymbol{Z}$: for each $i = 1, \ldots, n$, $j = 1, \ldots, k$,

$$\begin{aligned} p(Z_i = j|Z_{-i}, \boldsymbol{X}, \boldsymbol{p}, \boldsymbol{\phi}) &= p(Z_i = j|X_i = x_i, \boldsymbol{p}, \boldsymbol{\phi}) \\ &\propto p(Z_i = j)p(x_i|Z_i = j, \boldsymbol{p}, \boldsymbol{\phi}) \\ &= \frac{p_j \,\mathrm{N}(x_i|\phi_j, \sigma^2)}{\sum_{j=1}^k p_j \,\mathrm{N}(x_i|\phi_j, \sigma^2)}. \end{aligned}$$

- For $\boldsymbol{\phi}$: for each $j = 1, \ldots, k$

$$\begin{aligned} p(\phi_j|\boldsymbol{\phi}_{-j}, \boldsymbol{Z}, \boldsymbol{X}, \boldsymbol{p}) &= p(\phi_j|\boldsymbol{Z}, \{X_i = x_i \text{ such that } z_i = j\}) \\ &= \mathrm{N}\left(\phi_j \middle| \frac{\mu/\tau^2 + \sum x_i 1(z_i = j)/\sigma^2}{1/\tau^2 + n_j/\sigma^2}, \frac{1}{1/\tau^2 + n_j/\sigma^2}\right). \end{aligned}$$

  The first identity is due to conditional independence. The second identity is a standard posterior computation under a normal likelihood and a normal prior for the mean parameter (cf. Section 5). Here, $n_j = \sum_{i=1}^n 1(z_i = j)$, i.e., the number of data points are currently assigned to the mixture component j by means of having the label $z_i = j$.

- For $\boldsymbol{p}$:

$$\begin{aligned} p(\boldsymbol{p}|\boldsymbol{Z}, \boldsymbol{X}, \boldsymbol{\phi}) &= p(\boldsymbol{p}|\boldsymbol{Z}) \qquad \text{due to cond. indep.} \\ &\propto p(\boldsymbol{p})p(\boldsymbol{Z}|\boldsymbol{p}) \\ &\propto \prod_{j=1}^{k} p_j^{\alpha_j - 1} \times \prod_{j=1}^{k} p_j^{n_j} \\ &\propto \mathcal{D}(\boldsymbol{p}|\alpha_1 + n_1, \ldots, \alpha_k + n_k) \\ &= \mathcal{D}(\boldsymbol{p}|\boldsymbol{\alpha} + \boldsymbol{n}) \end{aligned}$$

  wherein the last line we use $\boldsymbol{n}$ to denote $(n_1, \ldots, n_k)$.

We make some comments

- The Gibbs updates for Z_i and $\boldsymbol{p}$ is can be viewed as the result of a "soft" (probabilistic) assignment of the cluster label for each of the data point x_i.

  Recall that in k-means clustering algorithm, there is a hard assignment of the cluster label associated with each data point. In the EM algorithm, this corresponds to the E-step, which updates the expectation of the parameters such as Z_i.

- The Gibbs update for ϕ_j is a probabilistic update of the cluster means. This is the direct counterpart of the M-step in the EM algorithm and the mean update step in k-means.

- Gibbs sampling is convenient but not the most efficient posterior computation technique. We may consider other forms of MCMC such as using Metropolis-Hastings algorithms, as we saw in Section 10. The wealth of posterior inference algorithms available is a hidden benefit of working with a rich Bayesian modeling framework. It is considerably harder to invent a deterministic counterpart of Metropolis-Hastings algorithms among frequentist approaches that must extend from the basic k-means and EM algorithms.

11.2 Infinite mixture models

As we said earlier, the salient feature of a nonparametric Bayesian approach is to allow infinitely many parameters to be present in the model.

Continuing with our present example of mixture modeling with normal components, an infinite mixture model admits the following density function

$$p(x|\boldsymbol{p}, \boldsymbol{\phi}) = \sum_{j=1}^{\infty} p_j f(x|\phi_j).$$

As before $f(x|\phi_j) = \mathrm{N}(x|\phi_j, \sigma^2)$ for some known σ^2, but here there are infinitely many parameters $(p_j, \phi_j)_{j=1}^{\infty}$.

An immediate question is: how do we specify a Bayesian prior on infinitely many parameters? Since ϕ_j's are unconstrained in the real line, we may again set the prior for these parameters as

$$\phi_1, \phi_2 \ldots \overset{iid}{\sim} G_0.$$

For instance, take $G_0 = \mathrm{N}(\mu, \tau^2)$.

The nontrivial issue lies in specifying the prior for $\boldsymbol{p} = (p_1, p_2, \ldots)$, which is now an infinite sequence satisfying the constraint that $p_j \geq 0$ and $\sum_j p_j = 1$.

Recall that if the sequence $\boldsymbol{p} = (p_1, \ldots, p_k)$ is a finite sequence, i.e., $k < \infty$, then we may use the Dirichlet distribution as a prior for $\boldsymbol{p} \in \Delta^{k-1}$, say $\boldsymbol{p} \sim \mathcal{D}(\boldsymbol{\alpha})$ for some $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_k) \in \mathbb{R}_+^k$. We need a generalization of the Dirichlet distribution that works for Δ^∞.

11.2.1 Dirichlet process prior

Recall a simple fact about the finite-dimensional Dirichlet distribution. If $k = 2$, then the Dirichlet distribution $\mathcal{D}_1((p_1, p_2)|\alpha_1, \alpha_2)$ reduces to the Beta distribution on the unit interval Beta$(p_1|\alpha_1, \alpha_2)$, because $p_2 = 1 - p1$.

With some moment of thought, it is possible to conceive the following distribution on the infinite sequence $\boldsymbol{p} = (p_1, p_2, \ldots)$ by constructing a random process of "stick-breaking" as following: take a stick of unit length, break it into two shorter pieces in a random fashion, one of which is assigned to be of length p_1, and the remaining part of length $1 - p_1$ is broken again randomly to obtain p_2, and so on. Whenever we break a piece of stick into two smaller pieces, we may take the proportions of the smaller pieces to be beta distributed.

To be precise, let $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots)$ be iid Beta$(1, \alpha)$. Define

$$p_1 = \beta_1, p_k = \prod_{i=1}^{k-1}(1 - \beta_i)\beta_k, \quad k = 2, 3, \ldots.$$

It is easy to check that the infinite sequence $\boldsymbol{p}$ constructed this way satisfies the constraint that $\sum_{k=1}^{\infty} p_k = 1$ almost surely.

We have just described a Dirichlet distribution on the infinite-dimensional probability simplex Δ^{∞}.

Collecting the above specifications gives us a definition of the famous Dirichlet process [16]

Definition 11.1. Let G_0 is a probability distribution on the real line and given an infinite i.i.d. sequence of random variables

$$\phi_1, \phi_2, \ldots \overset{iid}{\sim} G_0.$$

Let $\alpha > 0$ and given an infinite i.i.d. sequence of random variables

$$\beta_1, \beta_2, \ldots \overset{iid}{\sim} \text{Beta}(1, \alpha).$$

Set

$$p_1 = \beta_1, p_k = \prod_{i=1}^{k-1}(1 - \beta_i)\beta_k, \quad k = 2, 3, \ldots. \tag{62}$$

Define the discrete distribution on the real line

$$G := \sum_{j=1}^{\infty} p_j \delta_{\phi_j}$$

Then we say that G is a Dirichlet process on the real line. We write

$$G|\alpha, G_0 \sim \mathscr{D}(\alpha G_0). \tag{63}$$

□

[16] The Dirichlet process was first introduced by Thomas Ferguson. Definition 11.1, however, was given by Jayaram Sethuraman.

What we just defined is that G is a random variable taking values in the space of probability distributions on the real line, namely, $\mathscr{P}(\mathbb{R})$. The distribution from which the random G is generated, namely, $\mathscr{D}(\alpha G_0)$, is called a Dirichlet distribution, which generalizes the standard Dirichlet distribution on a finite-dimensional probability simplex to a distribution on the infinite-dimensional probability simplex Δ^∞.

Note that the distribution $\mathscr{D}(\alpha G_0)$ has two parameters: a positive scalar $\alpha > 0$, and G_0 is a distribution on the real line.

Conjugacy of Dirichlet prior with respect to i.i.d. sampling A remarkable property discovered by Thomas Ferguson is that the Dirichlet process provides a conjugate prior to i.i.d. sampling distributions. More precisely, if

$$\begin{aligned} G &\sim \mathscr{D}(\alpha G_0) \\ \theta_1, \ldots, \theta_n | G &\overset{i.i.d.}{\sim} G, \end{aligned}$$

then

$$G|\theta_1, \ldots, \theta_n \sim \mathscr{D}(\alpha G_0 + \sum_{i=1}^{n} \theta_i).$$

Back to our infinite mixture model setting

$$p(x|\boldsymbol{p}, \boldsymbol{\phi}) = \sum_{j=1}^{\infty} p_j f(x|\phi_j) \tag{64}$$

The distribution $G = \sum_{i=1}^{\infty} p_j \delta_{\phi_j}$ encapsulates all parameters for the infinite mixture model that we seek to estimate. We can rewrite the mixture model equivalently as

$$p(x|G) = \int f(x|\phi) G(\mathrm{d}\phi). \tag{65}$$

Eq. (65) gives us the view of infinite mixture model as a model parameterized by $G \in \mathscr{P}(\mathbb{R})$. G is called the *mixing distribution*, or *mixing measure* for the mixture model.

When the mixing distribution G is endowed with the Dirichlet prior given by Eq. (63):

$$G|\alpha, G_0 \sim \mathscr{D}(\alpha G_0)$$

we call our model *Dirichlet process mixture model*.

This is still a standard Bayesian formulation, although a nonparametric one, where the parameter of interest is the infinite dimensional $G \in \mathscr{P}(\mathbb{R})$.

Given an i.i.d. n-sample $X_1, \ldots, X_n|G \sim p(x|G)$, the immediate question of concern is that of posterior computation. How do we compute

$$p(G|X_1, \ldots, X_n)?$$

11.3 Posterior computation via slice sampling

The totality of all variables of interest include the observed data $\boldsymbol{X} = (X_1, \ldots, X_n)$, the mixing proportions $\boldsymbol{p} = (p_1, \ldots)$, atoms $\boldsymbol{\phi} = (\phi_1, \ldots)$. Moreover, $\boldsymbol{p}$ is constructed via the stick-breaking representation (62), which is based on variables $\boldsymbol{\beta} = (\beta_1, \ldots)$.

We shall make use of the auxiliary variable technique extensively. The first use is similar to the case of finite mixture that we saw in Section 11.1. According to the joint model,

- each data point X_i is associated with a mixture component label $Z_i \in \{1, 2, \ldots\}$.
- Given $\boldsymbol{p}$, $Z_i | \boldsymbol{p} \overset{iid}{\sim} \text{Cat}(\boldsymbol{p})$ for $i = 1, \ldots, n$.
- Given Z_i and all other variables, X_i is distributed according to $f(X_i | \phi_{Z_i})$.

Thus, we may write the joint model as

$$(\boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{Z}, \boldsymbol{X}) \sim \text{Beta}(1, \alpha)^{\infty} \times G_0^{\infty} \times \prod_{i=1}^{n} p_{Z_i} \times \prod_{i=1}^{n} f(X_i | \phi_{Z_i}). \tag{66}$$

The superscripts ∞ signify the infinitely many variables $\boldsymbol{\beta} = \{\beta_k\}_{k=1}^{\infty}$ and $\boldsymbol{\phi} = \{\phi_k\}_{k=1}^{\infty}$ present in the model.

We seek to devise a Markov chain that converges in distribution to the target stationary distribution which is the posterior of $\boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{Z}$ given data $\boldsymbol{X}$. The difficulty is apparent: there are an infinite number of variables to handle, which cannot possibly be sampled simultaneously. We use a technique known as "slice sampling".

Slice sampling involves the introduction of yet another set of auxiliary random variables, $\boldsymbol{u} := (u_1, \ldots, u_n)$ taking values in bounded intervals $(0, q_{z_i})$, where $i = 1, \ldots, n$ and $(q_j)_{j \geq 1}$ is a sequence of values in $(0, 1)$ either deterministically or randomly generated so that $\boldsymbol{q}$ tend to zero (certainly or almost surely).

In particular, for each i, given $\boldsymbol{q}$ we draw u_i from the uniform distribution on the interval $(0, q_{z_i})$.

Thus, the extended joint model takes the form

$$(\boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{u}, \boldsymbol{Z}, \boldsymbol{X}|\boldsymbol{q}) \sim \text{Beta}(1, \alpha)^{\infty} \times G_0^{\infty} \times \prod_{i=1}^{n} 1(u_i < q_{Z_i}) \frac{1}{q_{Z_i}} \times \prod_{i=1}^{n} p_{Z_i} \times \prod_{i=1}^{n} f(X_i|\phi_{Z_i}). \tag{67}$$

It is clear that integrating out all u_i in the joint distribution given by Eq. (67) leads to the joint distribution given by Eq. (66). Thus, it sufficient to construct a MC for the model given by Eq. (67).

What one gains in the introduction of auxiliary variables $\boldsymbol{u}$ is that, when $\boldsymbol{u}$ are conditioned on, we only need to choose labels Z_i from the *finite* set

$$H(u_i) := \{j \in \mathbb{N}_+ : q_j > u_i\}.$$

If one thinks of a bar graph in which the height of each bar represents the magnitude of q_j, $j = 1, \ldots$, then restricting the label Z_i to only $H(u_i)$ corresponds visually to "slicing" out the portion below the height u_i, and making only the bars higher than u_i to remain. Hence, the name "slice sampling".

Gibbs sampler for model (67)

- Sampling $\boldsymbol{u}$ given $\boldsymbol{\beta}, \boldsymbol{Z}, \boldsymbol{X}, \boldsymbol{q}$: for each $i = 1, \ldots, n$, draw $u_i \overset{indep}{\sim} \text{Uniform}[0, q_{Z_i}]$.
- Sampling $\boldsymbol{\beta}$ given $\boldsymbol{u}, \boldsymbol{Z}, \boldsymbol{\phi}, \boldsymbol{X}, \boldsymbol{q}$: Note that the variables $\boldsymbol{\beta}$ are relevant as far as the extent that they determines the variables p_j's. Moreover, the only variable p_j's of concern are those with indices j such that $j \in \cup_{i=1}^{n} H(u_i)$. Thus,

$$\begin{aligned}
p(\beta_j|\text{the rest}) &\propto (1-\beta_j)^{\alpha-1} \times \prod_{i: q_{Z_i} > u_i} p_{Z_i} \\
&\propto (1-\beta_j)^{\alpha-1} \times \prod_{i: q_{Z_i} > u_i} \left\{ \prod_{k=1}^{Z_i - 1} (1-\beta_k)\beta_{Z_i} \right\} \\
&\propto \beta_j^{\sum_{i=1}^{n} 1(Z_i = j; q_j > u_i)} (1-\beta_j)^{\alpha - 1 + \sum_{i=1}^{n} 1(Z_i > j; q_{Z_i} > u_i)} \\
&\propto \text{Beta}(1 + m_j, \alpha + \sum_{k>j} m_k),
\end{aligned}$$

where in the last line, we set $m_j := \sum_{i=1}^{n} 1(Z_i = j; u_i < q_j)$ for $j = 1, \ldots$, and note that $\sum_{i=1}^{n} 1(Z_i > j; q_{Z_i} > u_i) = \sum_{k>j} m_k$.

Clearly, in the above computation we only need to update for $j = 1, \ldots, K$ such that for all $k > K$, $m_k = 0$. K represents the upper bound of the number of "active" indices. K may change from one Gibbs iteration to the next.

- Sampling $\boldsymbol{\phi}$ given $\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{Z}, \boldsymbol{X}, \boldsymbol{q}$:

$$\begin{aligned} p(\phi_j|\text{the rest}) &\propto G_0(\mathrm{d}\phi_j)\prod_{i=1}^{n} f(X_i|\phi_{Z_i}) \\ &\propto \prod_{i:Z_i=j} \mathrm{N}(X_i|\phi_j,\sigma^2)G_0(\mathrm{d}\phi_j) \\ &\propto \mathrm{N}\left(\phi_j \middle| \frac{\mu/\tau^2+\sum x_i 1(z_i=j)/\sigma^2}{1/\tau^2+n_j/\sigma^2}, \frac{1}{1/\tau^2+n_j/\sigma^2}\right). \end{aligned}$$

  where $n_j = \sum_{i=1}^{n} 1(z_i = j)$, i.e., the number of data points are currently assigned to the mixture component j by means of having the label $z_i = j$. (Note that this step is similar to the sampling of the label in a finite mixture.)

- Sampling $\boldsymbol{Z}$ given $\boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{u}, \boldsymbol{X}, \boldsymbol{q}$: for $i = 1, \ldots, n$

$$p(Z_i = j|\text{the rest}) \propto 1(u_i < q_j)\frac{p_j}{q_j} f(X_i|\phi_j),$$

  for $j = 1, \ldots$.

  This is where we need to be careful since the support of Z_i is unbounded. Obviously, the above probability is positive only if $u_i < q_j$. If $\boldsymbol{q} \in \Delta^{\infty}$ (although this is not a strict requirement, more on this is below), then it suffices to update for all values $j = 1, \ldots$ up to the minimal index K where $1 - \sum_{k=1}^{K} q_k < \min_{i=1}^{n}\{u_i\}$.

  If we reach a new index k for which p_k and ϕ_k have not been generated, then we proceed by generating $\phi_k \sim G_0$, $\beta_k \sim \text{Beta}(1, \alpha)$, and letting $p_k = \prod_{i=1}^{k-1}(1-\beta_i)\beta_k = (1 - \sum_{i=1}^{k-1} p_i)\beta_k$.

- Sampling $\boldsymbol{q}$: If $\boldsymbol{q}$ is deterministically generated, then this step is not necessary (although the choice of this sequence may be critical to the mixing behavior of the underlying Markov chain). If $\boldsymbol{q}$ is randomly generated, there are several options

  - a simple method is to generate $\boldsymbol{q}$ independently of all over variables (e.g., via a fixed stick-breaking process). Then, we may update $\boldsymbol{q}$ after one or several iterations of the Gibbs updates for all other variables.
  - another approach is place an independent prior for $\boldsymbol{q}$: $q_j \sim$Uniform$(0, b_j)$ for $j = 1, 2, \ldots$ such that $b_j \downarrow 0$. Then the update for $\boldsymbol{q}$ can be achieved given $\boldsymbol{u}$ via the conditional distribution:

$$p(q_j|\text{the rest}) \propto q_j^{-n_j} 1(q_j > \max_{i:Z_i=j} u_i).$$

  - yet another approach is to take $\boldsymbol{q} := \boldsymbol{p}$, but then $\boldsymbol{q}$ is no longer independent of $\boldsymbol{\beta}$; the update of $\boldsymbol{\beta}$ may not have the conjugate form or an easily calculable form as given above.

Observe that the MCMC algorithm gradually and stochastically adds new components (β_j, ϕ_j) for $j = 1, 2, \ldots$ into the state space of the Markov chain. No upper bound on the number of components is required a priori!

11.4 Chinese restaurant process and another Gibbs sampler

Dirichlet processes have many other remarkable characterizations, which help us to understand them more deeply, while giving us additional ideas for computations. Next, we describe a Polya urn characterization of the Dirichlet processes.

Consider the following specification for a sequence of random variables which are i.i.d. draw from a Dirichlet process:

$$G|\alpha, G_0 \quad \sim \quad \mathscr{D}_{\alpha G_0} \tag{68}$$

$$\theta_1, \ldots, \theta_n|G \quad \overset{iid}{\sim} \quad G. \tag{69}$$

Note that given α and G_0, the random distribution G may be represented as $G = \sum_{k=1}^{\infty} p_j \delta_{\phi_j}$, where $\boldsymbol{p}$ and $\boldsymbol{\phi}$ are random variables given by Definition 11.1. Since $\theta_1, \theta_2, \ldots$ are a conditionally i.i.d. sequence, this is an exchangeable sequence of random variables.

We ask: what is the marginal distribution of the exchangeable sequence $\theta_1, \theta_2, \ldots$, which would be obtained if we integrate out the random G in the above specification?

Based on Definition 11.1 it is not difficult to verify that the joint distribution of the sequence $\theta_1, \theta_2, \ldots$ can be completely specified as follows:

$$\begin{aligned}
\theta_1 &\sim G_0, \\
\theta_2|\theta_1 &\propto \delta_{\theta_1} + \alpha G_0, \\
&\cdots \\
\theta_j|\theta_1, \ldots, \theta_{j-1} &\propto \sum_{k=1}^{j-1} \delta_{\theta_k} + \alpha G_0, \\
&\cdots
\end{aligned}$$

The sequence of random variables defined this way is generally known as a Pólya sequence. It makes explicit the clustering behavior of the collection of random variables $\theta_1, \theta_2, \ldots$ which are generated from a (random) Dirichlet process $G \sim \mathscr{D}(\alpha G_0)$: with positive probability each of the θ_j shares the same value as some of the other variables generated before it in the sequence.

This Pólya sequence has a tasty name, "the Chinese restaurant process". Consider the following imaginary Chinese restaurant, which receives an infinite sequence of customers labeled by $1, 2, \ldots$ with its infinitely many tables:

- customer 1 arrives, and sits by an arbitrary table there.
- the following customers $2, 3, \ldots$ arrive in sequence and choose their table according to the following rule: either one of the non-empty table is chosen with probability proportion to the current number of customers sitting at table; otherwise that customer chooses a new table with probability proportional to α
- for each table, a random dish is ordered i.i.d. from *menu* (distribution) G_0 for all to share.
  assign each θ_i to the dish that i is having.

Gibbs sampler based on the Pólya characterization The Dirichlet process mixture model can be expressed as follows. Recall the prior:

$$\begin{aligned} G|\alpha, G_0 &\sim \mathscr{D}_{\alpha G_0} \\ \theta_1, \ldots, \theta_n|G &\overset{iid}{\sim} G., \end{aligned}$$

which is combined with the likelihood specification: for $i = 1, \ldots, n$:

$$X_i|\theta_i \overset{indep}{\sim} f(X_i|\theta_i). \tag{70}$$

Latent variables $\theta_1, \ldots, \theta_n$ represent the parameter with each $X_1, \ldots, X_n$ are respectively associated. E.g., θ_i is the mean parameter for the mixture component X_i is associated with when use use $f(X_i|\theta_i) = \mathrm{N}(X_i|\theta_i, \sigma^2)$,

To implement a Gibbs sampler, we need to construct a Markov chain for $\{\theta_1, \ldots, \theta_n\}$ that converges to the target stationary distribution $\mathbb{P}(\theta_1, \ldots, \theta_n|\boldsymbol{X})$. For a Gibbs update, we need to compute the full conditional distribution for each θ_i given every other variables.

By the fact that $\theta_1, \ldots, \theta_n$ are a priori exchangeable, we may treat θ_i as the last element in the Pólya sequence (i.e., the last customer in the Chinese restaurant process). Thus,

$$\theta_i | \theta_{-i} \sim \sum_{j \neq i} \delta_{\theta_j} + \alpha G_0.$$

By Bayes' rule, and conditional independence, we have

$$\begin{aligned} p(\theta_i | \theta_{-i}, \boldsymbol{X}) &\propto p(\theta_i | \theta_{-i}) f(X_i | \theta_i) \\ &\propto \alpha f(X_i | \theta) G_0(d\theta) + \sum_{j \neq i} f(X_i | \theta_j) \delta_{\theta_j}. \end{aligned}$$

The above full conditional distribution is a mixture distribution: with probability proportional to $f(X_i|\theta_j)$ we set $\theta_i := \theta_j$, and with probability proportional to $\int \alpha f(X_i|\theta) G_0(d\theta)$ we draw $\theta_i \sim G_0$. The integration in question is available in closed form due to the normal-normal conjugacy between G_0 and f.

We see clearly in the Gibbs sampling step the types of move: one type of move is to select a cluster/table/dish for θ_i among the existing ones, and another type of move is to generate a new cluster/table/dish from the base distribution G_0. Thus, the number of clusters are also sampled as part of the Markov chain generation.

Summarizing, the Gibbs sampling algorithm consists of the following single line of code: For each MCMC step, do as follows:

(1) for $i = 1, \ldots, n$, draw θ_i given existing θ_{-i} and $\boldsymbol{X}$ by the full conditional distribution derived above.

This is only the simplest example of a Gibbs sampler based on the Pólya characterization of Dirichlet processes. Researchers have developed more sophisticated and efficient techniques based on Gibbs and Metropolis-Hastings sampling frameworks.

In this section offered a glimpse of Dirichlet process, which is just one of many powerful tools of Bayesian nonparametrics. For an expanded version of this short introduction, see also the lecture notes [Nguyen, 2015].

The materials in this course are adapted from materials created by David Blei, Yang Chen, Andrew Gelman, Scott Linderman, Long Nguyen, and the 3blue1brown channel.