Dirichlet Process Mixtures

- For density estimation, consider the DP mixture (DPM) model
  \[
y_i \sim N(\mu_i, \tau_i^{-1}), \quad \theta_i = (\mu_i, \tau_i) \sim P, \quad P \sim DP(\alpha P_0) \quad (*)
  \]

- Not immediate clear how to conduct posterior computation
- One strategy relies on marginalizing out \(P\) to obtain

  \[
  (\theta_i | \theta_1, \ldots, \theta_{i-1}) \sim \left(\frac{\alpha}{\alpha + i - 1}\right) P_0 + \sum_{j=1}^{i-1} \left(\frac{1}{\alpha + 1}\right) \delta_{\theta_j}.
  \]

DP prediction rule or Polya urn scheme (Blackwell & MacQueen, 73)
Chinese Restaurant Process

- Discreteness is clear from the form of the DP prediction rule
- Subjects will be grouped into clusters, with subjects in a cluster having the same $\theta_i$ value
- Chinese restaurant analogy - restaurant with $\infty$ tables
  1. 1st customer sits at table with dish $\theta_1^*$
  2. 2nd customer sits at first table with prob $\alpha/(1 + \alpha)$ or new table with prob $1/(1 + \alpha)$
  3. process encourages customers to sit at well occupied tables
Marginal Gibbs Sampler (Bush & MacEachern, 96)

- Let $\theta^* = (\theta_1^*, \ldots, \theta_k^*)$ denote the unique values of $\theta$
- Let $S_i = h$ if $\theta_i = \theta_h^*$ denote allocation of subject $i$ to cluster $h$
- Gibbs sampler alternates between
  1. Update the allocation $S$ by sampling from multinomial with

$$\Pr(S_i = h | -) \propto \begin{cases} n_h^{(-i)} N(y_i; \theta_h^*) & h = 1, \ldots, k^{(-i)} \\ \alpha \int N(y_i; \theta) dP_0(\theta) & h = k^{(-i)} + 1 \end{cases}$$

  2. Update the unique values $\theta^*$ by sampling

$$\left(\theta_h^* | - \right) = N(\mu_h^*; \hat{\mu}_h, \kappa_h \tau_h^{*,-1}) \text{Ga}(\tau_h^*; \hat{a}_{\tau_h}, \hat{b}_{\tau_h})$$

with parameters defined as in the finite mixture model case
Marginal Gibbs Sampler - Some Comments

- Only slightly more complicated the Gibbs sampling for finite mixture models
- # mixture components $k$ represented in the sample of $n$ subjects is unknown
- From the MCMC samples, we can estimate posterior distribution of $k$
- As subjects are added $k$ will increase stochastically
- To estimate the predictive density of $y_{n+1}$ use

\[ f(y) = \sum_{h=1}^{k} \left( \frac{n_h}{n + \alpha} \right) N(y; \theta^*_h) + \left( \frac{\alpha}{n + \alpha} \right) \int N(y; \theta) dP_0(\theta), \]

averaged over MCMC iterations after burn-in
Assume $y_i \sim N(\mu_i, \sigma^2)$ with $\mu_i \sim P$ & $P \sim \text{DP}(\alpha P_0)$

Marginalize out $P$ to obtain $k$ clusters in the $(\mu_1, \ldots, \mu_n)'$ vector

Let $A$ denote an $n \times k$ matrix with the $i$th row $a_i$ being a vector of zeros except for a one in the position corresponding to the cluster for subject $i$

Then, we obtain a linear model

$$y = A\mu^* + \epsilon, \quad \epsilon \sim N_n(0, \sigma^2 I_n),$$

with the elements of $\mu^*$ sampled iid from $P_0$
PX Gibbs Sampler

- Sampling $\mu^*, \sigma^2$ is trivial following standard approaches for updating in normal linear regression models.
- The key question is how to update the cluster allocation matrix $A$ efficiently.
- Kyung et al. propose a parameter-expansion (PX) approach:
  \[ q^{(t+1)} \sim \text{Diri}(n_1^{(t)} + 1, \ldots, n_k^{(t)} + 1, 1, \ldots, 1) \]
- Then, conditionally on $q^{(t+1)}$, update the rows of $A$ using Gibbs.
- Can be shown to dominate previous algorithms in operator norm & efficiency.
Avoiding Marginalization

- By marginalizing out the RPM $P$, we give up the ability to conduct inferences on $P$.
- By having approaches that avoid marginalization, we open the door to generalizations of DPMs.
- Stick-breaking representation (Sethuraman, 94),

$$
\theta_i \sim P = \sum_{h=1}^{\infty} V_h \prod_{l<h} (1 - V_l) \delta_{\Theta_h}, \quad V_h \overset{iid}{\sim} \text{beta}(1, \alpha), \quad \theta_h \overset{iid}{\sim} P_0.
$$

- Starting from unit probability stick,
  1. Break off a random piece ($V_1$) & allocate this to a random value ($\theta_1$)
  2. From the remaining $1 - V_1$, break off a proportion $V_2$ & allocate to $\theta_2$
  3. Repeat infinitely many times
Samples from the Dirichlet process with precision $\alpha$
Implications of Stick-Breaking

- For small $\alpha$, most of the probability is allocated to the first few components, favoring few latent classes
- Expected number of occupied components $\propto \alpha \log n$
- Weights $\pi_h$ decrease stochastically towards near zero rapidly in the index $h$
- Suggests truncation approximation (Muliere & Tardella, 98),

$$P = \sum_{h=1}^{N} V_h \prod_{l < h} (1 - V_l) \delta \Theta_h,$$

with $V_N = 1$ so that weights sum to one
Blocked Gibbs Sampler \((\text{Ishwaran & James, 01})\)

1. Update \(S_i \in \{1, \ldots, N\}\) by multinomial sampling with

\[
\Pr(S_i = h | -) = \frac{\pi_h N(y_i; \Theta_h)}{\sum_{l=1}^{N} \pi_l N(y_i; \Theta_l)}, \quad h = 1, \ldots, N.
\]

2. Update stick-breaking weight \(V_h, h = 1, \ldots, N - 1\), from

\[
\text{Be}\left(1 + n_h, \alpha + \sum_{l=h+1}^{N} n_l\right).
\]

3. Update \(\Theta_h, h = 1, \ldots, N\), exactly as in the finite mixture model.
Comments on Blocked Gibbs

- $N$ acts as an upper bound on the number of mixture components in the sample
- By choosing a large value, the approximation error should be small
- Possible to monitor this error during the MCMC
- Approximate inferences on functionals of $P$ are possible
- Slice (Walker, 07) & retrospective sampling (Papaspiliopoulos & Roberts, 08) approaches avoid truncation - exact block Gibbs (Papaspiliopoulos, 08) combine these approaches
Choosing the DP precision parameter

The DP precision parameter $\alpha$ plays a key role in controlling the prior on the number of clusters.

A number of strategies have been proposed in the literature -

1. Fix $\alpha$ at a small number to favor allocation to few clusters relative to the sample size - a commonly used default value is $\alpha = 1$
2. Assign a hyperprior (typically gamma) to $\alpha$ - refer to technical report by West (92) & recent article by Dorazio (09, JSPI, 139, 3384-3390)
3. Estimate $\alpha$ via empirical Bayes (Liu 96; McAulliffe et al. 06)
Choosing the base measure $P_0$

- Often the base measure is chosen for computational convenience to be conjugate.

- However, even in conjugate parametric families (e.g., normal-gamma) we can potentially improve flexibility by placing hyperparameters on the parameters in $P_0$.

- $P_0$ can be thought of as inducing the prior for the cluster locations - if these locations are too spread out ($P_0$ has very high variance) overly favors allocation to a single cluster.

- Through the use of hyperpriors, we can induce a $P_0$ with heavy-tails - e.g., multivariate $t$, allowing robustness.
Some comments on $P_0$

- It is crucial to consider the measurement scale of the data in choosing $P_0$
- The variance of $P_0$ is only meaningful relative to the measurement scale of the data
- A common approach is to standardize the data $y$ prior to analysis and then choose $P_0$ to be centered at zero with close to unit variance, possibly with heavy tails
- If we set unit variance & do not standardize, then how “flat” $P_0$ depends on the unit of measurement in the data - if we change from inches to miles, we may get completely different results
- Instead of standardizing the data, we can elicit the parameters in $P_0$ or put the scale in the hyperparameters
Some comments on $P_0$

- Although $\alpha$ has a more obvious impact on the posterior on the number of clusters, the choice of $P_0$ is often more important.
- In most applications, the data tend to be quite informative about $\alpha$ and even choosing a high variance prior tends to work well.
- There are empirical Bayes methods available to estimate $P_0$ nonparametrically - refer to McAulliffe et al. (2005) - seems to rely on there being sufficient numbers of clusters occupied.
Using R, simulate data from the following mixture of normals

\[ y_i \sim 0.1N(-1, 0.2) + 0.5N(0, 1) + 0.4N(1, 0.4), \]

\[ i = 1, \ldots, 100 \]

- Use the density function to obtain a frequentist estimate of the density & plot vs true density
- Run the finite mixture model Gibbs sampler for \( k = 10 \), \( a_h = \alpha/k \), \( \mu_0 = 0 \), \( \kappa = \alpha_T = b_T = \alpha = 1 \)
- Run the blocked Gibbs sampler for \( N = 10 \) & the same hyperparameter specification.
- Compare the resulting density estimates.