

Bayesian Statistics

Debdeep Pati
Florida State University

March 2, 2016

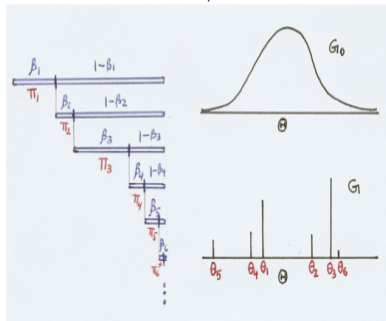
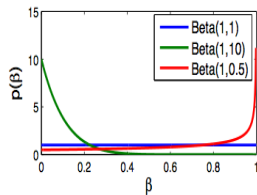
Stick Breaking construction for $G \sim DP(\cdot, \alpha, G_0)$

Stick-Breaking Formula

$$\pi_k = \beta_k \prod_{l=1}^{k-1} (1 - \beta_l), \beta_k \sim \text{Beta}(1, \alpha),$$

$$\theta_k^* \sim G_0, G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k^*}$$

Stick Breaking



Sketch of the proof (Sethuraman, 1994)

- ▶ Recall the posterior process

$$\left[\begin{array}{l} G \sim DP(\cdot \mid \alpha, G_0) \\ \theta \mid G \sim G \end{array} \right] \Leftrightarrow \left[\begin{array}{l} \theta \sim G_0 \\ G \mid \theta \sim DP\left(\cdot, \alpha + 1, \frac{\alpha G_0 + \delta_\theta}{\alpha + 1}\right) \end{array} \right]$$

- ▶ Consider a partition $(\theta, \Theta \setminus \theta)$ of Θ . We have

$$\begin{aligned} (G(\theta), G(\Theta \setminus \theta)) &\sim \text{Diri}\left\{(\alpha + 1) \frac{\alpha G_0 + \delta_\theta}{\alpha + 1}(\theta), (\alpha + 1) \frac{\alpha G_0 + \delta_\theta}{\alpha + 1}(\Theta \setminus \theta)\right\} \\ &= \text{Beta}(1, \alpha) \end{aligned}$$

- ▶ G has a point mass located at θ :

$$G = \beta \delta_\theta + (1 - \beta) G', \quad \beta \sim \text{Beta}(1, \alpha)$$

and G' is the renormalized probability measure with the point mass removed.

- ▶ What is G' ?

Sketch of the proof : What is G'

- ▶ Consider a further partition of $(\theta, A_1, \dots, A_K)$ of Θ .

$$\begin{aligned}(G(\theta), G(A_1), \dots, G(A_K)) &= (\beta, (1 - \beta)G'(A_1), \dots, (1 - \beta)G'(A_K)) \\ &\sim \text{Diri}(1, \alpha G_0(A_1), \dots, \alpha G_0(A_K))\end{aligned}$$

- ▶ Renormalizing

$$\begin{aligned}(G'(A_1), \dots, G'(A_K)) | \theta &= \text{Diri}(\alpha G_0(A_1), \dots, \alpha G_0(A_K)) \\ G' &\sim DP(\cdot, \alpha, G_0)\end{aligned}$$

Sketch of the proof

$$G \sim DP(\cdot, \alpha, G_0)$$

$$G = \beta_1 \delta_{\theta_1^*} + (1 - \beta_1) G_1$$

$$G = \beta_1 \delta_{\theta_1^*} + (1 - \beta_1)(\beta_2 \delta_{\theta_2^*} + (1 - \beta_2) G_2)$$

\vdots

$$G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k^*}$$

where $\pi_k = \beta \prod_{l=1}^{k-1} (1 - \beta_l)$, $\beta_k \sim \text{Beta}(1, \alpha)$, $\theta_k^* \sim G_0$.

Finite Mixture Models

- ▶ Finite mixture models are useful in a wide variety of settings, including density estimation, clustering, classification, etc
- ▶ Focus initially on problem in which $f(y)$, for $y \in \mathbb{R}$, is an unknown density function
- ▶ Finite mixture of Gaussians provides a flexible choice,

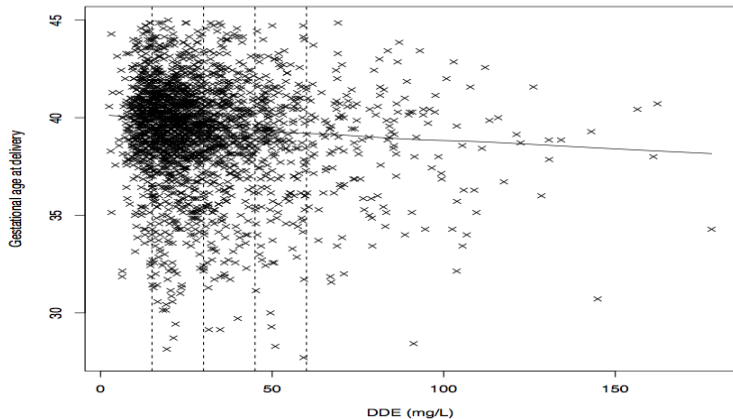
$$f(y) = \sum_{h=1}^k \pi_h N(y; \mu_h, \tau_h^{-1})$$

- ▶ It is well known that a mixture of normals can approximate any smooth density

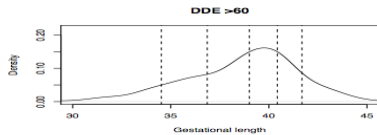
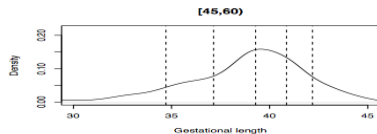
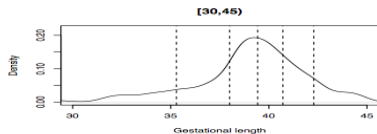
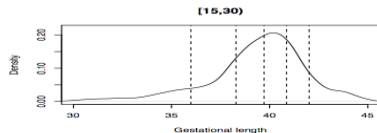
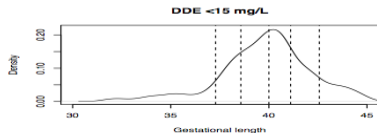
Application: Modeling length of gestation

- ▶ Preterm birth is a major public health problem leading to substantial mortality & short and long-term morbidity
- ▶ Preterm birth is typically defined as a delivery occurring prior to 37 weeks of completed gestation
- ▶ This cutoff is somewhat arbitrary & the shorter the length of gestation, the more adverse the associated health effects
- ▶ Appealing to model the distribution of gestational age at delivery as unknown & then allow predictors to impact this distribution

Gestational Length vs. DDE(mg/L)



Gestational Length Densities within DDE Categories



Comments on Gestational Length Data

- ▶ Data are non-Gaussian with a left skew
- ▶ Not straightforward to transform the data to approximate normality
- ▶ A different transformation would be needed within each DDE category
- ▶ First question: how to characterize gestational age at delivery distribution without considering predictors?

- ▶ Initially ignoring DDE
- ▶ Letting $y_i =$ gestational age at delivery for woman i ,

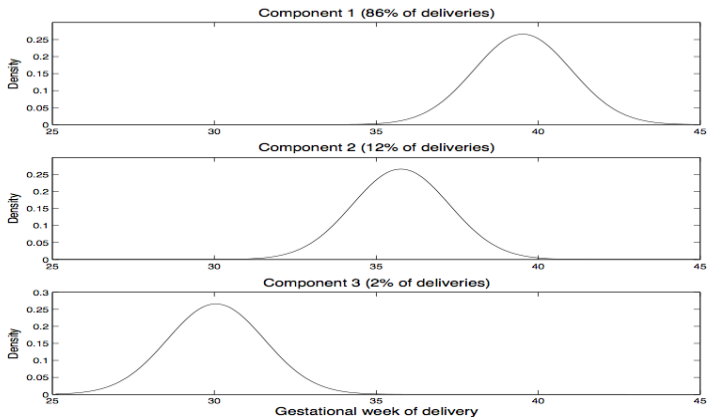
$$f(y_i) = \int N(y_i; \mu, \sigma^2) dG(\mu, \sigma^2),$$

where $G =$ mixture distribution for $\theta = (\mu, \sigma^2)$

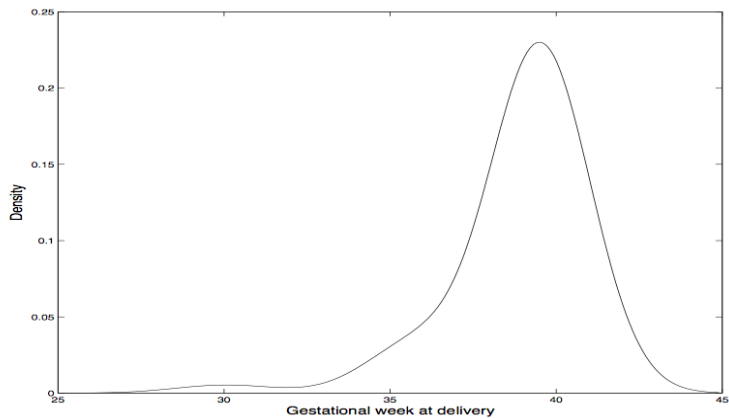
- ▶ Mixtures of normals can approximate any smooth density
- ▶ Finite location mixture with k components one possibility:

$$f(y) = \sum_{h=1}^k \pi_h N(y; \mu_h, \tau_h^{-1})$$

Mixture components for gestational age at delivery



Mixture-based density of gestational age at delivery



Some comments on mixture models

- ▶ $k = 3$ component mixture provides a good fit to gestational age at delivery data.
- ▶ Can be fit easily using the EM algorithm for maximum likelihood or Gibbs sampling for Bayesian inference.
- ▶ Ill focus on the Gibbs sampling approach here

- ▶ The finite mixture of normals can be equivalently expressed as

$$y_i \sim N(\mu_{S_i}; \tau_{S_i}^{-1}), S_i \sim \sum_{h=1}^k \pi_h \delta_h$$

δ_h = probability measure concentrated at the integer h ,
 $S_i \in \{1, 2, \dots, k\}$ indexes the mixture component for subject i , $i = 1, \dots, n$.

- ▶ A prior on $\pi = (\pi_1, \pi_2, \dots, \pi_k)$ and (μ_h, τ_h) , $h = 1, \dots, k$ is given by

$$\pi \sim \text{Dir}(a_1, a_2, \dots, a_k)$$
$$(\mu_h, \tau_h) \sim N(\mu_h; \mu_0, \kappa \tau_h^{-1}) \text{Ga}(\tau_h; a_\tau; b_\tau), h = 1, \dots, k$$

Posterior Computation in finite mixture models

- ▶ Update S_i from its multinomial conditional posterior with

$$Pr(S_i = h | -) = \frac{\pi_h \mathbf{N}(y_i; \mu_h, \tau_h^{-1})}{\sum_{l=1}^k \pi_l \mathbf{N}(y_i; \mu_l, \tau_l^{-1})}, h = 1, \dots, k$$

- ▶ Let $n_h = \#\{S_i = h, i = 1, \dots, n\}$ and $\bar{y}_h = \frac{1}{n_h} \sum_{i:S_i=h} y_i$ and update (μ_h, τ_h^{-1}) from its conditional posterior

$$(\mu_h, \tau_h^{-1} | -) = \mathbf{N}(\mu_h, \hat{\mu}_h, \hat{\kappa}_h \tau_h^{-1}) \text{Ga}(\tau_h, \hat{a}_{\tau_h}, \hat{b}_{\tau_h})$$

where

$$\hat{\kappa}_h = (\kappa^{-1} + n_h)^{-1}, \hat{\mu}_h = \hat{\kappa}(\kappa^{-1} \mu_0 + n_h \bar{y}_h), \hat{a}_{\tau_h} = a_\tau + \frac{n_h}{2},$$

$$\hat{b}_{\tau_h} = b_\tau + \frac{1}{2} \left\{ \sum_{i:S_i=h} (y_i - \bar{y}_h)^2 + \frac{n_h}{1 + \kappa n_h} (\bar{y}_h - \mu_0)^2 \right\}$$

- ▶ Update π as $(\pi | -) = \text{Dir}(a_1 + n_1, \dots, a_k + n_k)$.

- ▶ Gibbs sampler is trivial to implement
- ▶ Discarding a burn-in, monitor $f(y) = \sum_{h=1}^k \pi_h N(y; \mu_h, \tau_h^{-1})$ for a large number of iterations & a dense grid of y values
- ▶ Bayes estimate of $f(y)$ under squared error loss averages the samples
- ▶ Can also obtain 95% pointwise intervals for unknown density

Choosing the Dirichlet hyperparameters

- ▶ The choice of hyperparameters in the mixture model can have an important impact
- ▶ Focus initially on the choice of $(a_1, \dots, a_k)'$ in the Dirichlet prior
- ▶ A common choice is $a_1 = \dots = a_k = 1$, which seems “non-informative”.
- ▶ However, this is actually a poor choice in many cases, as it favors assigning roughly equal weights to the different components
- ▶ Ideally, we could choose k as an upper bound and choose hyperparameters, which favor a small number of components with relatively large weights

Finite Approximation to Dirichlet Process

- ▶ Ishwaran and Zarepour (2002), “Dirichlet prior sieves in finite normal mixtures,” *Statistica Sinica*, 12, 941-963) propose a finite approximation to the DP
- ▶ In particular, they propose letting
- ▶ $\pi \sim \text{Dir}(\alpha/k, \dots, \alpha/k)$ iid
- ▶ Assuming also that $\theta_h = (\mu_h, \tau_h) \sim P_0$, they show that

$$\lim_{k \rightarrow \infty} \sum_{h=1}^k \pi_h \delta_{\theta_h} \rightarrow DP(\alpha P_0)$$

- ▶ In addition, the posterior for the density is L_1 consistent if $\log k/n \rightarrow 0$.

- ▶ We can implement a finite mixture model analysis with a carefully chosen prior & sufficiently large k to obtain an accurate approximation to the DP mixture (DPM) model:
 - ▶ $f(y) = \int K(y; \theta) dP(\theta), P \sim DP(\alpha P_0)$
 - ▶ Here, $K(y; \theta)$ is a kernel parameterized by θ s - e.g., $K(y; \theta) = N(y; \mu, \tau^{-1})$ with $\theta = (\mu, \tau^{-1})$ for normal mixtures
 - ▶ P is now an unknown mixing measure
 - ▶ Hence, we no longer use the DP as a prior directly for the distribution of the data but instead use it for the mixture distribution

- ▶ The discreteness of the Dirichlet process is not a problem when it is used for a mixture distribution instead of directly for the data distribution
- ▶ In fact, in this setting the discreteness is appealing in leading to a simple representation of the mixture distribution that leads to clustering of the observations as a side effect
- ▶ Focusing on the finite approximation, $P \sim DP_k(\alpha, P_0)$, let

$$f(y) = \int N(y; \mu, \tau) dP(\mu, \tau)$$

- ▶ Note that induces a prior on f .