Bayesian Statistics

Debdeep Pati
Florida State University

March 21, 2017

Getting Started with Modeling

- As motivation, lets start with the relatively simple setting $y_i \sim f$ i.i.d
- The goal is to obtain a Bayes estimate of the density f
- From a frequentist perspective, a very common strategy is to rely on a simple histogram.
- Assume for simplicity we have pre-specified knots

$$\xi = (\xi_0, \xi_1, \ldots, \xi_k)',$$

$$\xi_0 < \xi_1 < \cdots < \xi_{k-1} < \xi_k \text{ and } y_i \in [\xi_0, \xi_k].$$

Bayesian Histograms

- The model for the density is as follows

$$f(y) = \sum_{h=1}^{k} 1(\xi_{h-1} < y \leq \xi_h) \frac{\pi_h}{(\xi_h - \xi_{h-1})}, y \in \mathbb{R}.$$

- To allow unknown numbers and locations of knots ξ, we can choose a prior for these quantities and use RJMCMC for posterior computation
- Focusing instead on fixed knots, we complete a Bayes specification with a prior for the probabilities

Dirichlet prior

- Assume a $Dirichlet(a_1, \ldots, a_k)$ prior for π,

$$\frac{\prod_{h=1}^{k} \Gamma(a_h)}{\Gamma(\sum_{h=1}^{k} a_h)} \prod_{h=1} \pi_h^{a_h - 1}$$

- The hyperparameter vector can be re-expressed as $a = \alpha\pi_0$, where $E(\pi) = \pi_0 = \{a_1/\sum_h a_h, \ldots, a_k/\sum_h a_h\}$ is the prior mean
- The posterior distribution of π is then calculated as

$$\begin{aligned} (\pi \mid y^n) &\propto \prod_{h=1}^{k} \pi_h^{a_h - 1} \prod_{i: y_i \in (\xi_{h-1}, \xi_h)} \frac{\pi_h}{\xi_h - \xi_{h-1}} \\ &\propto \prod_{h=1} \pi_h^{a_h + n_h - 1} \\ &\stackrel{\mathcal{D}}{=} Diri(a_1 + n_1, \ldots, a_k + n_k), \end{aligned}$$

where $n_h = \sum_i 1(\xi_{h-1} < y_i \leq \xi_h)$.

Simulation Experiment

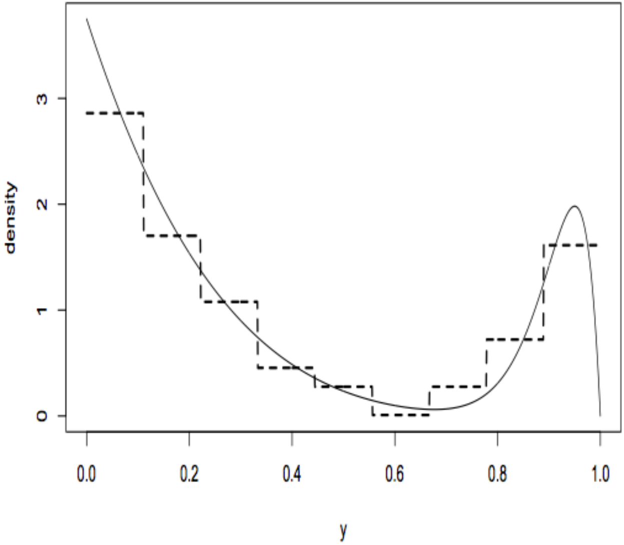
- To evaluate the Bayes histogram method, I simulated data from a mixture of two betas,

$$f(y) = 0.75\text{beta}(y; 1, 5) + 0.25\text{beta}(y; 20, 2).$$

 for $n = 100$ samples were obtained from this density

- Assuming data between $[0, 1]$ and choosing a 10 equally-spaced knots, we applied the Bayes histogram approach
- The true density and Bayes posterior mean are plotted on the next slide

Bayes Histogram Estimate for Simulation Example



Comments

- Procedure is really easy in that we have conjugacy
- Results very sensitive to knots & allowing free knots is computationally demanding
- In addition, even averaging over random knots we tend to get bumps in the density estimate as an artifact
- Allows prior information to be included in frequentist histogram estimates easily
- Dirichlet prior perhaps not best choice due to lack of smoothing across adjacent bins

Is this approach nonparametric?

- I would say no - we have a flexible parametric model
- Including free knots leads to a nonparametric specification in which any density can be accurately approximated & we can obtain large support
- The fixed knot Bayesian histogram approach does not have (full) weak support on the set of densities wrt to Lesbesgue measure.

The trouble with histograms?

- Histograms have the unappealing characteristics of bin sensitivity & approximating a smooth density with piecewise constants
- In addition, extending histograms to multiple dimensions & to include predictors is problematic due to an explosion of the number of bins needed
- To be realistic we need to account for uncertainty in the number & locations of bins, but this is a pain computationally
- Can we define a model that bypasses the need to explicitly specify bins?

Histograms & RPMs

- Suppose the sample space is Ω & we partition Ω into Borel subsets $B_1, \ldots, B_k$
- If $\Omega = \mathbb{R}$, then $B_1, \ldots, B_k$ are simply non-overlapping intervals partitioning the real line into a finite number of bins
- Letting P denote the unknown probability measure over $(\Omega, \mathcal{B})$, the probabilities allocated to the bins is

$$\{P(B_1), ..., P(B_k)\} = \left\{ \int_{B_1} f(y)dy, \ldots, \int_{B_k} f(y)dy \right\}$$

- If P is a random probability measure (RPM), then these bin probs are random variables

Dirichlet processes (Ferguson, 1973; 1974)

- As discussed last lecture, a simple conjugate prior for the bin probabilities corresponds to the Dirichlet distribution
- For example, we could let

$$\{P(B_1), ..., P(B_k)\} \sim Dir\{\alpha P_0(B_1), \ldots, \alpha P_0(B_k)\} \tag{1}$$

- P_0 is a "base" probability measure providing an initial guess at P & α is a prior concentration parameter
- Ferguson's idea: eliminate sensitivity to choice of $B_1, \ldots, B_k$ & induce a fully specified prior on P, through assuming (1) holds for all $B_1, \ldots, B_k$ & all k.

Dirichlet processes (Ferguson, 1973; 1974)

- For Ferguson's specification to be coherent, there must exist an RPM P such that the probs assigned to any measurable partition $B_1, \ldots, B_k$ by P is $Dir\{\alpha P_0(B_1), \ldots, \alpha P_0(B_k)\}$
- The existence of such a P can be shown by verifying the Kolmogorov consistency conditions
- The first Kolmogorov condition is automatic, since (1) is defined free of the order of the sets
- The remaining condition relates to coherence across different partitions - e.g, if we form a new partition by taking unions of some of the sets in $B_1, \ldots, B_k$ then the resulting probs assigned to this new partition must still be Dirichlet with the same form

Dirichlet process: a prior for the space of probability distributions

- A Dirichlet distribution is a distribution over the K-dimensional probability simplex:

$$\Delta_K = \{(\pi_1, \pi_2, \ldots, \pi_k) : \pi_k \geq 0, \sum_{k=1}^{K} \pi_k = 1\}$$

- We say $(\pi_1, \ldots, \pi_k)$ is Dirichlet distributed $(\lambda_1, \lambda_2, \ldots, \lambda_k)$ if

$$p(\pi_1, \ldots, \pi_k) = \frac{\Gamma(\sum_k \lambda_k)}{\prod_{k=1}^{K} \Gamma(\lambda_k)} \prod_{k=1}^{n} \pi_k^{\lambda_k - 1}$$

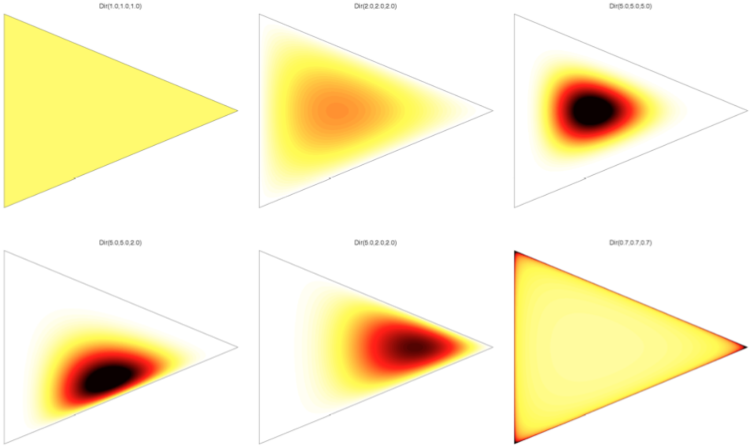
- Equivalent to normalizing a set of independent gamma variables

$$(\pi_1, \ldots, \pi_k) \stackrel{d}{=} \frac{1}{\sum_k \gamma_k}(\gamma_1, \ldots, \gamma_k)$$

$$\gamma_j \sim \text{Gamma}(\lambda_k, \beta)$$

Dirichlet distribution

Figure: Dirichlet distribution



Agglomerative & Decimative properties of DP

- Combining entries by their sum

$$
\begin{aligned}
(\pi_1, \ldots, \pi_K) &\sim \text{Diri}(\alpha_1, \ldots, \alpha_K) \\
(\pi_1, \ldots, \pi_i + \pi_j \ldots, \pi_K) &\sim \text{Diri}(\alpha_1, \ldots, \alpha_i + \alpha_j, \ldots \alpha_K)
\end{aligned}
$$

- Decimating one entry into two

$$
\begin{aligned}
(\pi_1, \ldots, \pi_K) &\sim \text{Diri}(\alpha_1, \ldots, \alpha_K) \\
(\tau_1, \tau_2) &\sim \text{Diri}(\alpha_i \beta_1, \alpha_i \beta_2) \\
(\pi_1, \ldots, \pi_i \tau_1, \pi_i \tau_2, \ldots, \pi_K) &\sim \text{Diri}(\alpha_1, \ldots, \alpha_i \beta_1, \alpha_i \beta_2, \ldots, \alpha_K)
\end{aligned}
$$

Existence of Dirichlet process

- $(B'_1, \ldots, B'_{k'})$ and $(B_1, \ldots, B_k)$ are measurable partitions
- $(B'_1, \ldots, B'_{k'})$ is a refinement of $(B_1, \ldots, B_k)$s with $B_1 = \cup_1^{r_1} B'_j, B_2 = \cup_{r_1+1}^{r_2} B'_j, \ldots B_k = \cup_{r_{k-1}+1}^{k'} B'_j$
- Then, the distribution of $P(B'_1), \ldots, P(B'_{k'})$ induces a distribution on

$$\sum_1^{r_1} P(B'_j), \sum_{r_1+1}^{r_2} P(B'_j), \cdots, \sum_{r_{k-1}+1}^{k'} P(B'_j)$$

  which is equivalent to the distribution of $P(B_1), \ldots, P(B_k)$.
- Ferguson shows this condition is sufficient for Kolmogorov consistency