STA 281: Nonparametric Methods
Lecture 2: Bayesian Histograms → Dirichlet Processes

David Dunson

Department of Statistical Science, Duke University

Tuesday Jan 19, 2010
The model for the density is as follows

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

with \( \pi = (\pi_1, \ldots, \pi_k)' \) an unknown probability vector.

To allow unknown numbers and locations of knots \( \xi \), 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.
The Dirichlet prior is appealing in having simple to interpret hyperparameters and in leading to a conjugate posterior distribution.

However, the Dirichlet prior cannot accommodate smoothing of adjacent bins in the histogram.

Appealing to favor $\pi_h \approx \pi_{h-1}$ (at least for equally-sized bins)

- Suppose the support is partitioned into $k$ intervals $B_1, \ldots, B_k$ of equal width,

  $$B_j = (\zeta_j - \frac{1}{2} W, \zeta_j + \frac{1}{2} W], \quad j = 1, \ldots, k,$$

  with midpoints $\zeta_1, \ldots, \zeta_k$ and width $W$

- Let $n_j$ denote the number of observations falling in interval $B_j$, with the probability of falling in interval $j$ being

  $$\pi_j = \int_{B_j} f(y) dy$$

- To allow dependence, we let $\pi_j = e^{\gamma_j} / \sum e^{\gamma_g}$, with $\gamma = (\gamma_1, \ldots, \gamma_k)'$ then assigned a multivariate normal prior

- The mean can be chosen to correspond to prior knowledge of the shape of the density (e.g., uniform or triangular)

- An AR-1 type dependence structure allows smoothing of adjacent bins
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 $\Omega$ & we partition $\Omega$ 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
  \[
  \{P(B_1), \ldots, P(B_k)\} = \left\{ \int_{B_1} f(y) dy, \ldots, \int_{B_k} f(y) dy \right\}.
  \]
- 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
  \[
  \{ P(B_1), \ldots, P(B_k) \} \sim \text{Diri}(\alpha P_0(B_1), \ldots, \alpha P_0(B_k)), \tag{1}
  \]
- \( P_0 \) is a “base” probability measure providing an initial guess at \( P \) & \( \alpha \) 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 \( k \).
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 $\text{Diri}(\alpha P_0(B_1), \ldots, \alpha P_0(B_k))$.

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.
Condition C

- $(B'_1, \ldots, B'_{k'})$ and $(B_1, \ldots, B_k)$ are measurable partitions
- $(B'_1, \ldots, B'_{k'})$ is a refinement of $(B_1, \ldots, B_k)$ with
  $B_1 = \bigcup_{j=1}^{r_1} B'_j$, $B_2 = \bigcup_{j=r_1+1}^{r_2} B'_j$, \ldots, $B_k = \bigcup_{j=r_{k-1}+1}^{k'} B'_j$
- Then, the distribution of $P(B'_1), \ldots, P(B'_{k'})$ induces a distribution on
  \[
  \left( \sum_{j=1}^{r_1} P(B'_j), \sum_{j=r_1+1}^{r_2} P(B'_j), \ldots, \sum_{j=r_{k-1}+1}^{k'} P(B'_j) \right)
  \]
  which is equivalent to the distribution of $P(B_1), \ldots, P(B_k)$
- Ferguson shows this condition is sufficient for Kolmogorov consistency
Heuristic proof of condition C

Note that $P(B'_1), \ldots, P(B'_k) \sim \text{Diri}(\alpha P_0(B'_1), \ldots, \alpha P_0(B'_k))$
can be induced by

$$Y_j = P(B'_j) = \frac{Z_j}{\sum_{i=1}^{k} Z_i}, \quad Z_j \sim \text{gamma}(\alpha P_0(B'_j), 1), j = 1, \ldots, k.$$ 

Hence, we have

$$P(B_h) = \sum_{r_h-1}^{r_h} P(B'_j) \quad \text{and}$$

$$P(B_h) = \frac{\sum_{r_h-1}^{r_h} Z_j}{\sum_{i=1}^{k} Z_i}, \quad \sum_{r_h-1}^{r_h} Z_j \sim \text{gamma}\left(\alpha \sum_{r_h-1}^{r_h} P_0(B'_j), 1\right).$$

It follows that

$P(B_1), \ldots, P(B_k) \sim \text{Diri}(\alpha P_0(B_1), \ldots, \alpha P_0(B_k))$
Moment properties of the DP

Let $P \sim \text{DP}(\alpha P_0)$ denote that the probability measure $P$ on $(\Omega, \mathcal{B})$ is assigned a Dirichlet process (DP) prior with scalar precision $\alpha > 0$ and base probability measure $P_0$.

From the definition of the Dirichlet process & properties of the Dirichlet, we have

$$P(B) \sim \text{beta}(\alpha P_0(B), \alpha\{1 - P_0(B)\}), \quad \text{for all } B \in \mathcal{B},$$

Hence, we have

$$\mathbb{E}\{P(B)\} = P_0(B), \quad \text{for all } B \in \mathcal{B},$$

so that the prior for $P$ is centered on $P_0$.

In addition, we have

$$\text{V}\{P(B)\} = \frac{P_0(B)\{1 - P_0(B)\}}{1 + \alpha}, \quad \text{for all } B \in \mathcal{B},$$

so that $\alpha$ is a precision parameter controlling the variance.
Large Support of the DP

Let $Q \in \mathcal{P}$ denote a fixed probability measure on $(\Omega, \mathcal{B})$

From proposition 3 in Ferguson (1973), for any positive integer $k$, measurable sets $B_1, \ldots, B_k$ and $\epsilon > 0$,

$$\Pr\{|P(B_i) - Q(B_i)| < \epsilon \text{ for } i = 1, \ldots, k\} > 0.$$ 

The topology of pointwise convergence corresponds to $P_n \rightarrow P$, if for every $B \in \mathcal{B}$, $P_n(B) \rightarrow P(B)$

Under this topology, the support of the DP contains all probability measures whose support is contained in the support of $P_0$
Conjugacy

- Let $P \sim \text{DP}(\alpha P_0)$ and let $y_i \overset{\text{iid}}{\sim} P$ (following standard practice in using $P$ to denote both the probability measure and its corresponding distribution).

- For any measurable partition $B_1, \ldots, B_k$, we have

\[
(P(B_1), \ldots, P(B_k) \mid y_1, \ldots, y_n) \sim \text{Diri}\left(\alpha P_0(B_1) + \sum_{i=1}^{n} 1(y_i \in B_1), \ldots, \alpha P_0(B_k) + \sum_{i=1}^{n} 1(y_i \in B_k)\right)
\]

- From this & the above development, it is straightforward to obtain

\[
(P \mid y_1, \ldots, y_n) \sim \text{DP}\left(\alpha P_0 + \sum_i \delta_{y_i}\right).
\]
The updated precision parameter is $\alpha + n$, so that $\alpha$ is in some sense a prior sample size.

The posterior expectation of $P$ is defined as

$$E\{P(B)\mid y^n\} = \left(\frac{\alpha}{\alpha + n}\right) P_0(B) + \left(\frac{n}{\alpha + n}\right) \sum_i \frac{1}{n}\delta_{y_i}.$$ 

Hence, the Bayes estimator of $P$ under squared error loss is the empirical measure with equal masses at the data points shrunk towards the base measure.
Bayesian Bootstrap

- Note that in the limit as $\alpha \to 0$ we obtain the posterior,

$$ (P \mid y^n) \sim \text{DP}(\sum_{i=1}^{n} \delta_{y_i}) $$

- This limiting posterior is known as the Bayesian bootstrap

- Samples from the Bayesian bootstrap correspond to discrete distributions supported at the observed data points with Dirichlet distributed weights

- Compared with the typical Efron bootstrap, the Bayesian bootstrap leads to smoothing of the weights
Posterior Asymptotics

- Refer to Ghosal (www4.stat.ncsu.edu/ sghosal/papers/BayesAsymp.pdf) for an excellent review of properties of the DP and theory.
- Under a DP prior with fixed $\alpha$ as $n \to \infty$, the posterior mean will correspond to the empirical probability measure.
- There is a rich literature on asymptotics of the empirical distribution - e.g., the empirical cdf is known to converge uniformly to the cdf of the true distribution.
- It can be shown that the posterior distribution of $P(B)$ converges to a degenerate mass at $P_t(B)$, with rate $n^{-1/2}$. 
Some (potentially) unappealing characteristics of DP

- The DP has a lack of smoothness, which is clear from the following example [on board]
- The DP also induces negative correlation between $P(B_1)$ and $P(B_2)$ for any two disjoint sets $B_1$ and $B_2$ - even if $B_1$ and $B_2$ are adjacent
- Realizations from the DP are almost surely discrete
- Note that $P$ is discrete iff $P(y : P(y) > 0) = 1$ - an obvious consequence of expression of DP as a normalized gamma process
Simulate data from a mixture of Gaussians,

\[ y_i \sim 0.25N(0, 0.5) + 0.75N(2, 1), \quad i = 1, \ldots, 100 \]

Assuming a Dirichlet process prior, obtain a Bayes estimate of the cdf & plot this estimate against the true cdf.

Repeat the simulation 100 times and summarize the results.

Comment on the performance of the Bayes estimator.