1 Bayesian histogram

As motivation, let us start with the 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.

$$\xi = (\xi_0, \xi_1, \ldots, \xi_k)^\prime,$$

where $\xi_0 < \xi_1 < \cdots < \xi_{k-1} < \xi_k$ and $y_i \in [\xi_0, \xi_k]$. 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 $\xi$, 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. Assume a Dirichlet$(a_1, \ldots, a_k)$ prior for $\pi$,

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

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 $\pi$ is then calculated as

$$\begin{align*}
(\pi | y^n) &\propto \prod_{h=1}^{k} \pi_h^{a_h-1} \prod_{i: y_i \in (\xi_{h-1}, \xi_h)} \frac{\pi_h}{\xi_h - \xi_{h-1}} \\
&\propto \prod_{h=1}^{k} \pi_h^{a_h + n_h - 1} \\
&\overset{D}{=} \text{Dirichlet}(a_1 + n_1, \ldots, a_k + n_k),
\end{align*}$$

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

1.1 Simulation Experiment

To evaluate the Bayes histogram method, let’s simulated data from a mixture of two betas,

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

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 in Figure 1.1.

Clearly, this procedure is really easy in that we have conjugacy. However, the 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. Dirichlet prior is perhaps not best choice due to lack of smoothing across adjacent bins. We have a flexible parametric model, but the approach is not nonparametric. 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.

**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?

### 2 Random probability measures

**Goal:** Let $(\Omega, \mathcal{B})$ denote the probability space. The objective is to define a probability on $\{P(B) : B \in \mathcal{B}\}$.

Suppose the sample space is $\Omega$ & we partition $\Omega$ into Borel subsets $B_1, \ldots, B_k$. For example, if $\Omega = \mathbb{R}$, $B_1, \ldots, B_k$ 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. A simple conjugate prior for the bin probabilities corresponds to the Dirichlet distribution. For example, we
could let
\[ \{P(B_1), \ldots, P(B_k)\} \sim \text{Dirichlet}\{\alpha P_0(B_1), \ldots, \alpha P_0(B_k)\} \quad (1) \]

\(P_0\) is a “base” probability measure providing an initial guess at \(P\) & \(\alpha\) is a prior concentration parameter. Ferguson’s idea \([1, 2]\) is to 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 Dirichlet\{\(\alpha P_0(B_1), \ldots, \alpha P_0(B_k)\)\}.

2.1 Dirichlet distribution

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

which is equivalent to normalizing a set of independent gamma variables
\[ (\pi_1, \ldots, \pi_k) \overset{d}{=} \frac{1}{\sum_k \gamma_k} (\gamma_1, \ldots, \gamma_k) \]
\[ \gamma_j \sim \text{Gamma}(\lambda_k, \beta) \]

Figure 1: Dirichlet distribution
2.2 Agglomerative, Decimative and Renormalization properties of Dirichlet distribution

Dirichlet distribution is invariant to marginalization

\[
\left(\pi_1, \ldots, \pi_K\right) \sim \text{Diri}(\alpha_1, \ldots, \alpha_K)
\]
\[
\left(\pi_1, \ldots, \pi_i + \pi_j, \ldots, \pi_K\right) \sim \text{Diri}(\alpha_1, \ldots, \alpha_i + \alpha_j, \ldots, \alpha_K).
\]

One can decimate a probability and still retain the distribution

\[
\left(\pi_1, \ldots, \pi_K\right) \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_K)
\]
\[
\left(\tau_1, \tau_2\right) \sim \text{Dirichlet}(\alpha_i\beta_1, \alpha_i\beta_2), \quad \beta_1 + \beta_2 = 1
\]
\[
\left(\pi_1, \ldots, \pi_i\tau_1, \pi_i\tau_2, \ldots, \pi_K\right) \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_i\beta_1, \alpha_i\beta_2, \ldots, \alpha_K)
\]

Finally, the Dirichlet distribution satisfies renormalization rule. If \(\left(\pi_1, \ldots, \pi_K\right) \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_K)\), then

\[
\left(\pi_2, \ldots, \pi_K\right) \sum_{k=2}^{K} \pi_k \sim \text{Dirichlet}(\alpha_2, \ldots, \alpha_K).
\]

2.3 Existence of a process with Dirichlet marginals (Dirichlet process) [1, 2]

The first Kolmogorov condition is automatic, since (1) is defined free of the order of the sets. Assume \((B_1', \ldots, B_k')\) and \((B_1, \ldots, B_k)\) are measurable partitions such that \((B_1', \ldots, B_k')\) is a refinement of \((B_1, \ldots, B_k)\) with \(B_1 = \bigcup_{r=1}^{r_1} B_1', B_2 = \bigcup_{r_1+1}^{r_2} B_2', \ldots, B_k = \bigcup_{r_{k-1}+1}^{r_k} B_k'\). Then, the distribution of \(P(B_1'), \ldots, P(B_k')\) induces a distribution on

\[
\sum_{1}^{r_1} P(B_1'), \sum_{r_1+1}^{r_2} P(B_2'), \ldots, \sum_{r_{k-1}+1}^{r_k} P(B_k')
\]

which has the same form of distribution for \(P(B_1), \ldots, P(B_k)\). Ferguson shows this condition is sufficient for Kolmogorov consistency. This ensures there exists a probability measure on \(\{P(B) : B \in \mathcal{B}\}\). 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 \(P_0\).

2.4 Moment properties of the DP

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

\[
P(B) \sim \text{Beta}[\alpha P_0(B), \alpha(1 - P_0(B))], \text{ for all } B \in \mathcal{B}.
\]

Hence, we have \(E\{P(B)\} = P_0(B)\), for all \(B \in \mathcal{B}\), so that the prior for \(P\) is centered on \(P_0\). In addition, we have

\[
\text{Var}\{P(B)\} = \frac{P_0(B)(1 - P_0(B))}{1 + \alpha}, \text{ for all } B \in \mathcal{B},
\]

so that \(\alpha\) is a precision parameter controlling the variance.
2.5 Conjugacy

Let $P \sim \text{DP}(\alpha, P_0)$ and let $y_i \mid P \sim P$ i.i.d (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

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

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

$$(P \mid y_1, \ldots, y_n) \sim \text{DP}\left(\alpha P_0 + \sum_{i}^{n} \delta_{y_i}\right)$$

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

$$E\{P(B) \mid y\} = \frac{\alpha}{\alpha + n} P_0(B) + \frac{n}{\alpha + n} \sum_{i}^{n} \frac{1}{n} \delta_{y_i}$$

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.

2.6 Posterior Dirichlet process

$$\begin{bmatrix} G \sim \text{DP}(\cdot \mid \alpha, G_0) \ \theta \mid G \sim G \end{bmatrix} \iff \begin{bmatrix} \theta \sim G_0 \ G \sim \text{DP}(\cdot \mid \alpha + 1, \frac{aG_0 + \delta_{\theta}}{\alpha + 1}) \end{bmatrix}$$

2.7 Pólya Urn Scheme & Chinese Restaurant process

$$\theta' \mid \theta, G_0 = \int [\theta' \mid G][G \mid \theta]dG = \int G[G \mid \theta]dG = \frac{\alpha G_0 + \delta_{\theta}}{\alpha + 1}$$

$$\theta_n \mid \theta_1, \ldots, \theta_{n-1}, G_0 \sim \frac{\alpha G_0 + \sum_{i=1}^{n-1} \delta_{\theta_i}}{\alpha + n - 1}$$

This shows the clustering effect explicitly. Restaurant has infinitely many tables $k = 1, \ldots$. Tables have values $\theta_k$ drawn from $G_0$. Customers are indexed by $i = 1, \ldots$ with values $\phi_i$. $K =$ total number of occupied tables so far. $n =$ total number of customers so far. $n_k =$ number of customers seated at table $k$.

2.8 Relationship between CRP and DP

DP is a distribution over distributions whereas DP results in discrete distributions, so if you draw $n$ points you are likely to get repeated values. A DP induces a partitioning of the $n$ points e.g. $(134)(25) \iff \phi_1 = \phi_3 = \phi_4 \neq \phi_2 = \phi_5$ CRP is the corresponding distribution over partitions.
2.9 A stick breaking construction for $G \sim DP(\cdot, \alpha, G_0)$

This is by Sethuraman (1994).

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

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

2.10 Sketch of the proof of equivalence

Recall the posterior process

\[
\begin{align*}
G & \sim DP(\cdot | \alpha, G_0) \\
\theta | G & \sim G
\end{align*}
\]

\(\iff\)

\[
\begin{align*}
\theta & \sim G_0 \\
G \mid \theta & \sim DP(\cdot, \alpha + 1, \frac{\alpha G_0 + \delta\theta}{\alpha + 1})
\end{align*}
\]

Consider a partition \((\theta, \Theta \setminus \theta)\) of \(\Theta\). We have

\[
(G(\theta), G(\Theta \setminus \theta)) \sim \text{Diri}\left\{(\alpha + 1)\frac{\alpha G_0 + \delta\theta}{\alpha + 1}(\theta), (\alpha + 1)\frac{\alpha G_0 + \delta\theta}{\alpha + 1}(\Theta \setminus \theta)\right\}
\]

\[= \text{Beta}(1, \alpha)\]

\(G\) has a point mass located at \(\theta\):

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

and \(G'\) is the renormalized probability measure with the point mass removed. Let us know find \(G'\).

Consider a further partition of \((\theta, A_1, \ldots, A_K)\) of \(\Theta\).

\[
(G(\theta), G(A_1), \ldots, G(A_K)) = (\beta, (1 - \beta)G'(A_1), \ldots, (1 - \beta)G'(A_K))
\]

\[\sim \text{Diri}(1, \alpha G_0(A_1), \ldots, \alpha G_0(A_K))\]

Renomalizing

\[
(G'(A_1), \ldots, G'(A_K)) \mid \theta = \text{Diri}(\alpha G_0(A_1), \ldots, \alpha G_0(A_K))
\]

\[G' \sim DP(\cdot, \alpha, G_0)\]

Then

\[G \sim DP(\cdot, \alpha, G_0)\]

\[G = \beta_1 \delta_{\theta_1} + (1 - \beta_1)G_1\]

\[G = \beta_1 \delta_{\theta_1} + (1 - \beta_1)(\beta_2 \delta_{\theta_2} + (1 - \beta_2)G_2)\]

\[\vdots\]

\[G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k}\]

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

References
