Bayesian Analysis through Mixture Extensions

STA 215. Surya Tokdar

Beyond simple models

For the Bayesian paradigm, so far we have only used models consisting of simple exponential family distributions. In contrast, we did see classical statistics being used in a broader context (see notes from Feb 15 and 17). Such uses came through statistical procedures for which frequentist guarantees could be attached for much larger collection of models. These procedures are rarely based on the likelihood function w.r.t to the large model, usually because the ML approach breaks down when applied to such large collections.

What about Bayesian models and analysis with models that go beyond simple exponential family distribution? Can we keep using the likelihood function and apply Bayes theorem? How much of extension can we make and still be able to use the Bayesian paradigm? It turns out that we can indeed match extensions from the classical paradigm, and yet be able to use the likelihood function and thus obey the likelihood principle. However, this requires a careful construction of models and a careful choice of the prior distribution.

A standard way to extend a simple exponential family distribution model to a larger collection is to use the concept of mixtures. The collection of two-component normal mixture pdfs is much larger than the collection of normal pdfs. With more components we get larger collections and can accommodate richer variations in the data distribution. We can even go up to infinite component models. Such extensions are focus of this handout.

Mixtures

The use of exponential family distributions can sometimes be justified based on first principles. In many cases it is felt that the population consists of homogeneous subgroups. Observables from each subgroup can be adequately modeled with an exponential family distribution. But different subgroups can have very different parameter values.

Example (Sweden speed limit experiment)**.** R documentation on the dataset `Traffic` (in `MASS`).

> An experiment was performed in Sweden in 1961-2 to assess the effect of a speed limit on the motorway accident rate. The experiment was conducted on 92 days in each year, matched so that day j in 1962 was comparable to day j in 1961. On some days the speed limit was in effect and enforced, while on other days there was no speed limit and cars tended to be driven faster. The speed limit days tended to be in contiguous blocks.

The variables recorded were `year` (1961 or 1962), `day` (of year), `limit` (was it imposed?) and `y` (traffic accident count of that day).

Consider only those days when speed limit was imposed. On any of these days it is reasonable to model the accident count as a Poisson variable. But certainly there are may intrinsic differences between these days that would alter the rate (weather, holidays, etc.). One can think of two types of days: 'Up' and 'Down' with high and low susceptibility of traffic accidents respectively.

With K (in above $K = 2$) levels of the intrinsic factors, the extended model looks like:

$$f(y|\theta) = \prod_{j=1}^{n} \left[\sum_{l=1}^{K} \omega_l Poi(y_j|\lambda_l) \right]$$

where $\theta = (\omega_l, \lambda_l)_{l=1}^{K}$ with $\omega_l > 0$, $\sum_l \omega_l = 1$ and $\lambda_l > 0$.

Prior distribution on θ. There is no conjugate prior distribution for θ! A reasonable choice is

$$\pi(\theta) = Dir(\omega|a_1, \cdots, a_K) \prod_{l=1}^{K} Ga(\lambda_l|b_{1l}, b_{2l})$$

The Dirichlet distributions are a multivariate extension of the beta distributions and are conjugate to multinomial sampling densities. The gamma distributions, as we have seen before, are conjugate to the Poisson sampling distributions. We shall see shortly that these are semi-conjugate specifications.

Variable augmentation. Let z_j denote the level day j belongs to. Then the above model is equivalent to:

$$\begin{aligned} f(y|z, \theta) &= \prod_{j=1}^{n} Poi(y_j|\lambda_{z_j}) \\ \pi(z|\theta) &= \prod_{j=1}^{n} \mathsf{Multinomial}(z_j|K, \omega) \\ \pi(\theta) &= Dir(\omega|a_1, \cdots, a_K) \prod_{l=1}^{K} Ga(\lambda_l|b_{1l}, b_{2l}) \end{aligned}$$

The z_j's are called latent variables as they are not directly observable – for us, they are just unknown quantities and hence a part of the parameters. We shall look at the joint posterior distribution of (θ, z).

Posterior computation. The likelihood function $L_y(z, \theta)$ equals

$$L_y(z, \theta) = \text{const} \times \prod_{j=1}^{n} e^{-\lambda_{z_j}} \lambda_{z_j}^{y_j} = \text{const} \times \prod_{l=1}^{k} e^{-m_l \lambda_l} \lambda_l^{m_l \bar{y}_l}$$

where $m_l = \#\{j : z_j = l\}$ and $\bar{y}_l = \frac{1}{m_l}\sum_{j:z_j=k} y_j$. With the same notations we can also simplify $\pi(z|\theta)$ as

$$\pi(z|\theta) = \text{const} \times \prod_{j=1}^{n} \omega_{z_j} = \text{const} \times \prod_{l=1}^{l} \omega_l^{m_l}.$$

Finally, by definition,

$$\pi(\theta) = \text{const} \times \prod_{l=1}^{k} \omega_l^{a_k - 1} \times \prod_{l=1}^{k} \lambda_l^{b_{1l}-1} e^{-b_{2l}\lambda_l}.$$

Putting all these together we see

$$\pi(z, \theta|y) = \text{const} \times \prod_{l=1}^{k} \omega_l^{a_k + m_l - 1} \times \prod_{l=1}^{k} \lambda_l^{b_{1l} + m_l\bar{y}_l - 1} e^{-(b_{2l}+m_l)\lambda_l}$$

from which we can write down the following.

$$\begin{aligned}
\pi(\theta|z, y) &= \text{const} \times \pi(\theta, z|y) \quad \text{[seen as a function of only } \theta\text{]} \\
&= Dir(\omega|a_1 + m_1, \cdots, a_K + m_K) \times \prod_{l=1}^{K} Ga(\lambda_l|b_{1l} + m_l\bar{y}_l, b_{2l} + m_l) \\
\pi(z|\theta, y) &= \text{const} \times \pi(\theta, z|y) \quad \text{[seen as a function of only } z\text{]} \\
&= \text{const} \times \prod_{l=1}^{k} \omega_l^{m_l} \lambda_l^{m_l\bar{y}_l} e^{-m_l\lambda_l} \\
&= \text{const} \times \prod_{j=1}^{n} \omega_{z_j} \lambda_{z_j}^{y_j} e^{-\lambda_{z_j}} \\
&= \prod_{j=1}^{n} \mathsf{Multinomial}(z_j|K, \omega^{(j)})
\end{aligned}$$

where $\omega^{(j)} \propto (\omega_1\lambda_1^{y_j}e^{-\lambda_1}, \cdots, \omega_K\lambda_K^{y_j}e^{-\lambda_K})$, $j = 1, \cdots, n$.

Monte Carlo approximation and Gibbs sampling. In principle, any posterior summary of the form $\int h(\theta, z)\pi(\theta, z|y)d\theta dz$ can be approximated by a Monte Carlo average

$$\int h(\theta, z)\pi(\theta, z|y)d\theta dz \approx \frac{1}{M}\sum_{m=1}^{m} h(\theta^{(m)}, z^{(m)}) \tag{1}$$

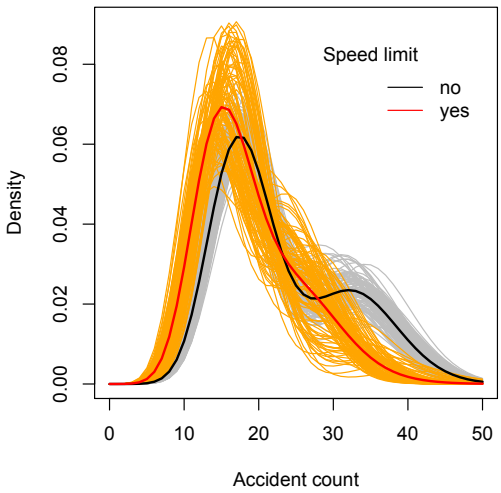
where $(\theta^{(1)}, z^{(1)}), \cdots, (\theta^{(M)}, z^{(M)})$ are M samples from the posterior distribution $\pi(\theta, z|y)$ (with M large). Such a sample can be collected by running a *Gibbs sampler* that creates a chain of realizations (θ^t, z^t), $t = 1, 2, \cdots$ as follows:

- Start with an initial guess (θ^0, z^0).
- For $t = 1, 2, \cdots$
  - Update z^{t-1} to z^t by sampling from $\pi(z|\theta = \theta^{t-1}, y)$
  - Then update θ^{t-1} to θ^t by sampling from $\pi(\theta|z = z^t, y)$.

The first B iterations of the chain are ignored (with B not too small) and then the chain is tracked at every r-th iteration to collect a sample: $(\theta^{(m)}, z^{(m)}) = (\theta^{B+mr}, z^{B+mr})$, $m = 1, \cdots, M$. Theoretical justification behind (1) can be found in various books on Monte Carlo computations. My favorite is Jun Liu's book titled *Monte Carlo Strategies in Scientific Computing*.

This can be taken one step further to make prediction about any future observable $Y^* \sim f^*(y^*|\theta^*)$. For example if we are interested in $\int h(y^*) f^*(y^*|y) dy^*$ then we can first generate samples $y^{*(m)}$ of Y^* from $f^*(y^*|\theta = \theta^{(m)})$ and then use the Monte Carlo average $\frac{1}{M}\sum_{m=1}^{M} h(y^{*(m)})$. Sampling $y^{*(m)}$ may first require sampling $z^* \sim \mathsf{Multinomial}(z^*|K, \omega^{(m)})$.

Example (Traffic data analysis)**.** For each group (`limit = yes` and `limit = no`) I used the above mixture model with: $K = 2$, $a_1 = 2, a_2 = 1$, $b_{1l} = 2$, $b_{2l} = 1/10$, $l = 1, 2$. For each data separately, I ran a Gibbs sampler for 2000 iterations, discarded the first 1000, and saved every 10th iteration of the remaining chain, giving a posterior sample of size $M = 100$. For each sampled $\theta^{(m)} = (\omega^{(m)}, \lambda^{(m)})$ I sampled a $z^{*(m)} \sim \mathsf{Multinomial}(K, \omega^{(m)})$ and then sampled $y^{*(m)} \sim Poi(\lambda^{(m)}_{z^{*(m)}})$. The resulting $(y^{*(1)}, \cdots, y^{*(M)})$ represent samples of a future observable $Y^* \sim f^*(y^*|\theta)$ from the posterior predictive $f^*(y^*|y) = \int f^*(y^*|\theta)\pi(\theta, z|y) d\theta dz$. The figure below shows $f^*(y^*|\theta^*)$ for θ^* drawn from the posterior for each of the two datasets.



Finally I compared the posterior predictive draws for the two data (speed limit yes/no) and calculated approximate the posterior probability that a speed limit day has less accidents than a no speed limit day. This probability turns out to be 65%.

Of course, this could be done with other choices of the parameter a_1, a_2 etc., or even with a completely different specification of the model. My purpose was to show that we can go beyond simple conjugate models and use interesting mixture extensions of simple exponential family distributions.

Infinite mixtures

Letting $g(y_i|\lambda)$ denote the Poisson pmf, the above model is same as: $Y_j \overset{\text{IID}}{\sim} f(y_j|\theta)$ with $\theta = (\omega, \lambda)$ where

$$f(y_j|\theta) = \sum_{l=1}^{K} \omega_l g(y_j|\lambda_l).$$

Of course this could be defined for any pmf/pdf $g(y_i|\lambda)$ suitable for modeling Y_j's, with a suitable parameter $\lambda_l \in \Lambda_g$. Taking this idea to the limit, we could also define an infinite mixture model

$$f(y_j|\theta) = \sum_{l=1}^{\infty} \omega_l g(y_j|\lambda_l)$$

with $\theta = (\omega, \lambda)$ where $\omega = (\omega_1, \omega_2, \cdots)$ is a countable sequence of non-negative numbers ω_l that add up to unity, and $\lambda = (\lambda_1, \lambda_2, \cdots)$ is a countable sequence of $\lambda_l \in \Lambda_g$. How do we construct a prior distribution on such an infinite dimensional quantity θ?

To start with we could specify a prior distribution on $\lambda = (\lambda_1, \lambda_2, \cdots)$ as

$$\lambda_l \overset{\text{IID}}{\sim} \pi_g \tag{2}$$

for some suitable pdf π_g on Λ_g. Usually, π_g would be a good candidate for a prior pdf for the basic model $Y_j \overset{\text{IID}}{\sim} g(y_j|\lambda)$, $\lambda \sim \pi_g$. One way to specify a prior distribution on $\omega = (\omega_1, \omega_2, \cdots)$ is to set:

$$\omega_1 = \beta_1, \omega_l = \beta_l \prod_{j<l} (1 - \beta_j), l = 2, 3, \cdots, \text{ where } \beta_1, \beta_2, \cdots \overset{\text{IID}}{\sim} Be(1, a) \tag{3}$$

for some $a > 0$. With such a definition one has $\sum_{l=1}^{k} \omega_l = 1 - \prod_{l=1}^{k}(1 - \beta_l)$ which tends to 1 almost surely as $k \to \infty$.

Notice that what we have defined above is a random $\theta = (\omega, \lambda)$ with realizations in the desired parameter space (i.e., $\omega_l \geq 0$, $\sum_{l=1}^{\infty} \omega_l = 1$ and $\lambda_l \in \Lambda_g$). However, we cannot write down functional formula for a prior pdf $\pi(\theta)$, simply because such pdfs do not exist in infinite dimensional spaces [they may exist, strictly speaking, but usually don't do in a useful way.] But we can still talk about the posterior distribution of θ given observed data $Y = y$. Although this can be done in a formal way invoking measure theory, for practical purposes it would be more useful to look at a different equivalent representation.

Dirichlet process

Let $\theta = (\omega, \lambda)$ be a random element defined as in the previous section. The elements of θ could be used to define a discrete probability measure P on Λ_g with atoms $\lambda_1, \lambda_2, \cdots$ bearing weights $\omega_1, \omega_2, \cdots$. That is,

$$P(\cdot) = \sum_{l=1}^{\infty} \omega_l \delta_{\theta_l}(\cdot),$$

or in other words, to any subset $B \subset \Lambda_g$ the measure P assigns probability $P(B) = \sum_{l=1}^{\infty} \omega_l I(\theta_l \in B)$.

With ω and λ as defined in the previous section, the probability measure P defines a random discrete measure on Λ_g and is said to have a Dirichlet process distribution with precision a and base measure π_g, denoted $\mathrm{DP}(a, \pi_g)$.

Note that for any θ we can define a unique discrete measure P as above. The reverse map is also unique but up to permutations of the indices l (i.e., up to label switching). So properties of the random θ can be studied through properties of the random P. A key property of P, which also justifies the name Dirichlet process is as follows.

Theorem 1. *$P \sim \mathrm{DP}(a, \pi_g)$ if and only if for any finite partition $B_1, \cdots, B_k$ of Λ_g (i.e., $B_i \cap B_j = \emptyset$ if $i \neq j$ and $B_1 \cup \cdots \cup B_k = \Lambda_g$), the random probability vector $(P(B_1), \cdots, P(B_k))$ is distributed according to $Dir(a\pi_g(B_1), \cdots, a\pi_g(B_k))$ where $\pi_g(B_l)$ denotes the probability π_g assigns to the set B_l.*

This result in fact uniquely characterizes a $P \sim \mathrm{DP}(a, \pi_g)$ and was actually used to give the original definition of a Dirichlet process (by Ferguson, 1973). But the constructive definition we saw in the previous section is more direct. A proof is beyond our scope, but you can read Ghosh and Ramamoorthi (2003). An equally important result about Dirichlet process is the following conjugacy property.

Theorem 2. *If $P \sim \mathrm{DP}(a, \pi_g)$ and $Z_1, \cdots, Z_n \overset{IID}{\sim} P$ then $P|(Z_1 = z_1, \cdots, Z_n = z_n) \sim \mathrm{DP}(a + n, \frac{a}{a+n}\pi_g + \frac{n}{a+n}F_{z,n})$ where $F_{z,n} = \frac{1}{n}\sum_{i=1}^{n} \delta_{z_i}$ is the discrete distribution on Λ_g with atoms $z_1, \cdots, z_n$ all bearing the same weight $1/n$.*

Proof. It suffices to prove for $n = 1$. Larger values of n can be handled by induction. So assume $P \sim \mathrm{DP}(a, \pi_g)$ and suppose $Z_1 \sim P$. Given $Z_1 = z_1$, we want to show $P \sim \mathrm{DP}(a+1, \frac{a}{a+1}\pi_g + \frac{1}{a+1}\delta_{z_1})$. By the earlier theorem, it suffices to show that for any partition $B_1, \cdots, B_k$ of Λ_g, the conditional distribution of $(P(B_1), \cdots, P(B_k))$ given $Z_1 = z_1$ is $Dir(a\pi_g(B_1) + I(z_1 \in B_1), \cdots, a\pi_g(B_k) + I(z_1 \in B_k))$

Given a partition $B_1, \cdots, B_k$ and an $z_1 \in \Lambda_g$, there is a partition element B_l containing z_1. Without loss of generality we can assume this element is B_1. Let $U \subset B_1$ be any set containing z_1 and let π_U denote the conditional distribution of $(P(B_1), \cdots, P(B_k))$ given $Z_1 \in U$. It suffices to show that

$$\lim_{U \downarrow \{z_1\}} \pi_U = Dir(a\pi_g(B_1) + 1, a\pi_g(B_1), \cdots, a\pi_g(B_k)). \tag{4}$$

Now, if $U \subset B_1$ then we can split B_1 into $B_{10} = U$ and $B_{11} = B_1 \setminus U$ to form a new partition $B_{10}, B_{11}, B_2, \cdots, B_k$. By simple conjugacy of multinomial and Dirichlet distributions we have the conditional distribution of $(P(B_{10}), P(B_{11}), P(B_2), \cdots, P(B_k))$ given $Z_1 \in U$ to be $Dir(a\pi_g(B_{10}) + 1, a\pi_g(B_{11}), a\pi_g(B_2), \cdots, a\pi_g(B_k))$. Therefore, becuase $P(B_1) = P(B_{10}) + P(B_{11})$, the conditional distribution of $(P(B_1), \cdots, P(B_k))$ given $Z_1 \in U$ is $\pi_U = Dir(a\pi_g(B_1) + 1, a\pi_g(B_1), \cdots, a\pi_g(B_k))$. This leads to (4). □

A third result, which combines the above two, has much practical implications.

Theorem 3. *If* $P \sim \mathrm{DP}(a, \pi_g)$ *and* $Z_1, Z_2, \cdots \overset{IID}{\sim} P$ *then marginally,* $Z_1 \sim \pi_g$ *and* $Z_{n+1}|(Z_1 = z_1, \cdots, Z_n = z_n) \sim \frac{a}{a+n}\pi_g + \frac{n}{a+n}F_{z,n}$.

Proof. Again suffices to prove for $n = 1$, because the other cases follow from this once you invoke Theorem 2. For Z_1 and any set $B \subset \Lambda_g$,

$$\Pr(Z_1 \in B) = \mathbb{E}[\Pr(Z_1 \in B|P)] = \mathbb{E}[P(B)] = \pi_g(B)$$

because, from Theorem 1, $P(B) \sim Be(a\pi_g(B), a\{1 - \pi_g(B)\})$. So $Z_1 \sim \pi_g$. □

Back to infinite mixture model

Now consider the mixture model $Y_j \overset{\text{IID}}{\sim} f(y_j|\theta) = \sum_{l=1}^{\infty} \omega_l g(y_j|\lambda_l)$ with $\theta = (\omega, \lambda)$ as defined in (2)-(3). As in the finite mixture case, introduce latent variables $\zeta_1, \cdots, \zeta_n$ with ζ_j giving λ_l of the component l that y_j belongs to. The infinite mixture model is equivalent to:

$$\begin{aligned} Y_j &\overset{\text{IND}}{\sim} g(y_j|\zeta_j) \\ \zeta_1, \cdots, \zeta_n &\overset{\text{IID}}{\sim} P \\ P &\sim \mathrm{DP}(a, \pi_g) \end{aligned}$$

where P is the discrete measure associated with $\theta = (\omega, \lambda)$. Now, we can use Theorem 3 and rewrite the above as:

$$\begin{aligned} f(y \mid \zeta) &= \prod_{j=1}^{n} g(y_j|\zeta_j) \\ \pi(\zeta) &= \prod_{j=1}^{n} \left[\frac{a}{a+j-1}\pi_g(\zeta_j) + \frac{1}{a+j-1}\sum_{l<j} \delta_{\zeta_l}(\zeta_j) \right] \end{aligned}$$

This representation deals with only a n-dimensional parameter $\zeta = (\zeta_1, \cdots, \zeta_n)$, the difficulties of infinity are all gone! The prior $\pi(\zeta)$ above is actually symmetric in $\zeta_1, \cdots, \zeta_n$ (label switching does not alter the prior probability of the whole configuration). In fact for any $i \in \{1, \cdots, n\}$ we have

$$\pi(\zeta_i|\zeta_{-i}) = \frac{a}{a+n-1}\pi_g(\zeta_i) + \frac{1}{a+n-1}\sum_{l \neq i} \delta_{\zeta_l}(\zeta_i).$$

This is again a consequence of Theorem 3 because Z_i's are exchangeable in the statement of the theorem. In above ζ_{-i} denotes the vector of ζ_l, $l = 1, \cdots, n$ without including ζ_i.

Conditional posterior and Gibbs sampling. The above conditional prior formula immediately gives

$$\begin{aligned}
\pi(\zeta_i|\zeta_{-i}, y) &= \text{const} \times \prod_{j=1}^{n} g(y_j|\zeta_j) \times \left[\frac{a}{a+n-1}\pi_g(\zeta_i) + \frac{1}{a+n-1}\sum_{l\neq i}\delta_{\zeta_l}(\zeta_i)\right] \\
&= \text{const} \times g(y_i|\zeta_i) \times \left[\frac{a}{a+n-1}\pi_g(\zeta_i) + \frac{1}{a+n-1}\sum_{l\neq i}\delta_{\zeta_l}(\zeta_i)\right] \\
&= \text{const} \times \left[\frac{a}{a+n-1}g(y_i|\zeta_i)\pi_g(\zeta_i) + \frac{1}{a+n-1}\sum_{l\neq i} g(y_i|\zeta_l)\delta_{\zeta_l}(\zeta_i)\right] \\
&= q_{i0}\pi_g(\zeta_i|y_i) + \sum_{l\neq i} q_{il}\delta_{\zeta_l}(\zeta_i),
\end{aligned}$$

with

$$\pi_g(\zeta_i|y_i) = \frac{g(y_i|\zeta_i)\pi_g(\zeta_i)}{\int_{\Lambda_g} g(y_i|z)\pi_g(z)dz}, \quad q_{i0} = ca\int_{\Lambda_g} g(y_i|z)\pi_g(z)dz, \quad q_{il} = cg(y_i|\zeta_l), l \neq i$$

where c is a constant that makes $q_{i0}, q_{i1}, \cdots, q_{i,i-1}, q_{i,i+1}, \cdots, q_{i,n}$ sum up to unity.

Note that $\pi(\zeta_i|\zeta_{-i}, y)$ is a mixed distribution, with $n-1$ atoms ζ_l, $l \neq i$ carrying weights q_{il} and the rest of the weight q_{i0} is distributed according to the pdf $\pi_g(\zeta_i|y_i)$. To sample from this distribution, we first draw an index l from $\{0, 1, \cdots, i-1, i+1, \cdots, n\}$ with probabilities proportional to $(q_{i0}, q_{i1}, \cdots, q_{i,i-1}, q_{i,i+1}, \cdots, q_{i,n})$. If $l \neq 0$ then we set $\zeta_i = \zeta_l$. Otherwise, we draw ζ_i from $\pi_g(\zeta_i|y_i)$.

With the conditional posterior pdfs in place, we can now do a Gibbs sampling to generate samples of ζ from the joint posterior $\pi(\zeta|y)$. We start with an initial configuration $\zeta^0 = (\zeta_1^0, \cdots, \zeta_n^0)$ and then for $t = 1, 2, \cdots$ cycle through $i = 1, \cdots, n$ to update ζ_i^{t-1} to ζ_i^t sampled from $\pi(\zeta_i|\zeta_{-i} = (\zeta_1^t, \cdots, \zeta_{i-1}^t, \zeta_{i+1}^{t-1}, \cdots, \zeta_n^{t-1}), y)$. B initial iterations are discarded and then every r-th one is stored to form a sample $\zeta^{(m)}$, $m = 1, \cdots, M$.

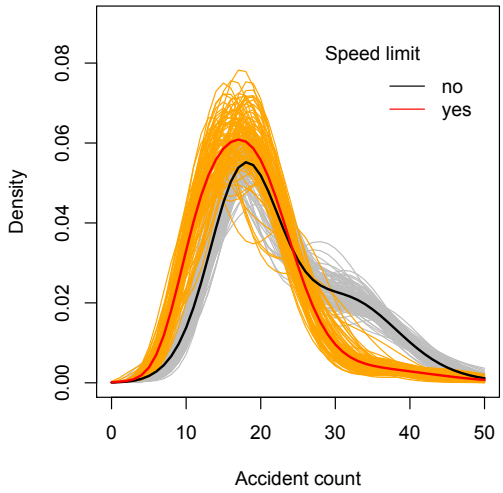
Example (Traffic data analysis)**.** For each group (`limit = yes` and `limit = no`) I used the above DP mixture model with: $g(y_i|\zeta_i) = Poi(y_i|\zeta_i)$, $a = 1$ and $\pi_g = Ga(b_1 = 2, b_2 = 1/10)$. For this choice of conjugate pair of $g(y_i|\zeta_i)$ and $\pi_g(\zeta_i)$, we get

$$\pi_g(\zeta_i|y_i) = Ga(b_1 + y_i, b_2 + 1), \quad \int_0^\infty g(y_i|z)\pi_g(z)dz = NBin\left(y_i|b_1, \frac{b_2}{b_2+1}\right)$$

the last expression denoting the negative-binomial density at y_i with size b_1 (number of successes desired) and probability $b_2/(1+b_2)$ (success probability in each trial).

For each data separately, I ran a Gibbs sampler for 2000 iterations, discarded the first 1000, and saved every 10th iteration of the remaining chain, giving a posterior sample of size $M = 100$.

To sample a future observable Y^* from the model, first note that we can write $Y^* \sim g(y^*|\zeta^*)$ with $\zeta^*|\zeta \sim \frac{a}{a+n}\pi_g(\zeta^*) + \frac{1}{a+n}\sum_{l=1}^{n}\delta_{\zeta_l}(\zeta^*)$. And so, for each sampled $\zeta^{(m)}$ I first generated a $l^{*(m)} \sim \mathsf{Multinomial}(n+1, (\frac{1}{a+n}, \cdots, \frac{1}{a+n}, \frac{a}{a+n}))$ and then sampled $y^{*(m)} \sim Poi(\zeta^{(m)}_{l^{*(m)}})$ if $l^{*(m)} \leq n$, otherwise sampled $y^{*(m)} \sim NBin(b_1, b_2/(b_2+1))$. The resulting $(y^{*(1)}, \cdots, y^{*(M)})$ represent samples of a future observable Y^* from the model. The figure below shows $f^*(y^*|\zeta)$ for $\zeta = (\zeta_1, \cdots, \zeta_n)$ drawn from the posterior for each of the two datasets.



Then I compared the posterior predictive draws for the two data (speed limit yes/no) and calculated the posterior probability that a speed limit day has less accidents than a no speed limit day. This probability turns out to be 68%.

References

Ferguson, T. (1973). Bayesian analysis of some nonparametric problems. *Annals of Statistics 1*, 209–230.

Ghosh, J. K. and R. V. Ramamoorthi (2003). *Bayesian Nonparametrics*. Springer-Verlag.