NONPARAMETRIC DENSITY ESTIMATION USING A MIXTURE OF ORDER-STATISTIC DISTRIBUTIONS

MARK FISHER

Abstract. This paper presents a Bayesian nonparametric model for predictive density estimation that incorporates order-statistic distributions into a Dirichlet Process Mixture (DPM) model. In particular, the kernel is the density of the $j$-th order statistic given a sample size of $k$ from a pre-specified continuous distribution. By requiring the prior distribution for $j$ to be uniform conditional on $k$, the prior predictive distribution is made to equal the pre-specified distribution. The model is completed by specifying a prior distribution for $k$, which plays the role of a precision parameter, and a prior distribution for the concentration parameter, which affects amount of clustering through the distribution of the stick-breaking weights. The model presented in this paper can be interpreted as a more flexible version of that in Petrone (1999) “Bayesian density estimation using Bernstein polynomials.”
1. INTRODUCTION

This paper presents a Bayesian nonparametric model for predictive density estimation that incorporates order-statistic distributions into a Dirichlet Process Mixture (DPM) model. In particular, the kernel is the density of the $j$-th order statistic given a sample size of $k$ from a pre-specified continuous distribution. The prior distribution for $j$ is uniform conditional on $k$, and consequently the pre-specified distribution is the prior predictive distribution. The model is completed by specifying a prior distribution for $k$. (A prior for the concentration parameter that affects the distribution of the stick-breaking weights may be specified as well.)

This paper is about density estimation using a mixture of beta distributions. (A simple extension allows for two-dimensional density estimation.) Here are the main features: (1) the approach to inference is Bayesian; (2) the parameters of the beta distributions are restricted to the natural numbers; (3) the potential number of mixture components is unbounded; and (4) the prior predictive density is part of the specification. As illustrations, the model is applied to a number a standard data sets in one- and two-dimensions.

In addition, I show how to apply the model to latent variables via what I call indirect density estimation. (In this context I introduce the distinction between generic and specific cases.) To illustrate this technique, I apply this estimation technique to compute the density of unobserved success rates that underly the observations from binomial experiments and/or units. The results are compared with those generated by an alternative model that has appeared in the literature.

The model is related to the Bernstein polynomial model introduced by Petrone (1999a). To make the comparison, let the prior predictive distribution be the uniform distribution over the unit interval. Petrone’s model mixes over Bernstein polynomials of different degrees, where each Bernstein polynomial is comprised of a complete set of Bernstein basis polynomials. By contrast, the model presented here takes a multi-resolution approach, mixing directly over the basis polynomials themselves of every degree. The potential number of mixture components is unbounded.

The model is first applied to observed data, and then applied to latent variables via indirect density estimation. In this latter setting the model becomes a DPM model for the prior. In the context of binomial observations and latent success rates, this prior is compared with the Dirichlet process (DP) prior found in the literature. (The DP prior can be viewed as a special case of the DPM prior.)

Related literature. For asymptotic properties of random Bernstein polynomials, see Petrone (1999b) and Petrone and Wasserman (2002). For a multivariate extension of Petrone’s model, see Zhao et al. (2013). Other related literature includes Kottas (2006), Trippa et al. (2011), and Quintana et al. (2009). Liu (1996) presents a related model in which the latent success rates for binomial observations have a Dirichlet Process (DP) prior.

A closely-related paper is Canale and Dunson (2016) which presents a multi-scale approach using Bernstein polynomials. See Appendix C for a comparison of their model with what is presented here.

Outline. Section 2 presents the model. Section 3 describes a Markov chain Monte Carlo (MCMC) sampler. Section 4 presents empirical results. Section 5 extends the model to the latent-variable case. Section 6 presents additional empirical results.

---

1Petrone’s model is discussed briefly in Section B.

2See also Canale (2017).
2. The model

Given \( n \) observations \( x_{1:n} = (x_1, \ldots, x_n) \), the object of interest is the predictive distribution for the next observation:

\[
p(x_{n+1}|x_{1:n}).
\]

Assume

\[
x_i \overset{iid}{\sim} p(\cdot|\psi) \quad \text{for } i = 1, 2, \ldots,
\]

where \( \psi \) is an unobserved parameter. Then the predictive distribution can be expressed as

\[
p(x_{n+1}|x_{1:n}) = \int p(x_{n+1}|\psi) p(\psi|x_{1:n}) \, d\psi,
\]

where \( p(\psi|x_{1:n}) \) is the posterior distribution for \( \psi \), which can be expressed in terms of the likelihood \( p(x_{1:n}|\psi) = \prod_{i=1}^{n} p(x_i|\psi) \) and the prior distribution \( p(\psi) \):

\[
p(\psi|x_{1:n}) \propto p(x_{1:n}|\psi) p(\psi).
\]

The model is completed by specifying \( p(x_i|\psi) \) and \( p(\psi) \).

The predictive distribution \( p(x_{n+1}|x_{1:n}) \) summarizes what is known about \( x_{n+1} \) given the observations \( x_{1:n} \). The parameter \( \psi \) is a conduit through which information flows from \( x_{1:n} \) to \( x_{n+1} \). Additional insight into the nature of the predictive distribution is provided in Appendix A where it is compared and contrasted with a different object of interest.

**Specification.** Let \( f(x_i|\theta_c) \) denote a probability density function for \( x_i \in \mathbb{R} \) conditional on a parameter \( \theta_c \). The density \( f(\cdot|\theta_c) \) is called the kernel. Let

\[
p(x_i|\psi) = \sum_{c=1}^{\infty} w_c f(x_i|\theta_c),
\]

where \( \psi = (w, \theta) \) and \( w = (w_1, w_2, \ldots) \) is an infinite collection of nonnegative mixture weights that sum to one and \( \theta = (\theta_1, \theta_2, \ldots) \) is a corresponding collection of mixture-component parameters. The structure of the prior for \( \psi \) is

\[
p(\psi) = p(w) p(\theta) = p(w) \prod_{c=1}^{\infty} p(\theta_c),
\]

where \( p(\theta_c) \) is called the base distribution.

**Prior for the mixture weights.** The prior for \( w \) is given by

\[
w \sim \text{Stick}(\alpha),
\]

where \( \text{Stick}(\alpha) \) denotes the stick-breaking distribution given by\(^3\, 4\)

\[
w_c = v_c \prod_{\ell=1}^{c-1} (1 - v_\ell) \quad \text{where } v_c \overset{iid}{\sim} \text{Beta}(1, \alpha).
\]

---

\(^3\)The specification adopted here is equivalent to a Dirichlet Process Mixture (DPM) model. The model can easily accommodate other stick-breaking priors. See Ishwaran and James (2001) for a general treatment of stick-breaking priors.

\(^4\)Start with a stick of length one. Break off the fraction \( v_1 \) leaving a stick of length \( 1 - v_1 \). Then break off the fraction \( v_2 \) of the remaining stick leaving a stick of length \( (1 - v_1)(1 - v_2) \). Continue in this manner. Alternative stick-breaking distributions can be constructed by changing the distribution for \( v_c \).
The parameter $\alpha$ is called the *concentration parameter*; it controls the rate at which the weights decline on average. In particular, the weights decline geometrically in expectation:

$$E[w_c|\alpha] = \alpha^{-1} (1 + \alpha)^{-c}. \quad (2.9)$$

Note $E[w_1|\alpha] = 1/(1 + \alpha)$ and $E[\sum_{c=n+1}^{\infty} w_c|\alpha] = (\alpha/(1 + \alpha))^{n}$.

If $\alpha$ is small, then the first few weights will dominate and only a small number of mixture components will be consequential; by contrast if $\alpha$ is large, then a large number of mixture components will consequential.$^5$

**Prior for the concentration parameter.** The concentration parameter plays an important role in determining the flexibility of the prior for a given finite sample size $n$. As such, it may be important to allow the data to help determine its magnitude.

One such prior distribution for the concentration parameter is given by

$$p(\alpha) = \text{Log-Logistic}(\alpha|1,1) = \frac{1}{(1+\alpha)^2}. \quad (2.10)$$

This distribution does not have a finite mean; its median equals one.

**The kernel and the base distribution.** The model described thus far is standard. The novelty lies in the combination of the kernel and the base distribution. Let $Q(x)$ denote the cumulative distribution function (CDF) for a continuous random variable defined on the real line. Let $q(x) = Q'(x)$ denote the associated density function. The kernel is given by$^6$

$$f(x|\theta_c) = f(x_j|\theta_c) = \text{Beta}(Q(x_j)|j_c, k_c - j_c + 1) q(x_j), \quad (2.11)$$

where $\theta_c = (j_c, k_c)$. One may verify that $f(\cdot|j,k)$ is the density for the $j$-th order statistic given a sample size of $k$ from the distribution $Q$. Thus $p(x_i|\psi)$ is an infinite-order mixture of order-statistic densities. The base distribution is given by $p(\theta_c) = p(j_c, k_c) = p(j_c|k_c) p(k_c)$, where

$$k_c \sim P_k \quad (2.12a)$$

$$j_c|k_c \sim \text{Uniform}\{1, \ldots, k_c\}, \quad (2.12b)$$

for some distribution $P_k$ over the positive integers.

Given the prior independence of $w$ and $\theta$, the prior predictive distribution for $x_i$ is

$$p(x_i) = \int f(x_i|\theta_c) p(\theta_c) d\theta_c = q(x_i), \quad (2.13)$$

where the last equality follows from the adding-up property of order-statistic distributions:

$$\sum_{j_c=1}^{k_c} f(x_j|j_c, k_c) p(j_c|k_c) = \frac{1}{k_c} \sum_{j_c=1}^{k_c} f(x_j|j_c, k_c) = q(x_i). \quad (2.14)$$

The adding-up property says that an equally-weighted mixture of all $k_c$ of the order distributions equals $Q$. This may be seen as follows. Suppose one makes $k_c$ independent draws

---

$^5$In order to understand the nature of the “concentration” that lends $\alpha$ its name, define the random probability distribution $G = \sum_{c=1}^{\infty} w_c \delta_{\theta_c}$, where $\delta_{\theta_c}$ is a point mass located at $\theta_c$. The randomness of $G$ follows from $w \sim \text{Stick}(\alpha)$ and $\theta_c \overset{iid}{\sim} H$, where $H$ is the base distribution. In other words, $G$ is distributed according to a Dirichlet Process: $G \sim \text{DP}(\alpha, H)$. As a consequence, the mean of $G$ is $H$: $E[G] = H$. At one extreme for $\alpha$, $\lim_{\alpha \to 0} G = \delta_{\theta_1}$ (where $\theta_1 \sim H$), which maximizes the variation of $G$ around $H$. At the other extreme, $\lim_{\alpha \to \infty} G \equiv H$. Thus, $\alpha$ controls the concentration of $G$ around $H$.

$^6$Beta$x|a,b) = x^{a-1} (1 - x)^{b-1}/B(a,b)$, where $B(a,b)$ is the beta function.
from \(Q\) and sorts them from smallest to largest. If one then chooses one of the sorted draws at random, the effect of the sorting is bypassed and the distribution of the choice is \(Q\) itself.

In summary, the model is completely determined by specifying the prior for \(\alpha\), the predictive distribution \(q(x_i)\), and the prior distribution for \(k_c\).

Both \(p(\alpha)\) and \(p(k_c)\) affect the variation of \(p(x_i|\psi)\) around \(q(x_i)\). The concentration parameter works through its effect on the weights, while \(k_c\) works through its effect on the kernel (via the base distribution\(^7\)). Variation can be completely removed by either channel. Focusing on the concentration parameter, in the limit as \(\alpha \to \infty\), no finite collection of weights dominates and consequently, regardless of the distribution for \(k_c\),\(^8\)

\[
\lim_{\alpha \to \infty} p(x_i|\psi) = q(x_i). \quad (2.15)
\]

On the other hand, if \(\text{Pr}[k_c = 1] = 1\), then every mixture component equals the predictive distribution regardless of the weights:

\[
p(x_i|\psi) = \sum_{c=1}^{\infty} w_c \text{Uniform}(Q(x_i)|0,1) q(x_i) = q(x_i). \quad (2.16)
\]

Going the other direction, a prior that puts most of its weight on large values of \(k_c\) will enhance the variation. In order to examine the extreme case, it is convenient to change variables from \((j_c, k_c)\) to \((\phi_c, k_c)\) where \(\phi_c = Q^{-1}(j_c/k_c)\). Note that \(j_c/k_c \approx \text{Uniform}(0,1)\) as \(k_c \to \infty\) and therefore \(\phi_c \approx Q\) as \(k_c \to \infty\). For fixed \(\phi_c\), the kernel collapses to a point mass located at \(\phi_c\):

\[
\lim_{k_c \to \infty} f(x_i|j_c = Q(\phi_c), k_c, k_c) = \delta(x_i - \phi_c), \quad (2.17)
\]

where \(\delta(\cdot)\) denotes the Dirac delta function. Thus, in the limit the DPM model becomes the Dirichlet Process (DP) model where \(\theta_c = \phi_c\) with base distribution \(Q\).\(^9\)

In the empirical section I will adopt the following prior for \(k_c\):

\[
k_c - 1 \sim \text{Geometric}(\xi), \quad (2.18)
\]

where \(\xi \in (0,1)\). Given this distribution, \(p(k_c = 1) = \xi\) and \(E[k_c] = 1/\xi\).

**Features of the prior.** It may be useful to understand some features of the prior. The prior encodes both a willingness to learn (via dependence) and open-mindedness (via flexibility).

**Dependence.** Dependence in the prior among the elements of \(x_{1:n}\) is the key to the ability to learn about \(x_{n+1}\) from \(x_{1:n}\). Without it there is no learning. We now examine how this dependence is structured within the prior by focusing on the joint prior distribution for \((x_1, x_2)\):

\[
p(x_1, x_2) = q(x_1) q(x_2) c(Q(x_1), Q(x_2)), \quad (2.19)
\]

where \(c(u_1, u_2)\) is a copula density for \((u_1, u_2) \in [0,1]^2\).

In order to derive the copula, first note that \(\sum_{c=1}^{\infty} w_c^2\) is the probability that \(x_1\) and \(x_2\) share the same component and recall \(E[\sum_{c=1}^{\infty} w_c^2|\alpha] = 1/(1+\alpha)\). Moreover, given the prior for \(\alpha\) [see (2.10)], the unconditional probability that \(x_1\) and \(x_2\) share the same component is

\[
\int_0^\infty \frac{1}{1+\alpha} p(\alpha) \, d\alpha = \int_0^\infty \frac{1}{(1+\alpha)^3} \, d\alpha = \frac{1}{2}. \quad (2.20)
\]

\(^7\)This formulation of the DPM allows one to change the base distribution without changing the predictive distribution.

\(^8\)As noted in Footnote 5, \(\lim_{\alpha \to \infty} G = H\).

\(^9\)This limiting case is revisited in Appendix E in the context of indirect density estimation.
Therefore, after integrating out \(w\), \(\alpha\), and \(j_c\), we have

\[
c(u_1, u_2) = \sum_{k_c=1}^{\infty} p(k_c) c(u_1, u_2|k_c),
\]

(2.21)

where \(c(u_1, u_2|k_c)\) is an order-statistic-based copula density that depends on \(k_c\).\(^{10}\) In particular,

\[
c(u_1, u_2|k_c) := \frac{1}{2} + \frac{1}{2} \sum_{j=1}^{k_c} \frac{1}{k_c} \prod_{i=1}^{2} \text{Beta}(u_i|j_c, k_c - j_c + 1)
\]

\[
= \frac{1}{2} + \frac{k_c}{2} \left( (1 - u_1)(1 - u_2) \right)^{k_c-1} \text{2F1}\left( 1 - k_c, 1 - k_c; 1; \frac{u_1 u_2}{(1 - u_1)(1 - u_2)} \right),
\]

(2.22)

where \(\text{2F1}\) is the hypergeometric function. Note \(c(u_1, u_2|k_c = 1) = 1\). For \(k_c > 1\), \(c(u_1, u_2|k_c)\) provides positive dependence between \(u_1\) and \(u_2\); the strength of the dependence increases with \(k_c\). Note that \(p(x_1, x_2|k_c) = q(x_1) q(x_2) c(Q(x_1), Q(x_2)|k_c)\). See Figure 1 for a plot of \(p(x_1, x_2|k_c = 10)\) assuming \(Q = \text{Uniform}(0, 1)\) and \(Q = \mathcal{N}(0, 1)\).

\(^{10}\text{See Baker (2008) for a treatment of copulas generated via order-statistics.}\)
prior predictive can be approximated by
\[ p(x_i) \approx \frac{1}{R} \sum_{r=1}^{R} p(x_i|\psi^{(r)}). \] (2.23)

For a subset of the draws, plot \( p(x_i|\psi^{(r)}) \) to examine the amount and sort of variation. In Figure 2, we display ten draws of the density \( p(x_i|\psi) \) given \( Q(x) = x \).

**Multi-dimensional predictive density.** The extension of the model to \( d \)-dimensional observations is straightforward. Let \( x_i = (x_{i1}, \ldots, x_{id}) \), \( \theta_c = (\theta_{c1}, \ldots, \theta_{cd}) \), where \( \theta_{c\ell} = (j_{c\ell}, k_{c\ell}) \) for \( \ell = 1, \ldots, d \). Define
\[
Q(x_i) := (Q_1(x_{i1}), \ldots, Q_d(x_{id}))
\]
\[
q(x_i) := \prod_{\ell=1}^{d} q_{\ell}(x_{i\ell}),
\]
(2.24a)
(2.24b)
where \( q_{\ell}(x_{i\ell}) \) is the PDF for the marginal prior predictive distribution for \( x_{i\ell} \). Let the kernel be given by
\[
f(x_i|\theta_c) = \prod_{\ell=1}^{d} \text{Beta}(Q_{\ell}(x_{i\ell})|j_{c\ell}, k_{c\ell} - j_{c\ell} + 1) \cdot q_{\ell}(x_{i\ell}).
\] (2.25)

Note the local independence in the kernel. A model with local dependence given \( d = 2 \) is described in Appendix D.

Let the base distribution be given by \( p(\theta_c) = \prod_{\ell=1}^{d} p(\theta_{c\ell}) \), where \( p(\theta_{c\ell}) = p(k_{c\ell})/k_{c\ell} \) and where \( k_{c\ell} \sim P_{k_{c\ell}} \). Consequently,
\[ p(x_i) = q(x_i).
\] (2.26)
The priors for $w$ and $\alpha$ are unchanged.

Simple adaptations of the sampling scheme described in Section 3 allow one to make draws of $\psi$ in the $d$-dimensional case.

3. MCMC SAMPLER

The model is equivalent to a Dirichlet Process Mixture (DPM) model. As such it may be computed via any number of existing algorithms. For example, it is possible to use Algorithm 2 in Neal (2000) even though the base distribution is not conjugate relative to the kernel (see below).

However, the simplest algorithm to describe and implement is the blocked Gibbs sampler described in Gelman et al. (2014, pp. 552–553). This sampler relies on approximating $p(x_i|\psi)$ with a finite sum: Choose $m$ large enough to make $(\alpha/(1 + \alpha))^m$ close enough to zero and set $v_m = 1$.

This sampler uses the classification variables $z_1:n = (z_1, \ldots, z_n)$, where $z_i = c$ signifies $x_i$ is assigned to cluster $c$. The Gibbs sampling scheme involves cycling through the following full conditional posterior distributions:

\begin{align*}
  p(z_1:n|x_1:n, w, \theta, \alpha) &= \prod_{i=1}^{n} p(z_i|x_i, w, \theta) \\
  p(w|x_1:n, z_1:n, \theta, \alpha) &= p(w|z_1:n, \alpha) \\
  p(\theta|x_1:n, z_1:n, w, \alpha) &= \prod_{c=1}^{m} p(\theta_c|x^c) \\
  p(\alpha|x_1:n, z_1:n, w, \theta) &= p(\alpha|z_1:n),
\end{align*}

where $x^c$ is the collection of observations for which $z_i = c$.

The conditional distribution for $z_i$ is characterized by

\begin{equation}
  p(z_i = c|x_1:n, w, \theta) \propto w_c f(x_i|\theta_c),
\end{equation}

for $c = 1, \ldots, m$. Let $n_c$ denote the multiplicity of $c$ in $z_1:n$ (i.e., the number of times $c$ occurs in $z_1:n$). Note $\sum_{c=1}^{m} n_c = n$. The weights $w$ can be updated by updating the stick-breaking weights $v$ via

\begin{equation}
  v_c|z_1:n \sim \text{Beta}(1 + n_c, \alpha + \sum_{c'=c+1}^{m} n_{c'})
\end{equation}

for $c = 1, \ldots, m - 1$. The cluster parameters, $\theta_c|x^c$, are updated as in a finite mixture model (using the Metropolis–Hastings scheme, for example), with the parameters for the unoccupied components (for which $n_c = 0$) sampled directly from the prior $p(\theta_c)$.

Regarding the concentration parameter, note that

\begin{equation}
  p(\alpha|z_1:n) \propto p(z_1:n|\alpha) p(\alpha) \propto \frac{\alpha^h \Gamma(\alpha)}{\Gamma(n + \alpha)} p(\alpha),
\end{equation}

where $h$ is the number of occupied clusters (i.e., clusters for which $n_c > 0$). Draws from $p(\alpha|z_1:n)$ may be made using the Metropolis–Hastings scheme.\footnote{Let $\zeta = \alpha/(1 + \alpha)$. Then $p(\zeta) = 1$ and
\[\log(p(\zeta|z_1:n)) = h \log\left( \frac{\zeta}{1 - \zeta} \right) + \log\Gamma\left( \frac{\zeta}{1 - \zeta} \right) - \log\Gamma\left( \frac{\zeta}{1 - \zeta} + n \right).\]}

\textbf{DENSITY ESTIMATION USING ORDER-STATISTIC DISTRIBUTIONS}
Given draws \( \{ \psi^{(r)} \}_{r=1}^R \) from \( p(\psi|x_{1:n}) \), the posterior predictive distribution can be approximated via

\[
p(x_{n+1}|x_{1:n}) \approx \frac{1}{R} \sum_{r=1}^R p(x_{n+1}|\psi^{(r)}) = \frac{1}{R} \sum_{r=1}^R \sum_{c=1}^m w_c^{(r)} f(x_{n+1}|\theta_c^{(r)}).
\] (3.5)

**A smoother approximation.** It is possible to obtain a lower-variance approximation to the generic distribution by integrating out the mixture weights and the cluster coefficients for the unoccupied clusters. The indices for the occupied clusters are given by \( \mathcal{C} = \{ c : c \in z_{1:n} \} \). Integrating out \( w \) given the classifications produces

\[
E[w_c|z_{1:n}, \alpha] = \frac{n_c}{n + \alpha} \quad \text{for } c \in \mathcal{C}
\] (3.6a)

\[
E\left[ \sum_{c \notin \mathcal{C}} w_c | z_{1:n}, \alpha \right] = \frac{\alpha}{n + \alpha}.
\] (3.6b)

In addition, for each \( c \notin \mathcal{C} \) we can use \( p(\theta_c) \) to integrate out \( \theta_c \), thereby replacing \( f(x_{n+1}|\theta_c) \) with \( q(x_{n+1}) \). Consequently, the generic distribution can be expressed conditionally as\(^{12}\)

\[
p(x_{n+1}|z_{1:n}, \{ \theta_c \}_{c \in \mathcal{C}}, \alpha) = \sum_{c \in \mathcal{C}} \frac{n_c}{n + \alpha} f(x_{n+1}|\theta_c) + \frac{\alpha}{n + \alpha} q(x_{n+1}).
\] (3.7)

The approximation is given by

\[
p(x_{n+1}|x_{1:n}) \approx \frac{1}{R} \sum_{r=1}^R p(x_{n+1}|z_{1:n}, \{ \theta_c^{(r)} \}_{c \in \mathcal{C}(r)}, \alpha^{(r)})
\]

\[
= \frac{1}{R} \sum_{r=1}^R \left( \sum_{c \in \mathcal{C}(r)} \frac{n_c^{(r)}}{n + \alpha^{(r)}} f(x_{n+1}|\theta_c^{(r)}) \right) + q(x_{n+1}) \frac{1}{R} \sum_{r=1}^R \frac{\alpha^{(r)}}{n + \alpha^{(r)}}.
\] (3.8)

**Transformation to the unit interval.** Note that \( Q \) defines a mapping from the support of the prior predictive distribution to the unit interval. It is convenient to carry out the computation in terms of the transformed observations

\[
\tilde{x}_i = Q(x_i).
\] (3.9)

The kernel for the transformed observations is

\[
\tilde{f}(\tilde{x}_i|\theta_c) = \text{Beta}(\tilde{x}_i|j_c, k_c - j_c + 1).
\] (3.10)

The rest of the model is unchanged. Given draws \( \{ (\psi^{(r)}, z_{1:n}^{(r)}, \alpha^{(r)}) \}_{r=1}^R \) conditioned on \( \tilde{x}_{1:n} \), we may compute the predictive distribution for \( x_{n+1} \) directly from (3.5) or (3.8).

**Algorithm 2 in Neal (2000).** Algorithm 2 in Neal (2000) may be used even though the prior is not conjugate. This is possible because (i) the prior predictive distribution is known (indeed, it is part of the specification of the model) and (ii) it is possible to draw \( \theta_c|\tilde{x}_i \) (to populate a newly created cluster). Regarding (ii), note

\[
p(j_c | k_c, x_i) = \frac{f(x_i|j_c, k_c)}{p(x_i)} = \frac{p(j_c | k_c, x_i) p(k_c)}{p(x_i)} = p(j_c | k_c, x_i) p(k_c),
\] (3.11)

\(^{12}\) This representation is associated with the Chinese Restaurant Process. It plays a central role in some samplers such as Algorithm 2 in Neal (2000).
where \( p(x_i) = q(x_i) \) and
\[
p(j_c|k_c, x_i) = \frac{\text{Beta}(Q(x_i)|j_c, k_c - j_c + 1)}{k_c} = \text{Binomial}(j_c - 1|k_c - 1, Q(x_i)).
\]
In other words, \( k_c \sim p(k_c) \) and \((j_c - 1) \sim \text{Binomial}(k_c - 1, Q(x_i))\).

### 4. Investigation: Part I

In this section I apply the model to a number of applications and investigate the performance: the Nassau County school enrollment data, the galaxy data, the Buffalo snowfall data, and the Old Faithful data (in two dimensions). [The prior predictive distributions are all flat. I intend to redo the estimation with other priors.]

Unless otherwise noted, \( \xi = 1/200 \). With this setting, the prior mean for \( k_c \) equals 200 and the prior standard deviation equals \( \sqrt{200 \times 199} \approx 199.5 \). The 90\% highest prior density (i.e., probability mass) region runs from \( k_c = 1 \) to \( k_c = 460 \).

As noted above, the prior for \( \alpha \) is given by \( p(\alpha) = \frac{1}{1 + \alpha^2} \) so that the prior median for \( \alpha \) equals one.

#### Nassau County school enrollment data.
These data have been used by Simonoff (1996) as a test bed for density estimation on the unit interval (illustrating boundary bias problems) and recently used by Geenens (2014) and Wen and Wu (2014). The data are the proportion of white student enrollment in 56 school districts in Nassau County (Long Island, New York), for the 1992–1993 school year. A total of 50,500 draws were made, with the first 500 discarded and 1,000 draws retained (every 50th) from the remaining 50,000. The predictive distribution is shown in Figure 5.

#### Galaxy data.
Figure 6 shows the quasi-Bernstein predictive density for the galaxy data with support over the interval \([5, 40]\). A total of 50,500 draws were made, with the first 500 discarded and 1,000 draws retained (every 50th) from the remaining 50,000.

#### Buffalo snowfall data.
Figure 7 shows the quasi-Bernstein predictive density for the galaxy data with support over the interval \([0, 150]\). A total of 50,500 draws were made, with the first 500 discarded and 1,000 draws retained (every 50th) from the remaining 50,000.

The density in Figure 7 is substantially smoother than what is produced by many alternative models which typically display three modes. In the current model, fixing \( \alpha = 5 \) will produce three modes, but this value for \( \alpha \) is deemed unlikely according the model when we learn about \( \alpha \). The posterior median for \( \alpha \) is about 0.31. The posterior probability of \( \alpha \geq 5 \) is about 20 times lower than the prior probability. (Increasing \( \alpha \) also has the effect of increasing the probability of new cluster, which in turn has the effect of increasing the predictive density at the boundaries of the region. For example, the predictive density increase by roughly a factor of 10 at \( x_{n+1} = 150 \).)

With this data set there is a strong (posterior) relation between \( \alpha \) and \( \tilde{k}_c \). The posterior median of \( \tilde{k}_c \) equals about 10 given \( \alpha < 1 \), but it equals about 140 given \( \alpha \geq 1 \).

#### Old Faithfull data.
Here we examine the Old Faithful data, which comprises 272 observations of pairs composed of eruption time (the duration of the current eruption in minutes) and waiting time (the amount of time until the subsequent eruption in minutes). Figure 8 shows a scatter plot of the data, a contour plot of the joint predictive distribution, and two line plots of conditional expectations computed from the joint distribution. The distribution was given positive support over the region \([1, 5.75] \times [35, 105]\). The distribution is distinctly bimodal.
5. INDIRECT DENSITY ESTIMATION FOR LATENT VARIABLES

Up to this point I have assumed that $x_{1:n}$ was observed and the goal of inference was the posterior predictive distribution $p(x_{n+1}|x_{1:n})$. In this section, I now suppose $x_{1:n}$ is latent and instead $Y_{1:n} = (Y_1, \ldots, Y_n)$ is observed, where $Y_i$ may be a vector of observations. To accommodate this situation, let

$$p(Y_i|x_i)$$

(5.1)

denote the sampling distribution for $Y_i$ given the latent variable $x_i$. The form of the density $p(Y_i|x_i)$ will depend on the specific application. Nuisance parameters may have been integrated out to obtain $p(Y_i|x_i)$. Conditional on the observation $Y_i$, one may interpret the likelihood $p(Y_i|x_i)$ as a noisy signal for $x_i$. Assume the joint likelihood is given by

$$p(Y_{1:n}|x_{1:n}) = \prod_{i=1}^{n} p(Y_i|x_i).$$

(5.2)

In this setting, the object of interest is

$$p(x_{n+1}|Y_{1:n}) = \int p(x_{n+1}|x_{1:n}) p(x_{1:n}|Y_{1:n}) dx_{1:n}$$

$$= \int p(x_{n+1}|\psi) p(\psi|Y_{1:n}) d\psi.$$  

(5.3)

The right-hand side of the first line in (5.3) expresses the distribution in terms of latent variable density estimation while the second line expresses the distribution directly in terms of the DPM. I will refer to $p(x_{n+1}|Y_{1:n})$ as the generic distribution, because it applies to any “$x$” for which there is as yet not direct signal (i.e., no observation “$Y$”). Note that $x_{n+1}$ does not appear in the likelihood (5.2) and is therefore not identified. The identified latent variables will have specific distributions that incorporate their specific signals: $p(x_i|Y_{1:n})$ for $i = 1, \ldots, n$.

The sampler works as before with the additional step of drawing $x_{1:n}|Y_{1:n}, \psi$ for each sweep of the sampler. Referring to (5.3), the generic distribution can be approximated by

$$p(x_{n+1}|Y_{1:n}) \approx \frac{1}{R} \sum_{r=1}^{R} p(x_{n+1}|\psi^{(r)}),$$

(5.4)

where $\{\psi^{(r)}\}_{r=1}^{R}$ are draws from $p(\psi|Y_{1:n})$. Referring to (3.7), the generic distribution can also be expressed as

$$p(x_{n+1}|Y_{1:n}) \approx \frac{1}{R} \sum_{r=1}^{R} p(x_{n+1}|z_{1:n}^{(r)}, \theta^{(r)}, \alpha^{(r)}),$$

(5.5)

where the draws $\{(z_{1:n}^{(r)}, \theta^{(r)}, \alpha^{(r)})\}_{r=1}^{R}$ are from the posterior given $Y_{1:n}$.

This assumption is solely for expositional simplicity. Any nuisance parameters may be retained and sampled.
We now turn to the specific distributions. Let $\theta_i$ denote $\theta_{z_i}$. Since the joint likelihood factors [see (5.2)], the full conditional posterior for $x_i$ reduces to the posterior for $x_i$ in isolation (conditional on $\theta_i$):

$$p(x_i|Y_{1:n}, x_{1:n}^{-i}, \psi, z_{1:n}) = p(x_i|Y_i, \theta_i), \quad (5.6)$$

where

$$p(x_i|Y_i, \theta_i) = \frac{p(Y_i|x_i) f(x_i|\theta_i)}{\int p(Y_i|x_i) f(x_i|\theta_i) d\theta_i}. \quad (5.7)$$

The posterior distributions of the specific cases can be approximated with histograms of the draws $\{x_i^{(r)}\}_{r=1}^R$ from the posterior. However, one can adopt a Rao–Blackwellization approach (as was done with the generic case) and obtain a lower variance approximation. In particular,

$$p(x_i|Y_{1:n}) = \int p(x_i|Y_i, \theta_i) p(\theta_i|Y_{1:n}) d\theta_i \approx \frac{1}{R} \sum_{r=1}^R p(x_i|Y_i, \theta_i^{(r)}). \quad (5.8)$$

Note $\theta_i^{(r)}$ is short-hand notation for $\theta_c^{(r)}$ where $c = z_i^{(r)}$.

**Transformation to the unit interval.** Again it is convenient to transform $x_{1:n}$ to the unit interval via $\hat{x}_i = Q(x_i)$ and use the transformed kernel (3.10). In order to resample $\hat{x}_i$, we can express the likelihood as

$$p(Y_i|\hat{x}_i) = p(Y_i|x_i)|_{x_i = Q^{-1}(\hat{x}_i)}. \quad (5.9)$$

**Sharing, shrinkage, and pooling.** Additional perspective on the role of the concentration parameter $\alpha$ can be seen through its effect on the extent to which observations share mixture components (also known as clusters). When $\alpha$ is small, $w_c$ is dominated by a few large values and consequently the amount of “sharing” is large. In the limit as $\alpha \to 0$, there is only one cluster, which amounts to complete sharing. By contrast, as $\alpha \to \infty$, the individual weights $w_c \to 0$ and each observation occupies its own cluster and there is no sharing.

One may interpret the model in terms of partial sharing of the parameters. Whenever a parameter $\theta_c$ is shared among cases, the associated coefficients ($x_i \in x^c$) are shrunk toward a common value. Complete sharing, therefore, implies global shrinkage, while partial sharing implies local shrinkage which allows for multiple modes to exist simultaneously.

Gelman et al. (2014) and Gelman and Hill (2007), discuss three types of pooling: no pooling, complete pooling, and partial pooling. The no-pooling model corresponds to the no-sharing prior and the partial-pooling model corresponds to the one-component complete sharing prior (global shrinkage). The complete-pooling model is a special case of the one-component complete sharing prior with the added restriction that all of the $x_i$ are the same (complete local shrinkage).

See Table 1 for the complete set of relationships.

6. **Investigation: Part II**

In this section I apply the model of indirect density estimation to a number of applications and investigate the performance. Rat tumor data, baseball data, and other data.
Table 1. Sharing, shrinkage, and pooling. Sharing is controlled by the concentration parameter $\alpha$. Complete sharing produces global shrinkage (to a single cluster). Local shrinkage is controlled by $k_c$, which determines the precision of the kernel. Complete local shrinkage identifies the cases in a given cluster (i.e., all cases in a given cluster have the same value). The Dirichlet Process (DP) and Dirichlet Process Mixture (DPM) are nonparametric priors.

<table>
<thead>
<tr>
<th>Local Shrinkage (controlled by $k_c$)</th>
<th>Sharing (controlled by $\alpha$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>complete ($k_c = \infty$)</td>
<td>complete ($\alpha = 0$)</td>
</tr>
<tr>
<td>partial</td>
<td>partial pooling</td>
</tr>
<tr>
<td>none ($k_c = 1$)</td>
<td>partial pooling</td>
</tr>
<tr>
<td></td>
<td>complete pooling</td>
</tr>
<tr>
<td></td>
<td>DP</td>
</tr>
<tr>
<td></td>
<td>no pooling</td>
</tr>
<tr>
<td></td>
<td>DPM</td>
</tr>
<tr>
<td></td>
<td>no pooling</td>
</tr>
<tr>
<td></td>
<td>no pooling</td>
</tr>
</tbody>
</table>

Rat tumor data. The rat tumor data is composed of the results from 71 studies. The number of rats per study varied from ten to 52. The rat tumor data are described in Table 5.1 in Gelman et al. (2014) and repeated for convenience in Table 2 (although the data are displayed in a different order). The data are plotted in Figure 4. This plot brings out certain features of the data that are not evident in the table. There are 59 studies for which the total number of rats is less than or equal to 35 and more than half of these studies (32) have observed tumor rates less than or equal to 10%. By contrast, none of the other 12 studies has an observed tumor rate less than or equal to 10%.

The posterior distribution for the generic case is shown in Figure 10. The posterior distributions for the specific cases are shown in Figure 11. This latter figure can be compared with Figure 5.4 in Gelman et al. (2014) to show the differences in the results obtained by the more general approach presented here.

Baseball batting skill. This example is inspired by the example in Efron (2010) which in turn draws on Efron and Morris (1975). We are interested in the ability of baseball players to generate hits. We do not observe this ability directly; rather we observe the outcomes (successes and failures) of a number of trials for a number of players. In this example $T_i$ is the number of “at-bats” and $s_i$ is the number of “hits” for player $i$. See Figure 3 for the data. [The analysis in not complete.]

Thumbtack data. The thumbtack data are shown in Table 3. The posterior distribution for the generic success rate is displayed in Figure 12.

The posterior distribution for the generic success rate given the alternative model is shown in Figure 14.

Appendix A. A different object of interest

In order to help clarify the nature of the predictive distribution (2.3), it may be useful to contrast it with a different object of interest. In particular, one could be interested in estimating the unknown (density) function $g(x) = p(x|\psi)$.

A frequentist estimate might be $\hat{g}(x) = p(x|\hat{\psi})$, where $\hat{\psi}$ is an estimate of $\psi$ such as the maximum likelihood estimate:

$$\hat{\psi} = \arg\max_{\psi} p(x_{1:n}|\psi).$$

(A.1)
Uncertainty regarding the estimate $\hat{g}(x)$ could be characterized by the sampling variation in $x_{1:n}$. For example, variation in $x_{1:n}$ would induce variation in $\hat{\psi}$ and consequently in $\hat{g}(x)$ as follows:

$$x_{1:n}^{(r)} \sim p(x_{1:n}|\psi) \quad \text{and} \quad x_{1:n}^{(r)} \rightarrow \hat{\psi}^{(r)} \rightarrow p(x|\hat{\psi}^{(r)}),$$

where $x_{1:n}^{(r)}$ denotes a draw from the sampling distribution (possibly approximated using a bootstrap approach) and $\hat{\psi}^{(r)}$ denotes the corresponding estimate computed from that draw.

From the Bayesian perspective, variation in $p(x|\psi)$ flows from variation in $\psi$ according to the posterior distribution for $\psi$:

$$\psi^{(r)} \sim p(\psi|x_{1:n}) \quad \text{and} \quad \psi^{(r)} \rightarrow p(x|\psi^{(r)}),$$

where $\psi^{(r)}$ denotes a draw from the posterior distribution (possibly approximated using an MCMC approach). A Bayesian estimate of the function $g(x)$ might compute the average of $p(x|\psi)$ with respect to the posterior distribution for $\psi$:

$$\tilde{g}(x) = \int p(x|\psi) p(\psi|x_{1:n}) d\psi.$$  \hfill (A.4)

Although the Bayesian estimate $\tilde{g}(x)$ has the same representation as the predictive distribution $p(x_{n+1}|x_{1:n})$ [see (2.3)], the two objects are fundamentally different. The predictive distribution is not an estimate of $g(x)$; rather, it is a summary of what is known about $x_{n+1}$ based on the observations $x_{1:n}$ — it is the distribution that would be used to make a decision that depends on $x_{n+1}$. Variation in $\psi$ due to its posterior distribution plays a role in constructing the predictive distribution, but such variation is not relevant once $\psi$ has been integrated out.

Nevertheless, there are two sorts of variation that are of interest regarding the predictive distribution. The first sort of variation involves the sensitivity of $p(x_{n+1}|x_{1:n})$ to the prior distribution $p(\psi)$. The second sort of variation relates to the possibility of delaying the decision until more observations are acquired: Perhaps additional observations can change the predictive distribution in a way that increases the expected utility of the decision by more than the associated cost from delay and acquisition.

For example, the utility of a decision that depends on $x_{n+2}$ may depend on $x_{n+1}$ through

$$p(x_{n+2}|x_{1:n}, x_{n+1}) = \int p(x_{n+2}|\psi) p(\psi|x_{1:n}, x_{n+1}) d\psi,$$  \hfill (A.5)

in which case $p(x_{n+1}|x_{1:n})$ will be used to compute the optimal decision. However, if $x_{n+1}$ provides very little information about $\psi$, then $p(\psi|x_{1:n}, x_{n+1}) \approx p(\psi|x_{1:n})$ and

$$p(x_{n+2}|x_{1:n}, x_{n+1}) \approx p(x_{n+2}|x_{1:n}).$$  \hfill (A.6)

Consequently, the decision will be essentially independent of $x_{n+1}$ and there will be little benefit from the additional observation.

**Appendix B. Comparison with Petrone’s model**

Petrone’s model can be expressed as follows:

$$p(x_i|\pi_k, k) = \sum_{j=1}^{k} \pi_{jk} \text{Beta}(x_i|j, k - j + 1),$$  \hfill (B.1)
where \( \pi_k = (\pi_{1k}, \ldots, \pi_{kk}) \) is the vector of mixture weights such that \( \pi_{jk} \geq 0 \) and \( \sum_{j=1}^{k} \pi_{jk} = 1 \). In Petrone’s model, the prior is given by

\[
\pi_k | k \sim \text{Dirichlet}(\alpha_k) \tag{B.2a}
\]

\[
k \sim P_k, \tag{B.2b}
\]

where \( \alpha_k = (\alpha_{1k}, \ldots, \alpha_{kk}) \).\(^{14}\)

Compare (B.1)–(B.2) with (2.5) and (2.11)–(2.12), letting \( Q(x) = x \). Petrone’s model is an average of finite-order mixtures, while the model in this paper is an infinite-order mixture model. Any finite mixture in my model can be represented exactly in Petrone’s model.

However, my model is a more parsimonious version of Petrone’s, in that it can represent the same functional forms with fewer parameters. I suggest my model is more efficient.

In any even, the two models can be compared in terms of their marginal likelihoods. The marginal likelihoods can be computed from the sequence of predictive distributions:

\[
p(x_{1:n} | A_m) = p(x_1) \prod_{i=2}^{n} p(x_i | x_{1:i-1}, A_m), \tag{B.3}
\]

where \( A_m \) stands for the assumptions of model \( m \). Let \( m = 1 \) indicate Petrone’s model and let \( m = 2 \) indicate my model.

**Appendix C. Comparison with Canale and Dunson**

Canale and Dunson (2016) present a model with close similarities to the model in this paper. The model in this paper is

\[
p(x_i | \cdot) = \sum_{c=1}^{\infty} w_c \text{Beta}(x_i | j_c, k_c - j_c + 1) \tag{C.1a}
\]

where

\[
w_c = v_c \prod_{\ell=1}^{c-1} (1 - v_{\ell}) \tag{C.1b}
\]

\[
v_c \overset{iid}{\sim} \text{Beta}(1, \alpha) \tag{C.1c}
\]

\[
k_c - 1 \overset{iid}{\sim} \text{Geometric}(\xi) \tag{C.1d}
\]

\[
j_c | k_c \sim \text{Uniform}\{1, \ldots, k_c\} \tag{C.1e}
\]

(The distributions for \( w \) and \( k_c \) may be easily changed if the situation calls for it.)

By contrast, the model of Canale and Dunson (2016) is

\[
p(x_i | \cdot) = \sum_{s=0}^{\infty} \sum_{h=1}^{2^s} \pi_{s,h} \text{Beta}(x_i | h, 2^s - h + 1), \tag{C.2a}
\]

where

\[
\pi_{s,h} = S_{s,h} \prod_{r<s} (1 - S_{r,g_{shr}}) T_{shr} \tag{C.2b}
\]

\[
S_{s,h} \sim \text{Beta}(1, \alpha) \tag{C.2c}
\]

\[
R_{s,h} \sim \text{Beta}(b, b) \tag{C.2d}
\]

\(^{14}\)Petrone re-expresses the model in such a way as to facilitate sampling from the posterior.
DENSITY ESTIMATION USING ORDER-STATISTIC DISTRIBUTIONS

where \( g_{shr} = \lceil h/2^{s-r} \rceil \) is the node traveled through at scale \( r \) on the way to node \( h \) at scale \( s \), \( T_{shr} = R_{r,g_{shr}} \) is the right daughter of node \((r, g_{shr})\), and \( T_{shr} = 1 - R_{r,g_{shr}} \) if \((r+1, g_{shr}+1)\) is the left daughter of \((r, g_{shr})\).

The marginal prior distribution for node \((s, h)\) is given by

\[ s \sim \text{Geometric}(1/(1 + a)) \quad \text{(C.3)} \]
\[ h|s \sim \text{Uniform}(\{1, \ldots, 2^s\}). \quad \text{(C.4)} \]

There are two ways in which the models differ. First, Canale and Dunson (2016) restrict the set of \( \{k\} \) to powers of two \((k = 2^s)\). Second, the way in which clustering is modeled is different. In Canale and Dunson (2016) clustering is affected by the choice of \( b \). However, the effect of \( b \) is dominated by the effect of \( a \). As the authors say Hyperpriors can be chosen for \( a \) and \( b \) to allow the data to inform about these tuning parameters; we find that choosing the hyperprior for \( a \) is particularly important, with \( b = 1 \) as a default.

APPENDIX D. LOCAL DEPENDENCE VIA COPULA

In this section I generalize the two-dimensional density model (presented in Section 2) to include local dependence, which is introduced via a copula.\(^\text{15}\) The Farlie–Gumbel–Morgenstern (FGM) copula is easy to work with because a flat prior for its copula parameter produces a flat prior predictive density over the unit square. However, the potential dependence is somewhat limited.

The FGM copula is given by

\[ C(u_1, u_2 | \tau) := u_1 u_2 (1 - \tau (1 - u_1)(1 - u_2)) \quad \text{where } -1 \leq \tau \leq 1. \quad \text{(D.1)} \]

Note that \( C(u_1, u_2 | \tau) \) is the CDF for \((u_1, u_2) \in [0, 1]^2\). The PDF is

\[ c(u_1, u_2 | \tau) := 1 + \tau (2 u_1 - 1)(2 u_2 - 1). \quad \text{(D.2)} \]

The marginal densities are flat: \( p(u_\ell | \tau) = 1_{[0,1]}(u_\ell) \) for \( \ell = 1, 2 \). The correlation between \( u_1 \) and \( u_2 \) is \( \tau/3 \). Setting \( \tau = 0 \) delivers independence.

The CDF for the beta distribution is the regularized incomplete beta function:

\[ I_x(a, b) = \int_0^x \text{Beta}(t|a, b) \, dt. \quad \text{(D.3)} \]

Let the joint CDF for a single observation \( x_i = (x_{i1}, x_{i2}) \) be given by

\[ F(x_i | \theta_c) := C(\zeta_{ic1}, \zeta_{ic2}, \tau_c), \quad \text{(D.4)} \]

where \( \theta_c = (j_{c1}, j_{c2}, k_{c1}, k_{c2}, \tau_c) \) and

\[ \zeta_{ic} := I_{x_{ic}}(j_{ic}, k_{ic} - j_{ic} + 1). \quad \text{(D.5)} \]

Then the joint density is given by

\[ f(x_i | \theta_c) = c(\zeta_{ic1}, \zeta_{ic2} | \tau) \prod_{\ell=1}^2 \text{Beta}(x_{i\ell} | j_{ic\ell}, k_{ic\ell} - j_{ic\ell} + 1). \quad \text{(D.6)} \]

\(^{15}\) The generalization to higher dimensions is straightforward.
Let the prior for $\theta_c$ be given by

$$p(\theta_c) = \frac{p(k_{c1}) p(k_{c2}) p(\tau_c)}{k_{c1} k_{c2}},$$

where

$$p(\tau_c) = \frac{1}{2} 1_{[-1,1]}(\tau_c).$$

The expectation with respect to $\tau_c$ produces (conditional) independence between $x_{i1}$ and $x_{i2}$:

$$p(x_i | j_c, k_c) = \int_{-1}^{1} f(x_i | \theta_c) p(\tau_c) d\tau_c = \prod_{\ell=1}^{2} \text{Beta}(x_{i\ell} | j_c, k_{c\ell} - j_{c\ell} + 1).$$

Given the conditional priors for $j_{c1}$ and $j_{c2}$, this prior produces a flat prior predictive for any $p(k_{c1})$ and $p(k_{c1})$:

$$p(x_i) = 1_{[0,1]}^2(x_i).$$

**Algorithm 2 in Neal (2000).** Algorithm 2 in Neal (2000) may used because (i) the prior predictive is known [see (D.10)] and (ii) it is possible to draw $\theta_c | x_i$. Regarding (ii), note

$$p(\theta_c | x_i) = \frac{f(x_i | \theta_c) p(\theta_c)}{p(x_i)} \left( \frac{2}{\prod_{\ell=1}^{2} p(k_{c\ell})} \frac{2}{\prod_{\ell=1}^{2} p(j_{c\ell} | x_{i\ell}, k_{c\ell})} \frac{c(\zeta_{ic1}, \zeta_{ic2} | \tau_c)}{2} \right).$$

Thus, we can make a draw form the joint posterior by first drawing $k_c$ from its prior distribution, then drawing $j_c$ from its distribution conditional on $k_c$, and finally drawing $\tau_c$ from its conditional distribution.

**Appendix E. Binomial likelihood**

An special case of some interest is when the data are binomial in nature:

$$p(Y_i | x_i) = \text{Binomial}(s_i | T_i, x_i),$$

where $T_i$ is the number of trials, $s_i$ is the number of successes, and $x_i$ is the latent probability of success. In this case, the conditional posterior for a specific case is

$$p(x_i | Y_i, \theta_{zi}) = \text{Beta}(x_i | j_{zi} + s_i, (k_{zi} - j_{zi} + 1) + T_i - s_i),$$

which can be used in (5.6) for sampling and in (5.8) for smoothing. In this setting both specific and generic posterior distributions are mixtures of beta distributions.

---

\[16\] With the binomial likelihood it is possible to analytically integrate out $x_1:n$ as shown in Appendix E. However, we do not follow that procedure in our estimation.
Integrate out the success rates. If the data are observations from binomial experiments and the prior predictive is the uniform distribution, then the prior is a mixture of beta distributions and it is possible to integrate out the unobserved success rates.

For this purpose, assume \( q(x_i) = \text{Uniform}(0, 1) \). Then there is a closed-form expression for the likelihood of \( \theta_c = (j_c, k_c) \) in terms of the observations:

\[
p(s_i|T_i, \theta_c) = p(s_i|T_i, j_c, k_c) = \int_0^1 \text{Binomial}(s_i|T_i, x_i) \text{Beta}(x_i|j_c, k_c - j_c + 1) \, dx_i
\]

\[
= \frac{(T_i)_{k_c} (s_i + j_c - 1)! (T_i + k_c - s_i - j_c)!}{(j_c - 1)! (k_c - j_c)! (T_i + k_c)!}
\]  

(E.3)

Note that

\[
p(s_i|T_i, k_c) = \sum_{j_c=1}^{k_c} \frac{p(s_i|T_i, j_c, k_c)}{k_c} = \frac{1}{T_i + 1},
\]  

(E.4)

which is independent of \( k_c \). Therefore, \( p(s_i|T_i, k_c) = p(s_i|T_i) = 1/(T_i + 1) \).

We can again use Algorithm 2 from Neal (2000) to make draws from the posterior distribution since the prior predictive distribution is known and

\[
p(j_c, k_c|T_i, s_i) = \frac{p(s_i|T_i, j_c, k_c) p(j_c, k_c)}{p(s_i|T_i)}
\]

\[
= \left\{ \frac{(T_i + 1) \text{Beta-Binomial}(s_i|j_c, k_c - j_c + 1, T_i)}{k_c} \right\} p(k_c|I)
\]  

(E.5)

\[
= \text{Beta-Binomial}(j_c - 1|s_i + 1, T_i - s_i + 1, k_c - 1) p(k_c),
\]

where

\[
(j_c - 1)|(T_i, s_i, k_c) \sim \text{Beta-Binomial}(s_i + 1, T_i - s_i + 1, k_c - 1),
\]  

(E.6)

with the proviso that \( j_c = 1 \) if \( k_c = 1 \).

Limiting case of DP. In Section 2, it was suggested that the limiting case as \( k_c \to \infty \) produced a Dirichlet Process (DP) model, where

\[
p(x_i|\psi) = \sum_{c=1}^{\infty} w_c \delta(x_i - \phi_c),
\]  

(E.7)

where \( \psi = (w, \phi) \). Here we apply this model to the case of indirect density estimation.

It is convenient to integrate out \( x_i \):

\[
p(Y_i|\phi_c) = \int p(Y_i|x_i) \delta(x_i - \phi_c) \, dx_i = p(Y_i|x_i)|_{x_i=\phi_c}.
\]  

(E.8)

Then

\[
p(Y_i|\psi) = \int p(Y_i|x_i) p(x_i|\psi) \, dx_i = \sum_{c=1}^{\infty} w_c p(Y_i|\phi_c),
\]  

(E.9)

where \( w \sim \text{Stick}(\alpha) \) and \( \phi_c \sim Q \). In (E.9), \( p(Y_i|\phi_c) \) plays the role of the kernel; however, the form of \( p(Y_i|\phi_c) \) depends on \( p(Y_i|x_i) \) which in turn depends on the observations (and possibly on other aspects of the likelihood).
We now combine the DP prior with the Binomial likelihood.\footnote{See Greenberg (2013) and Geweke et al. (2011) for textbook treatments of the DP with a binomial likelihood.} Referring to (E.1) and (E.8),
\[
p(Y_i|\phi_c) = \text{Binomial}(s_i|T_i, \phi_c).
\]
We can classify observations according to likelihood.

We can classify observations according to $z_c \propto w_c p(Y_i|\phi_c)$ and utilize the rest of the sampler described in Section 3 as well. If we assume
\[
q(\phi_c) = \text{Beta}(\phi_c|a_0, b_0),
\]
then updating the cluster parameters conditional on the classifications is straightforward:
\[
p(\phi_c|Y_{1:n}, z_{1:n}) = \text{Beta}(\phi_c|\tilde{a}_c, \tilde{b}_c),
\]
where
\[
\tilde{a}_c := a_0 + \sum_{\ell \in I_c} s_\ell \quad \text{and} \quad \tilde{b}_c := b_0 + \sum_{\ell \in I_c} T_\ell - s_\ell.
\]
where $I_c = \{i : z_i = c\}$. There is no change in the way the updates are computed for $w$ or for $\alpha$.

The specific distributions can be approximated via $\{x_i^{(r)}\}_{r=1}^R$ where $x_i^{(r)} = \phi_{z_i}^{(r)}$ and the generic distribution can be approximated via $\{x^{(r)}_{n+1}\}_{r=1}^R$ where
\[
x^{(r)}_{n+1} \sim \sum_{c=1}^m w_c^{(r)} \delta_{\phi_c^{(r)}}.
\]
If we assume (E.11), then smoother approximations are readily available. First consider the specific case. Since $x_i = \phi_{z_i}$, we have
\[
p(x_i|Y_{1:n}, z_{1:n}) = p(\phi_{z_i}|Y_{1:n}, z_{1:n}) = \text{Beta}(\phi_{z_i}|\tilde{a}_{z_i}, \tilde{b}_{z_i}) = \text{Beta}(x_i|\tilde{a}_{z_i}, \tilde{b}_{z_i}),
\]
where $\tilde{a}_{z_i}$ and $\tilde{b}_{z_i}$ are given in (E.13) with $c = z_i$. Therefore,
\[
p(x_i|Y_{1:n}) \approx \frac{1}{R} \sum_{r=1}^n p(x_i|Y_{1:n}, z_{1:n}^{(r)}) = \frac{1}{R} \sum_{r=1}^R \text{Beta}(x_i|\tilde{a}_{z_i}^{(r)}, \tilde{b}_{z_i}^{(r)}).
\]
Turning to the generic case, we have
\[
p(x_{n+1}|Y_{1:n}, z_{1:n}, w) = \int p(x_{n+1}|w, \phi) p(\phi|Y_{1:n}, z_{1:n}) d\phi
\]
\[
= \sum_{c=1}^m w_c \text{Beta}(x_{n+1}|\tilde{a}_c, \tilde{b}_c).
\]
Therefore,
\[
p(x_{n+1}|Y_{1:n}) = \int \int p(x_{n+1}|Y_{1:n}, z_{1:n}, w) p(z_{1:n}, w|Y_{1:n}) dz_{1:n} dw
\]
\[
\approx \frac{1}{R} \sum_{r=1}^R \sum_{c=1}^m w_c^{(r)} \text{Beta}(x_{n+1}|\tilde{a}_c^{(r)}, \tilde{b}_c^{(r)}).
\]
In the spirit of (3.7), an even smoother approximation is
\[
p(x_{n+1}|Y_{1:n}) \approx \frac{1}{R} \sum_{r=1}^{R} \sum_{c \in C(r)} \frac{n_c^{(r)}}{n + \alpha^{(r)}} \text{Beta}(x_{n+1}|\tilde{a}_c^{(r)}, \tilde{b}_c^{(r)}) + \frac{\alpha^{(r)}}{n + \alpha^{(r)}} q(x_{n+1}). \tag{E.19}
\]

**Model comparison.** The likelihood of the model using the DP prior may be compared with the likelihood of the model using the more general DPM prior. Let \( M \) denote the model. Then
\[
p(Y_{1:n}|M) = p(Y_1|M) \prod_{i=2}^{n} p(Y_i|Y_{1:i-1}, M), \tag{E.20}
\]
where
\[
p(Y_i|Y_{1:i-1}, M) = \int p(Y_i|x_i) p(x_i|Y_{1:i-1}, M) \, dx_i. \tag{E.21}
\]

In general, these integrals can be computed via numerical quadrature.

**References**


Table 2. Rat tumor data: 71 studies (rats with tumors/total number of rats).

<table>
<thead>
<tr>
<th>00/17</th>
<th>00/18</th>
<th>00/18</th>
<th>00/19</th>
<th>00/19</th>
<th>00/19</th>
<th>00/19</th>
<th>00/20</th>
<th>00/20</th>
</tr>
</thead>
<tbody>
<tr>
<td>00/20</td>
<td>00/20</td>
<td>00/20</td>
<td>00/20</td>
<td>00/20</td>
<td>01/20</td>
<td>01/20</td>
<td>01/20</td>
<td>01/20</td>
</tr>
<tr>
<td>01/19</td>
<td>01/19</td>
<td>01/18</td>
<td>01/18</td>
<td>02/25</td>
<td>02/24</td>
<td>02/23</td>
<td>01/10</td>
<td>02/20</td>
</tr>
<tr>
<td>02/20</td>
<td>02/20</td>
<td>02/20</td>
<td>02/20</td>
<td>02/20</td>
<td>05/49</td>
<td>02/19</td>
<td>05/46</td>
<td>03/27</td>
</tr>
<tr>
<td>02/17</td>
<td>07/49</td>
<td>07/47</td>
<td>03/20</td>
<td>03/20</td>
<td>02/13</td>
<td>09/48</td>
<td>04/20</td>
<td>04/20</td>
</tr>
<tr>
<td>04/20</td>
<td>04/20</td>
<td>04/20</td>
<td>04/20</td>
<td>04/20</td>
<td>10/50</td>
<td>10/48</td>
<td>04/19</td>
<td>04/19</td>
</tr>
<tr>
<td>04/19</td>
<td>05/22</td>
<td>11/46</td>
<td>12/49</td>
<td>05/20</td>
<td>05/20</td>
<td>06/23</td>
<td>05/19</td>
<td>06/22</td>
</tr>
<tr>
<td>04/14</td>
<td>06/20</td>
<td>06/20</td>
<td>06/20</td>
<td>16/52</td>
<td>15/47</td>
<td>15/46</td>
<td>09/24</td>
<td></td>
</tr>
</tbody>
</table>

Figure 3. Baseball data: 18 players with 45 at-bats each.

Table 3. The thumbtack data set: 320 instances of binomial experiments with 9 trials each. The results are summarized in terms of the number of experiments that have a given number of successes.

<table>
<thead>
<tr>
<th>No. of successes</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of experiments</td>
<td>0</td>
<td>3</td>
<td>13</td>
<td>18</td>
<td>48</td>
<td>47</td>
<td>67</td>
<td>54</td>
<td>51</td>
<td>19</td>
<td>320</td>
</tr>
<tr>
<td>Frequency (percent)</td>
<td>0</td>
<td>1</td>
<td>4</td>
<td>6</td>
<td>15</td>
<td>15</td>
<td>21</td>
<td>17</td>
<td>16</td>
<td>6</td>
<td>≈ 100</td>
</tr>
</tbody>
</table>
**Figure 4.** Rat tumor data: 71 studies. Number of studies (1 to 7) proportional to area of dot. Number of rats with tumors (0 to 16) indicated by contour lines. There are 59 studies for which the total number of rats is less than or equal to 35 and more than half of these studies (32) have observed tumor rates less than or equal to 10%. By contrast, none of the other 12 studies has an observed tumor rate less than or equal to 10%.

**Figure 5.** School data: Posterior predictive distribution.
Figure 6. Galaxy data: quasi-Bernstein predictive density with support over the interval $[5, 40]$ and a rug plot of the data.

Figure 7. Buffalo snowfall data: quasi-Bernstein predictive density with support over the interval $[0, 150]$ and a rug plot of the data.
Figure 8. Old Faithful data: Contours for posterior predictive density with support over \([1, 5.75] \times [35, 105]\). Lowest contour is at the level of the uniform prior \((\approx 0.003)\). Contour spacing above the lowest contour is \(\approx 0.006\). Data are shown as dots and conditional expectations are shown as thicker lines.

Figure 9. Old Faithful data: Marginal distributions for eruption time and waiting time computed from joint distribution.
Figure 10. Posterior distribution for generic rat tumor rate.

Figure 11. Posterior medians and 95% highest posterior density regions of rat tumor rates. Darker lines indicate multiple observations. Compare with Figure 5.4 in Gelman et al. (2014).
Figure 12. Thumbtack data: Posterior distribution for the generic probability of success.

Figure 13. Thumbtack data: Posterior distributions for the specific probabilities of success, computed for each exchangeable group with a common number of success, 1 through 9.
Figure 14. Thumbtack data: Posterior distribution for generic success rate given the alternative model compared with the distribution given the main model.