DENSITY ESTIMATION USING A MIXTURE OF ORDER-STATISTIC DISTRIBUTIONS

MARK FISHER

Preliminary and incomplete.

ABSTRACT. This paper presents a Bayesian nonparameteric 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 continuous distribution Q. The model fixes the prior distribution for j to be uniform conditional on k [i.e., $p(j|k) = 1/k$] with the consequence that the prior predictive distribution equals Q (regardless of the prior distributions for k or the mixture weights). The parameter k controls the precision of the kernel. In the limit as $k \to \infty$, the kernel collapses to a Dirac delta function and the model specializes to a Dirichlet Process (DP) model with base distribution Q for the remaining parameter (a rescaling of j/k that preserves location). The model is completely determined by the prior predictive distribution Q, the prior distribution for the precision parameter k, and the prior distribution for the mixture weights.

The model presented in this paper may be interpreted as a more flexible version of that in Petrone (1999) "Bayesian density estimation using Bernstein polynomials."

Date: 08:34 December 13, 2018. *Filename*: `multi-resolution`.

JEL Classification. C11, C14.

Key words and phrases. Bayesian nonparametric density estimation, Bernstein polynomials, Dirichlet process mixture model, order statistics.

1. Introduction

This paper presents a Bayesian nonparameteric 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 may be compared with those generated by an alternative model that has appeared in the literature. (The alternative model is based on a a prior that is shown to be a special case of the prior presented here.)

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. See Appendix A for a formal comparison of the two models.

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.[1] See Appendix B 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 extends the model to the latent-variable case. Sections 6 and 7 present empirical results.

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:

$$p(x_{n+1}|x_{1:n}). \tag{2.1}$$

[1] See also Canale (2017).

Assume

$$x_i \overset{\text{iid}}{\sim} p(\,\cdot\,|\psi) \qquad \text{for } i = 1, 2, \ldots, \tag{2.2}$$

where ψ is an unobserved parameter. Then the predictive distribution can be expressed as

$$p(x_{n+1}|x_{1:n}) = \int p(x_{n+1}|\psi)\, p(\psi|x_{1:n})\, d\psi, \tag{2.3}$$

where $p(\psi|x_{1:n})$ is the posterior distribution for ψ, 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 $p(\psi)$:

$$p(\psi|x_{1:n}) \propto p(x_{1:n}|\psi)\, p(\psi). \tag{2.4}$$

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 ψ 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 C 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 θ_c. The density $f(\,\cdot\,|\theta_c)$ is called the *kernel*. Let

$$p(x_i|\psi) = \sum_{c=1}^{\infty} w_c\, f(x_i|\theta_c), \tag{2.5}$$

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 *mixure-component parameters*. The structure of the prior for ψ is

$$p(\psi) = p(w)\, p(\theta) = p(w) \prod_{c=1}^{\infty} p(\theta_c), \tag{2.6}$$

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

It remains to specify (i) the prior for the weights (which is relatively standard) and (ii) the kernel and the base distribution (wherein the main novelty resides).

Prior for the mixture weights. The prior for w is given by

$$w \sim \mathsf{Stick}(\alpha), \tag{2.7}$$

where $\mathsf{Stick}(\alpha)$ denotes the stick-breaking distribution given by[2,3]

$$w_c = v_c \prod_{\ell=1}^{c-1} (1 - v_\ell) \qquad \text{where } v_c \overset{\text{iid}}{\sim} \mathsf{Beta}(1, \alpha). \tag{2.8}$$

The parameter α is called the *concentration parameter*; it controls the rate at which the weights decline on average. In particular, the weights decline geometrically in expectation:

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

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

[2]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.

[3]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.

If α is small, then the first few weights will dominate and only a small number of mixture components will be consequential; by contrast if α is large, then a large number of mixture components will consequential.[4]

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. This will be done by specifying a prior for α.

The kernel and the base distribution. The novelty lies in the combination of the kernel and the base distribution. Consider a distribution Q for a continuous random variable defined on the real line. Let $Q(x)$ denote its cumulative distribution function (CDF) and let $q(x) = Q'(x)$ denote the associated density function. The kernel is given by[5]

$$f(x_i|\theta_c) = f(x_i|j_c, k_c) = \mathsf{Beta}\big(Q(x_i)|j_c, k_c - j_c + 1\big)\, q(x_i), \tag{2.10}$$

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

$$k_c \sim P_k \tag{2.11a}$$

$$j_c|k_c \sim \mathsf{Uniform}(\{1, \ldots, k_c\}), \tag{2.11b}$$

for some distribution P_k over the positive integers.

Given the prior independence of w and θ, the prior predictive distribution for x_i is

$$p(x_i) = \int f(x_i|\theta_c)\, p(\theta_c)\, d\theta_c = q(x_i), \tag{2.12}$$

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

$$\sum_{j_c=1}^{k_c} f(x_i|j_c, k_c)\, p(j_c|k_c) = \frac{1}{k_c} \sum_{j_c=1}^{k_c} f(x_i|j_c, k_c) = q(x_i). \tag{2.13}$$

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 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 the choice is drawn from Q.

Summary. In summary, the model is completely determined by specifying (*i*) the predictive distribution Q, (*ii*) the prior distribution for k_c, and (*iii*) the prior for α.

---

[4] In order to understand the nature of the "concentration" that lends α its moniker, define the random probability distribution $G = \sum_{c=1}^{\infty} w_c\, \delta_{\theta_c}$, where δ_{θ_c} is a point mass located at θ_c. The randomness of G follows from $w \sim \mathsf{Stick}(\alpha)$ and $\theta_c \overset{\text{iid}}{\sim} H$, where H is the base distribution. In other words, G is distributed according to a Dirichlet Process: $G \sim \mathsf{DP}(\alpha, H)$. As a consequence, the mean of G is H, $E[G] = H$, and the concentration of G around H is controlled by α. At one extreme for α, $\lim_{\alpha\to 0}\ G = \delta_{\theta_1}$ (where $\theta_1 \sim H$), which maximizes the variation of G around H. At the other extreme there is no variation at all: $\lim_{\alpha\to\infty}\ G = H$.

[5] $\mathsf{Beta}(x|a, b) = x^{a-1}\,(1-x)^{b-1}/B(a, b)$, where $B(a, b)$ is the beta function.

Priors adopted in the empirical section. In the empirical section I will adopt the following prior for k_c:

$$k_c - 1 \sim \mathsf{Geometric}(\xi), \tag{2.14}$$

where $\xi \in (0,1)$. Given this distribution, $p(k_c = 1) = \xi$ and $E[k_c] = 1/\xi$. In addition, I will adopt the following prior for α:

$$p(\alpha) = \mathsf{Log\text{-}Logistic}(\alpha|1,1) = \frac{1}{(1+\alpha)^2}. \tag{2.15}$$

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

$$\int_0^\infty \left(\frac{\alpha}{1+\alpha}\right)^m p(\alpha)\, d\alpha = \frac{1}{1+m}. \tag{2.16}$$

Variation around the prior predictive. Both $p(\alpha)$ and $p(k_c)$ affect the variation of $p(x_i|\psi)$ around $q(x_i)$. The concentration parameter works though its effect on the weights, while k_c works through its effect on the kernel (via the base distribution[6]). 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,[7]

$$\lim_{\alpha\to\infty}\ p(x_i|\psi) = q(x_i). \tag{2.17}$$

Turning to the effect of k_c, if $\Pr[k_c = 1] = 1$, then every mixture component equals the prior predictive distribution regardless of the weights:

$$p(x_i|\psi) = \sum_{c=1}^{\infty} w_c\, \mathsf{Uniform}(Q(x_i)|0,1)\, q(x_i) = q(x_i). \tag{2.18}$$

To better appreciate the role of k_c, let $\mathsf{Q} = \mathsf{Uniform}(0,1)$ so that $Q(x) = x$ and $q(x) = 1$. In this case the variance of $x_i \sim \mathsf{Beta}(j_c, k_c - j_c + 1)$ is

$$\frac{j_c\,(k_c - j_c + 1)}{(k_c+1)^2\,(k_c+2)}. \tag{2.19}$$

Averaging over j_c, the prior expectation of the variance is

$$\frac{1}{6\,(k_c+1)}. \tag{2.20}$$

As k_c gets large, the variance of the kernel goes to zero. In the limit, the kernel becomes a Dirac delta function and the model becomes a Dirichlet Process (DP) model. This feature holds for any Q. See Section 5.

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).

[6]This formulation of the the DPM allows one to change the base distribution without changing the predictive distribution.

[7]As noted in Footnote 4, $\lim_{\alpha\to\infty}\ G = H$, and H delivers the prior predictive distribution which has been shown to be Q.

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):

$$p(x_1, x_2) = q(x_1)\, q(x_2)\, c\big(Q(x_1), Q(x_2)\big), \tag{2.21}$$

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 α [see (2.15)], the unconditional probability that x_1 and x_2 share the same component is

$$\int_0^{\infty} \frac{1}{1+\alpha}\, p(\alpha)\, d\alpha = \int_0^{\infty} \frac{1}{(1+\alpha)^3}\, d\alpha = \frac{1}{2}. \tag{2.22}$$

Therefore, after integrating out w, α, and j_c, we have

$$c(u_1, u_2) = \sum_{k_c=1}^{\infty} p(k_c)\, c(u_1, u_2 | k_c), \tag{2.23}$$

where $c(u_1, u_2|k_c)$ is an order-statistic-based copula density that depends on k_c.[8] In particular,

$$c(u_1, u_2|k_c) := \frac{1}{2} + \frac{1}{2} \sum_{j=1}^{k_c} \frac{1}{k_c} \prod_{i=1}^{2} \mathsf{Beta}(u_i | j_c, k_c - j_c + 1) = \frac{1}{2} + \frac{\widetilde{c}(u_1, u_2|k_c)}{2}, \tag{2.24}$$

where

$$\widetilde{c}(u_1, u_2|k_c) = k_c \left((1-u_1)\,(1-u_2)\right)^{k_c - 1}\, {}_2F_1\left(1-k_c, 1-k_c; 1; \frac{u_1\, u_2}{(1-u_1)\,(1-u_2)}\right) \tag{2.25}$$

and where ${}_2F_1$ is the hypergeometric function. Note $\widetilde{c}(u_1, u_2|k_c = 1) = 1$ which in turn implies $c(u_1, u_2|k_c = 1) = 1$. For $k_c > 1$, $\widetilde{c}(u_1, u_2|k_c)$ provides positive dependence between u_1 and u_2; the strength of the dependence increases with k_c. In particular, the correlation between u_1 and u_2 according to $\widetilde{c}(u_1, u_2|k_c)$ is $(k_c - 1)/(k_c + 1)$. 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 $\mathsf{Q} = \mathsf{Uniform}(0,1)$ and $\mathsf{Q} = \mathsf{N}(0,1)$.

Open-mindedness. An open-minded prior allows for substantial variation around the prior predictive distribution. The prior predictive distribution is given by $q(x_i)$. Variation around it can be examined as follows. Make draws $\{\psi^{(r)}\}_{r=1}^{R}$ from the prior, where $\psi^{(r)} \overset{\text{iid}}{\sim} p(\psi)$. The prior predictive can be approximated by

$$p(x_i) \approx \frac{1}{R} \sum_{r=1}^{R} p(x_i | \psi^{(r)}). \tag{2.26}$$

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 $Q(x) = x$.

[8] See Baker (2008) for a treatment of copulas generated via order-statistics.

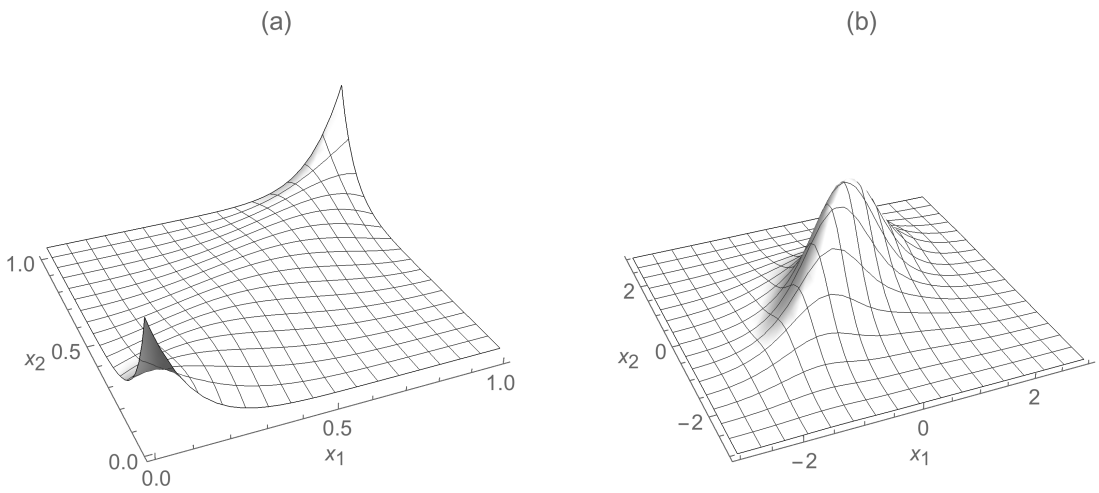

FIGURE 1. Plots of $p(x_1, x_2|k_c = 10)$ assuming (a) $\mathsf{Q} = \mathsf{Uniform}(0, 1)$ and (b) $\mathsf{Q} = \mathsf{N}(0, 1)$.

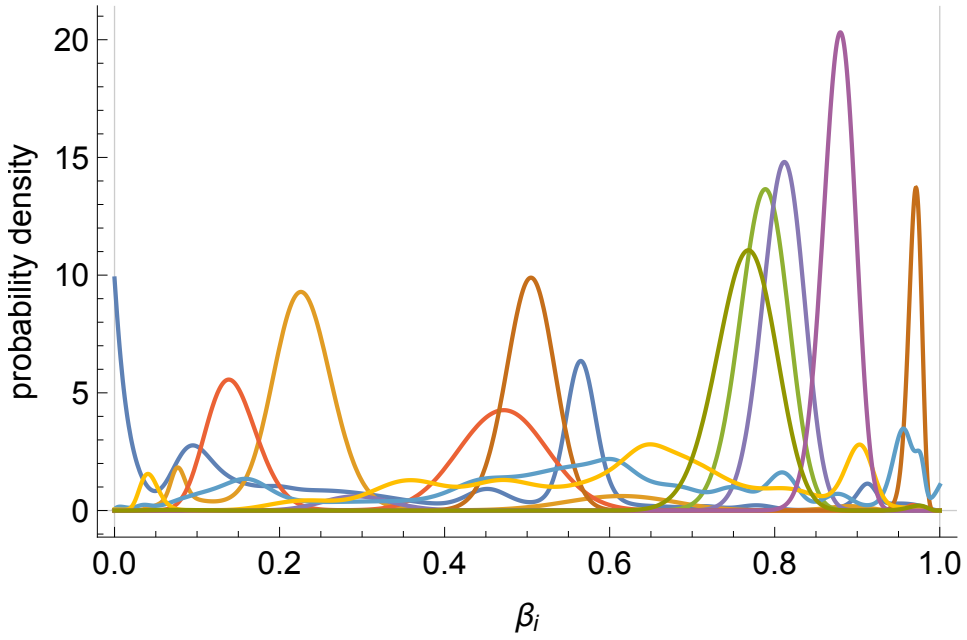

FIGURE 2. Illustrating one aspect of an open-minded prior: $p(x_i|\psi)$ is plotted for each of ten draws from $p(\psi)$, assuming $\mathsf{Q} = \mathsf{Uniform}(0, 1)$.

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

$$Q(x_i) := \big(Q_1(x_{i1}), \ldots, Q_d(x_{id})\big) \tag{2.27a}$$

$$q(x_i) := \prod_{\ell=1}^{d} q_\ell(x_{i\ell}), \tag{2.27b}$$

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

$$f(x_i|\theta_c) = \prod_{\ell=1}^{d} \mathsf{Beta}\big(Q_\ell(x_{i\ell})|j_{c\ell}, k_{c\ell} - j_{c\ell} + 1\big)\, q_\ell(x_{i\ell}). \tag{2.28}$$

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 $k_{c\ell} \sim P_{k_\ell}$. Consequently,

$$p(x_i) = q(x_i). \tag{2.29}$$

The priors for w and α are unchanged.

Simple adaptations of the sampling scheme described in Section 3 allow one to make draws of ψ 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 $\big(\alpha/(1+\alpha)\big)^m$ 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:

$$p(z_{1:n}|x_{1:n}, w, \theta, \alpha) = \prod_{i=1}^{n} p(z_i|x_i, w, \theta) \tag{3.1a}$$

$$p(w|x_{1:n}, z_{1:n}, \theta, \alpha) = p(w|z_{1:n}, \alpha) \tag{3.1b}$$

$$p(\theta|x_{1:n}, z_{1:n}, w, \alpha) = \prod_{c=1}^{m} p(\theta_c|x^c) \tag{3.1c}$$

$$p(\alpha|x_{1:n}, z_{1:n}, w, \theta) = p(\alpha|z_{1:n}), \tag{3.1d}$$

where x^c is the collection of observations for which $z_i = c$. Let $\mathcal{I}_c = \{i : z_i = c\}$ and let $n_c = |\mathcal{I}_c|$. Note $\sum_{c=1}^{m} n_c = n$. We say that a cluster is not occupied if $\mathcal{I}_c = \varnothing$ (i.e., if $n_c = 0$).

The conditional distribution for z_i is characterized by

$$p(z_i = c|x_{1:n}, w, \theta) \propto w_c\, f(x_i|\theta_c), \tag{3.2}$$

for $c = 1, \ldots, m$. The weights w can be updated by updating the stick-breaking weights v via

$$v_c | z_{1:n} \sim \mathsf{Beta}\big(1 + n_c, \alpha + \textstyle\sum_{c'=c+1}^{m} n_{c'}\big) \tag{3.3}$$

for $c = 1, \ldots, m-1$.

Regarding the concentration parameter, note that

$$p(\alpha | z_{1:n}) \propto p(z_{1:n} | \alpha)\, p(\alpha) \propto \frac{\alpha^h\, \Gamma(\alpha)}{\Gamma(n+\alpha)}\, p(\alpha), \tag{3.4}$$

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.

Sampler for θ_c. The sampler for $\theta_c | x^c$ comprises three cases that depend on n_c. In the first two cases the draws are made directly from the posterior distribution. The third case involves a Metropolis–Hastings step.

First, if $n_c = 0$, then draw (j_c, k_c) from its prior. Second, if $n_c = 1$, then follow this scheme: Draw k_c from its prior and then draw j_c according to

$$j_c - 1 \sim \mathsf{Binomial}\big(k_c - 1, Q(x_i)\big), \tag{3.5}$$

where x_i denotes the sole occupant of cluster c. The justification for this scheme is as follows:

$$p(j_c, k_c | x_i) = \frac{f(x_i | j_c, k_c)\, p(j_c | k_c)\, p(k_c)}{p(x_i)} = p(j_c | k_c, x_i)\, p(k_c), \tag{3.6}$$

where $p(x_i) = q(x_i)$ and

$$p(j_c | k_c, x_i) = \frac{\mathsf{Beta}\big(Q(x_i) | j_c, k_c - j_c + 1\big)}{k_c} = \mathsf{Binomial}\big(j_c - 1 | k_c - 1, Q(x_i)\big). \tag{3.7}$$

Third, if $n_c \geq 2$, then adopt a Metropolis–Hastings scheme. Consider the following proposal conditional on $\theta_c = (j_c, k_c)$:

$$k_c' - 1 \sim \mathsf{Poisson}(k_c) \tag{3.8}$$

$$j_c' - 1 \sim \mathsf{Binomial}\big(k_c' - 1, \overline{\widehat{x}^c}\big), \tag{3.9}$$

where $\overline{\widehat{x}^c}$ is the sample mean of the transformed observations $\widehat{x}_i = Q(x_i)$ for $x_i \in x^c$. Let

$$q(\theta_c' | \theta_c, \widehat{x}) = \mathsf{Poisson}(k_c' - 1 | k_c)\, \mathsf{Binomial}(j_c' - 1 | k_c' - 1, \widehat{x}). \tag{3.10}$$

Then

$$\theta_c^{(r+1)} = \begin{cases} \theta_c' & \mathcal{M}_c^{(r)} \geq u^{(r+1)} \\ \theta_c^{(r)} & \text{otherwise} \end{cases}, \tag{3.11}$$

where $u^{(r+1)} \sim \mathsf{Uniform}(0, 1)$ and

$$\mathcal{M}_c^{(r)} = \frac{p(\theta_c' | x^c)}{p(\theta_c^{(r)} | x^c)} \times \frac{q\big(\theta_c^{(r)} | \theta_c', \overline{\widehat{x}^c}^{(r)}\big)}{q\big(\theta_c' | \theta_c^{(r)}, \overline{\widehat{x}^c}^{(r)}\big)}. \tag{3.12}$$

Neal's Algorithm 2. 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 (as just shown) to draw $\theta_c | x_i$ to populate a newly-created cluster.

Transformation to the unit interval. Note that $x \mapsto Q(x)$ 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

$$\widehat{x}_i = Q(x_i). \tag{3.13}$$

The kernel for the transformed observations is

$$\widehat{f}(\widehat{x}_i|\theta_c) = \mathsf{Beta}(\widehat{x}_i|j_c, k_c - j_c + 1). \tag{3.14}$$

The rest of the model is unchanged. Given draws $\{(\psi^{(r)}, z_{1:n}^{(r)}, \alpha^{(r)})\}_{r=1}^R$ conditioned on $\widehat{x}_{1:n}$, we may compute the predictive distribution for x_{n+1} directly from (3.15) or (3.18) below.

Posterior predictive distribution. Given draws $\{\psi^{(r)}\}_{r=1}^R$ from $p(\psi|x_{1:n})$, the posterior predictive distribution can be approximated via

$$p(x_{n+1}|x_{1:n}) \approx \frac{1}{R}\sum_{r=1}^{R} p(x_{n+1}|\psi^{(r)}) = \frac{1}{R}\sum_{r=1}^{R}\sum_{c=1}^{m} w_c^{(r)}\, f(x_{n+1}|\theta_c^{(r)}). \tag{3.15}$$

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

$$E[w_c|z_{1:n}, \alpha] = \frac{n_c}{n+\alpha} \qquad \text{for } c \in \mathcal{C} \tag{3.16a}$$

$$E\Big[\sum_{c\notin\mathcal{C}} w_c \,\Big|\, z_{1:n}, \alpha\Big] = \frac{\alpha}{n+\alpha}. \tag{3.16b}$$

In addition, for each $c \notin \mathcal{C}$ we can use $p(\theta_c)$ to integrate out θ_c, thereby replacing $f(x_{n+1}|\theta_c)$ with $q(x_{n+1})$. Consequently, the generic distribution can be expressed conditionally as[9]

$$p(x_{n+1}|z_{1:n}, \{\theta_c\}_{c\in\mathcal{C}}, \alpha) = \sum_{c\in\mathcal{C}} \frac{n_c}{n+\alpha}\, f(x_{n+1}|\theta_c) + \frac{\alpha}{n+\alpha}\, q(x_{n+1}). \tag{3.17}$$

The approximation is given by

$$\begin{aligned} p(x_{n+1}|x_{1:n}) &\approx \frac{1}{R}\sum_{r=1}^{R} p(x_{n+1}|z_{1:n}^{(r)}, \{\theta_c^{(r)}\}_{c\in\mathcal{C}^{(r)}}, \alpha^{(r)}) \\ &= \frac{1}{R}\sum_{r=1}^{R}\Bigg(\sum_{c\in\mathcal{C}^{(r)}} \frac{n_c^{(r)}}{n+\alpha^{(r)}}\, f(x_{n+1}|\theta_c^{(r)})\Bigg) + q(x_{n+1})\,\frac{1}{R}\sum_{r=1}^{R}\frac{\alpha^{(r)}}{n+\alpha^{(r)}}. \end{aligned} \tag{3.18}$$

4. 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

$$p(Y_i|x_i) \tag{4.1}$$

[9]This representation is associated with the Chinese Restaurant Process. It plays a central role in some samplers such as Algorithm 2 in Neal (2000).

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 $p(Y_i|x_i)$.[10] 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

$$p(Y_{1:n}|x_{1:n}) = \prod_{i=1}^{n} p(Y_i|x_i). \tag{4.2}$$

In this setting, the object of interest is

$$\begin{aligned} p(x_{n+1}|Y_{1:n}) &= \int p(x_{n+1}|x_{1:n})\, p(x_{1:n}|Y_{1:n})\, dx_{1:n} \\ &= \int p(x_{n+1}|\psi)\, p(\psi|Y_{1:n})\, d\psi. \end{aligned} \tag{4.3}$$

The right-hand side of the first line in (4.3) expresses the distribution in terms of latent variable density estimation: $p(x_{n+1}|x_{1:n})$ is what we would calculate if $x_{1:n}$ were observd and $p(x_{1:n}|Y_{1:n})$ is what is known about $x_{1:n}$ given what is actually observed. The second line in (4.3) 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 (4.2) 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 $x_{1:n}$ for each sweep of the sampler.[11] Let θ_i denote θ_{z_i}. Since the joint likelihood factors [see (4.2)], the full conditional posterior for x_i reduces to the posterior for x_i in isolation (conditional on θ_i):

$$p(x_i|Y_{1:n}, x_{1:n}^{-i}, \psi, z_{1:n}) = p(x_i|Y_i, \theta_i), \tag{4.4}$$

where

$$p(x_i|Y_i, \theta_i) = \frac{p(Y_i|x_i)\, f(x_i|\theta_i)}{\int p(Y_i|x_i)\, f(x_i|\theta_i)\, d\theta_i}. \tag{4.5}$$

Again it is convenient to transform $x_{1:n}$ to the unit interval via $\widehat{x}_i = Q(x_i)$ and use the transformed kernel (3.14). Then

$$p(\widehat{x}_i|Y_i, \theta_i) \propto p(Y_i|\widehat{x}_i)\, \widehat{f}(\widehat{x}_i|\theta_i), \tag{4.6}$$

where

$$p(Y_i|\widehat{x}_i) = p(Y_i|x_i)|_{x_i = Q^{-1}(\widehat{x}_i)}. \tag{4.7}$$

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,

$$\begin{aligned} p(x_i|Y_{1:n}) &= \int p(x_i|Y_i, \theta_i)\, p(\theta_i|Y_{1:n})\, d\theta_i \\ &\approx \frac{1}{R} \sum_{r=1}^{R} p(x_i|Y_i, \theta_i^{(r)}). \end{aligned} \tag{4.8}$$

[10] This assumption is solely for expositional simplicity. Any nuisance parameters may be retained and sampled.

[11] In some cases it is possible to integrate $x_{1:n}$ out analytically. See Section 5 and Appendix E for examples.

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

Referring to (4.3), the generic distribution can be approximated by

$$p(x_{n+1}|Y_{1:n}) \approx \frac{1}{R}\sum_{r=1}^{R} p(x_{n+1}|\psi^{(r)}), \tag{4.9}$$

where $\{\psi^{(r)}\}_{r=1}^R$ are draws from $p(\psi|Y_{1:n})$. Referring to (3.17), the generic distribution can also be expressed as

$$p(x_{n+1}|Y_{1:n}) \approx \frac{1}{R}\sum_{r=1}^{R} p(x_{n+1}|z_{1:n}^{(r)}, \theta^{(r)}, \alpha^{(r)}), \tag{4.10}$$

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

Sharing, shrinkage, and pooling. The ways in which α and k_c affect the variation of $p(x_i|\psi)$ around the prior predictive distribution $q(x_i)$ were discussed in Section 2. In that section, it was assumed that $x_{1:n}$ is observed. At this point it is convenient to make explicit the effect of the concentration parameter on the notion of *sharing* (which could have been done in Section 2). In particular, the concentration parameter α has an effect on the extent to which observations share mixture components (also known as clusters). When α is small, w 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.

In the current section, by contrast, it is assumed that $x_{1:n}$ is latent. Consequently, the prior will have an effect on the posterior distribution of $x_{1:n}$. In this setting, it makes sense to talk additionally about *shrinkage* and *pooling*.

One may interpret the model in terms of partial sharing of the parameters. Whenever a parameter θ_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 on page 12 for the complete set of relationships. Table 1 refers to the Dirichlet Process (DP) model which is described in Section 5.

5. Dirichlet Process (DP) model as special case

In this section we examine the effect of large values of k_c. A prior that puts most of its weight on large values of k_c will enhance the variation. In order to examine the limiting case, it is convenient to change variables from (j_c, k_c) to (ϕ_c, k_c) where $\phi_c = Q^{-1}(j_c/k_c)$. Note that j_c/k_c converges in distribution to $\mathsf{Uniform}(0,1)$ as $k_c \to \infty$. Therefore, ϕ_c converges in distribution to Q. The kernel can be expressed in terms of this parameterization:

$$\widetilde{f}(x_i|\phi_c, k_c) = f(x_i|j_c = Q(\phi_c)\,k_c, k_c) = \mathsf{Beta}(Q(x_i)|Q(\phi_c)\,k_c, k_c - Q(\phi_c)\,k_c + 1)\,q(x_i). \tag{5.1}$$

TABLE 1. **Sharing, shrinkage, and pooling.** Sharing is controlled by the concentration parameter α. 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). There is no pooling if either $\alpha = \infty$ or $k_c = 1$. The Dirichlet Process (DP) and Dirichlet Process Mixture (DPM) are nonparametric priors.

Local Shrinkage (controlled by k_c)	Sharing (controlled by α)		
	complete ($\alpha = 0$)	partial	none ($\alpha = \infty$)
complete ($k_c = \infty$)	complete pooling	DP	no pooling
partial	partial pooling	DPM	no pooling
none ($k_c = 1$)	no pooling	no pooling	no pooling

The kernel collapses to a point mass located at ϕ_c as k_c gets large:

$$\lim_{k_c \to \infty} \widetilde{f}(x_i|\phi_c, k_c) = \delta(x_i - \phi_c), \tag{5.2}$$

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. We can express this limiting case as

$$p(x_i|\psi) = \sum_{c=1}^{\infty} w_c\, \delta(x_i - \phi_c). \tag{5.3}$$

Indirect density estimation. Here we apply the DP to the case of indirect density estimation. (Some authors refer to this as a "DPM model.")

It is convenient to integrate out x_i:

$$p(Y_i|\psi) = \int p(Y_i|x_i)\, p(x_i|\psi)\, dx_i = \sum_{c=1}^{\infty} w_c\, p(Y_i|\phi_c), \tag{5.4}$$

where

$$p(Y_i|\phi_c) = \int p(Y_i|x_i)\, \delta(x_i - \phi_c)\, dx_i = p(Y_i|x_i)|_{x_i = \phi_c}. \tag{5.5}$$

In (5.4), $p(Y_i|\phi_c)$ 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).

Recal $w \sim \mathsf{Stick}(\alpha)$ and $\phi_c \overset{\text{iid}}{\sim} \mathsf{Q}$. We can classify observations according to $z_c \propto w_c\, p(Y_i|\phi_c)$. Draws of w and α are unchanged. Regarding draws of ϕ_c, note that

$$p(\phi_c|Y_{1:n}, z_{1:n}) \propto q(\phi_c) \prod_{i \in \mathcal{I}_c} p(Y_i|\phi_c). \tag{5.6}$$

If $q(\phi_c)$ is a conjugate prior, then there is a closed-form expression for $p(\phi_c|Y_{1:n}, z_{1:n})$.

The specific distributions can be approximated via $\{x_i^{(r)}\}_{r=1}^R$ where $x_i^{(r)} = \phi_{z_i^{(r)}}^{(r)}$ and the generic distribution can be approximated via $\{x_{n+1}^{(r)}\}_{r=1}^R$ where

$$x_{n+1}^{(r)} \sim \sum_{c=1}^{m} w_c^{(r)}\, \delta_{\phi_c^{(r)}}. \tag{5.7}$$

Smoother approximations may be obtained as follows. For the specific case, since $x_i = \phi_{z_i}$, we have $p(x_i|Y_{1:n}, z_{1:n}) = p(\phi_{z_i} = x_i|Y_{1:n}, z_{1:n})$, so that

$$p(x_i|Y_{1:n}) \approx \frac{1}{R}\sum_{r=1}^{R} p\Big(\phi_{z_i^{(r)}} = x_i \,|\, Y_{1:n}, z_{1:n}^{(r)}\Big). \tag{5.8}$$

A smoother approximation for the generic distribution can be based on

$$\begin{aligned} p(x_{n+1}|Y_{1:n}, z_{1:n}, w) &= \int p(x_{n+1}|w,\phi)\, p(\phi|Y_{1:n}, z_{1:n})\, d\phi \\ &= \sum_{c=1}^{m} \int w_c\, \delta(x_{n+1} - \phi_c)\, p(\phi_c|Y_{1:n}, z_{1:n})\, d\phi_c \\ &= \sum_{c=1}^{m} w_c\, p(\phi_c = x_{n+1}|Y_{1:n}, z_{1:n}). \end{aligned} \tag{5.9}$$

Therefore,

$$p(x_{n+1}|Y_{1:n}) \approx \frac{1}{R}\sum_{r=1}^{R}\sum_{c=1}^{m} w_c^{(r)}\, p(\phi_c = x_{n+1}|Y_{1:n}, z_{1:n}^{(r)}). \tag{5.10}$$

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

$$p(Y_{1:n}|M) = p(Y_1|M)\prod_{i=2}^{n} p(Y_i|Y_{1:i-1}, M), \tag{5.11}$$

where

$$p(Y_i|Y_{1:i-1}, M) = \int p(Y_i|x_i)\, p(x_i|Y_{1:i-1}, M)\, dx_i. \tag{5.12}$$

In general these integrals can be computed via numerical quadrature.

Binomial data. We now illustrate the use of the DP prior for indirect density estimation in conjunction with the Binomial likelihood.[12]

Let

$$p(Y_i|x_i) = \mathsf{Binomial}(s_i|T_i, x_i), \tag{5.13}$$

where T_i is the number of trials, s_i is the number of successes, and x_i is the probability of success. Referring to (5.5),

$$p(Y_i|\phi_c) = \mathsf{Binomial}(s_i|T_i, \phi_c). \tag{5.14}$$

We can classify observations according to $z_i \propto w_c\, \mathsf{Binomial}(s_i|T_i, \phi_c)$ and utilize the sampler described in Section 3 for w and α.

We now turn to ϕ_c. If we assume

$$q(\phi_c) = \mathsf{Beta}(\phi_c|a_0, b_0), \tag{5.15}$$

[12]See Greenberg (2013) and Geweke et al. (2011) for textbook treatments of the DP with a binomial likelihood. These treatments differ in two ways from what is presented here. First they use a marginalized version of the model for sampling based on the Chinese Restaurant Process (CRP), along the lines of Algorithm 2 in Neal (2000). Second, this paper provides smoother posterior distributions for both specific and generic cases.

then

$$p(\phi_c|Y_{1:n}, z_{1:n}) = \mathsf{Beta}(\phi_c|A_c, B_c), \tag{5.16}$$

where

$$A_c := a_0 + \sum_{\ell \in \mathcal{I}_c} s_\ell \qquad \text{and} \qquad B_c := b_0 + \sum_{\ell \in \mathcal{I}_c} T_\ell - s_\ell. \tag{5.17}$$

We can sample ϕ_c from (5.16).

Specific and generic distributions. Since $x_i = \phi_{z_i}$ [and referring to (5.16)], we have

$$p(x_i|Y_{1:n}, z_{1:n}) = p(\phi_{z_i} = x_i|Y_{1:n}, z_{1:n}) = \mathsf{Beta}(x_i|A_{z_i}, B_{z_i}), \tag{5.18}$$

where A_{z_i} and B_{z_i} are given in (5.17) with $c = z_i$. Therefore,

$$p(x_i|Y_{1:n}) \approx \frac{1}{R}\sum_{r=1}^{n} p(x_i|Y_{1:n}, z_{1:n}^{(r)}) = \frac{1}{R}\sum_{r=1}^{R} \mathsf{Beta}(x_i|A_{z_i}^{(r)}, B_{z_i}^{(r)}). \tag{5.19}$$

Turning to the generic case, we have

$$p(x_{n+1}|Y_{1:n}) \approx \frac{1}{R}\sum_{r=1}^{R}\sum_{c=1}^{m} w_c^{(r)}\, \mathsf{Beta}(x_{n+1}|A_c^{(r)}, B_c^{(r)}). \tag{5.20}$$

In the spirit of (3.17), a smoother approximation is

$$p(x_{n+1}|Y_{1:n}) \approx \frac{1}{R}\sum_{r=1}^{R}\sum_{c \in \mathcal{C}^{(r)}} \frac{n_c^{(r)}}{n + \alpha^{(r)}}\, \mathsf{Beta}(x_{n+1}|A_c^{(r)}, B_c^{(r)}) + \frac{\alpha^{(r)}}{n + \alpha^{(r)}}\, \mathsf{Beta}(x_{n+1}|a_0, b_0). \tag{5.21}$$

(This is a weighted average of the specific distributions mixed with the prior predictive distribution.)

6. 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. Recall the prior has three components: the prior predictive distribution $q(x_i)$, the prior for k_c, the prior for α.

The prior predictive distributions are all flat. (It may be useful to redo the estimation with other prior predictive distributions.)

Regarding the prior for k_c, 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 α is given by $p(\alpha) = 1/(1+\alpha)^2$ so that the prior median for α 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 3.

Galaxy data. Figure 4 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 5 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 5 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 α is deemed unlikely according the model when we learn about α. The posterior median for α is about 0.31. The posterior probability of $\alpha \geq 5$ is about 20 times lower than the prior probability. (Increasing α 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 α and k_c. The posterior median of 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 6 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. Figure 7 shows the marginal predictive distributions computed from the joint distribution (along with rug plots of the data).

7. Investigation: Part II

In this section I apply the model of indirect density estimation to a number of applications, including rat tumor data, some baseball data, and the thumbtack data.

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 9. 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 8 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. Comparison with Petrone's model

Petrone's model can be expressed as follows:

$$p(x_i|\pi_k, k) = \sum_{j=1}^{k} \pi_{jk}\, \mathsf{Beta}(x_i|j, k-j+1), \tag{A.1}$$

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

$$\pi_k|k \sim \mathsf{Dirichlet}(\alpha\, \underline{\pi}_k) \tag{A.2a}$$

$$k \sim P_k, \tag{A.2b}$$

where $\alpha\, \underline{\pi}_k = (\alpha\, \underline{\pi}_{1k}, \ldots, \alpha\, \underline{\pi}_{kk})$ and where $\underline{\pi}_k \in \Delta^{k-1}$.

Compare (A.1)–(A.2) with (2.5) and (2.10)–(2.11), 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 event, the two models can be compared in terms of their marginal likelihoods. The marginal likelihoods can be computed from the sequence of predictive distributions:

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

where A_m stands for the assumptions of model m. Let $m = 1$ indicate Petrone's model and let $m = 2$ indicate my model.

Reformulation. Petrone reformulates the collection of Dirichlet distributions indexed by k in terms of a single underlying Dirichlet process (DP). This reformulation facilitates sampling by removing the effect of k on the "dimension" of the parameter, thereby obviating the need for inference methods such as reversible jump MCMC [see Green (1995)].

Let $G \sim \mathsf{DP}(\alpha, H)$ where the support of the base distribution H is $\mathcal{B} = (0, 1]$. G has the following representation: $G = \sum_{c=1}^{\infty} \omega_c\, \delta_{\xi_c}$, where $\omega \sim \mathsf{Stick}(\alpha)$, $\xi_c \overset{\text{iid}}{\sim} H$, and δ_x is a point mass located at x.

Let $\mathcal{B}_k = \{\mathcal{B}_{jk}\}_{j=1}^{k}$ denote a partition of $\mathcal{B}$ where

$$\mathcal{B}_{jk} = \left(\frac{j-1}{k}, \frac{j}{k}\right]. \tag{A.4}$$

Then, by the central property of DPs,

$$\big(G(\mathcal{B}_{1k}), \ldots, G(\mathcal{B}_{kk})\big) \sim \mathsf{Dirichlet}\big(\alpha\, H(\mathcal{B}_{1k}), \ldots, \alpha\, H(\mathcal{B}_{kk})\big), \tag{A.5}$$

where $H(\mathcal{B}_{jk}) = \int_0^1 1(\xi \in \mathcal{B}_{jk})\, dH(\xi)$ and

$$G(\mathcal{B}_{jk}) = \sum_{c=1}^{\infty} 1\big(\xi_c \in \mathcal{B}_{jk}\big)\, \omega_c. \tag{A.6}$$

Letting $\pi_{jk} = G(\mathcal{B}_{jk})$ and subject to the substantive restriction $\underline{\pi}_{jk} = H(\mathcal{B}_{jk})$, we obtain (A.2a). As an example, let H be the uniform distribution. Then $H(\mathcal{B}_{jk}) = 1/k$.

The classification of observation i is facilitated via a latent variable $\zeta_i \sim H$. Define $J(\zeta, k) = \lceil \zeta\, k \rceil$, where $\lceil x \rceil$ is the "ceiling" of x (i.e., the smallest integer greater than or equal to x). Let $z_i^k = J(\zeta_i, k) \in \{1, \ldots, k\}$. Also let $s_k = (s_{1k}, \ldots, s_{kk})$, where

$$s_{jk} = \sum_{i=1}^{n} 1\big(J(\zeta_i, k) = j\big). \tag{A.7}$$

Note $\sum_{j=1}^{k} s_{jk} = n$.

Given this setup (and assuming the base distribution H is uniform), a Gibbs sampler can be based on the following full conditional distributions:

$$p(z_i^k = j | k, \pi_{jk}, x_i) \propto \pi_{jk}\, \mathsf{Beta}(x_i | j, k - j + 1) \tag{A.8a}$$

$$p(\zeta_i | z_i^k = j, k) = \mathsf{Uniform}\big(\zeta_i | (j-1)/k, j/k\big) \tag{A.8b}$$

$$p(k | \zeta_{1:n}) \propto p(k) \prod_{i=1}^{n} \mathsf{Beta}(x_i | J(\zeta_i, k), k - J(\zeta_i, k) + 1) \tag{A.8c}$$

$$p(\pi_k | k, \zeta_{1:n}) = \mathsf{Dirichlet}(s_k + \alpha/k). \tag{A.8d}$$

Given R draws from the posterior distribution, the predictive distribution can be approximated by [see (A.1)]

$$p(x_{n+1} | x_{1:n}) \approx \frac{1}{R} \sum_{r=1}^{R} p(x_{n+1} | (\pi_k)^{(r)}, k^{(r)}) \tag{A.9}$$

For one-dimensional Old Faithful data, Petrone sets $K = 230$. If this were applied to the two-dimensional data, the number of components would be $K_1 \times K_2 = 230^2 = 52,900$, which makes this model effectively infeasible.

Sampling α. This is all conditional on α. Consequently, we can provide α with a prior and and sample it as well. The conditional posterior for α is

$$p(\alpha | k, s_k) \propto p(\alpha)\, \Gamma(n + \alpha) \prod_{j=1}^{k} \frac{\pi_{jk}^{s_{jk} + \alpha/k - 1}}{\Gamma(s_{jk} + \alpha/k)}. \tag{A.10}$$

However, there are issues with evaluating the likelihood given the small magnitude of some of the elements of π_k.

Appendix B. 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

$$p(x_i | -) = \sum_{c=1}^{\infty} w_c\, \mathsf{Beta}(x_i | j_c, k_c - j_c + 1) \tag{B.1a}$$

where

$$w_c = v_c \prod_{\ell=1}^{c-1} (1 - v_\ell) \tag{B.1b}$$

$$v_c \overset{\text{iid}}{\sim} \mathsf{Beta}(1, \alpha) \tag{B.1c}$$

$$k_c - 1 \overset{\text{iid}}{\sim} \mathsf{Geometric}(\xi) \tag{B.1d}$$

$$j_c | k_c \sim \mathsf{Uniform}(\{1, \ldots, k_c\}). \tag{B.1e}$$

(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

$$p(x_i|-) = \sum_{s=0}^{\infty} \sum_{h=1}^{2^s} \pi_{s,h} \, \mathsf{Beta}(x_i | h, 2^s - h + 1), \tag{B.2a}$$

where

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

$$S_{s,h} \sim \mathsf{Beta}(1, a) \tag{B.2c}$$

$$R_{s,h} \sim \mathsf{Beta}(b, b) \tag{B.2d}$$

and

> 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 $(r+1, g_{shr+1})$ 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

$$s \sim \mathsf{Geometric}\big(1/(1+a)\big) \tag{B.3}$$

$$h | s \sim \mathsf{Uniform}(\{1, \ldots, 2^s\}). \tag{B.4}$$

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 C. A different object of interest

In order to help clarify the nature of the predictive distribuition (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 $\widehat{g}(x) = p(x|\widehat{\psi})$, where $\widehat{\psi}$ is an estimate of ψ such as the maximum likelihood estimate:

$$\widehat{\psi} = \operatorname*{argmax}_{\psi} \; p(x_{1:n}|\psi). \tag{C.1}$$

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

$$x_{1:n}^{(r)} \sim p(x_{1:n}|\psi) \qquad \text{and} \qquad x_{1:n}^{(r)} \to \widehat{\psi}^{(r)} \to p(x|\widehat{\psi}^{(r)}), \tag{C.2}$$

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

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

$$\psi^{(r)} \sim p(\psi|x_{1:n}) \qquad \text{and} \qquad \psi^{(r)} \to p(x|\psi^{(r)}), \tag{C.3}$$

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 ψ:

$$\widetilde{g}(x) = \int p(x|\psi)\, p(\psi|x_{1:n})\, d\psi. \tag{C.4}$$

Although the Bayesian estimate $\widetilde{g}(x)$ 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 ψ due to its posterior distribution plays a role in *constructing* the predictive distribution, but such variation is not relevant once ψ 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 that 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

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

in which case $p(x_{n+2}|x_{1:n+1})$ will be used to compute the optimal decision. However, if x_{n+1} provides very little additional information about ψ, then $p(\psi|x_{1:n}, x_{n+1}) \approx p(\psi|x_{1:n})$ and

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

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

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. 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 density is given by[13]

$$c(u_1, u_2|\tau) := 1 + \tau\,(2\,u_1 - 1)\,(2\,u_2 - 1) \qquad \text{where } -1 \leq \tau \leq 1 \tag{D.1}$$

for $(u_1, u_2) \in [0,1]^2$. 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: $c(u_1, u_2|\tau = 0) = 1$. In addition, if

$$p(\tau) = \mathsf{Uniform}(\tau|-1, 1) = \frac{1}{2}\,1_{[-1,1]}(\tau), \tag{D.2}$$

then the expectation with respect to τ produces independence as well:

$$p(u_1, u_2) = \int_{-1}^{1} c(u_1, u_2|\tau)\,p(\tau)\,d\tau = 1. \tag{D.3}$$

Prior predictive distribution. Let

$$f_\ell(x_{i\ell}|j_{c\ell}, k_{c\ell}) = \mathsf{Beta}\big(Q_\ell(x_{i\ell})|j_{c\ell}, k_{c\ell} - j_{c\ell} + 1\big)\,q_\ell(x_{i\ell}) \tag{D.4}$$

and let $F_\ell(x_{i\ell}|j_{c\ell}, k_{c\ell})$ denote the associated CDF. Then let the joint kernel be given by

$$f(x_i|\theta_c) = c\big(F_1(x_{i1}|j_{c1}, k_{c1}), F_2(x_{i2}|j_{c2}, k_{c2})\,|\,\tau_c\big)\prod_{\ell=1}^{2} f_\ell(x_{i\ell}|j_{c\ell}, k_{c\ell}), \tag{D.5}$$

where $\theta_c = (j_{c1}, j_{c2}, k_{c1}, k_{c2}, \tau_c)$. Let the prior for θ_c be given by

$$p(\theta_c) = \frac{p(k_{c1})\,p(k_{c2})\,p(\tau_c)}{k_{c1}\,k_{c2}}, \tag{D.6}$$

where the prior for τ_c is given in (D.2). Then for any $p(k_{c1})$ and $p(k_{c1})$ we have

$$p(x_i) = \prod_{\ell=1}^{2} q_\ell(x_{i\ell}). \tag{D.7}$$

Algorithm 2 in Neal (2000). Algorithm 2 in Neal (2000) may used because (i) the prior predictive is known [see (D.7)] and (ii) it is possible to draw $\theta_c|x_i$. Regarding (ii), note

$$p(\theta_c|x_i) = \frac{f(x_i|\theta_c)\,p(\theta_c)}{p(x_i)} = p(k_c|x_i)\,p(j_c|x_i, k_c)\,p(\tau_c|x_i, j_c, k_c), \tag{D.8}$$

where

$$p(k_c|x_i) = \prod_{\ell=1}^{2} p(k_{c\ell}) \tag{D.9}$$

$$p(j_c|x_i, k_c) = \prod_{\ell=1}^{2} \frac{f_\ell(x_{i\ell}|j_{c\ell}, k_{c\ell})}{q_\ell(x_{i\ell})\,k_{c\ell}} = \prod_{\ell=1}^{2} \mathsf{Binomial}\big(j_{c\ell} - 1|k_{c\ell} - 1, Q_\ell(x_{i\ell})\big) \tag{D.10}$$

$$p(\tau_c|x_i, j_c, k_c) = \frac{1}{2}\,c\big(F_1(x_{i1}|j_{c1}, k_{c1}), F_2(x_{i2}|j_{c2}, k_{c2})\,|\,\tau_c\big). \tag{D.11}$$

[13]It is interesting to note that the FGM copula can be expressed in terms of a mixture of order-statistic-based copulas described on page 5. In particular, let

$$\widehat{c}(u_1, u_2|w) = w\,\widetilde{c}(u_1, u_2|2) + (1 - w)\,\widetilde{c}(1 - u_1, u_2|2) = 1 + (2\,w - 1)\,(1 - 2\,u_1)\,(1 - 2\,u_2).$$

Thus $\widehat{c}(u_1, u_2|w = (\tau + 1)/2) = c(u_1, u_2|\tau)$.

Thus, we can make a draw from 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 τ_c from its conditional distribution.

Appendix E. Binomial likelihood

A special case of some interest is when the data are binomial in nature and the object of interest is latent:

$$p(Y_i|x_i) = \mathsf{Binomial}(s_i|T_i, x_i), \tag{E.1}$$

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

$$p(x_i|Y_i, \theta_{z_i}) = \mathsf{Beta}(x_i|j_{z_i} + s_i, (k_{z_i} - j_{z_i} + 1) + T_i - s_i), \tag{E.2}$$

which can be used in (4.4) for sampling and in (4.8) for smoothing.

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

$$\begin{aligned} p(s_i|T_i, \theta_c) = p(s_i|T_i, j_c, k_c) &= \int_0^1 \mathsf{Binomial}(s_i|T_i, x_i)\, \mathsf{Beta}(x_i|j_c, k_c - j_c + 1)\, dx_i \\ &= \mathsf{Beta\text{-}Binomial}(s_i|j_c, k_c - j_c + 1, T_i) \\ &= \binom{T_i}{s_i} \frac{k_c!\,(s_i + j_c - 1)!\,(T_i + k_c - s_i - j_c)!}{(j_c - 1)!\,(k_c - j_c)!\,(T_i + k_c)!}. \end{aligned} \tag{E.3}$$

Note that

$$p(s_i|T_i, k_c) = \sum_{j_c=1}^{k_c} \frac{p(s_i|T_i, j_c, k_c)}{k_c} = \frac{1}{T_i + 1}, \tag{E.4}$$

which is independent of k_c. Therefore, $p(s_i|T_i) = 1/(T_i + 1)$, the uniform distribution over $s_i \in \{0, \ldots, T_i\}$.

We can again use Algorithm 2 from Neal (2000) to make draws from the posterior distribution since the prior predictive distribution is known and when $n_c = 1$

$$\begin{aligned} p(j_c, k_c|T_i, s_i) &= \frac{p(s_i|T_i, j_c, k_c)\, p(j_c, k_c)}{p(s_i|T_i)} \\ &= \left\{ \frac{(T_i + 1)\, \mathsf{Beta\text{-}Binomial}(s_i|j_c, k_c - j_c + 1, T_i)}{k_c} \right\} p(k_c) \\ &= \mathsf{Beta\text{-}Binomial}(j_c - 1|s_i + 1, T_i - s_i + 1, k_c - 1)\, p(k_c). \end{aligned} \tag{E.5}$$

Thus $k_c \sim p(k_c)$ and

$$j_c - 1 \sim \mathsf{Beta\text{-}Binomial}(s_i + 1, T_i - s_i + 1, k_c - 1). \tag{E.6}$$

This case when $n_c = 1$ suggests the a proposal for a Metropolis–Hastings scheme when $n_c \geq 2$:

$$k_c' - 1 \sim \mathsf{Poisson}(k_c) \tag{E.7}$$

$$j_c' - 1 \sim \mathsf{Beta\text{-}Binomial}(\overline{s^c} + 1, \overline{T^c} - \overline{s^c} + 1, k_c' - 1), \tag{E.8}$$

where $\overline{s^c}$ and $\overline{T^c}$ are sample means computed from the observations in cluster c.

References

Baker, R. (2008). An order-statistics-based method for constructing multivariate distributions with fixed marginals. *Journal of Multivariate Analysis 99*, 2312–2327.

Canale, A. (2017). msBP: An R package to perform Bayesian nonparametric inference using multiscale Bernstein polynomials mixtures. *Journal of Statistical Software 78*(6).

Canale, A. and D. B. Dunson (2016). Multiscale Bernstein polynomials for densities. *Statistica Sinica 26*(3), 1175–1195.

Efron, B. (2010). *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*. Cambridge University Press.

Efron, B. and C. Morris (1975). Data analysis using Stein's estimator and its generalizations. *Journal of the American Statistical Association 70*, 311–319.

Geenens, G. (2014). Probit transformation for kernel density estimation on the unit interval. *Journal of the American Statistical Association 109*(505), 346–358.

Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin (2014). *Bayesian data analysis* (Third ed.). CRC Press.

Gelman, A. and J. Hill (2007). *Data analsis using regression and multilevel/hierarchical models*. Cambridge University Press.

Geweke, J., G. Koop, and H. van Dijk (2011). *The Oxford Handbook of Bayesian Econometrics*. Oxford University Press.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika 82*, 711–132.

Greenberg, E. (2013). *Introduction to Bayesian Econometrics* (Second ed.). Cambridge University Press.

Ishwaran, H. and L. F. James (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association 96*, 161–173.

Kottas, A. (2006). Dirichlet process mixtures of beta distributions, with applications to density and intensity estimation. In *Workshop on Learning with Nonparametric Bayesian Methods*, 23rd International Conference on Machine Learning (ICML).

Liu, J. S. (1996). Nonparametric hierarchical Bayes via sequential imputations. *The Annals of Statistics 24*(3), 911–930.

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics 9*, 249–265.

Petrone, S. (1999a). Bayesian density estimation using Bernstein polynomials. *The Canadian Journal of Statisitcs 27*, 105–102.

Petrone, S. (1999b). Random Bernstein polynomials. *Scandinavian Journal of Statistics 26*, 373–393.

Petrone, S. and L. Wasserman (2002). Consistency of Bernstein polynomial posteriors. *Journal of the Royal Statistical Society. Series B (Statistical Methodology) 63*, 79–100.

Quintana, F. A., M. F. J. Steel, and J. T. A. S. Ferreira (2009). Flexible univariate continous distributions. *Bayesian Analysis 4*(4), 497–522.

Simonoff, J. S. (1996). *Smoothing Methods in Statistics*. Springer.

Trippa, L., P. Bulla, and S. Petrone (2011). Extended Bernstein prior via reinforced urn processes. *Annals of the Institute of Statistical Mathematics 63*, 481–496.

Wen, K. and X. Wu (2014). An improved transformation-based kernel estimator of densities on the unit interval. *Journal of the American Statistical Association 110*(510), 773–783. Accepted. DOI: 10.1080/01621459.2014.969426.

Zhao, Y., M. C. Ausín, and M. P. Wiper (2013). Bayesian multivariate Bernstain polynomial density estimation. Statistics and Econometrics Series 11 Working Paper 13-12, Universidad Carlos III de Madrid.

Federal Reserve Bank of Atlanta, Research Department, 1000 Peachtree Street N.E., Atlanta, GA 30309–4470

Email address: mark.fisher@atl.frb.org

URL: http://www.markfisher.net

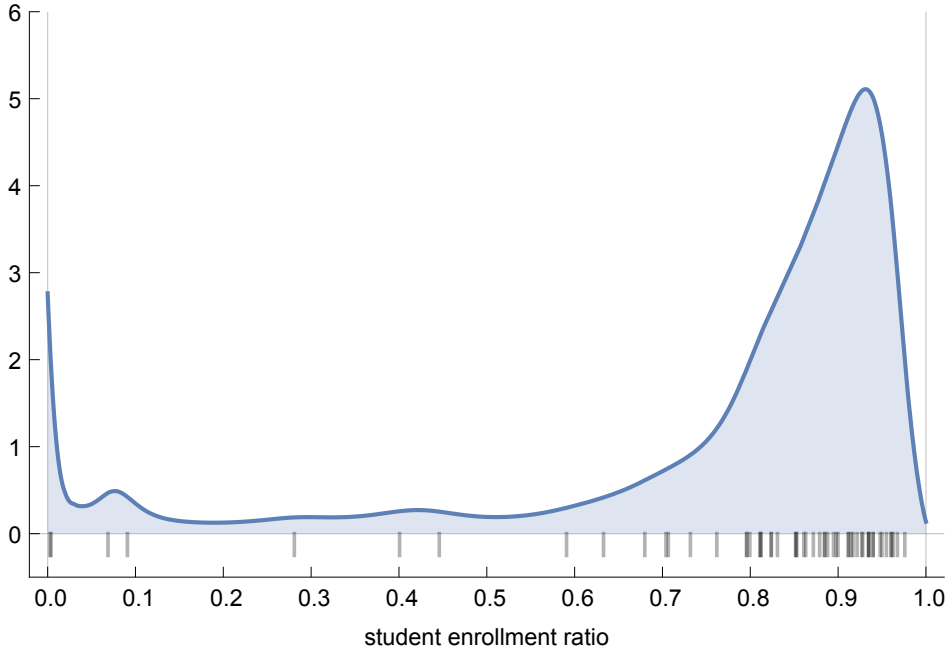

FIGURE 3. School data: Posterior predictive distribution.

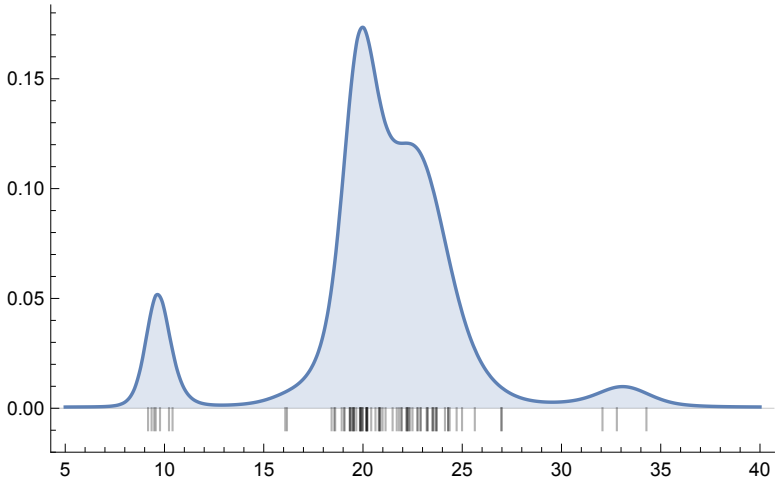

FIGURE 4. Galaxy data: quasi-Bernstein predictive density with support over the interval $[5, 40]$ and a rug plot of the data.

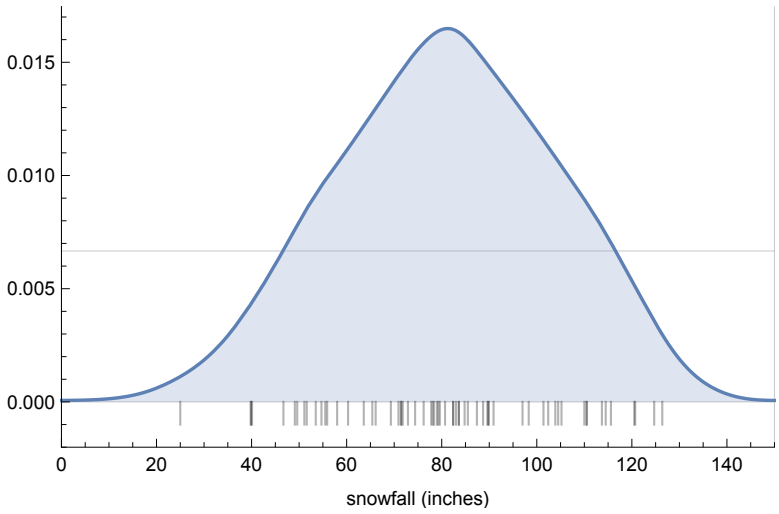

FIGURE 5. Buffalo snowfall data: quasi-Bernstein predictive density with support over the interval $[0, 150]$ and a rug plot of the data.

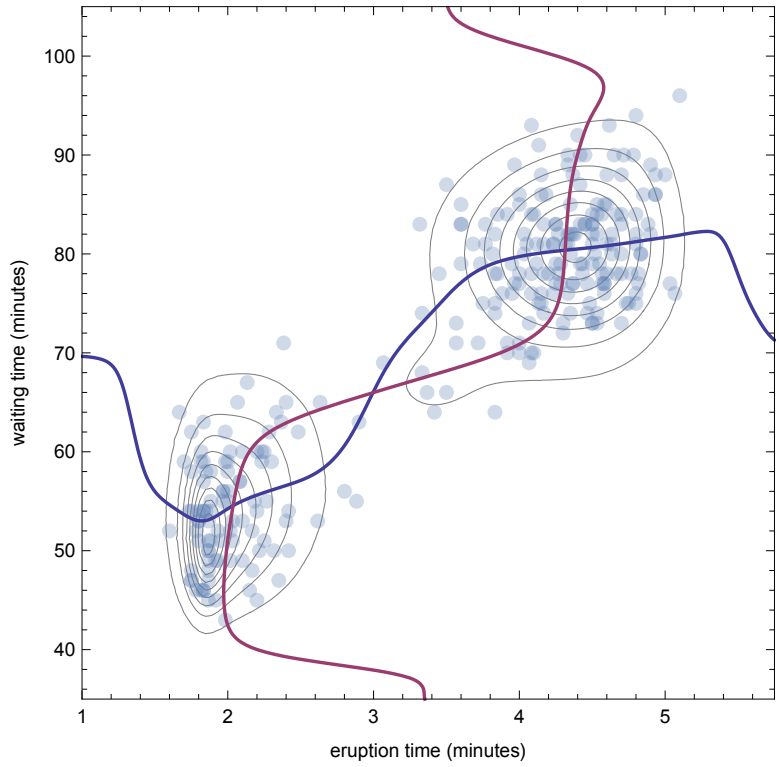

FIGURE 6. Old Faithful data: Contours for posterior predictive density with support over $[1, 5.75] \times [35, 105]$. Lowest contour is at the level of the uniform prior (≈ 0.003). Contour spacing above the lowest contour is ≈ 0.006. Data are shown as dots and conditional expectations are shown as thicker lines.

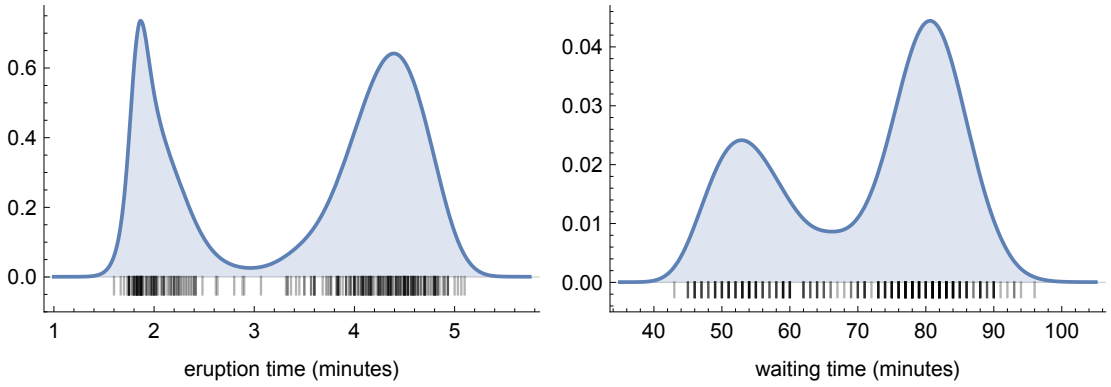

FIGURE 7. Old Faithful data: Marginal distributions for eruption time and waiting time computed from joint distribution.

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

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

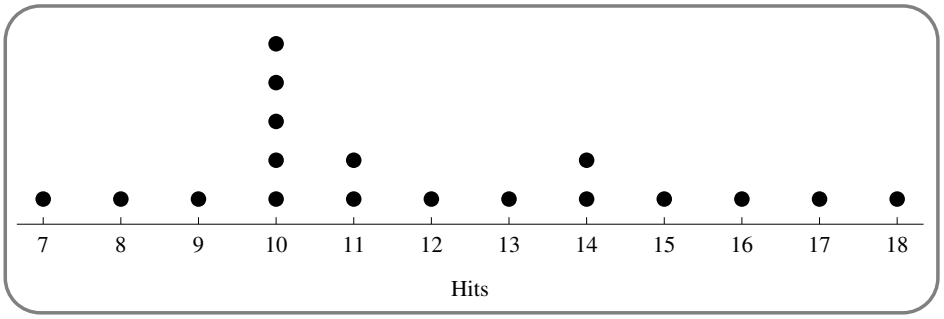

FIGURE 8. Baseball data: 18 players with 45 at-bats each.

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.

No. of successes	0	1	2	3	4	5	6	7	8	9	Total
No. of experiments	0	3	13	18	48	47	67	54	51	19	320
Frequency (percent)	0	1	4	6	15	15	21	17	16	6	≈ 100

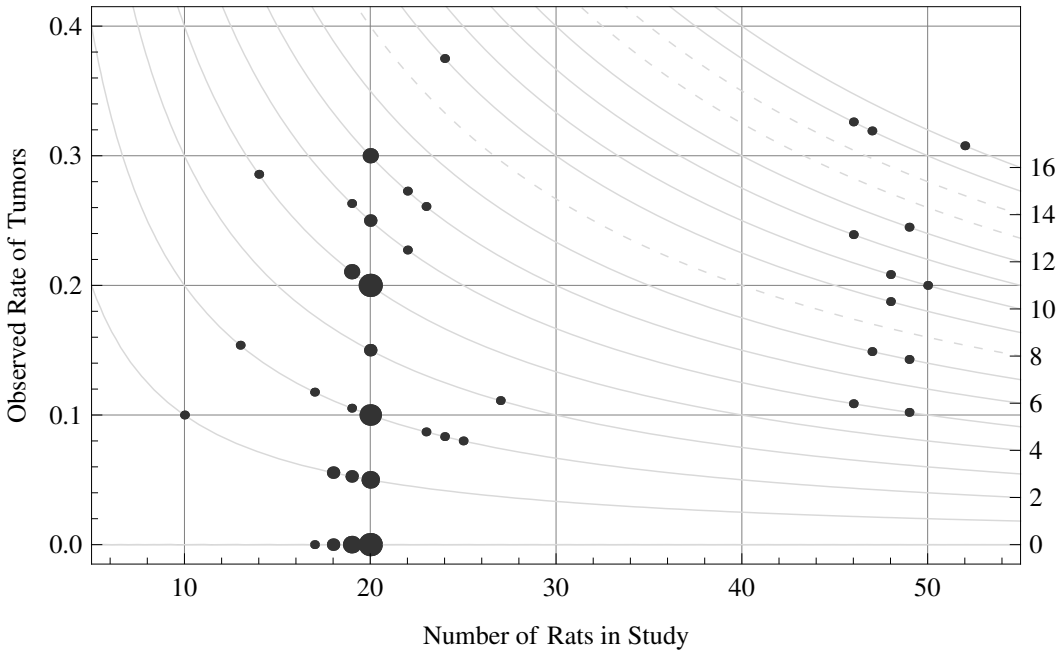

FIGURE 9. Rat tumor data: 71 studies. Number of studies (1 to 7) proportional to area of dot. Number of rats with tumors (0 to 16) indicated by contour lines. There are 59 studies for which the total number of rats is less than or equal to 35 and more than half of these studies (32) have observed tumor 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%.

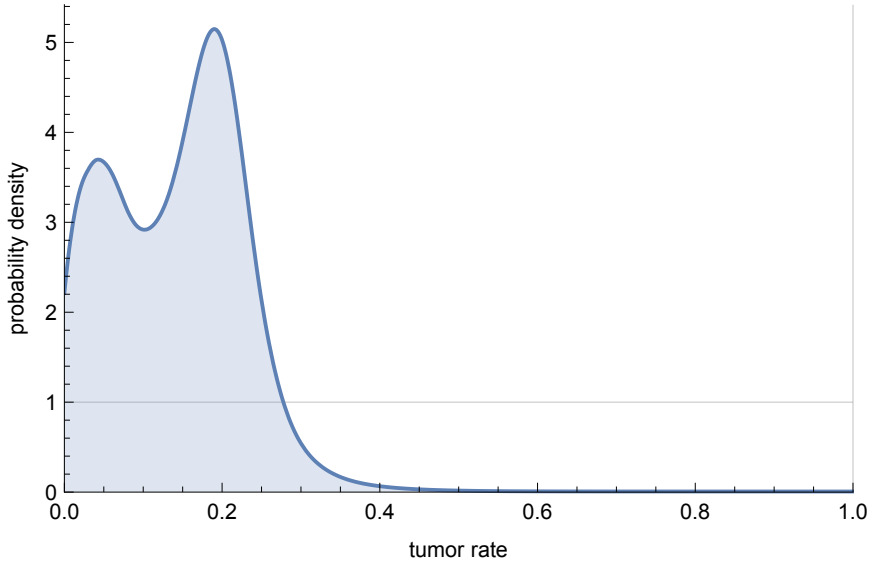

FIGURE 10. Posterior distribution for generic rat tumor rate.

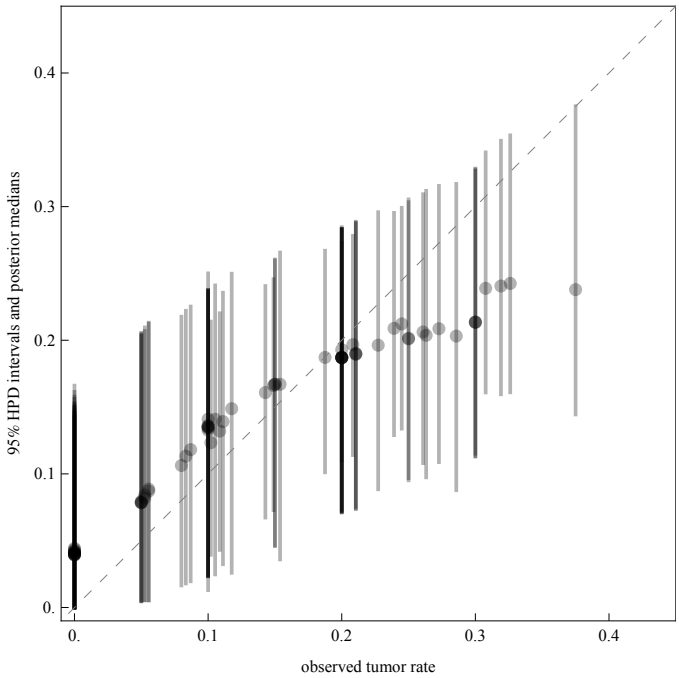

FIGURE 11. Posterior medians and 95% highest posterior density regions of rat tumor rates. Darker lines indicate multiple observations. Compare with Figure 5.4 in Gelman et al. (2014).

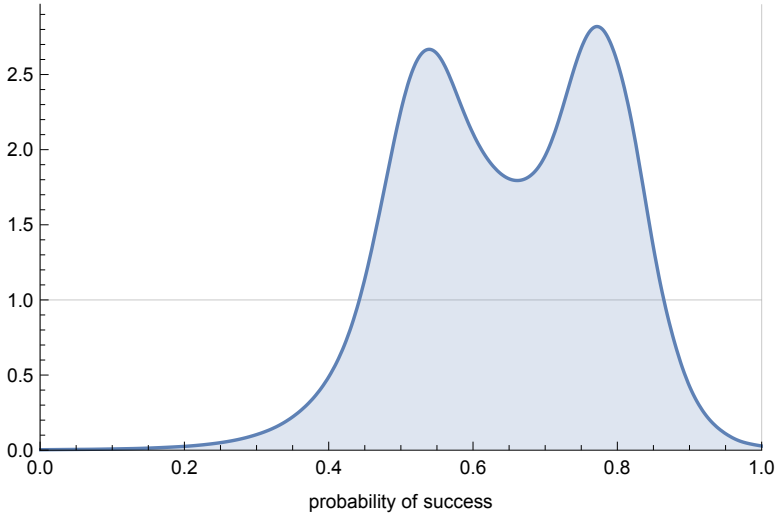

FIGURE 12. Thumbtack data: Posterior distribution for the generic probability of success.

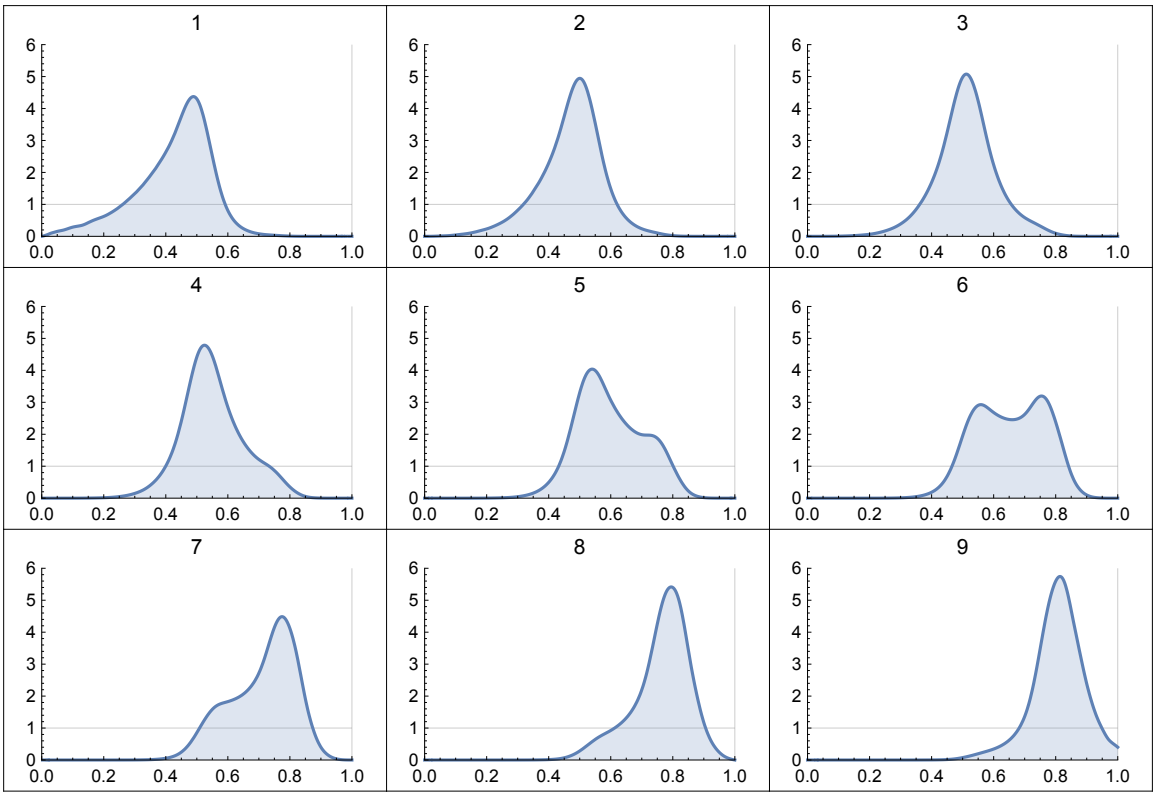

FIGURE 13. Thumbtack data: Posterior distributions for the specific probabilities of success, computed for each exchangeable group with a common number of success, 1 through 9.

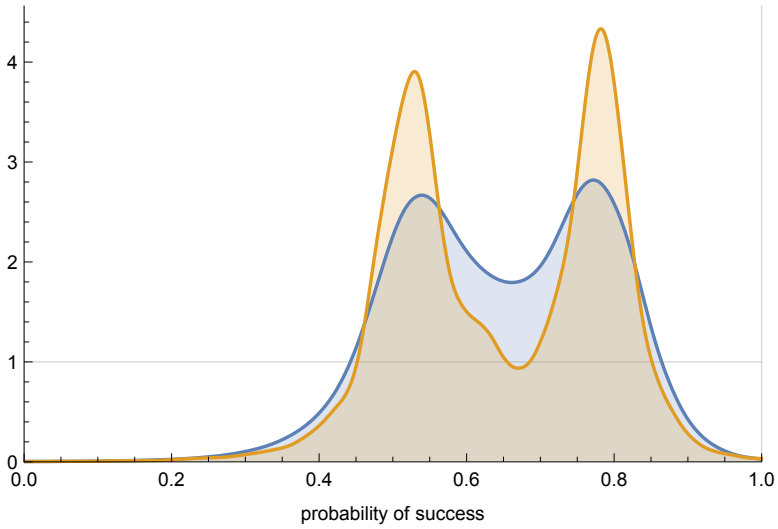

FIGURE 14. Thumbtack data: Posterior distribution for generic success rate given the alternative model compared with the distribution given the main model.