Communications for Statistical Applications and Methods
2016, Vol. 23, No. 6, 445–466

http://dx.doi.org/10.5351/CSAM.2016.23.6.445
Print ISSN 2287-7843 / Online ISSN 2383-4757

Nonparametric Bayesian methods: a gentle introduction and overview

Steven N. MacEachern[1,a]

[a]Department of Statistics, The Ohio State University, USA

Abstract

Nonparametric Bayesian methods have seen rapid and sustained growth over the past 25 years. We present a gentle introduction to the methods, motivating the methods through the twin perspectives of consistency and false consistency. We then step through the various constructions of the Dirichlet process, outline a number of the basic properties of this process and move on to the mixture of Dirichlet processes model, including a quick discussion of the computational methods used to fit the model. We touch on the main philosophies for nonparametric Bayesian data analysis and then reanalyze a famous data set. The reanalysis illustrates the concept of admissibility through a novel perturbation of the problem and data, showing the benefit of shrinkage estimation and the much greater benefit of nonparametric Bayesian modelling. We conclude with a too-brief survey of fancier nonparametric Bayesian methods.

Keywords: admissibility, dependent Dirichlet process, Dirichlet process, false consistency, Markov chain Monte Carlo, mixed model, shrinkage estimation

1. Motivation for nonparametric Bayesian methods

There are many ways to use data, whether experimental or observational, to better understand the world and to make better decisions. The Bayesian approach distinguishes itself from other approaches with two distinct sources of sound foundational support. The first is the theory of subjective probability, developed from a set of axioms that describe rational behavior. Subjective probability provides an alternative to the relative frequency definition of probability. Under subjective probability, individuals are free to have their own assessments of probabilities. This theory leads inexorably to Bayesian methods (Savage, 1954). The second source of support is decision theory which formalizes statistical inference as a decision problem. The combination of state-of-nature (parameter) and action (say, an estimate) yield a loss, and a good inference procedure (decision rule) leads to a small expected loss. Nearly all agree that inadmissible inference procedures are to be avoided. The complete class theorems show that the entire set of admissible inference procedures is comprised of Bayesian procedures and of procedures that are close to Bayesian in a technical sense (Berger, 1985). Procedures that are far from Bayesian can be useful, but they must be justified on special grounds–for example, our inability to discover a dominating procedure, our inability to implement the dominating procedure due to computational limitations, or to address robustness issues, perhaps due to the shortcomings of our formal mathematical model.

The twin perspectives of subjective probability and decision theory have convinced many that statistical inference should be driven by Bayesian methods. However, neither of these perspectives

[1] Department of Statistics, The Ohio State University, 281 W. Lane Ave., Columbus, OH 43210, USA.
E-mail: snm@stat.osu.edu

Published 30 November 2016 / journal homepage: http://csam.or.kr

describes *how* to conduct a sound Bayesian analysis. Surely, we have all seen examples of poor Bayesian analyses, and so we seek guiding principles to help in the use of Bayesian methods.

1.1. Consistency

Consistency is a fundamental principle of statistical inference. The simplest description of consistency is for a sequence of increasingly large random samples arising from a distribution. With such data, our inference ideally settles on the true distribution. This is captured through the usual definitions of consistency from classical statistics, where a consistent estimator of the parameter governing the distribution converges to the true parameter value in some sense.

The traditional definition of consistency has its parallel for subjective probabilists. Here, for consistency, the posterior distribution concentrates in each arbitrarily small neighborhood of the true parameter value. That is, for any given ϵ-neighborhood of the true θ, say $\mathcal{N}_\epsilon(\theta_T)$, the posterior probability of the neighborhood tends to 1. For decision-theoreticians, with mild conditions on the loss function, consistency becomes a statement that the posterior expected loss converges to the minimum loss given θ_T.

Bayesian purists have argued that Bayesian methods are always consistent–that the methods naturally lead to consistent estimation for any parameter value in the support of the prior distribution. While this is true under very mild conditions and with an appropriate definition of support, it begs an important question. If θ_T is omitted from the support of the prior distribution, the posterior distribution cannot concentrate at it, nor will one's loss typically converge to the minimum loss.

A sequence of coin flips illustrates this point in a simple, parametric setting.

$$\begin{aligned} \theta &\sim F \\ Y_i|\theta &\overset{ind}{\sim} \text{Bernoulli}(\theta), \end{aligned}$$

where the Bernoulli trials are, conditional on θ, independent. A Bayesian could well take as his prior distribution on θ a point mass at 0.5 corresponding to a coin known to be fair, or a uniform distribution on $[0.4, 0.6]$. Alternatively, the Bayesian could have a beta prior distribution, perhaps the uniform on $(0, 1)$ or the arc-sine distribution.

The posterior distributions corresponding to these prior distributions vary considerably. For data suggesting the coin is fair, all posteriors are consistent with the data. However, consider the behavior when the the sample proportion, $\hat{\theta} = n^{-1} \sum_{i=1}^{n} Y_i = 0.7$. For a large sample, each posterior distribution concentrates. The first at 0.5, the second near 0.6, and the others near 0.7. Limiting behavior follows the large-sample behavior–a closed-minded prior distribution may lead to inconsistent, poor inference; an open-minded prior distribution will eventually concentrate near θ_0, for any θ_0, and will eventually provide satisfactory inference.

Neither the development of subjective probability nor the decision theory-to-admissibility route to Bayes precludes the choice of a prior distribution with restricted support. A productive view is that the parameter θ has a fixed, but unknown value. The prior distribution expresses our uncertainty about this unknown value in the right language to describe uncertainty–the language of probability. Merely placing a distribution on θ does not, in any real sense, turn the fixed parameter into a random variable. Dennis Lindley famously referred to Cromwell's dictum, loosely paraphrased as "consider that you may be mistaken in your opinion", when arguing for full support of the prior distribution. Full support is one of the primary motivations for nonparametric Bayesian procedures.

1.2. False consistency

False consistency, focusing on the convergence of estimates or posterior distributions, provides a counterpoint to consistency. We say that an inference is falsely consistent if the inference is for something that has never been observed and the inference becomes degenerate.

A simple example illustrates the concept. Suppose that we observe a sequence of Y_i from the model

$$\begin{aligned} X_i &\overset{ind}{\sim} F \\ Y_i &= \text{round}(X_i), \end{aligned}$$

so that Y_i is X_i rounded to the nearest integer.

If one assumes that F represents a Normal(μ, σ^2) distribution, there are many ways to use the sequence of Y_i to estimate μ and σ^2, including those based on likelihood. Maximum likelihood estimates converge and a Bayesian posterior concentrates at these same values, provided the prior distribution has full support for μ and σ^2. In the restrictive context of the family of normal models, these estimates are consistent for (μ_T, σ_T^2). Outside of this family, the meaning of the parameters is not as clearly pinned down. Typically, if F has two moments and can generate at least three distinct values of Y_i, the distribution of Y_i has a unique, best-fitting normal distribution, and the estimate of (μ, σ^2) will converge to a well-defined value, say (μ_B, σ_B^2). Huber (1981) defines this last type of convergence as "Fisher consistency". This convergence argument applies, under mild conditions and with rich enough support for F, to finite dimensional parametric families. Typically, if Y_i admits $p + 1$ possible categories, p parameters can be identified.

Consider the conditional distribution of $X_i|(Y_i = 0)$. As the sample size grows and the estimates of μ and σ^2 converge to their limits, this conditional distribution also converges. Its limit, whether by a plug-in rule or by integration over a posterior distribution, is the conditional distribution of a normal(μ_B, σ_B^2) variate restricted to the interval $[-0.5, 0.5]$. In addition to convergence of the estimated conditional distribution, the stated uncertainty about the distribution vanishes. Standard errors from maximum likelihood go to zero, and the Bayesian posterior distribution concentrates in arbitrarily small neighborhoods.

Is this concentration of posterior good? or bad? For me, the answer depends upon the distribution from which the data arise. If the distribution is indeed normal, the limiting estimated distribution is correct; if not, the limiting estimated distribution will typically differ from the actual conditional distribution. The result is similar to the earlier Bernoulli example with the uniform prior distribution for θ on $[0.4, 0.6]$. If the true proportion θ_0 is in the range $[0.4, 0.6]$ the large sample result is good; if outside that range, the result is bad. The difference between the two settings is that in the Bernoulli example the conflict between the data and the large-sample inference is evident, while in the rounded-data example, we have no direct observations on the conditional distribution of $X_i|Y_i$. This lack of evidence prevents us from directly assessing the conflict between data and model, although in this instance we could examine the distribution of the Y_i for conflict with an underlying normal model.

Returning to Cromwell's dictum, it is certainly possible that the distribution of the X_i is non-normal, and a prior distribution with adequate support must acknowledge this fact. In this example, we might learn something about the conditional distribution of $X_i|Y_i$, but we cannot learn everything about it. The conditional distribution should not degenerate asymptotically.

To produce the desired behavior, we need a prior distribution with large support. The support should allow for consistent estimation of the distribution of Y_i, suggesting a high-dimensional prior distribution. The support should also be large enough, that, given the distribution for Y_i, there remains

a non-degenerate conditional distribution on the distribution of $X_i|Y_i$–that is, our model accurately reflects our uncertainty about the conditional distribution of $X_i|Y_i$. This is most naturally accomplished by the use of an infinite dimensional, nonparametric prior distribution.

2. The Dirichlet process

Nonparametric Bayesian methods take their name from the definition of nonparametric as not describable by a finite number of parameters.. The methods rely on models for which the effective number of parameters grows with the sample size. There are two main streams of Bayesian nonparametrics. One stream focuses on a function such as a regression function and replaces the traditional linear regression with a much more flexible regression. The prior distribution in this case is a probability model for "wiggly curves". We do not pursue this stream further here. The other stream focuses on probability distributions. It replaces the traditional parametric family of distributions with a much larger family. Instead of placing a prior distribution on a parametric family–that is, instead of writing a probability model for the parameters that govern the distribution, we write a probability model for the distributions themselves. To do so takes some technical work, with the first fully satisfying papers written in the early 1970s, and with antecedents stretching back at least to the 1950s.

The early boom in Bayesian nonparametrics was touched off by Ferguson's (1973) major paper which defines the Dirichlet process. He provides two basic properties that a prior distribution should have in order to qualify as both nonparametric and useful.

1. The prior distribution should have full support in some relevant space.

2. One should be able to perform the update from prior distribution to posterior distribution.

At the time of Ferguson's work, models were simpler and computational resources were limited. Full support for a distribution function translated to full support among distributions on $\mathcal{R}^p$, with support defined by the weak metric. The Dirichlet process is thus a probability model on distribution functions. The ability to move from prior distribution to posterior distribution equated to a conjugate model, allowing for an easy closed-form for the posterior.

The Dirichlet process plays a central role in Bayesian nonparametrics. In part, this is due to its being the first Bayesian nonparametric model to be developed; in part, this is due to its having several valuable representations, each of which lends itself to useful modelling. The remainder of this section describes constructions and basic properties of the Dirichlet process. For complete details, see the references.

2.1. The beta-binomial approach

The Dirichlet process is connected to the beta distribution, and through it to the multinomial distribution. The beta-binomial model has the form

$$\begin{aligned} p &\sim \text{beta}(\alpha, \beta) \\ Y_i|p &\overset{ind}{\sim} \text{Bernoulli}(p), \quad i = 1, \ldots, n, \end{aligned}$$

where $\alpha, \beta > 0$. The prior distribution for p has full support on the interval $(0, 1)$. The model is conjugate, and the posterior from a sample of size n is $p|Y_1, \ldots, Y_n \sim \text{beta}(\alpha + \sum Y_i, \beta + n - \sum Y_i)$.

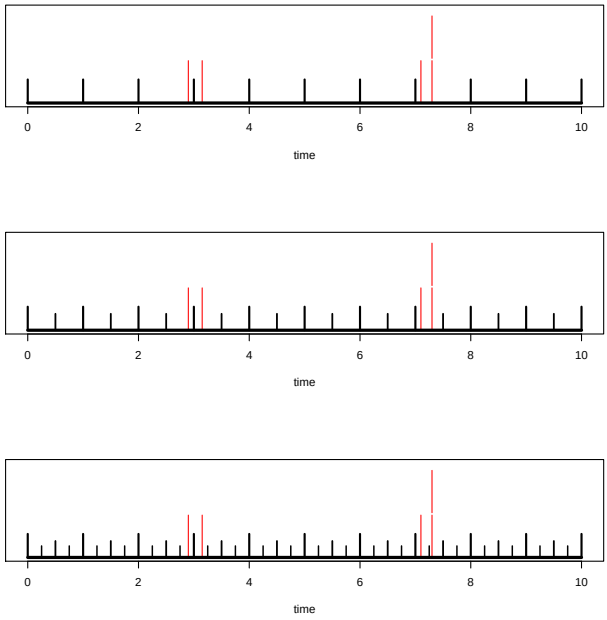

Figure 1: *The Dirichlet-multinomial. The figure shows the refinement of categories. Categories are indicated by the black lines, data by the red lines. At each level, we have a Dirichlet-multinomial model. Continued refinement leads to the Dirichlet process.*

The Dirichlet-multinomial model extends the beta-binomial model to more categories for the observable data.

$$p = (p_1, \ldots, p_k) \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_k)$$
$$Y_i | p \overset{ind}{\sim} \text{multinomial}(p),$$

where all $\alpha_i > 0$ for the prior density to exist. The posterior from a sample of size n is Dirichlet($\alpha_1 + \sum I(Y_i = 1), \ldots, \alpha_k + \sum I(Y_i = k)$). This conjugate update from prior to posterior is quick and has a simple closed form. If we focus only on the j^{th} category, we have a beta-binomial model with parameters α_j and $\sum_{i \neq j} \alpha_j$, and data $\sum_{i=1}^n I(Y_i = j)$, $\sum_{i=1}^n I(Y_i \neq j)$.

The categories of the multinomial distribution can be refined as needed, as shown in Figure 1 for data on an interval. For each partition, we write a Dirichlet-multinomial model.

There are two big questions about this refinement. The first is, if we pass to the limit, refining categories so that each pair of reals eventually fall into different categories, does a limiting "Dirichlet-multinomial" exist? The second is, if the limiting object exists, how do we work with it? Ferguson (1973) passed to the limit, answering both questions and constructing the Dirichlet process.

To keep the models consistent across refinements of the partition, we require the same probability statements about a fixed, coarser category given any partition. To ensure this self-consistency while staying withing the Dirichlet-multinomial framework, when a category is split, its parameter is divided and allocated to the pieces of the split. For example, if category 1 is partitioned into two subcategories, say 11 and 12, we require $\alpha_1 = \alpha_{11} + \alpha_{12}$. For ease of notation, the categories are renumbered after each refinement. Call the parameter vector α, regardless of dimension and introduce $M = \|\alpha\| = \sum \alpha_i$.

2.2. The gamma process

The refinement of partitions suggests the gamma process construction of the Dirichlet processs. Traditional distribution theory provides a link between beta distributions and gamma distributions. If $X \sim \text{gamma}(\alpha, c)$, $Y \sim \text{gamma}(\beta, c)$, and $X \perp Y$, then $X/(X+Y) \sim \text{beta}(\alpha, \beta)$. This relationship extends to the Dirichlet distribution. If $X_i \sim \text{gamma}(\alpha_i, c)$, $i = 1, \ldots, k$, then the X_i, scaled to sum to 1, follow a Dirichlet distribution.

The Dirichlet-multinomial refinement splits category parameters, while the refinement for the gammas splits the shape parameters in the same way. The gamma is an infinitely divisible distribution, and so the refinement is self-consistent through the limit.

Turning the problem around, we start with the limiting process and ask how we can match every Dirichlet-multinomial refinement. The limiting process is called a Dirichlet process. For a prior on distributions on the real line, we do so by replacing the parameter vector α with a finite, positive measure α. Set $M = \alpha(\mathcal{R})$. For any finite partition of the real line into measureable categories $\mathcal{R} = \cup_{i=1}^{k} C_i$, set $\alpha_i = \alpha(C_i)$ and we obtain the desired Dirichlet distribution. Self-consistency is ensured by construction.

The limiting model is written as

$$\begin{aligned} F &\sim \text{Dir}(\alpha) \\ \theta_i | F &\overset{ind}{\sim} F. \end{aligned} \tag{2.1}$$

The parameter α is the base measure of the Dirichlet process. Often, the base measure is recast as a scaled distribution function. For a distribution on the real line, define $F_0(t) = \alpha\{(-\infty, t]\}/M$. We may write $F \sim \text{Dir}(M \cdot F_0)$ in place of (2.1). We refer to F_0 as the base cdf and M as the mass of the base measure.

2.3. From prior to posterior

The construction of the Dirichlet process tells us how to find the posterior distribution. For the beta-binomial and Dirichlet-multinomial problems, the posterior is in the same family with parameter vector updated with category counts. Equivalently, we consider the independent gamma distributions with shape parameres updated with the category counts. Passing to the limit, we update the gamma process with a point mass at each observed value. The corresponding gamma process has measure $\alpha + \sum_{i=1}^{n} \delta_{Y_i}$. The corresponding posterior is a Dirichlet process with base measure $\alpha + \sum_{i=1}^{n} \delta_{Y_i}$.

The prior predictive distribution for the Model (2.1) is found via integration. For a measurable set A, partition $\mathcal{R}$ into A and A^C. Write the Dirichlet-multinomial (beta-binomial) model on these sets. The probability that $\theta_i \in A$ is, with somewhat sloppy notation,

$$P(\theta_i \in A) = E[F(A)] = \frac{\alpha(A)}{\alpha(\mathcal{R})} = F_0(A).$$

The predictive distribution for an observation from this model is the base cdf F_0.

The posterior predictive distribution of (2.1) is easily obtained.

$$\begin{aligned} F | \theta_1, \ldots, \theta_n &\sim \text{Dir}\left(\alpha + \sum_{i=1}^{n} \delta_{\theta_i}\right), \\ \theta_j | F, \theta_1, \ldots, \theta_n &\sim F, \ j > n. \end{aligned}$$

The conjugate pair of prior and likelihood leads to a posterior in the same family. The posterior predictive distribution follows the same form, with updated base measure, $\alpha = MF_0 + \sum_{i=1}^{n} \delta_{\theta_i}$. Denoting the empirical cdf of $\theta_1, \ldots, \theta_n$ by $\hat{F}_n$, the posterior predictive distribution is

$$\frac{M}{M+n} F_0 + \frac{n}{M+n} \hat{F}_n.$$

2.4. Blackwell, MacQueen, and the Polya urn scheme

Blackwell and MacQueen (1973) present a distinctly different view of the Dirichlet process. They focus on the successive predictive distributions for a sequence of observations drawn from F. Since we have results for the Dirichlet process, we make use of it. The first draw follows the prior predictive disribution, F_0. To obtain the distribution of the second draw, we update our prior Dirichlet process, moving to the posterior Dirichlet process with base measure $\alpha + \delta_{\theta_1}$. The (conditional) predictive distribution for the second draw rescales this, either sampling from the base cdf with probability propotional to M or sampling from the empirical cdf with probability proportional to 1. Proceeding inductively, to draw $\theta_{n+1}|(\theta_1, \ldots, \theta_n)$, the draw is either from the base cdf with probability $M/(M+n)$ or from the empirical cdf of the first n draws with probability $n/(M+n)$.

Blackwell and MacQueen view this as an urn scheme. The initial draw is taken from an urn containing a rainbow of colors. When the draw (color) is observed, it is returned to the urn along with an extra ball of the same color. The draws proceed sequentially, always according to the same rule. It is easy to see that these draws lead to the same joint distribution on the observable sequence of θ_i, and hence that they describe the same process. This leads to the simplified description of the conditional distribution of a draw. Noting that some θ_i are drawn from the base cdf and others are set equal to previously drawn values, we introduce the ' * ' notation. Among $\theta_1, \ldots, \theta_n$, there have been k draws from the base measure, leading to k groups, or clusters. The clusters are of size $n_1, \ldots, n_k$ with common values $\theta_1^*, \ldots, \theta_k^*$. Making use of this notation, the conditional distribution of θ_{n+1} is

$$\theta_{n+1}|(\theta_1, \ldots, \theta_n) \sim \begin{cases} \delta_{\theta_i^*}, & \text{w.p. } \dfrac{n_i}{M+n}, \ i = 1, \ldots, k, \\ F_0, & \text{w.p. } \dfrac{M}{M+n}, \end{cases}$$

where w.p. stands for "with probability".

The Polya urn scheme yields an easy proof of the discreteness of the Dirichlet process. From the above description, we see that the probability that θ_{n+1} is a value never before seen is $M/(M+n)$ if the base cdf is continuous, and smaller if the base cdf has a discrete component. Taking the limit as $n \to \infty$ sends this probability to 0. Consequently, the probability that F has a discrete component is 0.

2.5. Smoothing the Dirichlet process

The discreteness of the Dirichlet process poses problems for its use as a model for observable data. The most commonly used models for discrete data, such as those in an exponential family, have the same support for all distributions in the family. This provides an element of robustness for likelihood-based analyses, as a single observation does not zero out any parameter values and so has a limited impact on inference. In contrast, two F drawn independently from a Dirichlet process with continuous base measure are singular with respect to one another with probability one. This follows from the rule for generating draws from the two distributions. Consider θ_1, drawn from the first distribution.

$\theta_1 \sim F_0$. Similarly, the first draw from the second distribution independently follows F_0. These two draws differ with probability one, leading to the conclusion.

To alleviate this problem of discreteness, the Dirichlet process is used as a latent stage in a hierarchical model. The distribution F is smoothed, much like a kernel density estimate smooths the empirical distribution. The mixture of Dirichlet processes (MDP) model is written as follows.

$$\begin{aligned} F &\sim \text{Dir}(MF_0) \\ \theta_i|F &\overset{ind}{\sim} F, \quad i = 1, 2, \ldots, \\ Y_i|\theta_i &\overset{ind}{\sim} H, \quad i = 1, 2, \ldots. \end{aligned} \tag{2.2}$$

The distribution H has density $h(\cdot|\theta)$, with parameter θ. This model is formally a countable mixture distribution, with the Dirichlet process prior on the mixture. For this reason, some call this model a Dirichlet mixture model. We retain the original terminology dating to Antoniak (1974) who provides results for the MDP model. His results apply both to this type of smoothed model, where the θ_i are observed through the filter of the likelihood for Y_i and to models where there is a mixture over base measures. Similarly, we retain the original notation where α represents the base measure rather than its mass.

One specific MDP model is a conjugate location mixture of normal distributions, with $F_0 = \text{normal}(\mu_0, \tau^2)$ and $H = \text{normal}(\theta, \sigma^2)$. This model provides a Bayesian analog of classical density estimators (Escobar and West, 1995; Lo, 1984). A second model is a conjugate mixture of binomials model, with $F_0 = \text{beta}(\alpha, \beta)$ and $H = \text{binomial}(n, \theta)$. This model has been used for the compound decision problem (Berry and Christensen, 1979), and we revisit it in Section 4.

The MDP model is often embedded as part of a larger hierarchical model. A prior distribution may be placed on the parameters governing the base cdf, the mass parameter, or jointly on the two. Prior distributions are placed on the parameters in the smoothing kernel $h(\cdot)$, as the σ^2 in the normal model described above, and so on. In all, the MDP model serves as a valuable component in a hierarchical model, providing a prior distribution with full support over a space which has been judged relevant by the analyst.

2.6. Sethuraman's construction

Sethuraman (1994) provided an alternative construction of the Dirichlet process which follows from the Polya urn scheme and the internal consistency of a Bayesian model. Under his construction, the distribution F is built from two independent sequences of draws. One sequence gives the locations of atoms of mass in the discrete mixture. The second sequence determines the amount of mass associated with the atoms. Specifically, the sequences of draws are

$$\begin{aligned} \theta_i^* &\overset{ind}{\sim} F_0, \quad i = 1, 2, \ldots, \\ V_i &\overset{ind}{\sim} \text{beta}(1, M), \quad i = 1, 2, \ldots. \end{aligned}$$

The V_i are used to construct a set of point masses which sum to 1 by the rule $W_i = V_i \prod_{j<i}(1 - V_j)$. The resulting countable, discrete distribution is constructed as

$$F = \sum_{i=1}^{\infty} w_i \delta_{\theta_i^*}.$$

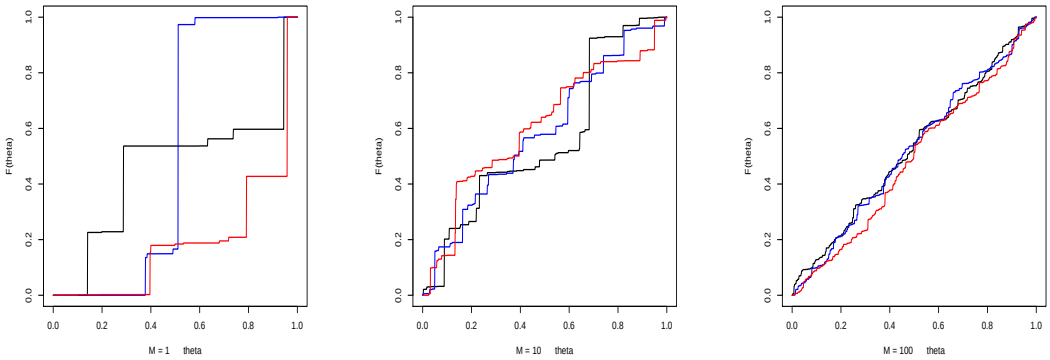

Figure 2: *The Dirichlet process. The three panels show draws from the Dirichlet process with uniform F_0 and mass parameters of* 1, 10, *and* 100. *The larger mass parameters produce smaller jumps in the realized F. The discreteness of the realized F is apparent from the jumps.*

A stream of data arising from F comes in the usual fashion, with $\theta_i|F \overset{iid}{\sim} F$.

This construction directly targets F as a countable discrete distribution, much as Blackwell and MacQueen's construction via an urn scheme. In doing so, it relaxes some of the technical assumptions made in Ferguson's (1973) constructions of the Dirichlet process. The distribution F no longer needs to be a distribution on $\mathcal{R}^p$ for some p, but can exist in a more general space.

2.7. Interpretation of parameters

The parameter of the Dirichlet process is variously given as the measure, α, or the mass of the measure and a cdf, M and F_0. Early work followed the Dirichlet-multinomial approach and took the mass parameter as a "prior sample size". This interpretation is passable when data are observed directly from F. Limiting models investigate a sequence of prior distributions where F_0 is held fixed and $M \to 0$. After updating with a fixed data set, there is a corresponding limit of posterior distributions. These converge to the empirical cdf when data are observed exactly, leading to the Bayesian bootstrap (Kim and Lee, 2003; Newton and Raftery, 1994; Rubin, 1981). This relationship has also been exploited to develop the connection between Bayesian survival analysis and the Kaplan-Meier estimator (Susarla and Van Ryzin, 1976). Johnson and Christensen (1986) note difficulties with this interpretation and suggest a remedy. This interpretation certainly breaks down when data are observed through the smoothing of the MDP model. Figure 2 shows several draws from the Dirichlet process with a variety of mass parameters.

Bush *et al.* (2010) consider the limiting behavior of the MDP model when $M \to 0$ while F_0 is held fixed. For smoothing kernels $H(\cdot|\theta)$ that are mutually absolutely continuous as θ varies, the limiting posterior distribution concentrates on the event where there is a single cluster containing all of the observations. These authors then create a "noninformative" version of the MDP model. Lee *et al.* (2014) pursue this idea further, suggesting the interpretability of local mass where M is tied to the scale of F_0. This is particularly important when the base cdf is deliberately chosen to be overdispersed.

It should be noted that Bayes factors for the MDP model (and indeed for directly observed data) are problematic. Ferguson (1973) cautions the reader about the impact of discreteness on tests. Xu *et al.* (2012) note that it would be difficult to believe that the MDP model truly captures one's subjective

beliefs, and suggest a partial data update before computing the Bayes factor. Kessler *et al.* (2014) make the prior more meaningful with a clever adjustment. A productive view of the MDP model is that it provides a useful modelling tool which, when updated with a sufficient amount of data, produces a sensible posterior distribution.

3. Computation

Computational methods for the mixture of Dirichlet process model are well developed. Escobar's (1988) landmark dissertation work (also Escobar, 1994) developed a Markov chain Monte Carlo (MCMC) strategy for fitting the models in a paper that predates Gelfand and Smith's (1990) development of the Gibbs sampler for general Bayesian inference. The centerpiece of this computational strategy is to set up a Markov chain with the posterior distribution as its limiting distribution, and to then obtain a realization of the Markov chain.

The Gibbs sampler relies on a sequence of draws from conditional distributions. The sampler itself works on a marginalized version of the model, removing the infinite-dimensional F and working with the finite-dimensional joint distribution of θ. The conditional distributions follow smoothly from the Polya urn scheme. For the basic MDP model, these are most easily described for the last parameter value, θ_n.

The conditional distribution for $\theta_n|(\theta_1, \ldots, \theta_{n-1}, Y_1, \ldots, Y_n)$ may be thought of in two parts. The first is the conditional prior distribution for $\theta_n|(\theta_1, \ldots, \theta_{n-1})$ and the second is the likelihood from $Y_n|\theta_n$, where irrelevant terms have been dropped. The two parts combine to form the conditional posterior distribution, described in the * notation.

$$\theta_n|\theta_{<n} = \theta_i^*, \quad \text{w.p.p. } n_i^- h(y_n|\theta_i^*), \ i = 1, \ldots, k^- \tag{3.1}$$

$$\sim F^*, \quad \text{w.p.p. } M \int h(y_n|\theta) f_0(\theta) d\theta. \tag{3.2}$$

The superscript $-$ indicates that θ_n is dropped before number of clusters and cluster sizes are computed. In the event that a new draw of a θ_i^* is taken, as in (3.2), the new draw is from the single-case posterior distribution where the prior F_0 is updated with data Y_n. Here, w.p.p. stands for "with probability proportional to".

This early Gibbs sampler suffers from poor mixing. Once placed, a value of θ_i^* persists until all observations desert it. This may take many iterates, especially when the data suggest that the θ_i contain clusters of substantial size. Two techniques alleviate this difficulty in mixing. The first is marginalization (MacEachern, 1994), wherein the θ_i are integrated out of the model for the MCMC routine and only generated as needed for particular inferences. The θ_i can be marginalized, provided integrals of the form (3.2) can be evaluated, as is the case for conjugate models. Conjugate models with marginalization are often used when stretching computation to the limit. The second technique to handle poor mixing is reparameterization during portions of the MCMC (Bush and MacEachern, 1996). Here, an additional stage can be added to the MCMC algorithm where $\theta_i^*|(Y_1, \ldots, Y_n), i = 1, \ldots, k$ are generated. This remixing step moves the values of the θ_i^* without having to empty clusters.

There have been many advances since these early samplers. Escobar and West (1995) show how to place a distribution on M and sample from it. MacEachern and Müller (1998) and Neal (2000) provide algorithms for non-conjugate situations where the integral in (3.2) does not have a closed form. Jain and Neal (2004, 2007), and Dahl (2003) develop split-merge moves which, when appended to a basic sampler, allow the Markov chain to move more freely, resulting in better mixing. Guha (2008) develops targeted proposals to improve mixing. Walker (2007) and Kalli *et al.* (2011) create slice

samplers which have proven effective for many problems.

The work on MCMC methods for the MDP model spilled over to impact MCMC methods for parametric Bayesian models. The early examples of marginalization and reparameterization gave rise to an understanding of the value of multiple reparameterizations of a model within an MCMC cycle. These techniques implicitly involve parameter expansion and marginalization to improve mixing. They provide a clear view of the benefits of non-identifiability in MCMC (MacEachern, 2007). These innovations all flow from the multiple representations of the Dirichlet process described in the preceding section.

Additional computational strategies have been explored. Sequential importance sampling (Liu, 1996; MacEachern *et al.*, 1999), or particle filtering provides an alternative to MCMC. With these methods, an importance sampling envelope is created as one passes through the data. Once again, marginalization and remixing via reparameterization improve the technique.

Current computational work is focused on the development of algorithms that scale up to much bigger data sets. Three themes are showing their value across a broad range of techniques. First, a willingness to accept analytic approximation as a necessary tradeoff for scale and speed. Second, a deliberate oversimplification of the model to admit conjugate forms which often serves the dual purpose of enabling fast computation and improving approximations. Third, a truncation of a complex model with estimates plugged in for some parameters to avoid the time needed for exploration of the full posterior distribution.

One strategy which has proven effective for scaling algorithms is to find an algorithm which is "exact" for a narrowly defined problem and to then apply a version of the algorithm in other cases. In the MDP setting, a redistribution of mass algorithm (MacEachern, 1988) and an associated direct sampling algorithm (Kim, 1999; Kuo and Smith, 1992) allow for exact evalution of and simulation from the posterior distribution for survival data. Newton and Zhang (1999) proposed a variant on this method to quickly approximate the posterior in non-survival settings. Martin and Tokdar (2009) pursue this approach and establish asymptotic properties. In a similar vein, variational methods (Blei and Jordan, 2006) are exact in certain simple settings, but become approximations in more complex settings. They stand our for providing a quick, scalable approximation to the posterior. The challenges of variational methods include choice of an implementation and understanding the accuracy of the approximation. Data-splitting techniques seek to perform calculations with some level of parallelization, thereby allowing larger data sets. Jara's DPPackage (Jara *et al.*, 2011) provides an easy means to try out the MDP and related models.

4. Applying the model

There are two main perspectives for use of the MDP model. These perspectives mimic general developments in the Bayesian community since the 1970s. One follows the objective Bayesian philosophy, seeking to impose little structure on a problem and seeking to incorporate a minimum of subjective information into the analysis. Under this approach, the model is used for density estimation, and inferences are extracted from the density estimate. Lo (1984) espoused this philosophy, although its implementation in realistic problems was suspended until computational developments made the method practical. Escobar and West (1995) pursue this strategy, focusing on the one-sample problem in one or several dimensions.

Müller *et al.* (1996) exemplifies the density estimation-to-inference paradigm. These authors fit a mixture of multivariate normals (MDP model) in several dimensions and proceed to extract the regression of response on covariates as the conditional mean of one coordinate given the others. The

resulting regression is naturally nonlinear. In a similar fashion, one could extract the median regression or other summaries from the joint distribution.

Developments in density estimation include theoretical results as well as empirical implementation. Results current at the time are presented in Ghosh and Ramamoorthi (2003), while further developments include Ghosal and van der Vaart (2001) and Walker (2004). These authors lay the theoretical foundation for density estimation with the MDP model, providing a selection of results on consistency of the estimators and on rates of convergence.

Experience with the density estimators has produced improvements on the basic MDP model. Griffin (2010) modified the model by partitioning variance between the base measure and the smoothing kernel. Empirical evidence suggests that this partition improves density estimation. Bean *et al.* (2016) note difficulties with estimation of long-tailed and skewed densities and propose a transformation-based approach along the lines of classical work by Yang and Marron (1999).

The second perspective on use of the MDP model follows the development of the hierarchical model. It is driven by first determining the structure of the model and then filling in appropriate distributions for the various parts of the model. The Dirichlet process or MDP model typically appears in a portion of the model where the analyst desires full support. The literature contains many examples of the successful use of this strategy.

The most evident uses of the Dirichlet process or MDP model are for portions of the model devoted to what, in a classical analysis, would be termed random effects. Bush and MacEachern (1996) argue that full support for random effects means full support on the distributions from which independent and identically distributed draws are made–that is, the model is used to produce effects which are exchangeable for all n. In contrast, full support for a set of p fixed effects is full support on $\mathcal{R}^p$. This view provides a Bayesian version of the classical notions of fixed and random effects. This argument applies wherever random effects arise, for example in frailty models in survival analysis, general meta-analysis models, or situations where there is population heterogeneity. Kleinman and Ibrahim (1998) extend this argument to general linear mixed models.

For linear regression problems, Gelfand and Kottas (2002) provide median zero errors and describe a computational strategy for inference for essentially arbitrary functionals of the Dirichlet process. MacEachern and Guha (2011) explain how the posterior for regression coefficients can be more concentrated with an MDP model for the errors than with a normal model for the errors. Wang (2009) considers analogs of weighted least squares, distinguishing between models for a scale family, a convolution family, and families in between.

4.1. Admissibility, shrinkage, and modelling

The decision-theoretic motivation for nonparametric Bayesian methods relies on the concept of admissibility. Admissibility is, in turn, closely linked to the development of empirical Bayesian methods (Efron and Morris, 1975) which in many cases are popularly equated with "shrinkage". In this subsection, we revisit a famous data set, tinkering with the example to investigate the notions of admissibility and shrinkage and contrasting them with nonparametric Bayesian modelling.

Efron and Morris (1975) motivated empirical Bayesian methods with an example which is compelling to all who are familiar with baseball. They set up a prediction problem for players in the United States' major leagues. The data from which the predictions are to be made consist of the results of the first 45 at bats for a set of players, described in terms of their batting average (here, we take batting average to be the proportion "hits divided by at bats"). The goal is to predict, based only on these data, each player's batting average for the remainder of the season. The quality of a single forecast is judged by squared prediction error, and the quality of the collection of forecasts is judged

by sum of squared prediction error.

Baseball fans know much about batting averages. At the time, the league-wide average was about 0.250, with the highest averages in the mid 0.300 s. It would be very unlikely that any season-ending average would be as high as 0.400. Various players are good hitters (here, a high batting average) or are poor hitters, and the differences can be substantial. The data, consisting of all players with 45 at bats on a particular day, exclude pitchers who would have too few at bats. They do include one exceptionally good hitter, by design. Also, after a mere 45 at bats, there is substantial variation in the sample batting average. The observed averages of 0.156 to 0.400 show a clear excess of variation–all would agree that the large values should be pulled down, and the small values pushed up.

Formally, we begin with a too-simple model for the data, where $Y_i|\theta_i \overset{ind}{\sim}$ binomial(45, θ_i), with a similar, conditionally independent, model envisioned for the remainder of the season. This model ignores known effects, such as the quality of an opposing pitcher who is typically linked to several at bats, the baseball park in which the game is held, etc. However, creating an admittedly imperfect model is a necessary part of data analysis, and this model follows Efron and Morris.

The original focus was on the comparison of estimation techniques such as maximum likelihood or method of moments to empirical Bayesian estimation in the James-Stein style, or shrinkage estimation. Defining the remainder of the season batting average for player i as $\tilde{\theta}_i$, and using the generic $\hat{\theta}_i$ as an estimator, the squared prediction error for the player is $\text{SPE}_i = (\hat{\theta}_i - \tilde{\theta}_i)^2$. Accumulating these across all players yields the sum of squared prediction error. We expand this error about θ_i

$$\begin{aligned}\text{SSPE} &= \sum_{i=1}^{18} \left(\hat{\theta}_i - \tilde{\theta}_i\right)^2 \\ &= \sum_{i=1}^{18} \left[\left(\hat{\theta}_i - \theta_i\right)^2 + \left(\theta_i - \tilde{\theta}_i\right)^2\right],\end{aligned}$$

where the cross-product terms disappear due to the presumed independence of the observed data and the remainder of the season performance, conditional on the θ_i. The first terms describe the quality of the estimation method while the second terms represent a noise floor of pure prediction error which cannot be eliminated. In this data set, we estimate this noise floor with a plug-in method for the binomial model as $\sum_{i=1}^{18} \tilde{\theta}_i(1 - \tilde{\theta}_i)/n_i \approx 0.0119$. The n_i in this expression is number of at bats in the remainder of season for player i.

Our comparison begins with maximum likelihood estimation. The playerwise maximum likelihood estimate is $\hat{\theta}_{1,i} = Y_i/45$. This estimate results in $\text{SSPE}_1 = 0.0857$. A simple shrinkage estimator is based on components of variation. The target of shrinkage is the overall batting average, here equal to $\bar{\theta} = \sum_{i=1}^{18} \hat{\theta}_{1,i}/18$, since all players have 45 at bats. There are two components of variation – between players and within players. Within player variation is estimated to be $\hat{\sigma}^2 = \bar{\theta}(1 - \bar{\theta})/45$. This estimate is more stable than one allowing estimated variation to vary by player and has little bias as the binomial variance changes only modestly over reasonable values for θ_i. Total variation is estimated by $S^2_{\theta_i}$, the sample variance of the $\hat{\theta}_{1,i}$. Between player variation is estimated to be $\hat{\tau}^2 = S^2_{\theta_i} - \hat{\sigma}^2$. The shrinkage estimator for player i is a precision weighted average

$$\hat{\theta}_{2,i} = \frac{\hat{\sigma}^{-2}\hat{\theta}_{1,i} + \hat{\tau}^{-2}\bar{\theta}}{\hat{\sigma}^{-2} + \hat{\tau}^{-2}}.$$

This estimate results in $\text{SSPE}_2 = 0.0267$, showing a strong advantage for the shrinkage estimator relative to maximum likelihood.

Our third estimation strategy relies on a nonparametric Bayesian model, a straightforward mixture of Dirichlet processes in the style of Berry and Christensen (1979). The model is, with F_0 the uniform distribution on $(0, 1)$,

$$\begin{aligned} F &\sim \text{Dir}(F_0) \\ \theta_i|F &\overset{ind}{\sim} F \\ Y_i|\theta_i &\overset{ind}{\sim} \text{binomial}(45,\ \theta_i). \end{aligned} \tag{4.1}$$

The base measure for the Dirichlet process component of the model is Lebesgue measure on $(0, 1)$.

We fit the model in (4.1) with a marginalized Gibbs sampler, initializing the players in 18 distinct clusters. The Gibbs sampler is run for a burn-in period of 100 iterates, followed by 1,000 iterates for estimation. The run is short because the model is simple with only 18 players, and because the next portion of the example requires many repeated fits of this model to variations on the data. Estimation takes advantage of the known form of the conditional distributions, using Rao-Blackwellization to average conditional means. The resulting estimates, $\hat{\theta}_{3,i}$ lead to $\text{SSPE}_3 = 0.0251$, slightly bettering the shrinkage estimator and soundly beating maximum likelihood.

The substantial improvements over maximum likelihood in this example come as no surprise to those who follow baseball, as it is clear that the maximum likelihood estimates need to be adjusted in some fashion. However, admissibility suggests much more–namely that the type of strategy we pursue should bring benefits for any parameter vector $\theta = (\theta_1, \ldots, \theta_{18})^\top$. To investigate this, we consider a series of problems involving different parameter vectors θ. An extreme variation reverses one player, counting "outs" as successes rather than "hits". With the binomial model, this player's θ_i is effectively replaced with $1 - \theta_i$, the player's data becomes $45 - Y_i$, and the remainder of the season target is $1 - \tilde{\theta}_i$. With batting averages pushed well away from 0.500, this is a difficult problem for shrinkage estimation: shrinking this player's estimate toward the others will produce a poor estimate, and including this player in the estimate of $\bar{\theta}$ will tend to shrink the other player's estimates toward too-high values. In contrast, maximum likelihood is unaffected by the reversal.

The three approaches were fit with k players reversed, k ranging from 1 to 9. The symmetric nature of the problem implies that k reversals and $18 - k$ reversals lead to the same SSPE for all three of the approaches considered here. For each $k > 1$, 1,800 reversals were drawn from the set of possible reversals. The same reversals were used for each of the techniques.

Figure 3 summarizes the results of the study. The performance of maximum likelihood is unaffected by the reversals, and SSPE_1 remains constant at 0.0857. As expected, the shrinkage estimator performs more poorly than before, with SSPE_2 increasing as the number of reversals is increased. With 9 reversals, the player-wise estimates are shrunk to something near 0.500. The vertical lines in the figure extend from the 25^{th} to 75^{th} percentiles of the results, with the dot placed at the mean. Nevertheless, shrinkage still improves estimation, bringing a clear, if modest, benefit, even in these artificially difficult problems. The nonparametric Bayesian method shows strong performance across the entire range of reversals. While there is a small uptick in SSPE_3 as the number of reversals increases, the prediction errors remain remarkably constant.

The second panel in Figure 3 quantifies the benefits of shrinkage estimation and nonparametric Bayesian techniques relative to maximum likelihood. The figure shows the percentage decrease in SSPE above the noise floor, relative to maximum likelihood. The decreases for shrinkage estimation with no reversals and for the nonparametric Bayes method, with or without reversals, are extreme, holding steady at about 80%. Decreases for shrinkage estimation with reversals range from the mid teens to slightly above 20%.

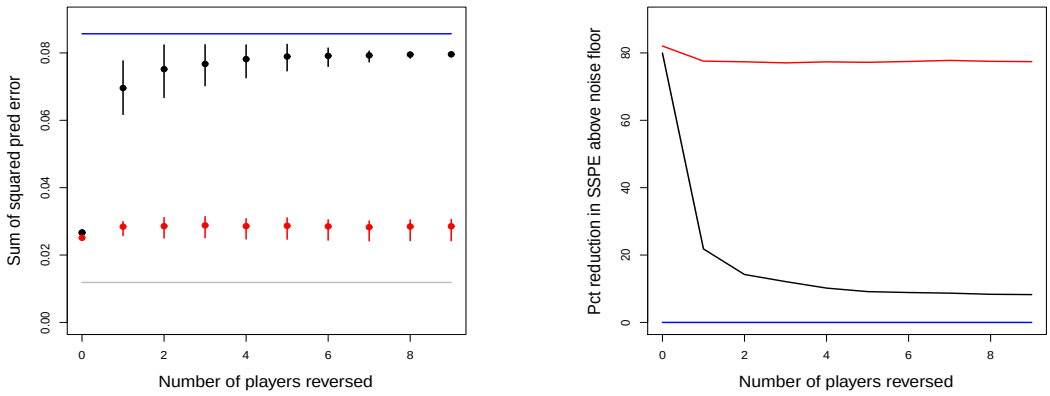

Figure 3: *Efron and Morris' baseball data (The left panel shows sum of squared prediction error for several methods, plotted against number of players reversed. The line at the top of the plot is for the player-wise maximum likelihood estimator; the black dots and lines are for an empirical Bayesian shrinkage estimator; the red dots and lines are for a mixture of Dirichlet processes model. The lines extend from the 25^{th} to 75^{th} percentiles of sum of squares, accounting for different reversals. The gray line at the bottom of the plot is an estimate of non-eliminable prediction error. The right panel provides an estimate of the reduction in non-eliminable risk when compared to the player-wise maximum likelihood estimator. Blue for the maximum likelihood estimator, black for the empirical Bayes estimator, and red for the nonparametric Bayes method.).*

The success of shrinkage estimation lies in its dual as a Bayes estimator. Although the usual description is in terms of shrinkage, the estimates exactly match posterior means from a set of conjugate beta-binomial models, one for each player. For the real data, the beta prior distribution is set from the data, with a prior mean of $\bar{\theta}$ and the shrinkage coefficient translating into the parameter values $\alpha \approx 100$ and $\beta \approx 277$. The large parameter values imply an approximate normal prior distribution for the players' θ_i. This strategy performs well as long as the prior distribution provides a good approximation to the apparent distribution of the θ_i. It breaks down when it does not, as in the case of several reversals. With the reversals, we retain the benefits of admissibility, but lose the much greater value of an effective model.

The nonparametric Bayesian approach focuses on modelling, and the prior distribution has full support for distributions on the interval. When there are no reversals, the data themselves drive the posterior distribution on the unknown F to a sensible place, leading to excellent estimates. When there are reversals, the prior distribution is flexible enough to adapt, picking out the fact that F has two (or perhaps more) modes. This "full support" aspect of the prior distribution translates to a better model which then leads to more effective estimates.

5. Beyond the Dirichlet process

The Dirichlet process and its extension to the MDP model are widely used. There are many additional Bayesian nonparametric methods which serve as prior distributions for an unknown distribution. Nearly all of them are suited to use as a component in a hierarchical model. Some focus on low-dimensional distributions, some on discrete problems, some on particular applications. These methods are often motivated by a shortcoming of the MDP model and seek to repair the shortcoming. Most have parallel tracks of applied modelling, computational strategies, and theoretical results. This section provides capsule descriptions of a few of the more prominent of these methods.

5.1. Focus on a single distribution

The early variations on the Dirichlet process addressed difficulties in implementation. Directly observed data are easier to handle than indirectly observed data. Coupling this with the tailfree nature of the Dirichlet process (Doksum, 1974) suggested modifications appropriate for survival analysis. Dykstra and Laud (1981) considered monotonicity of a hazard rate, developing the extended gamma process which is computationally conjugate for survival data consisting of exact event times and right censored event times. Hjort (1990) created the beta process which has become the mainstay of Bayesian survival analysis. Walker *et al.* (1999) developed theory and computation for models which include Dirichlet process as a special case and which increase the scope for modelling.

There are several ways to model continuous distributions. The Polya tree (Lavine, 1992; Mauldin *et al.*, 1992) generalizes the Dirichlet-multinomial construction of the Dirichlet process with the addition of a tree structure and can produce continuous distributions. The construction leads to mixtures of Polya trees (Hanson, 2006). A Gaussian process can be placed on the log-density, with adjustment so that the resulting density integrates to one (Lenk, 1988; Tokdar, 2007). Focusing on the unit interval, Petrone (1999) develops Bernstein polynomials to provide a density estimator.

Several variants on the Dirichlet process as a countable mixture distribution have been developed. Ishwaran and James (2001) pursue Pitman-Yor processes, adding an extra parameter to the Dirichlet process. James *et al.* (2005), Lijoi *et al.* (2005), and Regazzini *et al.* (2003) work with normalized processes. Lee *et al.* (2013) consider species sampling models.

5.2. Collections of distributions

The early work on nonparametric Bayesian methods focused on a single distribution. From this base, a variety of authors began to build through the standard progression of models seen in undergraduate and graduate coursework. The one-sample problem (a single distribution) leads to the k-sample problem (ANOVA) and regression. Along the way, the division of effects into fixed effects and random effects suggests the mixed model in both ANOVA and regression settings. The progression of models naturally segues into models for a collection of distributions. These models have attracted an enormous amount of attention since their initial development.

It would be a major undertaking to describe the work on collections of distributions and how they relate to data. A brief overview follows. Müller *et al.* (2004) present a model for k distributions. They describe a common component of size ϵ, shared by each distribution, with a unique remaining $1 - \epsilon$ for each distribution. Placing Dirichlet processes on the individual parts makes all of these distributions nonparametric. In dissertation work, Tomlinson (1998), advised by Escobar, develops a density-based version of ANOVA by constructing a hierarchy of nonparametric processes–one process at the top, latent stage of the model from which k lower stage distributions are drawn. These lower stage distributions are also nonparametric, resulting in a mixture of mixtures.

MacEachern (1999, 2000) developed dependent Dirichlet processes to address the regression problem in a nonparametric fashion. For a simple example, imagine the growth charts by which children's health is tracked in many countries. For each given age, percentiles for height, weight, and other physical characteristics are established. These percentiles show clear evidence of non-normality that would be difficult to capture with a low-dimensional model. For this snapshot at any given time, a nonparametric model is appropriate. Considering growth, the distributions are continuous in time. The technology that is needed is a mechanism for placing a probability distribution on a collection of nonparametric distributions which evolve in a continuous fashion as a covariate changes, here time. Four main properties were proposed: (i) large (full) support at any fixed covariate value, (ii) feasible

computation for updating prior to posterior, (iii) an understandable marginal distribution at each fixed covariate value, and (iv) continuity of the realized distributions.

The key to constructing a model with these properties is Sethuraman's construction of the Dirichlet process. For a simple version of the dependent Dirichlet process, the single-p model, we replace the draw of a scalar or vector θ_i^* with the realization of a stochastic process whose index set is the covariate space. An infinite sequence of these realizations replaces the infinite set of locations in the Dirichlet process. The draws used to construct the weights are as in Sethuraman's construction–an iid sample from the beta$(1, M)$ distribution. The result is a countable mixture of stochastic processes. At each covariate value x, the random distribution, say F_x, is a countable mixture. The marginal distribution of F_x is a Dirichlet process. Viewed in its entirety, the result is a measure-valued (or probability distribution-valued) stochastic process. These discrete distributions can then be smoothed in the same way that the Dirichlet process is smoothed to form the MDP model.

This recipe for constructing a prior distribution for a collection of distributions has considerable flexibility. The main ingredients are a covariate space which doubles as an index set for stochastic processes, a stochastic process for the locations (the θ_i^*), and a stochastic process for the variates used to construct the weights (the V_i). The index set is imbued with a topology. It need not be restricted to an interval, but can live in $\mathcal{R}^p$ or more general spaces. The locations need not be univariate, but can themselves be more complex objects. To yield marginal Dirichlet processes, the stochastic processes need only have the desired beta$(1, M_x)$ marginal, where the mass may vary with the covariate. The collection of distributions drawn from the dependent Dirichlet process inherit many properties from the stochastic processes used to construct the distributions.

The development of these processes expanded a line of work investigating the split between modelling and inference where modelling takes place in a high-dimensional space and inference in a much lower dimensional space (MacEachern, 2001; Walker and Gutiérrez-Pena, 1999). Recent developments stress the differences between inference in the traditional, low-dimensional parametric setting and in high-dimensional settings (Hahn and Carvalho, 2015; Lee and MacEachern, 2014).

Many variants on and special cases of the dependent Dirichlet process have been developed. MacEachern *et al.* (2001) develop a broader class of distributions which admit more flexible marginals. Gelfand *et al.* (2005) develop the spatial Dirichlet process. Griffin and Steel (2006) consider time series. Teh *et al.* (2006) write of the hierarchical Dirichlet process. Rodriguez *et al.* (2008) work with the nested Dirichlet process. De Iorio *et al.* (2004) write of the linear dependent Dirichlet process. Dunson *et al.* (2007) coined the term "density regression", Dunson and Park (2008) develop kernel stick-breaking priors, and Pennell and Dunson (2006) work with dynamic Dirichlet processes. Theoretical results accompany the models, as in Barrientos *et al.* (2012). Broderick *et al.* (2013) provides complete characterizations of many of these models.

Distinct collections of models have been developed for alternative data structures. To name but two, Griffiths and Ghahramani's (2011) Indian buffet process opened new territory for latent feature models. Orbanz and Roy (2015) pursue the full support argument and create deep results and effective models for network data.

6. Parting words

Only a few decades ago, Bayesian methods were regarded by many in the statistics community as a curiosity. They had strong theoretical support, but there was great difficulty in implementing the methods, especially beyond low-dimensional conjugate settings. With the development of the hierarchical model, modern computational strategies, and the widespread availability of complex data and

fast computing, the methods have proven their value in studies throughout the academic and corporate worlds.

Similarly, many Bayesians in the recent past considered nonparametric Bayesian methods to be a curiosity with strong theoretical support but great difficulty in implementation. The developments that have propelled the success of Bayesian methods have also propelled the success of nonparametric Bayesian methods. Furthermore, in our current data-rich environment, the typical parametric model can be falsified, with a clear need to use a more complex model. Rather than expanding the model slowly by adding a parameter or two at a time, nonparametric Bayesian methods jump directly to an infinite-dimensional parameter space. To make this jump successfully requires some experience with model building and some knowledge of these models in particular. Once this experience is gained, the empirical evidence of success, across a range of problems and by many practitioners, is overwhelming. With a wealth of data and sophisticated models and computation in place, nonparametric Bayesian approaches are assured of an interesting and active future.

This review has touched on only a small portion of the work on nonparametric Bayesian methods. The papers I refer to are tilted toward the topics I have chosen for the review and toward references I am more familiar with. Additional excellent reviews and book include Müller *et al.* (2015), Müller and Mitra (2013), and Müller and Quintana (2004), among others. There is a substantial literature outside of the usual statistical journals, especially in outlets favored by the machine learning community. A google search on limited keywords brings up well over 200,000 hits. Many more references will be found online.

Acknowledgements

This work was supported in part by the United States National Science Foundation grant number DMS-1613110. The views expressed in this work are not necessarily those of the NSF. This paper This paper closely follows a tutorial presented at the meeting of the Bayesian Statistics Section of the Korean Statistical Society on Jeju Island in 2016. The supplementary materials contain the slides from the tutorial.

References

Antoniak CE (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems, *Annals of Statistics*, **2**, 1152–1174.

Barrientos AF, Jara A, and Quintana FA (2012). On the support of MacEachern's dependent Dirichlet processes and extensions, *Bayesian Analysis*, **7**, 277–310.

Bean A, Xu X, and MacEachern SN (2016). Transformations and Bayesian density estimation. To appear in the *Electronic Journal of Statistics*, **10**, 3355–3373.

Berger JO (1985). *Statistical Decision Theory and Bayesian Analysis* (2nd ed), Springer-Verlag, New York.

Berry DA and Christensen R (1979). Empirical Bayes estimation of a binomial parameter via mixtures of Dirichlet processes, *Annals of Statistics*, **7**, 558–568.

Blackwell D and MacQueen JB (1973). Ferguson distributions via Polya urn schemes, *Annals of Statistics*, **1**, 353–355.

Blei DM and Jordan MI (2006). Variational inference for Dirichlet process mixtures, *Bayesian Analysis*, **1**, 121–143.

Broderick T, Pitman J, and Jordan MI (2013). Feature allocations, probability functions, and paintboxes, *Bayesian Analysis*, **8**, 801–836.

Bush CA, Lee J, and MacEachern SN (2010). Minimally informative prior distributions for nonparametric Bayesian analysis, *Journal of the Royal Statistical Society Series B (Statistical Methodology)*, **72**, 253–268.

Bush CA and MacEachern SN (1996). A semiparametric model for randomised block designs, *Biometrika*, **83**, 275–285.

Dahl DB (2003). An improved merge-split sampler for conjugate Dirichlet process mixture models, Department of Statistics, University of Wisconsin. Technical Report 1086.

De Iorio M, Müller P, Rosner G, and MacEachern SN (2004). An ANOVA model for dependent random measures, *Journal of the American Statistical Association*, **99**, 205–215.

Doksum K (1974). Tailfree and neutral random probabilities and their posterior distributions, *Annals of Probability*, **2**, 183–201.

Dunson DB and Park JH (2008). Kernel stick-breaking processes, *Biometrika*, **95**, 307–323.

Dunson DB, Pillai N, and Park JH (2007). Bayesian density regression, *Journal of the Royal Statistical Society Series B (Statistical Methodology)*, **69**, 163–183.

Dykstra RL and Laud P (1981). Bayesian nonparametric approach to reliability, *Annals of Statistics*, **9**, 356–367.

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

Escobar MD (1988). Estimating the means of several normal populations by estimating the distribution of the means (Doctoral dissertation), Yale University, New Haven, CT.

Escobar MD (1994). Estimating normal means with a Dirichlet process prior, *Journal of the American Statistical Association*, **89**, 268–277.

Escobar MD and West M (1995). Bayesian density estimation and inference using mixtures, *Journal of the American Statistical Association*, **90**, 577–588.

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

Gelfand AE and Kottas A (2002). A computational approach for full nonparametric Bayesian inference under Dirichlet process mixture models, *Journal of Computational and Graphical Statistics*, **11**, 289–305.

Gelfand AE, Kottas A, and MacEachern SN (2005). Bayesian nonparametric spatial modeling with Dirichlet process mixing, *Journal of the American Statistical Association*, **100**, 1021–1035.

Gelfand AE and Smith AFM (1990). Sampling-based approaches to calculating marginal densities, *Journal of the American Statistical Association*, **85**, 398–409.

Ghosal S and van der Vaart AW (2001). Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities, *Annals of Statistics*, **29**, 1233–1263.

Ghosh JK and Ramamoorthi RV (2003). *Bayesian Nonparametrics*, Springer, New York.

Griffin JE (2010). Default priors for density estimation with mixture models, *Bayesian Analysis*, **5**, 45–64.

Griffin JE and Steel MFJ (2006). Order-based dependent Dirichlet processes, *Journal of the American Statistical Association*, **101**, 179–194.

Griffiths TL and Ghahramani Z (2011). The Indian buffet process: an introduction and review, *Journal of Machine Learning Research*, **12**, 1185–1224.

Guha S (2008). Posterior simulation in the generalized linear mixed model with semiparametric random effects, *Journal of Computational and Graphical Statistics*, **17**, 410–425.

Hahn PR and Carvalho CM (2015). Decoupled shrinkage and selection in Bayesian linear models: a posterior summary perspective, *Journal of the American Statistical Association*, **110**, 435–448.

Hanson TE (2006). Inference for mixtures of finite Polya tree models, *Journal of the American Statistical Association*, **101**, 1548–1565.

Hjort NL (1990). Nonparametric Bayes estimators based on beta processes in models for life history data, *Annals of Statistics*, **18**, 1259–1294.

Huber PJ (1981). *Robust Statistics*, John Wiley & Sons, New York.

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

James LF, Lijoi A, and Prünster I (2005). Conjugacy as a distinctive feature of the Dirichlet process, *Scandinavian Journal of Statistics*, **33**, 105–120.

Jain S and Neal RM (2004). A split-merge Markov chain Monte Carlo procedure for the Dirichlet process mixture model, *Journal of Computational and Graphical Statistics*, **13**, 158–182.

Jain S and Neal RM (2007). Splitting and merging components of a nonconjugate Dirichlet process mixture model, *Bayesian Analysis*, **2**, 445–472.

Jara A, Hanson T, Quintana FA, Müller P, and Rosner GL (2011). DPpackage: Bayesian semi- and nonparametric modeling in R, *Journal of Statistical Software*, **40**, 1–30.

Johnson W and Christensen R (1986). Bayesian nonparametric survival analysis for grouped data, *Canadian Journal of Statistics*, **14**, 307–314.

Kalli M, Griffin JE, and Walker SG (2011). Slice sampling mixture models, *Statistics and Computing*, **21**, 93–105.

Kessler DC, Hoff PD, and Dunson DB (2014). Marginally specified priors for nonparametric Bayesian estimation, *Journal of the Royal Statistical Society B(Statistical Methodology)*, **77**, 35–58.

Kim Y (1999). Nonparametric Bayesian estimators for counting processes, *Annals of Statistics*, **27**, 562–588.

Kim Y and Lee J (2003). Bayesian bootstrap for proportional hazards models, *Annals of Statistics*, **31**, 1905–1922.

Kleinman KP and Ibrahim JG (1998). A semiparametric Bayesian approach to the random effects model, *Biometrics*, **54**, 921–938.

Kuo L and Smith AF (1992). Bayesian computations in survival models via the Gibbs sampler (with discussion). In JP Klein and PK Goel (Eds), *Survival Analysis: State of the Art* (pp. 11–24), Springer Netherlands, Dordrecht.

Lavine M (1992). Some aspects of Polya tree distributions for statistical modelling, *Annals of Statistics*, **20**, 1222–1235.

Lee J and MacEachern SN (2014). Inference functions in high dimensional Bayesian inference, *Statistics and Its Interface*, **7**, 477–486

Lee J, MacEachern SN, Lu Y, and Mills GB (2014). Local-mass preserving prior distributions for nonparametric Bayesian models, *Bayesian Analysis*, **9**, 307–330.

Lee J, Quintana FA, Müller P, and Trippa L (2013). Defining predictive probability functions for species sampling models, *Statistical Science*, **28**, 209–222.

Lenk PJ (1988). The logistic normal distribution for Bayesian, nonparametric, predictive densities, *Journal of the American Statistical Association*, **83**, 509–516.

Lijoi A, Mena RH, and Prünster I (2005). Hierarchical mixture modelling with normalized inverse-Gaussian priors, *Journal of the American Statistical Association*, **100**, 1278–1291.

Liu JS (1996). Nonparametric hierarchical Bayes via sequential imputations, *Annals of Statistics*, **24**, 910–930.

Lo AY (1984). On a class of Bayesian nonparametric estimates: I. Density estimates, *Annals of Statistics*, **12**, 351–357.

MacEachern SN (1988). Sequential Bayesian bioassay design (Doctoral dissertation), University of Minnesota, Minneapolis, MN.

MacEachern SN (1994). Estimating normal means with a conjugate style Dirichlet process prior, *Communications in Statistics - Simulation and Computation*, **23**, 727–741.

MacEachern SN (1999). Dependent nonparametric processes, in *American Statistical Association 1999 Proceedings of the Section on Bayesian Statistics*, Alexandria, VA, 50–55.

MacEachern SN (2000). *Dependent Dirichlet Processes*, The Ohio State University, Department of Statistics, Columbus, OH.

MacEachern SN (2001). Decision theoretic aspects of dependent nonparametric processes, in *In Bayesian Methods with Applications to Science, Policy, and Official Statistics*, (pp. 551–560), Eurostat, Luxembourg.

MacEachern SN (2007). Comment on article by Jain and Neal, *Bayesian Analysis*, **2**, 483–494.

MacEachern SN, Clyde M, and Liu JS (1999). Sequential importance sampling for nonparametric Bayes models: the next generation, *Canadian Journal of Statistics*, **27**, 251–267.

MacEachern SN and Guha S (2011). Parametric and semiparametric hypotheses in the linear model, *Canadian Journal of Statistics*, **39**, 165–180.

MacEachern SN, Kottas A, and Gelfand AE (2001). Spatial nonparametric Bayesian models, In *Proceedings of the 2001 Joint Statistical Meetings*, Atlanta, GA.

MacEachern SN and Müller P (1998). Estimating mixture of Dirichlet process models, *Journal of Computational and Graphical Statistics*, **7**, 223–238.

Martin R and Tokdar ST (2009). Asymptotic properties of predictive recursion: robustness and rate of convergence, *Electronic Journal of Statistics*, **3**, 1455–1472.

Mauldin RD, Sudderth WD, and Williams SC (1992). Polya trees and random distributions, *Annals of Statistics*, **20**, 1203–1221.

Müller P, Erkanli A, and West M (1996). Bayesian curve fitting using multivariate normal mixtures, *Biometrika*, **83**, 67–79.

Müller P and Mitra R (2013). Bayesian nonparametric inference: why and how, *Bayesian Analysis*, **8**, 1–35

Müller P and Quintana FA (2004). Nonparametric Bayesian data analysis, *Statistical Science*, **19**, 95–110.

Müller P, Quintana FA, Jara A, and Hanson T (2015). *Bayesian Nonparametric Data Analysis*, Springer, New York.

Müller P, Quintana FA, and Rosner G (2004). A method for combining inference across related nonparametric Bayesian models, *Journal of the Royal Statistical Society B (Statistical Methodology)*, **66**, 735–749.

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

Newton MA and Raftery AE (1994). Approximate Bayesian inference with the weighted likelihood bootstrap, *Journal of the Royal Statistical Society B (Methodological)*, **56**, 3–48.

Newton MA and Zhang Y (1999). A recursive algorithm for nonparametric analysis with missing data, *Biometrika*, **86**, 15–26.

Orbanz P and Roy DM (2015). Bayesian models of graphs, arrays, and other exchangeable random structures, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **37**, 437–461.

Pennell ML and Dunson DB (2006). Bayesian semiparametric dynamic frailty models for multiple event time data, *Biometrics*, **62**, 1044–1052.

Petrone S (1999). Bayesian density estimation using Bernstein polynomials, *Canadian Journal of*

Statistics, **27**, 105–126.

Regazzini E, Lijoi A, and Prünster I (2003). Distributional results for means of normalized random measures with independent increments, *Annals of Statistics*, **31**, 560–585.

Rodriguez A, Dunson DB, and Gelfand AE (2008). The nested Dirichlet process, *Journal of the American Statistical Association*, **103**, 1131–1154.

Rubin DB (1981). The Bayesian bootstrap, *Annals of Statistics*, **9**, 130–134.

Savage LJ (1954). *The Foundations of Statistics*, John Wiley & Sons, New York.

Sethuraman J (1994). A constructive definition of Dirichlet priors, *Statistica Sinica*, **4**, 639–650.

Susarla V and Van Ryzin J (1976). Nonparametric Bayesian estimation of survival curves from incomplete observations, *Journal of the American Statistical Association*, **71**, 897–902.

Teh YW, Jordan MI, Beal MJ, and Blei DM (2006). Hierarchical Dirichlet processes, *Journal of the American Statistical Association*, **101**, 1566–1581.

Tokdar ST (2007). Towards a faster implementation of density estimation with logistic Gaussian process priors, *Journal of Computational and Graphical Statistics*, **16**, 633–655.

Tomlinson GA (1998). Analysis of densities (Doctoral dissertation), University of Toronto, ON.

Walker SG (2004). New approaches to Bayesian consistency, *Annals of Statistics*, **32**, 2028–2043.

Walker SG (2007). Sampling the Dirichlet mixture model with slices, *Communications in Statistics - Simulation and Computation*, **36**, 45–54.

Walker SG, Damien P, Laud PW, and Smith AFM (1999). Bayesian nonparametric inference for random distributions and related functions, *Journal of the Royal Statistical Society B (Statistical Methodology)*, **61**, 485–527.

Walker SG and Gutiérrez-Pena E (1999). Robustifying Bayesian procedures, *Bayesian Statistics*, **6**, 685-710.

Wang Z (2009). Semiparametric Bayesian models extending weighted least squares (Doctoral dissertation), The Ohio State University, Columbus, OH.

Xu X, Lu P, MacEachern SN, and Xu R (2012). Calibrated Bayes factor for model comparison and prediction, Department of Statistics, The Ohio State University, Technical Report.

Yang L and Marron JS (1999). Iterated transformation-kernel density estimation, *Journal of the American Statistical Association*, **94**, 580–589.

Received October 22, 2016; Revised November 2, 2016; Accepted November 8, 2016