1

Bayesian nonparametric methods: motivation and ideas

Stephen G. Walker

It is now possible to demonstrate many applications of Bayesian nonparametric methods. It works. It is clear, however, that nonparametric methods are more complicated to understand, use and derive conclusions from, when compared to their parametric counterparts. For this reason it is imperative to provide specific and comprehensive motivation for using nonparametric methods. This chapter aims to do this, and the discussions in this part are restricted to the case of independent and identically distributed (i.i.d.) observations. Although such types of observation are quite specific, the arguments and ideas laid out in this chapter can be extended to cover more complicated types of observation. The usefulness in discussing i.i.d. observations is that the maths is simplified.

1.1 Introduction

Even though there is no physical connection between observations, there is a real and obvious reason for creating a dependence between them from a modeling perspective. The first observation, say X_1, provides information about the unknown density f from which it came, which in turn provides information about the second observation X_2, and so on. How a Bayesian learns is her choice but it is clear that with i.i.d. observations the order of learning should not matter and hence we enter the realms of *exchangeable* learning models. The mathematics is by now well known (de Finetti, 1937; Hewitt and Savage, 1955) and involves the construction of a prior distribution $\Pi(\mathrm{d}f)$ on a suitable space of density functions. The learning mechanism involves updating $\Pi(\mathrm{d}f)$ via Bayes' theorem, so that after n observations beliefs about f are now encapsulated in the posterior distribution, given by

$$\Pi(\mathrm{d}f|X_1,\ldots,X_n) = \frac{\prod_{i=1}^n f(X_i)\,\Pi(\mathrm{d}f)}{\int \prod_{i=1}^n f(X_i)\,\Pi(\mathrm{d}f)}$$

and this in turn provides information about the future observation X_{n+1} via the predictive density

$$f(X_{n+1}|X_1,\ldots,X_n) = \int f(X_{n+1})\,\Pi(\mathrm{d}f|X_1,\ldots,X_n).$$

From this it is easy to see that the prior represents what has been learnt about the unknown density function without the presence of any of the observations. Depending on how much is known at this point, that is with no observations, the strength of the prior ranges from very precise with a lot of information, to so-called noninformative or default priors which typically are so disperse that they are even improper (see e.g. Kass and Wasserman, 1996).

This prior distribution is a single object and is a prior distribution on a suitable space of density (or equivalent) functions. Too many Bayesians think of the notion of a likelihood and a prior and this can be a hindrance. The fundamental idea is the construction of random density functions, such as normal shapes, with random means and variances; or the infinite-dimensional exponential family, where probabilities are assigned to the infinite collection of random parameters. It is instructive to think of all Bayesians as constructing priors on spaces of density functions, and it is clear that this is the case. The Bayesian nonparametric statistician is merely constructing random density functions with unrestricted shapes.

This is achieved by modeling random density functions, or related functions such as distribution functions and hazard functions, using stochastic processes; Gaussian processes and independent increment processes are the two most commonly used. The prior is the law governing the stochastic process. The most commonly used is the Dirichlet process (Ferguson, 1973) which has sample paths behaving almost surely as a discrete distribution function. They appear most often as the mixing distribution generating random density functions: the so-called mixture of Dirichlet process model (Lo, 1984), which has many pages dedicated to it within this book. This model became arguably the most important prior for Bayesian nonparametrics with the advent of sampling based approaches to Bayesian inference, which arose in the late 1980s (Escobar, 1988).

The outline of this chapter is as follows. In Section 1.2 we consider the important role that Bayesian nonparametrics plays. Ideas for providing information for nonparametric priors are also discussed. Section 1.3 discusses how many of the practices and high-dimensional activities of Bayesians can be carried out coherently under the umbrella of the nonparametric model. The special case when the nonparametric prior is taken as the Bayesian bootstrap is considered. Section 1.4 discusses the importance of asymptotic studies. Section 1.5 is a direct consequence of recent consistency studies which put the model assumptions and true sampling assumptions at odds with each other. This section provides an alternative derivation of the Bayesian posterior distribution using loss functions; as such it is no less a rigorous approach to constructing a learning model than is the traditional approach using the Bayes theorem. So Section 1.5 can be thought of as "food for thought." Finally, Section 1.6 concludes with a brief discussion.

1.2 Bayesian choices

Many of the questions posed to the nonparametric methods are of the type "what if this and what if that?" referring to the possibility that the true density is normal or some other low-dimensional density and so using many parameters is going to be highly inefficient. In truth, it is these questions that are more appropriately directed to those who consistently use low-dimensional densities for modeling: "what if the model is not normal?"

However, there was a time, and not so long ago, in fact pre–Markov chain Monte Carlo, when Bayesian methods were largely restricted to a few parametric models, such as the normal, and the use of conjugate prior distributions. Box and Tiao (1973) was as deep as it got. It is therefore not surprising that in this environment, where only simple models were available, the ideas of model selection and model comparison took hold, for the want of something to do and a need to compare log–normal and Weibull distributions. Hence, such model assessments were vital, irrespective of any formal views one may have had about the theory of Bayesian methods (see Bernardo and Smith, 1994, Chapter 2). But it is not difficult to argue that Bayesian model criticism is unsound, and the word that is often used is *incoherent.*

To argue this point, let us keep to the realm of independent and identically distributed observations. In this case, the prior distribution is a probability measure on a space of density functions. This is true for all Bayesians, even those relying on the normal distribution, in which case the Bayesian is putting probability one on the shape of the density function matching those of the normal family.

There is more responsibility on the Bayesian: she gets more out in the form of a posterior distribution on the object of interest. Hence more care needs to be taken in what gets put into the model in the first place. For the posterior to mean anything it must be representing genuine posterior beliefs, solely derived by a combination of the data and prior beliefs via the use of the Bayes theorem. Hence, the prior used must genuinely represent prior beliefs (beliefs without data). If it does not, how can the posterior represent posterior beliefs? So a "prior" that has been selected post data via some check and test from a set of possible "prior" distributions cannot represent genuine prior beliefs. This is obvious, since no one of these "priors" can genuinely represent prior beliefs. The posterior distributions based on such a practice are meaningless.

The prior must encapsulate prior beliefs and be large enough to accommodate all uncertainties. As has been mentioned before, years back prior distributions could not be enlarged to accommodate such problems, and the incoherence of model (prior) selection was adopted for pragmatic reasons, see Box (1980). However, nowadays, it is quite straightforward to build large prior distributions and to

undertake prior to posterior analysis. How large a prior should be is a clear matter. It is large enough so that no matter what subsequently occurs, the prior is not checked. Hence, in may cases, it is only going to be a nonparametric model that is going to suffice.

If a Bayesian has a prior distribution and suspects there is additional uncertainty, there are two possible actions. The first is to consider an alternative prior and then select one or the other after the data have been observed. The second action is to enlarge the prior before observing the data to cover the additional uncertainty. It is the latter action which is correct and coherent.

Some Bayesians would argue that it is too hard a choice to enlarge the prior or work with nonparametric priors, particularly in specifying information or putting beliefs into nonparametric priors. If this is the case, though I do not believe it to be true, then it is a matter of further investigation and research to overcome the difficulties rather than to lapse into pseudo-Bayesian and incoherent practices.

To discuss the issue of pinning down a nonparametric prior we can if needed do this in a parametric frame of mind. For the nonparametric model one typically has two functions to specify which relate to $\mu_1(x) = \mathrm{E}f(x)$ and $\mu_2(x) = \mathrm{E}f^2(x)$. If it is possible to specify such functions then a nonparametric prior has typically been pinned down. Two such functions are easy to specify. They can, for example, be obtained from a parametric model, even the normal, in which case one would take

$$\mu_1(x) = \int \mathrm{N}(x|\theta, \sigma^2)\,\pi(\mathrm{d}\theta, \mathrm{d}\sigma)$$

$$\mu_2(x) = \int \mathrm{N}^2(x|\theta, \sigma^2)\,\pi(\mathrm{d}\theta, \mathrm{d}\sigma),$$

for some probability measure $\pi(\mathrm{d}\theta, \mathrm{d}\sigma)$. The big difference now is that a Bayesian using this normal model, i.e.

$$X \sim \mathrm{N}(\theta, \sigma^2) \quad \text{and} \quad (\theta, \sigma) \sim \pi(\theta, \sigma),$$

would be restricted to normal shapes, whereas the nonparametric Bayesian, whose prior beliefs about μ_1 and μ_2, equivalently $\mathrm{E}f(x)$ and $\mathrm{Var}f(x)$, coincide with the parametric Bayesian, has unrestricted shapes to work with.

A common argument is that it is not possible to learn about all the parameters of a nonparametric model. This spectacularly misses the point. Bayesian inference is about being willing and able to specify all uncertainties into a prior distribution. If one does not like the outcome, do not be a Bayesian. Even a parametric model needs a certain amount of data to learn anything reasonable and the nonparametric model, which reflects greater starting uncertainty than a parametric model, needs more data to overcome the additional starting uncertainty. But it is not right to wish away the prior uncertainty or purposefully to underestimate it.

1.3 Decision theory

Many of the Bayesian procedures based on incomplete priors (i.e. priors for which all uncertainty has not been taken into account) can be undertaken coherently (i.e. using a complete prior) using decision theory. Any selection of parametric models can be done under the umbrella of the complete prior. This approach makes extensive use of the utility function for assessing the benefit of actions (such as model selection etc.) when one has presumed a particular value for the correct but unknown density function. Let us consider an example. Which specific density from a family of densities indexed by a parameter $\theta \in \Theta$ is the best approximation to the data?

If the parametric family of densities is $\{f(x;\theta)\}$, then the first task is to choose a utility function which describes the reward in selecting θ, for the parameter space is the action space, when f is the true density. Basing this on a distance between densities seems appropriate here, so we can take

$$u(f, \theta) = -d(f(\cdot;\theta), f(\cdot)).$$

The prior is the nonparametric one, or the complete prior $\Pi(\mathrm{d}f)$, and so making decisions on the basis of the maximization of expected utility, the choice of θ is $\widehat{\theta}$ which maximizes

$$U_n(\theta) = -\int d(f(\cdot;\theta), f(\cdot))\, \Pi(\mathrm{d}f|X_1, \ldots, X_n).$$

An interesting special case arises when we take d to be based on the Kullback–Leibler divergence; that is $d(g, f) = \int g \log(g/f)$ in which case we would choose $\widehat{\theta}$ to maximize

$$\tilde{U}_n(\theta) = \int \log f(x;\theta)\, f_n(\mathrm{d}x)$$

where f_n is the nonparametric predictive density, given by

$$f_n(x) = \int f(x)\, \Pi(\mathrm{d}f|X_1, \ldots, X_n).$$

Furthermore, taking $\Pi(\mathrm{d}f)$ to be the Bayesian bootstrap (Rubin, 1981), so that f_n is the density with point mass $1/n$ at each of the data points, then

$$\tilde{U}_n(\theta) = n^{-1} \sum_{i=1}^{n} \log f(X_i;\theta)$$

and so $\widehat{\theta}$ is the maximum likelihood estimator.

There are many other types of lower dimensional decisions that can be made under the larger prior/posterior; see Gutièrrez-Peña and Walker (2005). As an example, suppose it is required to construct a probability on Θ space when the true posterior is $\Pi(\mathrm{d}f|X_1, \ldots, X_n)$. It is necessary to link up a random f from this

posterior with a random θ from Θ space. This can be done by taking θ to maximize $u(f, \theta)$. An interesting special case arises when the posterior is once again taken to be the Bayesian bootstrap in which case we can take

$$f_n(\mathrm{d}x) = \sum_{i=1}^{n} w_i \, \delta_{X_i}(\mathrm{d}x),$$

where the $(w_1, \ldots, w_n)$ are from a Dirichlet distribution with parameters all equal to 1. Therefore, a distribution on Θ space can be obtained by repeated simulation of the weights from the Dirichlet distribution and taking θ to maximize

$$\sum_{i=1}^{n} w_i \, \log f(X_i; \theta).$$

This is precisely the weighted likelihood bootstrap approach to Bayesian inference proposed by Newton and Raftery (1994).

To set up the scene for the next section, let us note that if a Bayesian is making such assessments on utilities, in order to undertake decision theory, then she must be willing to think about the true density function and that this comes from a set of possibilities. How is it possible to make such judgments while having discarded the notion of a true density function?

1.4 Asymptotics

Traditionally, Bayesians have shunned this aspect of statistical inference. The prior and data yield the posterior and the subjectiveness of this strategy does not need the idea of what happens if further data arise. Anyway, there was the theorem of Doob (1949), but like other Bayesian computations from the past, this theorem involves assuming that the final distribution of the observations depends explicitly on and is fully specified by the chosen prior distribution, that is

$$p(X_1, \ldots, X_n) = \int \prod_{i=1}^{n} f(X_i) \, \Pi(\mathrm{d}f).$$

It is unrealistic to undertake asymptotic studies, or indeed any other Bayesian studies, based on this assumption, since it is not true. Doob's theorem relies on this assumption. Even though one knows that this model is mathematically incorrect, it does serve as a useful learning model, as discussed earlier.

On the other hand, it is correct to assume the observations are independent and identically distributed from some true density function f_0 and to undertake the mathematics on this assumption. One is then asking that the posterior distribution accumulates in suitable neighborhoods of this true density function.

This exposes the Bayesian model as being quite different from the correct assumption. There is no conflict here in the discrepancy between the true assumption and the model assumption. The Bayesian model is about learning from observations in a way that the order in which they arrive does not matter (exchangeability). The first observation provides information about the true density function and this in turn provides information about the second observation and so on. The Bayesian writes down how this learning is achieved and specifically how an observation provides information about the true density function. In this approach one obviously needs to start with initial or prior information about the true density function.

In short, the Bayesian believes the data are i.i.d. from some true density function f_0 and then writes down an exchangeable learning model as to how they see the observations providing information about f_0.

So why is consistency important? The important point is that the prior, which fully specifies the learning model, is setting up the learning model. In a way it is doing two tasks. One is representing prior beliefs, learnt about f_0 before or without the presence of data, and the second is fully specifying the learning model. It is this latter task that is often neglected by subjective Bayesians.

Hence, the learning part of the model needs to be understood. With an unlimited amount of data the Bayesian must expect to be able to pin down the density generating her observations exactly. It is perfectly reasonable to expect that as data arrive the learning is going in the right direction and that the process ends up at f_0. If it does not then the learning model (prior) has not been set well, even though the prior might be appropriate as representing prior beliefs.

The basic idea is to ensure that

$$\Pi(d(f, f_0) > \epsilon | X_1, \ldots, X_n) \to 0 \text{ a.s. } F_0^\infty$$

where d is some measure of distance between densities. It is typically taken to be the Hellinger distance since this favors the mathematics. Conditions are assigned to Π to ensure this happens and involve a support condition and a further condition which ensures that the densities which can track the data too closely are given sufficiently low prior mass, see Chapter 2.

However, an alternative "likelihood," given by

$$l_n^{(\alpha)} = \prod_{i=1}^{n} f(X_i)^\alpha$$

for any $0 < \alpha < 1$ yields Hellinger consistency with only a support condition. Can this approach be justified? It possibly can. For consider a cumulative loss function approach to posterior inference, as in the next section.

2

The Dirichlet process, related priors and posterior asymptotics

Subhashis Ghosal

Here we review the role of the Dirichlet process and related prior distribtions in nonparametric Bayesian inference. We discuss construction and various properties of the Dirichlet process. We then review the asymptotic properties of posterior distributions. Starting with the definition of posterior consistency and examples of inconsistency, we discuss general theorems which lead to consistency. We then describe the method of calculating posterior convergence rates and briefly outline how such rates can be computed in nonparametric examples. We also discuss the issue of posterior rate adaptation, Bayes factor consistency in model selection and Bernshteĭn–von Mises type theorems for nonparametric problems.

2.1 Introduction

Making inferences from observed data requires modeling the data-generating mechanism. Often, owing to a lack of clear knowledge about the data-generating mechanism, we can only make very general assumptions, leaving a large portion of the mechanism unspecified, in the sense that the distribution of the data is not specified by a finite number of parameters. Such nonparametric models guard against possible gross misspecification of the data-generating mechanism, and are quite popular, especially when adequate amounts of data can be collected. In such cases, the parameters can be best described by functions, or some infinite-dimensional objects, which assume the role of parameters. Examples of such infinite-dimensional parameters include the cumulative distribution function (c.d.f.), density function, nonparametric regression function, spectral density of a time series, unknown link function in a generalized linear model, transition density of a Markov chain and so on. The Bayesian approach to nonparametric inference, however, faces challenging issues since construction of prior distribution involves specifying appropriate probability measures on function spaces where the parameters lie. Typically, subjective knowledge about the minute details of the distribution on these infinite-dimensional spaces is not available for nonparametric problems. A prior distribution is generally chosen based on tractability, computational convenience and desirable frequentist

behavior, except that some key parameters of the prior may be chosen subjectively. In particular, it is desirable that a chosen prior is spread all over the parameter space, that is, the prior has large topological *support*. Together with additional conditions, large support of a prior helps the corresponding posterior distribution to have good frequentist properties in large samples. To study frequentist properties, it is assumed that there is a true value of the unknown parameter which governs the distribution of the generated data.

We are interested in knowing whether the posterior distribution eventually concentrates in the neighborhood of the true value of the parameter. This property, known as *posterior consistency*, provides the basic frequentist validation of a Bayesian procedure under consideration, in that it ensures that with a sufficiently large amount of data, it is nearly possible to discover the truth accurately. Lack of consistency is extremely undesirable, and one should not use a prior if the corresponding posterior is inconsistent. However, consistency is satisfied by many procedures, so typically more effort is needed to distinguish between consistent procedures. The speed of convergence of the posterior distribution to the true value of the parameter may be measured by looking at the smallest shrinking ball around the true value which contains posterior probability nearly one. It will be desirable to pick up the prior for which the size of such a shrinking ball is the minimum possible. However, in general it is extremely hard to characterize size exactly, so we shall restrict ourselves only to the rate at which a ball around the true value can shrink while retaining almost all of the posterior probability, and call this the *rate of convergence* of the posterior distribution. We shall also discuss adaptation with respect to multiple models, consistency for model selection and *Bernshteĭn–von Mises theorems*.

In the following sections, we describe the role of the *Dirichlet process* and some related prior distributions, and discuss their most important properties. We shall then discuss results on convergence of posterior distributions, and shall often illustrate results using priors related to the Dirichlet process. At the risk of being less than perfectly precise, we shall prefer somewhat informal statements and informal arguments leading to these results. An area which we do not attempt to cover is that of Bayesian survival analysis, where several interesting priors have been constructed and consistency and rate of convergence results have been derived. We refer readers to Ghosh and Ramamoorthi (2003) and Ghosal and van der Vaart (2010) as general references for all topics discussed in this chapter.

2.2 The Dirichlet process

2.2.1 Motivation

We begin with the simplest nonparametric inference problem for an uncountable sample space, namely, that of estimating a probability measure (equivalently, a

c.d.f.) on the real line, with independent and identically distributed (i.i.d.) observations from it, where the c.d.f. is completely arbitrary. Obviously, the classical estimator, the empirical distribution function, is well known and is quite satisfactory. A Bayesian solution requires describing a random probability measure and developing methods of computation of the posterior distribution. In order to understand the idea, it is fruitful to look at the closest parametric relative of the problem, namely the multinomial model. Observe that the multinomial model specifies an arbitrary probability distribution on the sample space of finitely many integers, and that a multinomial model can be derived from an arbitrary distribution by grouping the data in finitely many categories. Under the operation of grouping, the data are reduced to counts of these categories. Let $(\pi_1, \ldots, \pi_k)$ be the probabilities of the categories with frequencies $n_1, \ldots, n_k$. Then the likelihood is proportional to $\pi_1^{n_1} \cdots \pi_k^{n_k}$. The form of the likelihood matches with the form of the finite-dimensional Dirichlet prior, which has density † proportional to $\pi_1^{c_1-1} \cdots \pi_k^{c_k-1}$. Hence the posterior density is proportional to $\pi_1^{n_1+c_1-1} \cdots \pi_k^{n_k+c_k-1}$, which is again a Dirichlet distribution.

With this nice conjugacy property in mind, Ferguson (1973) introduced the idea of a Dirichlet process – a probability distribution on the space of probability measures which induces finite-dimensional Dirichlet distributions when the data are grouped. Since grouping can be done in many different ways, reduction to a finite-dimensional Dirichlet distribution should hold under any grouping mechanism. In more precise terms, this means that for any finite measurable partition $\{B_1, \ldots, B_k\}$ of $\mathbb{R}$, the joint distribution of the probability vector $(P(B_1), \ldots, P(B_k))$ is a finite-dimensional Dirichlet distribution. This is a very rigid requirement. For this to be true, the parameters of the finite-dimensional Dirichlet distributions need to be very special. This is because the joint distribution of $(P(B_1), \ldots, P(B_k))$ should agree with other specifications such as those derived from the joint distribution of the probability vector $(P(A_1), \ldots, P(A_m))$ for another partition $\{A_1, \ldots, A_m\}$ finer than $\{B_1, \ldots, B_k\}$, since any $P(B_i)$ is a sum of some $P(A_j)$. A basic property of a finite-dimensional Dirichlet distribution is that the sums of probabilities of disjoint chunks again give rise to a joint Dirichlet distribution whose parameters are obtained by adding the parameters of the original Dirichlet distribution. Letting $\alpha(B)$ be the parameter corresponding to $P(B)$ in the specified Dirichlet joint distribution, it thus follows that $\alpha(\cdot)$ must be an additive set function. Thus it is a prudent strategy to let α actually be a measure. Actually, the countable additivity of α will be needed to bring in countable additivity of the random P constructed in this way. The whole idea can be generalized to an abstract Polish space.

† Because of the restriction $\sum_{i=1}^{k} \pi_i = 1$, the density has to be interpreted as that of the first $k-1$ components.

Definition 2.1 Let α be a finite measure on a given Polish space $\mathfrak{X}$. A random measure P on $\mathfrak{X}$ is called a Dirichlet process if for every finite measurable partition $\{B_1, \ldots, B_k\}$ of $\mathfrak{X}$, the joint distribution of $(P(B_1), \ldots, P(B_k))$ is a k-dimensional Dirichlet distribution with paramaeters $\alpha(B_1), \ldots, \alpha(B_k)$.

We shall call α the base measure of the Dirichlet process, and denote the Dirichlet process measure by $\mathcal{D}_\alpha$.

Even for the case when α is a measure so that joint distributions are consistently specified, it still remains to be shown that the random set function P is a probability measure. Moreover, the primary motivation for the Dirichlet process was to exploit the conjugacy under the grouped data setting. Had the posterior distribution been computed based on conditioning on the counts for the partitioning sets, we would clearly retain the conjugacy property of finite-dimensional Dirichlet distributions. However, as the full data are available under the setup of continuous data, a gap needs to be bridged. We shall see shortly that both issues can be resolved positively.

2.2.2 *Construction of the Dirichlet process*

Naive construction

At first glance, because joint distributions are consistently specified, viewing P as a function from the Borel σ-field $\mathscr{B}$ to the unit interval, a measure with the specified marginals can be constructed on the uncountable product space $[0, 1]^{\mathscr{B}}$ with the help of Kolmogorov's consistency theorem. Unfortunately, this simple strategy is not very fruitful for two reasons. First, the product σ-field on $[0, 1]^{\mathscr{B}}$ is not rich enough to contain the space of probability measures. This difficulty can be avoided by working with outer measures, provided that we can show that P is a.s. countably additive. For a given sequence of disjoint sets A_n, it is indeed true that $P(\cup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} P(A_n)$ a.s. Unfortunately, the null set involved in the a.s. statement is dependent on the sequence A_n, and since the number of such sequences is uncountable, the naive strategy using the Kolmogorov consistency theorem fails to deliver the final result.

Construction using a countable generator

To save the above construction, we need to work with a countable generating field $\mathscr{F}$ for $\mathscr{B}$ and view each probability measure P as a function from $\mathscr{F}$ to $[0, 1]$. The previously encountered measure theoretic difficulties do not arise on the countable product $[0, 1]^{\mathscr{F}}$.

Construction by normalization

There is another construction of the Dirichlet process which involves normalizing a *gamma process* with intensity measure α. A gamma process is an independent

increment process whose existence is known from the general theory of *Lévy processes*. The gamma process representation of the Dirichlet process is particularly useful for finding the distribution of the mean functional of P and estimating of the tails of P when P follows a Dirichlet process on $\mathbb{R}$.

2.2.3 Properties

Once the Dirichlet process is constructed, some of its properties are immediately obtained.

Moments and marginal distribution

Considering the partition $\{A, A^c\}$, it follows that $P(A)$ is distributed as Beta$(\alpha(A), \alpha(A^c))$. Thus in particular, $\mathrm{E}(P(A)) = \alpha(A)/(\alpha(A) + \alpha(A^c)) = G(A)$, where $G(A) = \alpha(A)/M$, a probability measure and $M = \alpha(\mathbb{R})$, the total mass of α. This means that if $X|P \sim P$ and P is given the measure $\mathcal{D}_\alpha$, then the marginal distribution of X is G. We shall call G the *center measure*. Also, observe that $\mathrm{Var}(P(A)) = G(A)G(A^c)/(M+1)$, so that the prior is more tightly concentrated around its mean when M is larger, that is, the prior is more precise. Hence the parameter M can be regarded as the *precision parameter*. When P is distributed as the Dirichlet process with base measure $\alpha = MG$, we shall often write $P \sim \mathrm{DP}(M, G)$.

Linear functionals

If ψ is a G-integrable function, then $\mathrm{E}(\int \psi \mathrm{d}P) = \int \psi \mathrm{d}G$. This holds for indicators from the relation $\mathrm{E}(P(A)) = G(A)$, and then standard measure theoretic arguments extend this sequentially to simple measurable functions, nonnegative measurable functions and finally to all integrable functions. The distribution of $\int \psi \mathrm{d}P$ can also be obtained analytically, but this distribution is substantially more complicated than beta distribution followed by $P(A)$. The derivation involves the use of a lot of sophisticated machinery. Interested readers are referred to Regazzini, Guglielmi and Di Nunno (2002), Hjort and Ongaro (2005), and references therein.

Conjugacy

Just as the finite-dimensional Dirichlet distribution is conjugate to the multinomial likelihood, the Dirichlet process prior is also conjugate for estimating a completely unknown distribution from i.i.d. data. More precisely, if $X_1, \ldots, X_n$ are i.i.d. with distribution P and P is given the prior $\mathcal{D}_\alpha$, then the posterior distribution of P given $X_1, \ldots, X_n$ is $\mathcal{D}_{\alpha + \sum_{i=1}^n \delta_{X_i}}$.† To see this, we need to show that for any measurable finite partition $\{A_1, \ldots, A_k\}$, the posterior distribution of $(P(A_1), \ldots, P(A_k))$

† Of course, there are other versions of the posterior distribution which can differ on a null set for the joint distribution.

given $X_1, \ldots, X_n$ is k-dimensional Dirichlet with parameters $\alpha(A_j) + N_j$, where $N_j = \sum_{i=1}^n 1\!\!1\{X_i \in A_j\}$, the count for A_j, $j = 1, \ldots, k$. This certainly holds by the conjugacy of the finite-dimensional Dirichlet prior with respect to the multinomial likelihood had the data been coarsened to only the counts $N_1, \ldots, N_k$. Therefore, the result will follow if we can show that the additional information contained in the original data $X_1, \ldots, X_n$ is irrelevant as far as the posterior distribution of $(P(A_1), \ldots, P(A_k))$ is concerned. One can show this by first considering a partition $\{B_1, \ldots, B_m\}$ finer than $\{A_1, \ldots, A_k\}$, computing the posterior distribution of $(P(B_1), \ldots, P(B_m))$ given the counts of $\{B_1, \ldots, B_m\}$, and marginalizing to the posterior distribution of $(P(A_1), \ldots, P(A_k))$ given the counts of $\{B_1, \ldots, B_m\}$. By the properties of finite-dimensional Dirichlet, this coincides with the posterior distribution of $(P(A_1), \ldots, P(A_k))$ given the counts of $\{A_1, \ldots, A_k\}$. Now making the partitions infinitely finer and applying the martingale convergence theorem, the final result is obtained.

Posterior mean

The above expression for the posterior distribution combined with the formula for the mean of a Dirichlet process imply that the posterior mean of P given $X_1, \ldots, X_n$ can be expressed as

$$\tilde{\mathbb{P}}_n = \mathrm{E}(P|X_1, \ldots, X_n) = \frac{M}{M+n} G + \frac{n}{M+n} \mathbb{P}_n, \tag{2.1}$$

a convex combination of the prior mean and the empirical distribution. Thus the posterior mean essentially shrinks the empirical distribution towards the prior mean. The relative weight attached to the prior is proportional to the total mass M, giving one more reason to call M the precision parameter, while the weight attached to the empirical distribution is proportional to the number of observations it is based on.

Limits of the posterior

When n is kept fixed, letting $M \to 0$ may be regarded as making the prior imprecise or noninformative. The limiting posterior, namely $\mathcal{D}_{\sum_{i=1}^n \delta_{X_i}}$, is known as the Bayesian bootstrap. Samples from the Bayesian bootstrap are discrete distributions supported at only the observation points whose weights are distributed according to the Dirichlet distribution, and hence the *Bayesian bootstrap* can be regarded as a resampling scheme which is smoother than Efron's bootstrap. On the other hand, when M is kept fixed and n varies, the asymptotic behavior of the posterior mean is entirely controlled by that of the empirical distribution. In particular, the c.d.f. of $\tilde{\mathbb{P}}_n$ converges uniformly to the c.d.f. of the true distribution P_0 and $\sqrt{n}(\tilde{\mathbb{P}}_n - P_0)$ converges weakly to a Brownian bridge process. Further, for any

set A, the posterior variance of $P(A)$ is easily seen to be $O(n^{-1})$ as $n \to \infty$. Hence Chebyshev's inequality implies that the posterior distribution of $P(A)$ approaches the degenerate distribution at $P_0(A)$, that is, the posterior distribution of $P(A)$ is consistent at P_0, and the rate of this convergence is $n^{-1/2}$. Shortly, we shall see that the entire posterior of P is also consistent at P_0.

Lack of smoothness

The presence of the point masses δ_{X_i} in the base measure of the posterior Dirichlet process gives rise to some peculiar behavior. One such property is the total disregard of the topology of the sample space. For instance, if A is a set such that many observations fall close to it but A itself does not contain any observed point, then the posterior mean of $P(A)$ is smaller than its prior mean. Thus the presence of observations in the vicinity does not enhance the assessment of the probability of a set unless the observations are actually contained there. Hence it is clear that the Dirichlet process is somewhat primitive in that it does not offer any smoothing, quite unlike the characteristic of a Bayes estimator.

Negative correlation

Another peculiar property of the Dirichlet process is negative correlation between probabilities of any two disjoint sets. For a random probability distribution, one may expect that the masses assigned to nearby places increase or decrease together, so the blanket negative correlation attached by the Dirichlet process may be disappointing. This again demonstrates that the topology of the underlying space is not considered by the Dirichlet process in its mass assignment.

Discreteness

A very intriguing property of the Dirichlet process is the discreteness of the distributions sampled from it, even when G is purely nonatomic. This property also has its roots in the expression for the posterior of a Dirichlet process. To see why this is so, observe that a distribution P is discrete if and only if $P(x : P\{x\} > 0) = 1$. Now, considering the model $X|P \sim P$ and P given $\mathcal{D}_\alpha$ measure, the property holds if

$$(\mathcal{D}_\alpha \times P)\{(P, x) : P\{x\} > 0\} = 1. \tag{2.2}$$

The assertion is equivalent to

$$(G \times \mathcal{D}_{\alpha+\delta_x})\{(x, P) : P\{x\} > 0\} = 1 \tag{2.3}$$

as G is the marginal of X and the conditional distribution of $P|X$ is $\mathcal{D}_{\alpha+\delta_X}$. The last relation holds, since the presence of the atom at x in the base measure of the posterior Dirichlet process ensures that almost all random P sampled from

the posterior process assigns positive mass to the point x. Thus the discreteness property is the consequence of the presence of an atom at the observation in the base measure of the posterior Dirichlet process.

The discreteness property of the Dirichlet process may be disappointing if one's perception of the true distribution is nonatomic, such as when it has a density. However, discreteness itself may not be an obstacle to good convergence properties of estimators, considering the fact that the empirical distribution is also discrete but converges uniformly to any true distribution.

Support

Even though only discrete distributions can actually be sampled from a Dirichlet process, the topological support of the Dirichlet measure $\mathcal{D}_\alpha$, which is technically the smallest closed set of probability one, could be quite big. The support is actually characterized as all probability measures P^* whose supports are contained in that of G, that is,

$$\text{supp}(\mathcal{D}_\alpha) = \{P^* : \text{supp}(P^*) \subset \text{supp}(G)\}. \tag{2.4}$$

In particular, if G is fully supported, like the normal distribution on the line, then trivially every probability measure is in the support of $\mathcal{D}_\alpha$. To see why (2.4) is true, first observe that any supported P^* must have $P^*(A) = 0$ if A is disjoint from the support of G, which implies that $G(A) = 0$ and so $P(A) = 0$ a.s. $[\mathcal{D}_\alpha]$. For the opposite direction, we use the fact that weak approximation will hold if probabilities of a fine partition are approximated well, and this property can be ensured by the nonsingularity of the Dirichlet distribution with positive parameters.

Self-similarity

Another property of the Dirichlet process which distinguishes it from other processes is the self-similarity property described as follows. Let A be any set with $0 < G(A) < 1$, which ensures that $0 < P(A) < 1$ for almost all Dirichlet process samples. Let $P|_A$ be the restriction of P to A, that is, the probability distribution defined by $P|_A(B) = P(A \cap B)/P(A)$, and similarly $P|_{A^c}$ is defined. Then the processes $\{P(A), P(A^c)\}$, $P|_A$ and $P|_{A^c}$ are mutually independent, and moreover $P|_A$ follows $\text{DP}(MG(A), G|_A)$. Thus the assertion says that at any given locality A, how mass is distributed within A is independent of how mass is distributed within A^c, and both mass distribution processes are independent of how much total mass is assigned to the locality A. Further, the distribution process within A again follows a Dirichlet process with an appropriate scale. The property has its roots in the connection between independent gamma variables and the Dirichlet distributed variable formed by their ratios: if $X_1, \ldots, X_k$ are independent gamma variables, then $X = \sum_{i=1}^k X_i$ and $(X_1/X, \ldots, X_k/X)$ are independent. The self-similarity property has many

interesting consequences, an important one being that a Dirichlet process may be generated by sequentially distributing mass independently to various subregions following a tree structure. The independence at various levels of allocation, known as the *tail-freeness* property, is instrumental in obtaining large weak support of the prior and weak consistency of posterior. In fact, the Dirichlet process is the only *tail-free process* where the choice of the partition does not play a role.

Limit types

When we consider a sequence of Dirichlet processes such that the center measures converge to a limit G, then there can be three types of limits:

(i) if the total mass goes to infinity, the sequence converges to the prior degenerate at G;
(ii) if the total mass goes to a finite nonzero number M, then the limit is DP(M, G);
(iii) if the total mass goes to 0, the limiting process chooses a random point from G and puts the whole mass 1 at that sampled point.

To show the result, one first observes that tightness is automatic here because of the convergence of the center measures, while finite dimensionals are Dirichlet distributions, which converge to the appropriate limit by convergence of all mixed moments. The property has implications in two different scenarios: the Dirichlet posterior converges weakly to the Bayesian bootstrap when the precision parameter goes to zero, and converges to the degenerate measure at P_0 as the sample size n tends to infinity, where P_0 is the true distribution. Thus the entire posterior of P is weakly consistent at P_0, and the convergence automatically strengthens to convergence in the Kolmogorov–Smirnov distance, much in the tone with the Glivenko–Cantelli theorem for the empirical distribution. The result is extremely intriguing in that no condition on the base measure of the prior is required; consistency holds regardless of the choice of the prior, even when the true distribution is not in the support of the prior. This is very peculiar in the Bayesian context, where having the true distribution in the support of the prior is viewed as the minimum condition required to make the posterior distribution consistent. The rough argument is that when the prior excludes a region, the posterior, obtained by multiplying the prior with the likelihood and normalizing, ought to exclude that region. In the present context, the family is undominated and the posterior is not obtained by applying the Bayes theorem, so the paradox is resolved.

Dirichlet samples and ties

As mentioned earlier, the Dirichlet process samples only discrete distributions. The discreteness property, on the other hand, is able to generate ties in the

observations and is extremely useful in clustering applications. More specifically, the marginal joint distribution of n observations $(X_1, \ldots, X_n)$ from P which is sampled from DP(M, G) may be described sequentially as follows. Clearly, $X_1 \sim G$ marginally. Now

$$X_2|P, X_1 \sim P \quad \text{and} \quad P|X_1 \sim \text{DP}\left(M+1, \frac{M}{M+1}G + \frac{1}{M+1}\delta_{X_1}\right), \tag{2.5}$$

which implies, after eliminating P, that $X_2|X_1 \sim \frac{M}{M+1}G + \frac{1}{M+1}\delta_{X_1}$, that is, the distribution of X_2 given X_1 can be described as duplicating X_1 with probability $1/(M+1)$ and getting a fresh draw from G with probability $M/(M+1)$. Continuing this argument to X_n given $X_1, \ldots, X_{n-1}$, it is clear that X_n will duplicate any previous X_i with probability $1/(M+n-1)$ and will obtain a fresh draw from G with probability $M/(M+n-1)$. Of course, many of the previous X_i are equal among themselves, so the conditional draw can be characterized as setting to θ_j with probability $n_j/(M+n-1)$, where the θ_j are distinct values of $\{X_1, \ldots, X_{n-1}\}$ with frequencies n_j respectively, $j = 1, \ldots, k$, and as before, a fresh draw from G with probability $M/(M+n-1)$:

$$X_n|X_1, \ldots, X_{n-1} \sim \begin{cases} \delta_{\theta_j} & \text{with probability } \frac{n_j}{M+n-1} \quad j = 1, \ldots, k \\ G & \text{with probability } \frac{M}{M+n-1}, \end{cases}$$

where k is the number of distinct observations in $X_1, \ldots, X_{n-1}$ and $\theta_1, \ldots, \theta_k$ are those distinct values. Also observe that, since $(X_1, \ldots, X_n)$ are exchangeable, the same description applies to any X_i given X_j, $j = 1, \ldots, i-1, i+1, \ldots, n$. This procedure, studied in Blackwell and MacQueen (1973), is known as the generalized Pólya urn scheme. This will turn out to have a key role in the development of Markov chain Monte Carlo (MCMC) procedures for latent variables sampled from a Dirichlet process, as in Dirichlet mixtures discussed shortly.

Because of ties in the above description, the number of distinct observations, the total number of fresh draws from G including the first, is generally much smaller than n. The probabilities of drawing a fresh observation at steps $1, 2, \ldots, n$ are $1, M/(M+1), \ldots, M/(M+n-1)$ respectively, and so the expected number of distinct values K_n is

$$\text{E}(K_n) = \sum_{i=1}^{n} \frac{M}{M+i-1} \sim M \log \frac{n}{M} \quad \text{as } n \to \infty. \tag{2.6}$$

Moreover, one can obtain the exact distribution of K_n, and its normal and Poisson approximation, quite easily. The logarithmic growth of K_n induces sparsity that is often used in machine learning applications.

Sethuraman stick-breaking representation

The Dirichlet process DP(M, G) also has a remarkable representation known as the Sethuraman (1994) representation:

$$P = \sum_{i=1}^{\infty} V_i \delta_{\theta_i}, \quad \theta_i \overset{\text{i.i.d.}}{\sim} G, \quad V_i = \left[\prod_{j=1}^{i-1}(1 - Y_j)\right] Y_i, \quad Y_i \overset{\text{i.i.d.}}{\sim} \text{Beta}(1, M). \tag{2.7}$$

Thus $P = Y_1\delta_{\theta_1} + (1 - Y_1)\sum_{i=2}^{\infty} V_i'\delta_{\theta_{i+1}}$, where $V_i' = [\prod_{j=2}^{i}(1 - Y_j)]Y_{i+1}$, so that

$$P =_d Y_1\delta_{\theta_1} + (1 - Y)P. \tag{2.8}$$

This distributional equation is equivalent to the representation (2.7), and can be used to derive various properties of the random measure defined by (2.7) and to generate such a process by MCMC sampling. The weights V_i attached to the points $\theta_1, \theta_2, \ldots$ respectively may be viewed as the result of breaking a stick of unit length randomly in infinite fragments as follows. First break the stick at a location $Y_1 \sim \text{Beta}(1, M)$ and assign the mass Y_1 to a random point $\theta_1 \sim G$. The remaining mass $(1 - Y_1)$ is the split in the proportion $Y_2 \sim \text{Beta}(1, M)$ and the net mass $(1 - Y_1)Y_2$ is assigned to a random point $\theta_2 \sim G$. This process continues infinitely many times to complete the assignment of the whole mass to countably many points. What is intriguing is that the resulting process is actually DP(M, G). To get a rough idea why this is so, recall that for any random distribution P and $\theta \sim P$, the prior for P is equal to the mixture of the posterior distribution $P|\theta$ where θ follows its marginal distribution. In the context of the Dirichlet process, this means that $\mathcal{D}_\alpha = \int \mathcal{D}_{\alpha+\delta_\theta} dG(\theta)$. Now if P is sampled from $\mathcal{D}_{\alpha+\delta_\theta}$, then $P\{\theta\} \sim \text{Beta}(1, M)$ assuming that α is nonatomic. Thus the random P has a point mass at θ of random magnitude distributed as $Y \sim \text{Beta}(1, M)$. With the remaining probability, P is spread over $\{\theta\}^c$, and $P|_{\{\theta\}^c} \sim \text{DP}(M, G)$ independently of $P\{\theta\}$ by the self-similarity property of the Dirichlet process, that is $P|_{\{\theta\}^c} =_d P$. This implies that the DP(M, G) satisfies the distributional equation (2.8), where $Y \sim \text{Beta}(1, M)$, $\theta \sim G$ and are mutually independent of P. The solution of the equation can be shown to be unique, so the process constructed through the stick-breaking procedure described above must be DP(M, G).

Sethuraman's representation of the Dirichlet process has far reaching significance. First, along with an appropriate finite stage truncation, it allows us to generate a Dirichlet process approximately. This is indispensable in various complicated applications involving Dirichlet processes, where analytic expressions are not available, so that posterior quantities can be calculated only by simulating them from their posterior distribution. Once a finite stage truncation is imposed, for computational purposes, the problem can be treated essentially as a parametric problem for which general MCMC techniques such as Metropolis–Hastings algorithms and

reversible jump MCMC methods can be applied. Another advantage of the sum representation is that new random measures can be constructed by changing the stick-breaking distribution from Beta(1, M) to other possibilities. One example is the *two-parameter Poisson–Dirichlet process* where actually the stick-breaking distribution varies with the stage. Even more significantly, for more complicated applications involving covariates, dependence can be introduced among several random measures which are marginally Dirichlet by allowing dependence in their support points θ, or their weights V or both.

Mutual singularity

There are many more interesting properties of the Dirichlet process, for example any two Dirichlet processes are mutually singular unless their base measures share same atoms; see Korwar and Hollander (1973). In particular, the prior and the posterior Dirichlet processes are mutually singular if the prior base measure is nonatomic. This is somewhat peculiar because the Bayes theorem, whenever applicable, implies that the posterior is absolutely continuous with respect to the prior distribution. Of course, the family under consideration is undominated, so the Bayes theorem does not apply in the present context.

Tail of a Dirichlet process

We end this section by mentioning the behavior of the tail of a Dirichlet process. Since $\mathrm{E}(P) = G$, one may think that the tails of G and the random P are equal on average. However, this is false as the tails of P are much thinner almost surely. Mathematically, this is quite possible as the thickness of the tail is an asymptotic property. The exact description of the tail involves long expressions, so we do not present it here; see Doss and Sellke (1982). However, it may be mentioned that if G is standard normal, the tail of $P(X > x)$ is thinner than $\exp[-e^{x^2/2}]$ for all sufficiently large x a.s., much thinner than the original Gaussian tail. In a similar manner, if G is standard Cauchy, the corresponding random P has finite moment generating functions, even though the Cauchy distribution does not even have a mean.

2.3 Priors related to the Dirichlet process

Many processes constructed using the Dirichlet process are useful as prior distributions under a variety of situations. Below we discuss some of these processes.

2.3.1 Mixtures of Dirichlet processes

In order to elicit the parameters of a Dirichlet process DP(M, G), as the center measure G is also the prior expectation of P, it is considered as the prior guess