JIRSS (2002)
Vol. 1, Nos. 1-2, pp 143-163

Bayesian Nonparametric and Parametric Inference

Stephen G. Walker

Department of Mathematical Sciences, University of Bath, Bath BA2 7AY, United Kingdom. (massgw@maths.bath.ac.uk)

Abstract. This paper reviews Bayesian Nonparametric methods and discusses how parametric predictive densities can be constructed using nonparametric ideas.

1 Introduction

I will start with what might seem an unfair comment, yet it is meant to be instructive. The topic I am writing about, "Bayesian nonparametric theory", should, in my opinion, be entitled "Bayesian theory". This name has however been reserved for "Bayesian *parametric* theory".

The goal of the Bayesian, in the first instance, is the construction of a prior probability distribution on a space of density functions. Exactly how large a set of densities the prior covers is the first choice to be made; the second is the shape of the prior. It is fair to say that the larger the set of densities, the harder it is to put a probability on

Received: October 2002
Key words and phrases: Consistency, decision theory, Dirichlet process, expected utility rule, Hellinger distance, Kullback-Leibler divergence, nonparametric prior, parametric predictive density.

it. Even as recently as 1982, Ghorai and Rubin concluded that "Even though the literature on nonparametric density estimation is large, the literature on Bayesian nonparametric estimation of the density function is relatively small. The reason is the lack of a suitable prior over the space of probability density functions."

Fortunately, this is no longer the case and there are a number of suitable and useful priors which are supported by all densities and, technically, supported by a set of densities with respect to a measure, such as the Lebesgue measure. In this case posterior distributions are derived via Bayes Theorem. Some priors on spaces of distribution functions are not supported by densities and this poses some problems. Much of the early work in Bayesian nonparametrics involved such priors, most notably the Dirichlet process (Ferguson, 1973) and neutral to the right priors (Doksum, 1974).

Recent work has focused on priors which are supported by densities. An extremely popular model is the mixture of Dirichlet process prior (Lo, 1984) constructed by taking a convolution of the Dirichlet process with a continuous kernel. These models have been exploited by Escobar and West (1995) and many others. See Dey et al. (1998) or Walker et al. (1999) for a review.

The parametric Bayesian constructs a prior on spaces of density functions by first selecting a parametric family of densities, indexed by a finite dimensional parameter θ, and then puts a prior distribution on the domain of θ. This obviously puts a prior distribution on densities. It is also obvious that the densities supported by such a procedure are going to be restricted. There is then the real possibility of being caught out by the data. The nonparametric Bayesian avoids such potential embarrassment by not restricting the shapes of densities in the prior.

There is another reason why putting probability one on a parametric family of densities is worrying. Probability one is a sure event and so the parametric Bayesian must believe that the true density is a member of the parametric family. This is rarely the case in practice. See Lindsey (1999) and Draper (1999) for more serious comments on these points.

1.1 History

Before the seminal paper of Ferguson (1973), who formally introduced the Dirichlet process, following work on the gamma process

by Ferguson and Klass (1972), there had been little work done on nonparametric priors. Perhaps the most well known at the time was the class of tailfree priors (Freedman, 1963; Fabius, 1964). These were the invention of Freedman (1963) who was concerned with prior distributions on the integers which gave rise to consistent posteriors. Further work on tailfree priors was done by Kraft (1964) and Kraft and van Eeden (1964) who were working with what modern day Bayesian nonparametric statisticians would recognise as Pólya trees. Dubins and Freedman (1965) and Freedman (1965) explored further the notion of random distribution functions on countable sets.

The most effective review of the early work in Bayes nonparametrics is given in Ferguson (1974). An alternative review is given in Lindley (1972).

The introduction of the Dirichlet process in Ferguson (1973) heralded the birth of modern day Bayes nonparametrics and arguably the Dirichlet process remains the most important and widely used (in various guises) nonparametric prior today. Ferguson (1973, 1974) was largely concerned with developing Bayes decision rules using the Dirichlet process, making comparisons with classic nonparametric procedures. Antoniak (1974) extended the ideas of Ferguson by considering Dirichlet mixtures. At around the same time, Doksum (1974) introduced the class of neutral priors, of which the Dirichlet process is a special case.

The seventies and eighties saw the introduction of a number of nonparametric priors. We will mention the work of Lo (1984), who introduced mixtures of Dirichlet processes, in which a continuous kernel is mixed with a Dirichlet process. This has turned out to be one of the most popular and successful of all nonparametric priors. Dykstra and Laud (1981) developed the idea of modelling the hazard rate function nonparametrically via the use of an independent increment process. Neutral priors make use of such processes to model the hazard function which are, with probability one, discrete; that is the hazard functions are increasing step functions. The model of Dykstra and Laud (1981) gives continuous hazard functions. A number of authors worked on Gaussian process priors, including Lenk (1989, 1991).

Nowadays, there are a number of nonparametric priors which have proved to be useful. The most widely used is undoubtedly the mixtures of Dirichlet process prior. Other popular models include Pólya trees and Gaussian process priors. See Dey et al. (1998) or Walker

et al. (1999) for recent reviews.

1.2 Preliminaries

The interpretation of "nonparametric" is "infinite dimensional". We can then differentiate between a countably infinite space, in which case we would consider priors which are supported by densities with respect to the counting measure, and something like the real line, in which case priors would be constructed to be supported by densities with respect to the Lebesgue measure, for example. It is the latter case that is the concern of this article.

In this section, notation and definitions will be introduced. The data will be denoted by $X_1, X_2 \ldots$ and are assumed to be independent and identically distributed from some unknown distribution function F. The prior on the space of distribution functions will be denoted by Π and a random distribution from Π will be denoted by F. If all such F are absolutely continuous with respect to the Lebesgue measure then we will write the corresponding density as f and will in this case use Π to also represent the prior on the space of density functions.

Let us for the moment assume the prior is supported by densities. The likelihood function is given by

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

While it is not possible to work around this likelihood from a Classical perspective, it is possible to work with from a Bayesian perspective. The posterior distribution of f is available via Bayes theorem and the posterior mass assigned to a set A of densities is given by

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

Technical details involve ensuring all objects are measurable and this we will assume. The numerator and denominator of $\Pi_n(A)$ are then understood in terms of expectations; that is, the denominator is

$$\mathrm{E}\{f(X_1) \ldots f(X_n)\}$$

The tough part of Bayesian nonparametrics is computing these expectations. In fact it is not easy and often sampling based methods

are required. The reason why Bayesian nonparametrics is currently seeing a revival is precisely because of the recent advances being made in sampling based Bayesian inference. This was pioneered by Tanner and Wong (1987) and later by Gelfand and Smith (1990), Smith and Roberts (1993) who introduced the Gibbs sampler to mainstream statistical awareness. In fact, the first Gibbs sampler appeared in the PhD Thesis of Michael Escobar, which was written in 1988 and was concerned with inference for normal means. One of his models involved the Dirichlet process and a solution to this inference problem only seems possible via the Gibbs sampler.

2 A nonparametric model based on the Dirichlet process

The Dirichlet process prior and related priors are the most widely known and widely used nonparametric priors on distribution functions. The Dirichlet process itself is remarkably simple to work with.

2.1 The Dirichlet process

The easiest way to think of the Dirichlet process is as a stochastic process which has sample paths behaving almost surely as a distribution function. When thinking about a stochastic process, the finite dimensional distributions are of interest and for a Dirichlet process these are Dirichlet distributions. In fact, for any partition $(B_1, \ldots, B_k)$ of Ω, the sample space,

$$(F(B_1), \ldots, F(B_k)) \sim \mathrm{Dir}\,(\alpha(B_1), \ldots, \alpha(B_k))$$

where $\alpha(\cdot)$ is a finite non-null measure.

The updating mechanism from prior to posterior based on an independent and identically distributed sample is available and it turns out that the posterior distribution is also a Dirichlet process. We will write

$$F \sim \mathrm{DP}\,\{\alpha(\cdot)\}$$

to denote that F is a random distribution from a Dirichlet process prior with parameter α. The posterior distribution based on an independent and identically distributed sample $(X_1, \ldots, X_n)$ is given

by

$$\mathrm{DP}\left(\alpha + \sum_{i=1}^{n} \delta_{X_i}\right)$$

where δ_X is the measure with point mass 1 at the location X. The well known Bayesian bootstrap (Rubin, 1982) follows by now taking α to be a null measure.

In the case of the Dirichlet process, there is no dominating measure. Each random distribution chosen from a Dirichlet process has a probability mass function yet no two random distributions will have jumps in the same place. This means that $\mathrm{E}\{f(X_1)\}$ does not make any sense and so Bayes Theorem is not available to obtain the posterior distribution. Other ideas are required; see Ferguson (1973) and Doksum (1974), for example.

Since $\alpha(\cdot)$ is a finite measure it is possible to write $\alpha(\cdot) = cG(\cdot)$ where $c > 0$ and G is a probability measure on Ω. Straightforward manipulation of the Dirichlet distribution gives

$$\mathrm{E}\, F = G$$

in the sense that $\mathrm{E}\, F(A) = G(A)$ for all sets A. In fact,

$$F(A) \sim \mathrm{be}\{\alpha(A), \alpha(\infty) - \alpha(A)\}$$

where 'be' denotes the beta distribution, and consequently

$$\mathrm{Var}\, F(A) = \frac{G(A)\{1 - G(A)\}}{c + 1}$$

From this we can see that the variance is fully determined by the shape of G. So it might be considered that G plays too large a part in the prior. To ensure the prior is not too concentrated about G, it might be chosen to make c small, forcing the variance to be as large as possible. This can be done, but there is a snag to this which needs to mentioned, and was discovered by Sethuraman and Tiwari (1982). They showed that as $c \to 0$, then $F \to_d F^*$ where a random F^* is a probability measure with point mass 1 at θ and $\theta \sim G$. Thus, for small c, the prior is putting a lot of probability on distributions which are highly discrete, having at least one big jump.

2.2 Mixtures of Dirichlet processes

Blackwell (1973) proved that a Dirichlet process is discrete and it is this discreteness which is not liked. The mixture of Dirichlet process

model avoids the discreteness and was first given attention by Lo (1984), who considered a randomly generated density function given by

$$f(x) = \int K(x;\theta)\,\mathrm{d}P(\theta)$$

where $K(\cdot;\theta)$ is a continuous kernel, and a density function for each θ, and P is a Dirichlet process. The constructive definition of the Dirichlet process (Sethuraman and Tiwari, 1982; Sethuraman, 1994) gives

$$P(\theta) = \sum_j \omega_j\,\mathbf{1}(\nu_j \leq \theta)$$

where $\{\omega_j\}_{j=1}^{\infty}$ are random weights whose distribution depends only on $\alpha(\Omega)$ and the $\{\nu_j\}$ are independent and identically distributed from $\alpha(\cdot)/\alpha(\Omega)$. So, if $K(x;\theta) = K_h(x-\theta)$, a normal kernel with variance h^2, then the model is easily seen to be a mixture of normal model. See Richardson and Green (1997) for an alternative distribution of weights and means and a more complicated sampling strategy, involving reversible jump Markov chain Monte Carlo.

The most common form of the mixture of Dirichlet process model is as a hierarchical model, and is given here;

$$X_i|\theta_i \sim K(\cdot;\theta_i) \quad i = 1,\ldots,n$$

$$\theta_1,\ldots,\theta_n|P \sim_{\text{iid}} P$$

$$P \sim \text{DP}\,\{c\,G(\cdot;\phi)\}$$

$$c \sim \pi_c(c) \quad \text{and} \quad \phi \sim \pi_\phi(\phi)$$

where ϕ are the parameters of G. This can be viewed as a more flexible version of the parametric hierarchical model, introduced by Lindley and Smith (1972);

$$X_i|\theta_i \sim K(\cdot;\theta_i) \quad i = 1,\ldots,n$$

$$\theta_1,\ldots,\theta_n|\phi \sim_{\text{iid}} G(\cdot;\phi)$$

$$\phi \sim \pi_\phi(\phi)$$

It is well known that P can be integrated out of the nonparametric model, see for example Blackwell and MacQueen (1973), leading to the revised, but equivalent, hierarchical model;

$$X_i|\theta_i \sim K(\cdot;\theta_i) \quad i = 1,\ldots,n$$

$$p(\theta_1,\ldots,\theta_n) = g(\theta_1;\phi)\prod_{i=2}^{n}\frac{c\,g(\theta_i;\phi)+\sum_{j=1}^{i-1}\delta_{\theta_j}(\theta_i)}{c+i-1}$$

$$c \sim \pi_c(c) \quad \text{and} \quad \phi \sim \pi_\phi(\phi)$$

Here g is the density function corresponding to G.

The posterior distribution is mathematically intractable and only sampling strategies are going to be able to obtain posterior summaries of interest. Kuo (1986) tried out an importance sampling scheme but the efficient algorithms that are currently in use are all based on MCMC methods.

2.3 Sampling the model

The clever idea of Escobar (1988) was to sample the posterior distribution of $(\theta_1,\ldots,\theta_n)$ using a Markov chain, which he constructed by successively sampling from the full conditional densities,

$$p(\theta_i|\theta_{-i},x_1,\ldots,x_n) \quad i=1,\ldots,n$$

where $\theta_{-i}=(\theta_1,\ldots,\theta_{i-1},\theta_{i+1},\ldots,\theta_n)$ and, for example,

$$p(\theta_1|\theta_{-1},x_1,\ldots,x_n) \propto K(x_1;\theta_1)\left\{c\,g(\theta_1)+\sum_{i=2}^{n}\delta_{\theta_i}(\theta_1;\phi)\right\}$$

This algorithm is what we know now to be a Gibbs sampler. Escobar (1988) used this idea to estimate the normal means. MacEachern (1994) developed an alternative sampling strategy which was still based on the Gibbs sampler. Current mixture of Dirichlet process sampling strategies are based on the idea of MacEachern (1994). A sample from $p(\theta_1,\ldots,\theta_n)$ defines a random variable S_n, defined on $\mathcal{S}_n=\{1,\ldots,n\}$; S_n is the number of clusters or the number of distinct elements in $\mathcal{S}_n$. The distinct θ_i are independent and identically distributed from g and denote these distinct elements by θ_j^*, $j=1,\ldots,S_n$. The scheme also defines a random variable $s=(s_1,\ldots,s_n)$ which indicates the cluster in which θ_i lies. That is, $s_i=j$ means that θ_i is in the jth cluster and $\theta_i=\theta_j^*$. Let k denote the number of clusters, and n_j, for $j=1,\ldots,k$, the number of cases in each cluster. Then let n_j^- and k^- denote the same things with case i removed. The basic algorithm is as follows:

1) Generate $(s_i, \theta_i)|s_{-i}, \theta_{-i}, x_i, c, \phi$ by taking $s_i = j$ with probability proportional to

$$q_j = n_j^- \, K(x_i, \theta_j^*)$$

or taking $s_i = k^- + 1$ with probability proportional to

$$q_0 = c \int K(x_i, \theta) \, g(\theta; \phi) \, \mathrm{d}\theta$$

If $s_i = k^{-1} + 1$ then a new location $\theta_i = \theta_k^*$ for the new cluster, with 1 element in it, is sampled from the density proportional to

$$K(x_i; \theta) \, g(\theta; \phi)$$

2) Generate $\theta_j^*|x_1, \ldots, x_n, k$ by taking θ_j^* from the density proportional to

$$\left\{ \prod_{s_i = j} K(x_i; \theta) \right\} g(\theta; \phi)$$

Step 1) is repeated for $i = 1, \ldots, n$ and step 2) for $j = 1, \ldots, k$.

3) Generate $\phi|\theta_1^*, \ldots, \theta_k^*, k$ from the density proportional to

$$\left\{ \prod_{j=1}^{k} g(\theta_j^*; \phi) \right\} \pi_\phi(\phi)$$

4) Generate $c|k$ from the density proportional to

$$c^k \frac{\Gamma(c)}{\Gamma(c+k)} \, \pi_c(c)$$

A good review is given in MacEachern (1998).

Difficulties arise with this Gibbs sampler and others like it if $K(x; \theta)$ and $g(\theta; \phi)$ do not form a conjugate pair. The integration needed to find q_0 will not be possible. Modifications to the algorithm required to cope with this problem are given in MacEachern and Müller (1998). The only tricky aspect in this case is sampling from the density proportional to $K(x; \theta) \, g(\theta; \phi)$, which is not surprising, as it is also required to be sampled from in an MCMC algorithm within the parametric framework (see Wakefield et al., 1994). Further work and references can be found in West et al. (1994), Müller et al. (1996), MacEachern (1998).

A sample from the predictive density function, say $x^* \sim f_n(\cdot)$, is easy to obtain. One simply takes $\theta^* \sim p(\cdot|\theta_1, \ldots, \theta_n)$ and then takes $x^* \sim K(\cdot; \theta^*)$. Here, $p(\cdot|\theta_1, \ldots, \theta_n)$ can be thought of as $p(\theta_{n+1}|\theta_1, \ldots, \theta_n)$.

3 Parametric Inference

It seems to me that there is a contradiction at the heart of Bayesian parametric inference. Probability one is put on a family of parametric densities and then more often than not, after the data has been observed, some sort of check is made to see whether the data is compatible with the choice of parametric model. Probability one was not meant after all. Of course, checking a parametric model is a prudent thing to do but clearly incompatible with probability one being assigned to the model a priori. If one knows that one is going to check the model after seeing the data, it is an internal contradiction in the inference process to assign probability one at the outset. See Lindsey (1999) for further remarks on this point.

A simple resolution to this problem is provided by assigning probability one to all densities. There is then no reason to check the model. However, parametric models are useful and the work of this section is concerned with how Bayesian parametric inference can be performed while prior distributions are nonparametric. The basic idea is to select a parametric model for parametric Bayesian inference while acknowledging the model is wrong. From a predictive perspective there is a straightforward way to selecting the best parametric model.

A parametric Bayesian model will provide a predictive density, say $p_n(x)$, which is given by

$$p_n(x) = \int f(x; \theta)\, \pi_n(\theta)\, \mathrm{d}\theta$$

where $\pi_n(\theta)$ is the posterior distribution for θ. The idea then is to select the model which has predictive closest, in some sense, to the predictive provided by the nonparametric model. In this respect it can be viewed as a *minimum distance estimation* idea. Such a nonparametric predictive density can be provided by the Mixture of Dirichlet Process model described in the previous section and let this predictive be denoted by

$$f_n(x) = \int f(x)\, \Pi_n(\mathrm{d}f)$$

If λ indexes the models under consideration, which could be either discrete or continuous, the preferred model, $\hat{\lambda}$, is the one which minimises

$$d(p_n(\cdot;\lambda), f_n(\cdot))$$

Here d represents a distance between density functions and candidates include the Hellinger distance,

$$d(f,g) = \left\{ \int (\sqrt{f} - \sqrt{g})^2 \right\}^{1/2}$$

and the Kullback–Leibler divergence,

$$d(f,g) = \int g \, \log(g/f)$$

The work presented in Gutièrrez-Peña and Walker (2001) used the Kullback–Leibler divergence though the derivation was via a decision theoretic approach using the logarithmic score utility function. The well known elements of a decision problem are as follows:

(1) a set of decisions; $\{\lambda \in \Lambda\}$.

(2) a set of states of nature; $\{f \in \mathcal{F}\}$ and $\mathcal{F}$ is the set of densities with respect to the Lebesgue measure, for example .

(3) a utility function $U(\lambda, f)$ evaluating the desirability of λ when f is the true density function.

(4) a probability distribution on the space of density functions representing beliefs in the true state of nature. In a Bayesian context, this probability is the prior Π in the no sample problem and is Π_n once the data $X^n = x^n$ has been observed.

With the Kullback–Leibler divergence, the utility of model λ assuming f to be the true density function is given by

$$U(f,\lambda) = \int \log p_n(x;\lambda) \, f(x) \, \mathrm{d}x$$

According to the decision theory, the best decision to take is the one maximising the posterior expectation of $U(f,\lambda)$, that is, maximise

$$\begin{aligned} U(\lambda) &= \int U(f,\lambda) \, \Pi_n(\mathrm{d}f) \\ &= \int \log p_n(x;\lambda) \, f_n(x) \, \mathrm{d}x \end{aligned}$$

This maximiser is obviously the minimiser of $d(p_n(\cdot;\lambda), f_n(\cdot))$ with d the Kullback–Leibler distance. See de Groot (1970) and Bernardo and Smith (1994) for more information about the Bayesian theory underpinning decision theory.

The Kullback–Leibler distance is therefore appealing to use. However there is a detail about the Kullback–Leibler divergence which is unappealing. The desired asymptotics are as follows. We desire $d(f_n, f_0) \to 0$ almost surely, where f_0 is the true density function. Unfortunately, if d is the Kullback–Leibler distance then the required sufficient conditions on Π in order for the asymptotics to hold are currently unknown. This is not the case if d is the Hellinger distance.

3.1 Hellinger consistency

Throughout this section we will assume all relevant unknowns are densities and, for simplicity, are densities with respect to the Lebesgue measure. Consequently we will only consider priors which are supported by such densities. Let Ω be the set of all densities with respect to the Lebesgue measure. A Hellinger neighbourhood of the density function g is given by

$$A_\varepsilon^c(g) = \{f \in \Omega : d(f,g) < \varepsilon\}$$

where $d(f,g)$ is the Hellinger distance between densities f and g, given by

$$d(f,g) = \left\{\int \left(\sqrt{f} - \sqrt{g}\right)^2\right\}^{1/2} = \left\{2\left(1 - \int \sqrt{f\,g}\right)\right\}^{1/2}$$

and use will be made of $h(f) = \frac{1}{2}d^2(f, f_0)$. If Π_n is the sequence of posterior distributions based on a sample of size n from f_0 (with distribution function F_0), the density function generating the observations, then Bayesian consistency is concerned with conditions to be imposed on the prior Π for which

$$\Pi_n(A_\varepsilon(f_0)) \to 0 \ \text{ almost surely } [F_0]$$

for all $\varepsilon > 0$. Here we write

$$\Pi_n(A) = \frac{\int_A R_n(f)\,\Pi(\mathrm{d}f)}{\int R_n(f)\,\Pi(\mathrm{d}f)}$$

where

$$R_n(f) = \prod_{i=1}^{n} f(X_i)/f_0(X_i)$$

and $X_1, X_2, \ldots$ are the data. The inclusion of $\prod_{i=1}^{n} f_0(X_i)$ in both numerator and denominator has reasons which will become clear later on.

For Hellinger consistency there are two conditions which a prior Π needs to possess. These are necessary conditions but to date are the most straightforward sufficient conditions. The first property has become a standard requirement for both weak (see Schwartz, 1965) and strong consistency. I will refer to it as the Kullback–Leibler property for Π and is given by

$$\Pi\{f : d_K(f, f_0) < \delta\} > 0$$

for all $\delta > 0$. Here $d_K(f, g) = \int g \log(g/f)$. Since f_0 is unknown, this Kullback–Leibler property can only be known to hold if $\Pi\{f : d_K(f, g) < \delta\} > 0$ for all $\delta > 0$ and all densities g.

The second property is one which all reasonable priors will possess. A prior Π is said to have property Q if

$$h(f_{nA(\varepsilon)}) > \varepsilon \text{ for all } n \text{ and for all } \varepsilon > 0 \text{ when } A(\varepsilon) = \{f : h(f) > \varepsilon\}.$$

Here f_{nA} is the predictive density based on the posterior distribution restricted to the set A. The idea behind property Q is that the predictive density based on a posterior restricted to the set A, which does not include any density closer than ε to f_0 in the Hellinger sense, can never itself get closer than a distance ε to f_0. This class of prior would seem to include all reasonable ones; in fact it would be disappointing to find a prior in regular use which did not have property Q.

Theorem 3.1.1. *If Π has the Kullback–Leibler property and property Q then Π_n is Hellinger consistent.*

Proof. Let $A = A_\varepsilon = \{f : h(f) > \varepsilon\}$. A key identity is given by

$$\int_A R_{n+1}(f)\,\Pi(\mathrm{d}f) = \frac{f_{nA}(X_{n+1})}{f_0(X_{n+1})} \int_A R_n(f)\,\Pi(\mathrm{d}f)$$

We then define

$$J_n = \sqrt{\int_A R_n(f)\,\Pi(\mathrm{d}f)}$$

so that

$$\mathrm{E}\,(J_{n+1}|\mathcal{F}_n) = J_n \int \sqrt{f_0\, f_{nA}} = J_n\,\{1 - h(f_{nA})\}$$

where $\mathcal{F}_n = \sigma(X_1, \ldots, X_n)$. The numerator for $\Pi_n(A)$ is J_n^2 and the above gives $J_n < \exp(-nd)$ almost surely for all large n and for all $d < -\log(1-\varepsilon)$. The denominator of $\Pi_n(A)$ is $I_n = \int R_n(f)\,\Pi(\mathrm{d}f)$ which with the Kullback–Leibler property is bounded below by $\exp(-nc)$ almost surely for all large n and for all $c > 0$. Then pick $c < d$.

While mathematically one requires both the Kullback–Leibler property and property Q to guarantee Hellinger consistency, practically the Kullback–Leibler property is sufficient. The prior would actually need to be quite strange for it not to have property Q.

See Barron et al. (1999) and Ghosal et al. (1999) for alternative sufficient conditions for priors to give rise to a Hellinger consistent sequence of posterior distributions.

3.2 Parametric prediction using f_n

For a sample of size n, we assume the parametric predictive density will be from a family of densities, say $p(\cdot;\lambda)$. The best parametric predictive density according to our criterion will be $p(\cdot;\widehat{\lambda})$ where $\widehat{\lambda}$ minimises $d(p(\cdot;\lambda), f_n(\cdot))$. We assume such a minimiser is unique, but in fact this does not really matter.

Gutièrrez-Peña and Walker (2001) use parametric predictive densities obtained via Bayes theorem. There is actually no need for this. The true predictive is given by f_n and for parametric prediction one only needs to select a parametric density from those competing densities which is closest to f_n.

Here f_n can be the predictive from the mixture of Dirichlet process model with the sufficient conditions to ensure the Hellinger consistency of f_n. See Ghosal et al. (1999) for these sufficient conditions. From now on we will write $p(\cdot;\lambda)$ as p_λ. Now, for any λ,

$$d(p_\lambda, f_0) \geq d(p_\lambda, f_n) - d(f_n, f_0)$$

and $d(f_n, f_0) \to 0$ almost surely and so

$$\limsup_n d(p_\lambda, f_n) \leq d(p_\lambda, f_0)$$

From the definition of $\widehat{\lambda}$ we have that

$$d(p_\lambda, f_n) \geq d(p_{\widehat{\lambda}}, f_n)$$

for all n, and hence

$$\limsup_n d(p_{\widehat{\lambda}}, f_0) \leq d(p_\lambda, f_0)$$

almost surely for all λ. Hence, eventually $\widehat{\lambda}$ sticks to those λ which minimise $d(p_\lambda, f_0)$ and if this minimiser, say λ_0 exists and is unique, then clearly $\widehat{\lambda} \to \lambda_0$ almost surely. This is precisely the required asymptotic result. Asymptotically, it is known that the best parametric model will be selected.

3.3 Minimising Hellinger distance

The aim is to minimise $d(p_\lambda, f_n)$ which is equivalent to maximising

$$U(\lambda) = \int \sqrt{p(x;\lambda)\, f_n(x)}\, \mathrm{d}x$$

This is not going to be easy an easy task in general. A general solution is to maximise an approximation to U based on sampling;

$$\widehat{U}(\lambda) = \frac{1}{N} \sum_{j=1}^{N} \sqrt{\frac{p(X_j;\lambda)}{f_n(X_j)}}$$

where the $\{X_j\}_{j=1}^N$ will be available as a Markov sample from f_n. Section 2 described how to obtain such a sample.

The maximisation can then be performed on

$$\widehat{U}(\lambda) = \sum_{j=1}^{N} \omega_j\, p^{1/2}(X_j;\lambda)$$

where

$$\omega_j = \frac{1}{N\, f_n^{1/2}(X_j)}$$

This should be straightforward to maximise with modern computing technology and software. The Newton-Raphson method is one possible approach and in fact there are many other routines available.

It should be pointed out that there is a lot of literature already about the subject of minimum Hellinger distance estimation. This is from a Classical perspective and the idea is to estimate the parameter λ from a parametric family by minimising

$$d(p_\lambda, g_n)$$

where g_n is typically a kernel density estimator. See Beran (1977) and Lindsay (1994), for example.

4 Further work and discussion

Here we collect some ideas for future research and discuss their merits.

4.1 Kullback-Leibler divergence

If instead of using the Hellinger distance, we use the Kullback-Leibler divergence, the desired asymptotics become

$$U_n(\widehat{\lambda}) \equiv \int \log\{p(x;\widehat{\lambda})\}\, f_n(x)\, \mathrm{d}x \to \int \log\{p(x;\lambda_0)\}\, f_0(x)\, \mathrm{d}x$$

almost surely, where λ_0 maximises $\int \log\{p(x;\lambda)\}\, f_0(x)\, \mathrm{d}x$ and $\widehat{\lambda}$ maximises $\int \log\{p(x;\lambda)\}\, f_n(x)\, \mathrm{d}x$. As mentioned previously, there are some nice features about using the Kullback-Leibler divergence. For example, if we take the Bayesian bootstrap, then

$$U_n(\lambda) \equiv \frac{1}{n}\sum_{i=1}^{n} \log\{p(x_i;\lambda)\}$$

and now $\widehat{\lambda}$ is the maximum likelihood estimator. If Λ is a finite set then under certain regularity conditions,

$$U_n(\widehat{\lambda}) \to \int \log\{p(x;\lambda_0)\}\, f_0(x)\, \mathrm{d}x$$

almost surely, where λ_0 maximises

$$\int \log\{p(x;\lambda)\}\, f_0(x)\, \mathrm{d}x$$

The asymptotics concerning the convergence highlighted in the previous paragraph has not, to my knowledge, been studied. Perhaps the most closely related work has been done by Berk (1965; 1970).

4.2 The decision space

A relevant question is what exactly is the decision space. This depends on the preferences of the experimenter and so there is no right or wrong set of decisions. On the other hand, there does not seem to be any reason why the set should depend on the data and this suggests the general set indexed by λ. The set Λ does not need to be specified in full at any fixed point in time.

An important question is when a model, say $p(x;\widehat{\lambda})$ is "good enough" as a predictive or whether a better model should be sought; that is, the enlargement of Λ. For example, one question could be whether the value of $U_n(\widehat{\lambda})$ is "acceptable", or whether the dimension of Λ could be increased leading to a new value of $U_n(\widehat{\lambda})$. Does the increase in dimension lead to a big enough increase in the utility? These are not easy questions to answer - work is in progress.

4.3 Discussion

The message trying to emanate from this paper is that Bayesian inference is not divided between parametric and nonparametric. Bayesians of both persuasions are placing prior distributions on sets of densities and it is the size of these sets which distinguishes the two camps. It seems to me that more faith and justification is required for those restricting the sets of densities to parametric families.

There also seems a blatant contradiction at the heart of Bayesian parametric inference involving the checking of models. This point has been raised before (Lindsey, 1999). However, it is also obvious that parametric models are useful and so a coherent construction of a parametric predictive density function needs to be considered. Logically, the prior must then be placed on a set of densities from which the experimenter is certain the true density function can be found, so that no model checking need take place.

The predictive density can then be decided via Bayesian decision theory, making use of a utility function on the joint decision and state of nature spaces. The precise density being the one from the decision set which maximises the posterior expected utility function.

Acknowledgements

The research of the author is supported by an EPSRC Advanced Research Fellowship.

References

Antoniak, C. E. (1974), Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. Ann. Statist., **2**,

1152-1174.

Barron, A., Schervish, M. J., and Wasserman, L. (1999), The consistency of distributions in nonparametric problems. Ann. Statist., **27**, 536–561.

Beran, R. (1977), Minimum Hellinger distance estimates for parametric models. Ann. Statist., **5**, 445–463.

Berk, R. H. (1965), Limiting behaviour of posterior distributions when the model is incorrect. Ann. Math. Statist., **37**, 51-58.

Berk, R. H. (1970), Consistency a posteriori. Ann. Math. Statist., **41**, 894–906.

Bernardo, J. M., Smith, A. F. M. (1994). Bayesian Theory. Wiley & Sons.

Blackwell, D. (1973), The discreteness of Ferguson selections. Ann. Statist., **1**, 356–358.

Blackwell, D., MacQueen, J. B. (1973), Ferguson distributions via Pólya-urn schemes. Ann. Statist., **1**, 353–355.

De Groot, M. (1970), Optimal Statistical Decisions. McGraw Hill Book Company.

Dey, D., Müller, P., and Sinha, D. (1998), Practical Nonparametric and Semiparametric Bayesian Statistics. New York: Springer.

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

Draper, D. (1999), Discussion of the paper "Bayesian nonparametric inference for random distributions and related functions" by Walker et al., J. Roy. Statist. Soc., B, **61**, 485–527.

Dubins, L. and Freedman, D. (1965), Random distribution functions. Bull. Amer. Math. Soc., **69**, 548–551.

Dykstra, R. L., Laud, P. W. (1981). A Bayesian nonparametric approach to reliability. Ann. Statist., **9**, 356–367.

Escobar, M. D. (1988), Estimating the means of several normal populations by nonparametric estimation of the distribution of the means. Unpublished Ph.D. dissertation, Department of Statistics, Yale University.

Escobar, M. D. (1994), Estimating normal means with a Dirichlet process prior. J. Am. Statist. Assoc., **89**, 268–277.

Escobar, M. D. and West, M. (1995), Bayesian density estimation and inference using mixtures. J. Am. Statist. Assoc., **90**, 577–588.

Fabius, J. (1964), Asymptotic behaviour of Bayes estimates. Ann. Math. Statist., **35**, 846–856.

Ferguson, T. S. and Klass, M. J. (1972), A representation of independent increment processes without Gaussian components. Ann. Math. Statist., **43**, 1634–1643.

Ferguson, T. S. (1973), A Bayesian analysis of some nonparametric problems. Ann. Statist., **1**, 209–230.

Ferguson, T. S. (1974), Prior distributions on spaces of probability measures. Ann. Statist., **2**, 615–629.

Freedman, D. A. (1963), On the asymptotic behaviour of Bayes estimates in the discrete case I. Ann. Math. Statist., **34**, 1386–1403.

Freedman, D. A. (1965), On the asymptotic behaviour of Bayes estimates in the discrete case II. Ann. Math. Statist., **36**, 454–456.

Gelfand, A. E. and Smith, A. F. M. (1990), Sampling based approaches to calculating marginal densities. J. Amer. Statist. Assoc., **85**, 398–409.

Ghorai, J. K. and Rubin, H. (1982), Bayes risk consistency of nonparametric Bayes density estimators. Austral. J. Statist., **24**, 51–66.

Ghosal, S., Ghosh, J. K., and Ramamoorthi, R. V. (1999), Posterior consistency of Dirichlet mixtures in density estimation. Ann. Statist., **27**, 143–158.

Gutierrèz-Peña, E. and Walker, S. G. (2001), A Bayesian predictive approach to model selection. J. Statist. Plann. Inf., **93**, 259–276.

Kraft, C. H. (1964), A class of distribution function processes which have derivatives. J. Appl. Probab., **1**, 385–388.

Kraft, C. H. and van Eeden, C. (1964), Bayesian bioassay. Ann. Math. Statist., **35**, 886–890.

Kuo, L. (1986), Computations of mixtures of Dirichlet processes. SIAM Journal of Scientific and Statistical Computing, **7**, 60–71.

Lenk, P. J. (1988), The logistic normal distribution for Bayesian, nonparametric, predictive densities. J. Amer. Statist. Assoc., **83**, 509–516.

Lenk, P. J. (1991), Towards a practicable Bayesian nonparametric density estimator. Biometrika, **78**, 531–543.

Lindley, D. V. (1972), Bayesian Statistics, a Review. SIAM, Philadelphia.

Lindley, D. V. and Smith, A. F. M. (1972), Bayes estimates for the linear model. J. Roy. Statist. Soc. B, **34**, 1–41.

Lindsay, B. G. (1994), Efficiency versus robustness: The case for minimum Hellinger distance and related methods. Ann. Statist., **22**, 1081–1114.

Lindsey, J. K. (1999), Some statistical heresies. The Statistician, **48**, 1–40.

Lo, A. Y. (1984), On a class of Bayesian density estimates: I. density estimates. Ann. Statist., **12**, 351–357.

MacEachern, S. N. (1994), Estimating normal means with a conjugate style Dirichlet process prior. Comm. Statist.: Simulation and Computation, **23**.

MacEachern, S. N., Mueller, P. (1998), Estimating mixtures of Dirichlet process models. J. Comp. Graph. Statist, **7**, 223–238.

MacEachern, S. N. (1998), Computational methods for mixture of Dirichlet process models. In Practical Nonparametric and Semiparametric Bayesian Statistics. Eds. D. Dey, P. Müller and D. Sinha, Springer-Verlag, pp 23–44.

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

Richardson, S., Green, P. J. (1997), On Bayesian analysis of mixtures and unknown number of components. J. Roy. Statist. Soc., B, **59**, 731–792.

Rubin, D. B. (1981), The Bayesian bootstrap. Ann. Statist., **9**, 130–134.

Schwartz, L. (1965), On Bayes procedures. Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete., **4**, 10–26.

Sethuraman, J. and Tiwari, R. (1982), Convergence of Dirichlet measures and the interpretation of their parameter. In Proceedings of the third Purdue symposium on statistical decision theory and related topics. Eds. S. S. Gupta and J. O. Berger, New York: Academic press.

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

Smith, A. F. M. and Roberts, G. O. (1993), Bayesian computations via the Gibbs sampler and related Markov chain Monte Carlo methods. J. Roy. Statist. Soc., B, **55**, 3–23.

Tanner, M. and Wong, W. H. (1987), The calculation of posterior distributions by data augmentation. J. Amer. Statist. Assoc., **82**, 528–550.

Walker, S. G., Damien P., Laud, P. W., and Smith, A. F. M. (1999), Bayesian nonparametric inference for random distributions and related functions (with discussion). J. Roy. Statist. Soc., B, **61**, 485–527.

Wakefield, J. C., Smith, A. F. M., Racine-Poon, A., and Gelfand, A. E. (1994), Bayesian analysis of linear and nonlinear population models using the Gibbs sampler. Applied Statistics, **43**, 201-221.

West, M., Müller, P., and Escobar, M. D. (1994), Hierarchical priors and mixture models with applications in regression and density estimation. In Aspects of Uncertainty: A tribute to D. V. Lindley, Eds: A. F. M. Smith and P. R. Freedman, London: Wiley & Sons.