Handbook of Statistics, Vol. 25
ISSN: 0169-7161

DOI 10.1016/S0169-7161(05)25013-7

13

Bayesian Methods for Function Estimation

Nidhan Choudhuri, Subhashis Ghosal and Anindya Roy

Abstract

Keywords: consistency; convergence rate; Dirichlet process; density estimation; Markov chain Monte Carlo; posterior distribution; regression function; spectral density; transition density

1. Introduction

Nonparametric and semiparametric statistical models are increasingly replacing parametric models, for the latter's lack of sufficient flexibility to address a wide variety of data. A nonparametric or semiparametric model involves at least one infinite-dimensional parameter, usually a function, and hence may also be referred to as an infinite-dimensional model. Functions of common interest, among many others, include the cumulative distribution function, density function, regression function, hazard rate, transition density of a Markov process, and spectral density of a time series. While frequentist methods for nonparametric estimation have been flourishing for many of these problems, nonparametric Bayesian estimation methods had been relatively less developed.

Besides philosophical reasons, there are some practical advantages of the Bayesian approach. On the one hand, the Bayesian approach allows one to reflect ones prior beliefs into the analysis. On the other hand, the Bayesian approach is straightforward in principle where inference is based on the posterior distribution only. Subjective elicitation of priors is relatively simple in a parametric framework, and in the absence of any concrete knowledge, there are many default mechanisms for prior specification. However, the recent popularity of Bayesian analysis comes from the availability of various Markov chain Monte Carlo (MCMC) algorithms that make the computation feasible with today's computers in almost every parametric problem. Prediction, which is sometimes the primary objective of a statistical analysis, is solved most naturally if one follows the Bayesian approach. Many non-Bayesian methods, including the maximum likelihood estimator (MLE), can have very unnatural behavior (such as staying on the boundary with high probability) when the parameter space is restricted, while

a Bayesian estimator does not suffer from this drawback. Besides, the optimality of a parametric Bayesian procedure is often justified through large sample as well as finite sample admissibility properties.

The difficulties for a Bayesian analysis in a nonparametric framework is threefold. First, a subjective elicitation of a prior is not possible due to the vastness of the parameter space and the construction of a default prior becomes difficult mainly due to the absence of the Lebesgue measure. Secondly, common MCMC techniques do not directly apply as the parameter space is infinite-dimensional. Sampling from the posterior distribution often requires innovative MCMC algorithms that depend on the problem at hand as well as the prior given on the functional parameter. Some of these techniques include the introduction of latent variables, data augmentation and reparametrization of the parameter space. Thus, the problem of prior elicitation cannot be separated from the computational issues.

When a statistical method is developed, particular attention should be given to the quality of the corresponding solution. Of the many different criteria, asymptotic consistency and rate of convergence are perhaps among the least disputed. Consistency may be thought of as a validation of the method used by the Bayesian. Consider an imaginary experiment where an experimenter generates observations from a given stochastic model with some value of the parameter and presents the data to a Bayesian without revealing the true value of the parameter. If enough information is provided in the form of a large number of observations, the Bayesian's assessment of the unknown parameter should be close to the true value of it. Another reason to study consistency is its relationship with robustness with respect to the choice of the prior. Due to the lack of complete faith in the prior, we should require that at least eventually, the data overrides the prior opinion. Alternatively two Bayesians, with two different priors, presented with the same data eventually must agree. This large sample "merging of opinions" is equivalent to consistency (Blackwell and Dubins, 1962; Diaconis and Freedman, 1986a, 1986b; Ghosh et al., 1994). For virtually all finite-dimensional problems, the posterior distribution is consistent (Ibragimov and Has'minskii, 1981; Le Cam, 1986; Ghosal et al., 1995) if the prior does not rule out the true value. This is roughly a consequence of the fact that the likelihood is highly peaked near the true value of the parameter if the sample size is large. However, for infinite-dimensional problems, such a conclusion is false (Freedman, 1963; Diaconis and Freedman, 1986a, 1986b; Doss, 1985a, 1985b; Kim and Lee, 2001). Thus posterior consistency must be verified before using a prior.

In this chapter, we review Bayesian methods for some important curve estimation problems. There are several good reviews available in the literature such as Hjort (1996, 2003), Wasserman (1998), Ghosal et al. (1999a), the monograph of Ghosh and Ramamoorthi (2003) and several chapters in this volume. We omit many details which may be found from these sources. We focus on three different aspects of the problem: prior specification, computation and asymptotic properties of the posterior distribution. In Section 2, we describe various priors on infinite-dimensional spaces. General results on posterior consistency and rate of convergence are reviewed in Section 3. Specific curve estimation problems are addressed in the subsequent sections.

hs25 v.2005/04/21 Prn:19/05/2005; 15:11 F:hs25013.tex; VTEX/VJ p. 3
aid: 25013 pii: S0169-7161(05)25013-7 docsubty: REV

2. Priors on infinite-dimensional spaces

A well accepted criterion for the choice of a nonparametric prior is that the prior has a large or full topological support. Intuitively, such a prior can reach every corner of the parameter space and thus can be expected to have a consistent posterior. More flexible models have higher complexity and hence the process of prior elicitation becomes more complex. Priors are usually constructed from the consideration of mathematical tractability, feasibility of computation, and good large sample behavior. The form of the prior is chosen according to some default mechanism while the key hyper-parameters are chosen to reflect any prior beliefs. A prior on a function space may be thought of as a stochastic process taking values in the given function space. Thus, a prior may be specified by describing a sampling scheme that generate random function with desired properties or they can be specified by describing the finite-dimensional laws. An advantage of the first approach is that the existence of the prior measure is automatic, while for the latter, the nontrivial proposition of existence needs to be established. Often the function space is approximated by a sequence of sieves in such a way that it is easier to put a prior on these sieves. A prior on the entire space is then described by letting the index of the sieve vary with the sample size, or by putting a further prior on the index thus leading to a hierarchical mixture prior. Here we describe some general methods of prior construction on function spaces.

2.1. Dirichlet process

Dirichlet processes were introduced by Ferguson (1973) as prior distributions on the space of probability measures on a given measurable space $(\mathfrak{X}, \mathcal{B})$. Let $M > 0$ and G be a probability measure on $(\mathfrak{X}, \mathcal{B})$. A Dirichlet process on $(\mathfrak{X}, \mathcal{B})$ with parameters (M, G) is a random probability measure P which assigns a number $P(B)$ to every $B \in \mathcal{B}$ such that

(i) $P(B)$ is a measurable $[0, 1]$-valued random variable;
(ii) each realization of P is a probability measure on $(\mathfrak{X}, \mathcal{B})$;
(iii) for each measurable finite partition $\{B_1, \ldots, B_k\}$ of $\mathfrak{X}$, the joint distribution of the vector $(P(B_1), \ldots, P(B_k))$ on the k-dimensional unit simplex has Dirichlet distribution with parameters $(k; MG(B_1), \ldots, MG(B_k))$.

(We follow the usual convention for the Dirichlet distribution that a component is a.s. 0 if the corresponding parameter is 0.) Using Kolmogorov's consistency theorem, Ferguson (1973) showed that a process with the stated properties exists. The argument could be made more elegant and transparent by using a countable generator of $\mathcal{B}$ as in Blackwell (1973). The distribution of P is also uniquely defined by its specified finite-dimensional distributions in (iii) above. We shall denote the process by $\mathrm{Dir}(M, G)$. If $(M_1, G_1) \neq (M_2, G_2)$ then the corresponding Dirichlet processes $\mathrm{Dir}(M_1, G_1)$ and $\mathrm{Dir}(M_2, G_2)$ are different, unless both G_1 and G_2 are degenerate at the same point. The parameter M is called the precision, G is called the center measure, and the product MG is called the base measure of the Dirichlet process. Note that

$$\mathrm{E}\big(P(B)\big) = G(B), \qquad \mathrm{var}\big(P(B)\big) = \frac{G(B)(1 - G(B))}{1 + M}. \tag{2.1}$$

hs25 v.2005/04/21 Prn:19/05/2005; 15:11 F:hs25013.tex; VTEX/VJ p. 4
aid: 25013 pii: S0169-7161(05)25013-7 docsubty: REV

Therefore, if M is large, P is tightly concentrated about G justifying the terminology. The relation (2.1) easily follows by the observation that each $P(B)$ is distributed as beta with parameters $MG(B)$ and $M(1-G(B))$. By considering finite linear combinations of indicator of sets and passing to the limit, it readily follows that (2.1) could be extended to functions, that is, $\mathrm{E}(\int \psi \, \mathrm{d}P) = \int \psi \, \mathrm{d}G$, and $\mathrm{var}(\int \psi \, \mathrm{d}P) = \mathrm{var}_G(\psi)/(1+M)$.

As $P(A)$ is distributed as beta $(MG(A), MG(A^c))$, it follows that $P(A) > 0$ a.s. if and only if $G(A) > 0$. However, this does not imply that P is a.s. mutually absolutely continuous with G, as the null set could depend on A. As a matter of fact, the two measures are often a.s. mutually singular.

If $\mathfrak{X}$ is a separable metric space, the topological support of a measure on $\mathfrak{X}$ and the weak[1] topology on the space $\mathfrak{M}(\mathfrak{X})$ of all probability measures on $\mathfrak{X}$ may be defined. The support of $\mathrm{Dir}(M, G)$ with respect to the weak topology is given by $\{P \in \mathfrak{M}(\mathfrak{X})\colon \mathrm{supp}(P) \subset \mathrm{supp}(G)\}$. In particular, if the support of G is $\mathfrak{X}$, then the support of $\mathrm{Dir}(M, G)$ is the whole of $\mathfrak{M}(\mathfrak{X})$. Thus the Dirichlet process can be easily chosen to be well spread over the space of probability measures. This may however look apparently contradictory to the fact that a random P following $\mathrm{Dir}(M, G)$ is a.s. discrete. This important (but perhaps somewhat disappointing) property was observed in Ferguson (1973) by using a gamma process representation of the Dirichlet process and in Blackwell (1973) by using a Polya urn scheme representation. In the latter case, the Dirichlet process arises as the mixing measure in de Finetti's representation in the following continuous analogue of the Polya urn scheme: $X_1 \sim G$; for $i = 1, 2, \ldots$, $X_i = X_j$ with probability $1/(M+i-1)$ for $j = 1, \ldots, i-1$ and $X_i \sim G$ with probability $M/(M+i-1)$ independently of the other variables. This representation is extremely crucial for MCMC sampling from a Dirichlet process. The representation also shows that ties are expected among $X_1, \ldots, X_n$. The expected number of distinct X's, as $n \to \infty$, is $M \log \frac{n}{M}$, which asymptotically much smaller than n. A simple proof of a.s. discreteness of Dirichlet random measure, due to Savage, is given in Theorem 3.2.3 of Ghosh and Ramamoorthi (2003).

Sethuraman (1994) gave a constructive representation of the Dirichlet process. If $\theta_1, \theta_2, \ldots$ are i.i.d. G_0, $Y_1, Y_2, \ldots$ are i.i.d. beta$(1, M)$, $V_i = Y_i \prod_{j=1}^{i-1}(1-Y_j)$ and

$$P = \sum_{i=1}^{\infty} V_i \delta_{\theta_i}, \tag{2.2}$$

then the above infinite series converges a.s. to a random probability measure that is distributed as $\mathrm{Dir}(M, G)$. It may be noted that the masses V_i's are obtained by successive "stick-breaking" with $Y_1, Y_2, \ldots$ as the corresponding stick-breaking proportions, and allotted to randomly chosen points $\theta_1, \theta_2, \ldots$ generated from G. Sethuraman's representation has made it possible to use the Dirichlet process in many complex problem using some truncation and Monte Carlo algorithms. Approximations of this type are discussed by Muliere and Tardella (1998) and Iswaran and Zarepour (2002a, 2002b). Another consequence of the Sethuraman representation is that if $P \sim \mathrm{Dir}(M, G)$, $\theta \sim G$ and $Y \sim \mathrm{beta}(1, M)$, and all of them are independent, then $Y\delta_\theta + (1-Y)P$ also has $\mathrm{Dir}(M, G)$ distribution. This property leads to important distributional equations for

[1] What we call weak is termed as weak star in functional analysis.

hs25 v.2005/04/21 Prn:19/05/2005; 15:11 F:hs25013.tex; VTEX/VJ p. 5
aid: 25013 pii: S0169-7161(05)25013-7 docsubty: REV

functionals of the Dirichlet process, and could also be used to simulate a Markov chain on $\mathfrak{M}(\mathfrak{X})$ with $\mathrm{Dir}(M, G)$ as its stationary distribution.

The Dirichlet process has a very important conditioning property. If A is set with $G(A) > 0$ (which implies that $P(A) > 0$ a.s.), then the random measure $P|_A$, the restriction of P to A defined by $P|_A(B) = P(B|A) = P(B \cap A)/P(A)$, is distributed as Dirichlet process with parameters $MG(A)$ and $G|_A$ and is independent of $P(A)$. The argument can be extended to more than one set. Thus the Dirichlet process locally splits into numerous independent Dirichlet processes.

A peculiar property of the Dirichlet process is that any two Dirichlet processes $\mathrm{Dir}(M_1, G_1)$ and $\mathrm{Dir}(M_2, G_2)$ are mutually singular if G_1, G_2 are nonatomic and $(M_1, G_1) \neq (M_2, G_2)$.

The distribution of a random mean functional $\int \psi \, \mathrm{d}P$, where ψ is a measurable function, is of some interest. Although, $\int \psi \, \mathrm{d}P$ has finite mean if and only if $\int |\psi| \, \mathrm{d}G < \infty$, P has a significantly shorter tail than that of G. For instance, the random P generated by a Dirichlet process with Cauchy base measure has all moments. Distributions of the random mean functional has been studied in many articles including Cifarelli and Regazzini (1990) and Regazzini et al. (2002). Interestingly the distribution of $\int x \, \mathrm{d}P(x)$ is G if and only if G is Cauchy.

The behavior of the tail probabilities of a random P obtained from a Dirichlet process is important for various purposes. Fristedt (1967) and Fristedt and Pruitt (1971) characterized the growth rate of a gamma process and using their result, Doss and Sellke (1982) obtained analogous results for the tail probabilities of P.

Weak convergence properties of the Dirichlet process are controlled by the convergence of its parameters. Let G_n weakly converge to G. Then

(i) if $M_n \to M > 0$, then $\mathrm{Dir}(M_n, G_n)$ converges weakly to $\mathrm{Dir}(M, G)$;
(ii) if $M_n \to 0$, then $\mathrm{Dir}(M_n, G_n)$ converges weakly to a measure degenerated at a random $\theta \sim G$;
(iii) if $M_n \to \infty$, then $\mathrm{Dir}(M_n, G_n)$ converges weakly to random measure degenerate at G.

2.2. Processes derived from the Dirichlet process

2.2.1. Mixtures of Dirichlet processes

The mixture of Dirichlet processes was introduced by Antoniak (1974). While eliciting the base measure using (2.1), it may be reasonable to guess that the prior mean measure is normal, but it may be difficult to specify the values of the mean and the variance of this normal distribution. It therefore makes sense to put a prior on the mean and the variance. More generally, one may propose a parametric family as the base measure and put hyper-priors on the parameters of that family. The resulting procedure has an intuitive appeal in that if one is a weak believer in a parametric family, then instead of using a parametric analysis, one may use the corresponding mixture of Dirichlet to robustify the parametric procedure. More formally, we may write the hierarchical Bayesian model $P \sim \mathrm{Dir}(M_\theta, G_\theta)$, where the indexing parameter $\theta \sim \pi$.

In semiparametric problems, mixtures of Dirichlet priors appear if the nonparametric part is given a Dirichlet process. In this case, the interest is usually in the posterior

hs25 v.2005/04/21 Prn:19/05/2005; 15:11 F:hs25013.tex; VTEX/VJ p. 6
aid: 25013 pii: S0169-7161(05)25013-7 docsubty: REV

distribution of the parametric part, which has a role much bigger than that of an indexing parameter.

2.2.2. Dirichlet mixtures

Although the Dirichlet process cannot be used as a prior for estimating a density, convoluting it with a kernel will produce smooth densities. Such an approach was pioneered by Ferguson (1983) and Lo (1984). Let Θ be a parameter set, typically a Euclidean space. For each θ, let $\psi(x,\theta)$ be a probability density function. A nonparametric mixture of $\psi(x,\theta)$ is obtained by considering $p_F(x) = \int \psi(x,\theta)\,\mathrm{d}F(\theta)$. These mixtures can form a very rich family. For instance, the location and scale mixture of the form $\sigma^{-1}k((x-\mu)/\sigma)$, for some fixed density k, may approximate any density in the L_1-sense if σ is allowed to approach to 0. Thus, a prior on densities may be induced by putting a Dirichlet process prior on the mixing distribution F and a prior on σ.

The choice of an appropriate kernel depends on the underlying sample space. If the underlying density function is defined on the entire real line, a location-scale kernel is appropriate. If on the unit interval, beta distributions form a flexible two parameter family. If on the positive half line, mixtures of gamma, Weibull or lognormal may be used. The use of a uniform kernel leads to random histograms. Petrone and Veronese (2002) motivated a canonical way of viewing the choice of a kernel through the notion of the Feller sampling scheme, and call the resulting prior a Feller prior.

2.2.3. Invariant Dirichlet process

The invariant Dirichlet process was considered by Dalal (1979). Suppose that we want to put a prior on the space of all probability measures symmetric about zero. One may let P follow $\mathrm{Dir}(M,G)$ and put $\overline{P}(A) = (P(A)+P(-A))/2$, where $-A = \{x\colon -x \in A\}$.[2] More generally, one can consider a compact group $\mathfrak{G}$ acting on the sample space $\mathfrak{X}$ and consider the distribution of $\overline{P}$ as the invariant Dirichlet process where $\overline{P}(A) = \int P(gA)\,\mathrm{d}\mu(g)$, μ stands for the Haar probability measure on $\mathfrak{G}$ and P follows the Dirichlet process.

The technique is particularly helpful for constructing priors on the error distribution F for the location problem $X = \theta + \epsilon$. The problem is not identifiable without some restriction on F, and symmetry about zero is a reasonable condition on F ensuring identifiability. The symmetrized Dirichlet process prior was used by Diaconis and Freedman (1986a, 1986b) to present a striking example of inconsistency of the posterior distribution.

2.2.4. Pinned-down Dirichlet

If $\{B_1,\ldots,B_k\}$ is a finite partition, called control sets, then the conditional distribution of P given $\{P(B_j) = w_j,\ j = 1,\ldots,k\}$, where P follows $\mathrm{Dir}(M,G)$ and $w_j \geqslant 0$, $\sum_{j=1}^{k} w_j = 1$, is called a pinned-down Dirichlet process. By the conditioning property of the Dirichlet process mentioned in the last subsection, it follows that the above

[2] Another way of randomly generating symmetric probabilities is to consider a Dirichlet process P on $[0,\infty)$ and unfold it to $\widetilde{P}$ on $\mathbb{R}$ by $\widetilde{P}(-A) = \widetilde{P}(A) = \frac{1}{2}P(A)$.

hs25 v.2005/04/21 Prn:19/05/2005; 15:11 F:hs25013.tex; VTEX/VJ p. 7
aid: 25013 pii: S0169-7161(05)25013-7 docsubty: REV

process may be written as $P = \sum_{j=1}^{k} w_j P_j$, where each P_j is a Dirichlet process on B_j. Consequently P is a countable mixture of Dirichlet (with orthogonal supports).

A particular case of pinned-down Dirichlet is obtained when one puts the restriction that P has median 0. Doss (1985a, 1985b) used this idea to put a prior for the semiparametric location problem and showed an inconsistency result similar to Diaconis and Freedman (1986a, 1986b) mentioned above.

2.3. Generalizations of the Dirichlet process

While the Dirichlet process is a prior with many fascinating properties, its reliance on only two parameters may sometimes be restrictive. One drawback of Dirichlet process is that it always produces discrete random probability measures. Another property of Dirichlet which is sometimes embarrassing is that the correlation between the random probabilities of two sets is always negative. Often, random probabilities of sets that are close enough are expected to be positively related if some smoothness is present. More flexible priors may be constructed by generalizing the way the prior probabilities are assigned. Below we discuss some of the important generalizations of a Dirichlet process.

2.3.1. Tail-free and neutral to the right process

The concept of a tail-free process was introduced by Freedman (1963) and chronologically precedes that of the Dirichlet process. A tail-free process is defined by random allocations of probabilities to sets in a nested sequence of partitions. Let $E = \{0, 1\}$ and E^m be the m-fold Cartesian product $E \times \cdots \times E$ where $E^0 = \emptyset$. Further, set $E^* = \bigcup_{m=0}^{\infty} E^m$. Let $\pi_0 = \{\mathfrak{X}\}$ and for each $m = 1, 2, \ldots$, let $\pi_m = \{B_\varepsilon\colon\ \varepsilon \in E^m\}$ be a partition of $\mathfrak{X}$ so that sets of π_{m+1} are obtained from a binary split of the sets of π_m and $\bigcup_{m=0}^{\infty} \pi_m$ be a generator for the Borel sigma-field on $\mathbb{R}$. A probability P may then be described by specifying all the conditional probabilities $\{V_\varepsilon = P(B_{\varepsilon 0}|B_\varepsilon)\colon\ \varepsilon \in E^*\}$. A prior for P may thus be defined by specifying the joint distribution of all V_ε's. The specification may be written in a tree form. The different hierarchy in the tree signifies prior specification of different levels. A prior for P is said to be tail-free with respect to the sequence of partitions $\{\pi_m\}$ if the collections $\{V_\emptyset\}, \{V_0, V_1\}, \{V_{00}, V_{01}, V_{10}, V_{11}\}, \ldots$, are mutually independent. Note that, variables within the same hierarchy need not be independent; only the variables at different levels are required to be so. Partitions more general than binary partitions could be used, although that will not lead to more general priors.

A Dirichlet process is tail-free with respect to any sequence of partitions. Indeed, the Dirichlet process is the only prior that has this distinguished property; see Ferguson (1974) and the references therein. Tail-free priors satisfy some interesting zero-one laws, namely, the random measure generated by a tail-free process is absolutely continuous with respect to a given finite measure with probability zero or one. This follows from the fact that the criterion of absolute continuity may be expressed as tail event with respect to a collection of independent random variables and Kolmogorov's zero-one law may be applied; see Ghosh and Ramamoorthi (2003) for details. Kraft (1964) gave a very useful sufficient condition for the almost sure absolute continuity of a tail-free process.

hs25 v.2005/04/21 Prn:19/05/2005; 15:11 F:hs25013.tex; VTEX/VJ p. 8
aid: 25013 pii: S0169-7161(05)25013-7 docsubty: REV

Neutral to the right processes, introduced by Doksum (1974), are also tail-free processes, but the concept is applicable only to survival distribution functions. If F is a random distribution function on the positive half line, then F is said to follow a neutral to the right process if for every k and $0 < t_1 < \cdots < t_k$, there exists independent random variables $V_1, \ldots, V_k$ such that the joint distribution of $(1 - F(t_1), 1 - F(t_2), \ldots, 1 - F(t_k))$ is same as that of the successive products $(V_1, V_1V_2, \ldots, \prod_{j=1}^{k} V_j)$. Thus a neutral to the right prior is obtained by stick breaking. Clearly the process is tail-free with respect to the nested sequence $\{[0, t_1], (t_1, \infty)\}, \{[0, t_1], (t_1, t_2], (t_2, \infty)\}, \ldots$ of partitions. Note that $F(x)$ may be written as $e^{-H(x)}$, where $H(\cdot)$ is a process of independent increments.

2.3.2. Polya tree process

A Polya tree process is a special case of a tail-free process, where besides across row independence, the random conditional probabilities are also independent within row and have beta distributions. To elaborate, let $\{\pi_m\}$ be a sequence of binary partition as before and $\{\alpha_\varepsilon\colon\ \varepsilon \in E^*\}$ be a collection of nonnegative numbers. A random probability measure P on $\mathbb{R}$ is said to possess a Polya tree distribution with parameters $(\{\pi_m\}, \{\alpha_\varepsilon\colon\ \varepsilon \in E^*\})$, if there exist a collection $\mathcal{Y} = \{Y_\varepsilon\colon\ \varepsilon \in E^*\}$ of random variables such that the following hold:

(i) The collection $\mathcal{Y}$ consists of mutually independent random variables;
(ii) For each $\varepsilon \in E^*$, Y_ε has a beta distribution with parameters $\alpha_{\varepsilon 0}$ and $\alpha_{\varepsilon 1}$;
(iii) The random probability measure P is related to $\mathcal{Y}$ through the relations

$$P(B_{\varepsilon_1 \cdots \varepsilon_m}) = \left(\prod_{j=1;\varepsilon_j=0}^{m} Y_{\varepsilon_1 \cdots \varepsilon_{j-1}} \right) \left(\prod_{j=1;\varepsilon_j=1}^{m} (1 - Y_{\varepsilon_1 \cdots \varepsilon_{j-1}}) \right), \quad m = 1, 2, \ldots,$$

where the factors are $Y_\emptyset$ or $1 - Y_\emptyset$ if $j = 1$.

The concept of a Polya tree was originally considered by Ferguson (1974) and Blackwell and MacQueen (1973), and later studied thoroughly by Mauldin et al. (1992) and Lavine (1992, 1994). The prior can be seen as arising as the de Finetti measure in a generalized Polya urn scheme; see Mauldin et al. (1992) for details.

The class of Polya trees contain all Dirichlet processes, characterized by the relation that $\alpha_{\varepsilon 0} + \alpha_{\varepsilon 1} = \alpha_\varepsilon$ for all ε. A Polya tree can be chosen to generate only absolutely continuous distributions. The prior expectation of the process could be easily written down; see Lavine (1992) for details. Below we consider an important special case for discussion, which is most relevant for statistical use. Consider $\mathfrak{X}$ to be a subset of the real line and let G be a probability measure. Let the partitions be obtained successively by splitting the line at the median, the quartiles, the octiles, and in general, binary quantiles of G. If $\alpha_{\varepsilon 0} = \alpha_{\varepsilon 1}$ for all $\varepsilon \in E^*$, then it follows that $\mathrm{E}(P) = G$. Thus G will have the role similar to that of the center measure of a Dirichlet process, and hence will be relatively easy to elicit. Besides, the Polya tree will have infinitely many more parameters which may be used to describe one's prior belief. Often, to avoid specifying too many parameters, a default method is adopted, where one chooses α_ε depending

hs25 v.2005/04/21 Prn:19/05/2005; 15:11 F:hs25013.tex; VTEX/VJ p. 9
aid: 25013 pii: S0169-7161(05)25013-7 docsubty: REV

only on the length of the finite string ε. Let a_m stand for the value of α_ε when ε has length m. The growth rate of a_m controls the smoothness of the Polya tree process. For instance, if $a_m = c2^{-m}$, we obtain the Dirichlet process, which generate discrete probabilities. If $\sum_{m=1}^{\infty} a_m^{-1} < \infty$ (for instance, if $a_m = cm^2$), then it follows from Kraft's (1964) result that the random P is absolutely continuous with respect to G. The choice $a_m = c$ leads to singular continuous distributions almost surely; see Ferguson (1974). This could guide one to choose the sequence a_m. For smoothness, one should choose rapidly growing a_m. One may actually like to choose according to one's prior belief in the beginning of the tree deviating from the above default choice, and let a default method choose the parameters at the later stages where practically no prior information is available. An extreme form of this will lead to partially specified Polya trees, where one chooses a_m to be infinity after a certain stage (which is equivalent to uniformly spreading the mass inside a given interval).

Although the prior mean distribution function may have a smooth Lebesgue density, the randomly sampled densities from a Polya tree are very rough, being nowhere differentiable. To overcome this difficulty, mixtures of a Polya tree, where the partitioning measure G involves some additional parameter θ with some prior, may be considered. The additional parameter will average out jumps to yield smooth densities; see Hanson and Johnson (2002). However, then the tail-freeness is lost and the resulting posterior distribution could be inconsistent. Berger and Guglielmi (2001) considered a mixture where the partition remains fixed and the α-parameters depend on θ, and applied the resulting prior to a model selection problem.

2.3.3. Generalized Dirichlet process

The k-dimensional Dirichlet distribution may be viewed as the conditional distribution of $(p_1, \ldots, p_k)$ given that $\sum_{j=1}^{k} p_j = 1$, where $p_j = \mathrm{e}^{-Y_j}$ and Y_j's are independent exponential variables. In general, if Y_j's have a joint density $h(y_1, \ldots, y_k)$, the conditional joint density of $(p_1, \ldots, p_{k-1})$ is proportional to $h(-\log p_1, \ldots, -\log p_k)p_k^{-1} \cdots p_k^{-1}$, where $p_k = 1 - \sum_{j=1}^{k-1} p_j$. Hjort (1996) considered the joint density of Y_j's to be proportional to $\prod_{j=1}^{k} \mathrm{e}^{-\alpha_j y_j} g_0(y_1, \ldots, y_k)$, and hence the resulting (conditional) density of $p_1, \ldots, p_{k-1}$ is proportional to $p_1^{\alpha_1 - 1} \cdots p_k^{\alpha_k - 1} g(p_1, \ldots, p_k)$, where $g(p_1, \ldots, p_k) = g_0(-\log p_1, \ldots, -\log p_k)$. We may put $g(p) = \mathrm{e}^{-\lambda \Delta(p)}$, where $\Delta(p)$ is a penalty term for roughness such as $\sum_{j=1}^{k-1}(p_{j+1} - p_j)^2$, $\sum_{j=2}^{k-1}(p_{j+1} - 2p_j + p_{j-1})^2$ or $\sum_{j=1}^{k-1}(\log p_{j+1} - \log p_j)^2$. The penalty term helps maintain positive correlation and hence "smoothness". The tuning parameter λ controls the extent to which penalty is imposed for roughness. The resulting posterior distribution is conjugate with mode equivalent to a penalized MLE. Combined with random histogram or passing through the limit as the bin width goes to 0, the technique could also be applied to continuous data.

2.3.4. Priors obtained from random series representation

Sethuraman's (1994) infinite series representation creates a lot of possibilities of generalizing the Dirichlet process by changing the distribution of the weights, the support points, or even the number of terms. Consider a random probability measure given by

hs25 v.2005/04/21 Prn:19/05/2005; 15:11 F:hs25013.tex; VTEX/VJ p. 10
aid: 25013 pii: S0169-7161(05)25013-7 docsubty: REV

$P = \sum_{i=1}^{N} V_i \delta_{\theta_i}$, where $1 \leqslant N \leqslant \infty$, $\sum_{i=1}^{N} V_i = 1$ and N may be given a further prior distribution. Note that the resulting random probability measure is almost surely discrete. Choosing $N = \infty$, θ_i's as i.i.d. G as in the Sethuraman representation, $V_i = Y_i \prod_{j=1}^{i-1}(1 - Y_j)$, where $Y_1, Y_2, \ldots$ are i.i.d. beta(a, b), Hjort (2000) obtained an interesting generalization of the Dirichlet process. The resulting process admits, as in the case of a Dirichlet process, explicit formulae for the posterior mean and variance of a mean functional.

From computational point of view, a prior is more tractable if N is chosen to be finite. To be able to achieve reasonable large sample properties, either N has to depend on the sample size n, or N must be given a prior which is infinitely supported. Given $N = k$, the prior on $(V_1, \ldots, V_k)$ is taken to be k-dimensional Dirichlet distribution with parameters $(\alpha_{1,n}, \ldots, \alpha_{k,n})$. The parameters θ_i's are usually chosen as in the Sethuraman's representation, that is i.i.d. G. Iswaran and Zarepour (2002a) studied convergence properties of these random measures. For the choice $\alpha_{j,k} = M/k$, the limiting measure is Dir(M, G). However, the commonly advocated choice $\alpha_{j,k} = M$ leads essentially to a parametric prior, and hence to an inconsistent posterior.

2.4. Gaussian process

Considered first by Leonard (1978), and then by Lenk (1988, 1991) in the context of density estimation, a Gaussian process may be used in a wider generality because of its ability to produce arbitrary shapes. The method may be applied to nonparametric regression where only smoothness is assumed for the regression function. The mean function reflects any prior belief while the covariance kernel may be tuned to control the smoothness of the sample paths as well as to reflect the confidence in the prior guess. In a generalized regression, where the function of interest has restricted range, a link function is used to map the unrestricted range of the Gaussian process to the desired one. A commonly used Gaussian process in the regression context is the integrated Wiener process with some random intercept term as in Wahba (1978). Choudhuri et al. (2004b) used a general Gaussian process prior for binary regression.

2.5. Independent increment process

Suppose that we want to put a prior on survival distribution functions, that is, distribution functions on the positive half line. Let $Z(t)$ be a process with independent nonnegative increments such that $Z(\infty)$, the total mass of Z, is a.s. finite. Then a prior on F may be constructed by the relation $F(t) = Z(t)/Z(\infty)$. Such a prior is necessarily neutral to the right. When $Z(t)$ is the gamma process, that is an independent increment process with $Z(t) \sim \text{gamma}(MG(t), 1)$, then the resulting distribution of P is Dirichlet process Dir(M, G).

For estimating a survival function, it is often easier to work with the cumulative hazard function, which needs only be positive. If $Z(t)$ is a process such that $Z(\infty) = \infty$ a.s., then $F(t) = 1 - \mathrm{e}^{-Z(t)}$ is a distribution function. The process $Z(t)$ may be characterized in terms of its Lévy measure $N_t(\cdot)$, and is called a Lévy process. Unfortunately, as $Z(t)$ necessarily increases by jumps only, $Z(t)$ is not the cumulative hazard function corresponding to $F(t)$. Instead, one may define $F(t)$ by the relation

hs25 v.2005/04/21 Prn:19/05/2005; 15:11 F:hs25013.tex; VTEX/VJ p. 11
aid: 25013 pii: S0169-7161(05)25013-7 docsubty: REV

$Z(t) = \int_0^t \mathrm{d}F(s)/(1 - F(s-))$. The expressions of prior mean and variance, and posterior updating are relatively straightforward in terms of the Lévy measure; see Hjort (1990) and Kim (1999). Particular choices of the Lévy measure lead to special priors such as the Dirichlet process, completely homogeneous process (Ferguson and Phadia, 1979), gamma process (Lo, 1982), beta process (Hjort, 1990), beta-Stacy process (Walker and Muliere, 1997) and extended beta process (Kim and Lee, 2001). Kim and Lee (2001) settled the issue of consistency, and provided an interesting example of inconsistency.

A disadvantage of modeling the process $Z(t)$ is that the resulting F is discrete. Dykstra and Laud (1981) considered a Lévy process to model the hazard rate. However, this approach leads only to monotone hazard functions. Nieto-Barajas and Walker (2004) replaced the independent increments process by a Markov process and obtained continuous sample paths.

2.6. Some other processes

One approach to putting a prior on a function space is to decompose a function into a basis expansion of the form $\sum_{j=1}^{\infty} b_j \psi_j(\cdot)$ for some fixed basis functions and then putting priors on b_j's. An orthogonal basis is very useful if the function space of interest is a Hilbert space. Various popular choices of such basis include polynomials, trigonometric functions, splines and wavelets among many others. If the coefficients are unrestricted, independent normal priors may used. Interestingly, when the coefficients are normally distributed, the prior on the random function is a Gaussian process. Conversely, a Gaussian process may be represented in this way by virtue of the Karhunen–Loévé expansion. When the function values are restricted, transformations should be used prior to a basis expansion. For instance, for a density function, an expansion should be raised to the exponential and then normalized. Barron et al. (1999) used polynomials to construct an infinite-dimensional exponential family. Hjort (1996) discussed a prior on a density induced by the Hermite polynomial expansion and a prior on the sequence of cumulants.

Instead of considering an infinite series representation, one may consider a series based on the first k terms, where k is deterministically increased to infinity with the sample size, or is itself given a prior that has infinite support. The span of the first k functions, as k tends to infinity, form approximating sieves in the sense of Grenander (1981). The resulting priors are recommended as default priors in infinite-dimensional spaces by Ghosal et al. (1997). In Ghosal et al. (2000), this idea was used with a spline basis for density estimation. They showed that with a suitable choice of k, depending on the sample size and the smoothness level of the target function, optimal convergence rates could be obtained.

If the domain is a bounded interval then the sequence of moments uniquely determines the probability measure. Hence a prior on the space of probability measures could be induced from that on the sequence of moments. One may control the location, scale, skewness and kurtosis of the random probability by using subjective priors on the first four moments. Priors for the higher-order moments are difficult to elicit, and some default method should be used.

hs25 v.2005/04/21 Prn:19/05/2005; 15:11 F:hs25013.tex; VTEX/VJ p. 12
aid: 25013 pii: S0169-7161(05)25013-7 docsubty: REV

Priors for quantiles are much easier to elicit than that for moments. One may put priors on all dyadic quantiles honoring the order restrictions. Conceptually, this operation is opposite to that of specifying a tree based prior such as the Polya tree or a tail-free process. Here masses are predetermined and the partitions are chosen randomly. In practice, one may put priors only for a finite number of quantiles, and then distribute the remaining masses uniformly over the corresponding interval. Interestingly, if the prior on the quantile process is induced from a Dirichlet process on the random probability, then the posterior expectation of a quantile (in the noninformative limit $M \to 0$) is seen to be a Bernstein polynomial smoother of the empirical quantile process. This leads to a quantile density estimator, which, upon inversion, leads to an automatically smoothed empirical density estimator; see Hjort (1996) for more details.

3. Consistency and rates of convergence

Let $\{(\mathcal{X}^{(n)}, \mathcal{A}^{(n)}, P_\theta^{(n)})\colon \theta \in \Theta\}$ be a sequence of statistical experiments with observations $X^{(n)}$, where the parameter set Θ is an arbitrary topological space and n is an indexing parameter, usually the sample size. Let $\mathcal{B}$ be the Borel sigma-field on Θ and Π_n be a probability measure on $(\Theta, \mathcal{B})$, which, in general, may depend on n. The posterior distribution is defined to be a version of the regular conditional probability of θ given $X^{(n)}$, and is denoted by $\Pi_n(\cdot|X^{(n)})$.

Let $\theta_0 \in \Theta$. We say that the posterior distribution is consistent at θ_0 (with respect to the given topology on Θ) if $\Pi_n(\cdot|X^{(n)})$ converges weakly to δ_{θ_0} as $n \to \infty$ under $P_{\theta_0}^{(n)}$-probability, or almost surely under the distribution induced by the parameter value θ_0. If the latter makes sense, it is a more appealing concept.

The above condition (in the almost sure sense) is equivalent to checking that except on a θ_0-induced null set of sample sequences, for any neighborhood U of θ_0, $\Pi_n(U^c|X^{(n)}) \to 0$. If the topology on Θ is countably generated (as in the case of a separable metric space), this reduces to $\Pi_n(U^c|X^{(n)}) \to 0$ a.s. under the distribution induced by θ_0 for every neighborhood U. An analogous conclusion holds for consistency in probability. Henceforth we work with the second formulation.

Consistency may be motivated as follows. A (prior or posterior) distribution stands for one's knowledge about the parameter. Perfect knowledge implies a degenerate prior. Thus consistency means weak convergence of knowledge towards the perfect knowledge as the amount of data increases.

Doob (1948) obtained a very general result on posterior consistency. Let the prior Π be fixed and the observations be i.i.d. Under some mild measurability conditions on the sample space (a standard Borel space will suffice) and model identifiability, Doob (1948) showed that the set of all $\theta \in \Theta$ where consistency does not hold is Π-null. This follows by the convergence of the martingale $\mathrm{E}I(\theta \in B|X_1, \ldots, X_n)$ to $\mathrm{E}I(\theta \in B|X_1, X_2, \ldots) = I(\theta \in B)$. The condition of i.i.d. observations could be replaced by the assumption that in the product space $\Theta \times \mathcal{X}^\infty$, the parameter θ is $\mathcal{A}^\infty$-measurable. Statistically speaking, the condition holds if there is a consistent estimate of some bimeasurable function of θ.

hs25 v.2005/04/21 Prn:19/05/2005; 15:11 F:hs25013.tex; VTEX/VJ p. 13
aid: 25013 pii: S0169-7161(05)25013-7 docsubty: REV

The above result should not however create a false sense of satisfaction as the Π-null set could be very large. It is important to know at which parameter values consistency holds. Indeed, barring a countable parameter space, Doob's (1948) is of little help. On the other hand, Doob's (1948) theorem implies that consistency holds at a parameter point whenever there is a prior point mass there.

Freedman (1963) showed that merely having positive Π-probability in a neighborhood of θ_0 does not imply consistency at that point.

EXAMPLE 1. Let $\Theta = \mathfrak{M}(\mathbb{Z}_+)$, the space of all discrete distribution on positive integers with the total variation distance on Θ. Let θ_0 be the geometric distribution with parameter $1/4$. There exists a prior Π such that every neighborhood of θ_0 has positive probability under Π, yet

$$\Pi(\theta \in U | X_1, \ldots, X_n) \to 1 \quad \text{a.s. } \big[\theta_0^\infty\big], \tag{3.1}$$

where U is any neighborhood of θ_1, the geometric distribution with parameter $3/4$.

Indeed, the following result of Freedman (1963) shows that the above example of inconsistency is somewhat generic in a topological sense.

THEOREM 1. *Let $\Theta = \mathfrak{M}(\mathbb{Z}_+)$ with the total variation distance on it, and let $\mathfrak{M}(\Theta)$ be the space of all priors on Θ with the weak topology. Put the product topology on $\Theta \times \mathfrak{M}(\Theta)$. Then*

$$\Big\{(\theta, \Pi) \in \Theta \times \mathfrak{M}(\Theta)\colon \limsup_{n\to\infty} \Pi(\theta \in U | X_1, \ldots, X_n) = 1 \quad \forall U \text{ open}, \ U \neq \emptyset\Big\} \tag{3.2}$$

is the complement of a meager set.[3]

Thus, Freedman's (1963) result tells us that except for a relatively small collection of pairs of (θ, Π), the posterior distribution wanders aimlessly around the parameter space. In particular, consistency will not hold at any given θ. While this result cautions us about naive uses of Bayesian methods, it does not mean that Bayesian methods are useless. Indeed, a pragmatic Bayesian's only aim might be to just be able to find a reasonable prior complying with one's subjective belief (if available) and obtaining consistency at various parameter values. There could be plenty of such priors available even though there will be many more that are not appropriate. The situation may be compared with the role of differentiable functions among the class of all continuous functions. Functions that are differentiable at some point form a small set in the same sense while nowhere differentiable functions are much more abundant.

From a pragmatic point of view, useful sufficient conditions ensuring consistency at a given point is the most important proposition. Freedman (1963, 1965) showed that for estimation of a probability measure, if the prior distribution is tail-free, then (a suitable

[3] A meager set is one which can be written as a countable union of closed sets without any interior points, and is considered to be topologically small.

hs25 v.2005/04/21 Prn:19/05/2005; 15:11 F:hs25013.tex; VTEX/VJ p. 14
aid: 25013 pii: S0169-7161(05)25013-7 docsubty: REV

version of) the posterior distribution is consistent at any point with respect to the weak topology. The idea behind this result is reducing every weak neighborhood to a Euclidean neighborhood in some finite-dimensional projection using the tail-free property.

Schwartz (1965), in a celebrated paper, obtained a general result on consistency. Schwartz's (1965) theorem requires a testing condition and a condition on the support of the prior.

Consider i.i.d. observations generated by a statistical model indexed by an abstract parameter space Θ admitting a density $p(x,\theta)$ with respect to some sigma-finite measure μ. Let $K(\theta_1,\theta_2)$ denote the Kullback–Leibler divergence $\int p(x,\theta_1)\log(p(x,\theta_1)/p(x,\theta_2))\,\mathrm{d}\mu(x)$. We say that $\theta_0 \in \Theta$ is in the Kullback–Leibler support of Π, we write $\theta_0 \in \mathrm{KL}(\Pi)$, if for every $\varepsilon > 0$, $\Pi\{\theta\colon K(\theta_0,\theta) < \varepsilon\}$. As the Kullback–Leibler divergence is asymmetric and not a metric, the support may not be interpreted in a topological sense. Indeed, a prior may have empty Kullback–Leibler support even on a separable metric space.

THEOREM 2. *Let $\theta_0 \in U \subset \Theta$. If there exists $m \geqslant 1$, a test function $\phi(X_1,\ldots,X_m)$ for testing $H_0\colon \theta = \theta_0$ against $H\colon \theta \in U^c$ with the property that* $\inf\{\mathrm{E}_\theta\phi(X_1,\ldots,X_m)\colon \theta \in U^c\} > \mathrm{E}_{\theta_0}\phi(X_1,\ldots,X_m)$ *and $\theta_0 \in \mathrm{KL}(\Pi)$, then $\Pi\{\theta \in U^c|X_1,\ldots,X_n\} \to 0$ a.s. $[P_{\theta_0}^\infty]$.*

The importance of Schwartz's theorem cannot be overemphasized. It forms the basic foundation of Bayesian asymptotic theory for general parameter spaces. The first condition requires existence of a strictly unbiased test for testing the hypothesis $H_0\colon \theta = \theta_0$ against the complement of a neighborhood U. The condition implies the existence of a sequence of tests $\Phi_n(X_1,\ldots,X_n)$ such that probabilities of both the type I error $\mathrm{E}_{\theta_0}\Phi_n(X_1,\ldots,X_n)$ and the (maximum) type II error $\sup_{\theta\in U^c}\mathrm{E}_\theta(1-\Phi_n(X_1,\ldots,X_n))$ converges to zero exponentially fast. This existence of test is thus only a size restriction on the model and not a condition on the prior. Writing

$$\Pi\big(\theta \in U^c|X_1,\ldots,X_n\big) = \frac{\int_{U^c}\prod_{i=1}^n \frac{p(X_i,\theta)}{p(X_i,\theta_0)}\,\mathrm{d}\Pi(\theta)}{\int_\Theta \prod_{i=1}^n \frac{p(X_i,\theta)}{p(X_i,\theta_0)}\,\mathrm{d}\Pi(\theta)}, \tag{3.3}$$

this condition is used to show that for some $c > 0$, the numerator in (3.3) is smaller than e^{-nc} for all sufficiently large n a.s. $[P_{\theta_0}^\infty]$. The condition on Kullback–Leibler support is a condition on the prior as well as the model. The condition implies that for all $c > 0$, $\mathrm{e}^{nc}\int_\Theta \prod_{i=1}^n \frac{p(X_i,\theta)}{p(X_i,\theta_0)}\,\mathrm{d}\Pi(\theta) \to \infty$ a.s. $[P_{\theta_0}^\infty]$. Combining these two assertions, we obtain the result of the theorem. The latter assertion follows by first replacing Θ by the subset $\{\theta\colon K(\theta_0,\theta) < \varepsilon\}$, applying the strong law of large numbers to the integrand and invoking Fatou's lemma. It may be noted that θ_0 needs to be in the Kullback–Leibler support, not merely in the topological support of the prior for this argument to go through. In practice, the condition is derived from the condition that θ_0 is in the topological support of the prior along with some conditions on "nicety" of $p(x,\theta_0)$.

The testing condition is usually more difficult to satisfy. In finite dimension, the condition usually holds. On the space of probability measures with the weak topology on it, it is also not difficult to show that the required test exists; see Theorem 4.4.2

hs25 v.2005/04/21 Prn:19/05/2005; 15:11 F:hs25013.tex; VTEX/VJ p. 15
aid: 25013 pii: S0169-7161(05)25013-7 docsubty: REV

of Ghosh and Ramamoorthi (2003). However, in more complicated problems or for stronger topologies on densities (such as the variation or the Hellinger distance), the required tests do not exist without an additional compactness condition. Le Cam (1986) and Birgé (1983) developed an elegant theory of existence of uniformly exponentially powerful tests. However, the theory applies provided that the two hypotheses are convex. It is therefore helpful to split U^c into small balls for which required tests exist. If Θ is compact, the number of balls needed to cover U^c will be finite, and hence by taking the maximum of the resulting tests, the required test for testing $\theta = \theta_0$ against $\theta \in U^c$ may be obtained. However, the compactness condition imposes a severe restriction.

By a simple yet very useful observation, Barron (1988) concluded that it suffices that Φ_n satisfy

$$\sup_{\theta \in U^c \cap \Theta_n} \mathrm{E}_\theta\bigl(1 - \Phi_n(X_1, \ldots, X_n)\bigr) < a\,\mathrm{e}^{-bn} \tag{3.4}$$

for some constants $a, b > 0$ and some "sieve" $\Theta_n \subset \Theta$, provided that it can be shown separately that

$$\Pi\bigl(\theta \in \Theta_n^c | X_1, \ldots, X_n\bigr) \to 0 \quad \text{a.s. } \bigl[P_{\theta_0}^\infty\bigr]. \tag{3.5}$$

By a simple application of Fubini's theorem, Barron (1988) concluded that (3.5) is implied by a condition only on the prior probability, namely, for some $c, d > 0$, $\Pi(\theta \in \Theta_n^c) \leqslant c\,\mathrm{e}^{-nd}$. Now one may choose each Θ_n to be compact. However, because of dependence on n, one needs to estimate the number of balls required to cover Θ_n. From the same arguments, it follows that one needs to cover the sieve Θ_n with a maximum of e^{nc} balls, which is essentially a restriction on the covering number of the sieve Θ_n. The remaining part Θ_n^c, which may be topologically much bigger receives only a negligible prior probability by the given condition. It is interesting to note that unlike in sieve methods in non-Bayesian contexts, the sieve is merely a technical device for establishing consistency; the prior and the resulting Bayes procedure is not influenced by the choice of the sieve. Moreover, the sieve can be chosen depending on the accuracy level defined by the neighborhood U.

Barron's (1988) useful observation made it possible to apply Schwartz's ideas to prove posterior consistency in noncompact spaces as well. When the observations are i.i.d., one may take the parameter θ to be the density p itself. Let p_0 stand for the true density of each observation. Exploiting this idea, for a space $\mathcal{P}$ of densities, Barron et al. (1999) gave a sufficient condition for posterior consistency in Hellinger distance $d_H(p_1, p_2) = (\int (p_1^{1/2} - p_2^{1/2})^2)^{1/2}$ in terms of a condition on bracketing Hellinger entropy[4] a sieve $\mathcal{P}_n \subset \mathcal{P}$. Barron et al. (1999) used brackets to directly bound the likelihood ratios uniformly in the numerator of (3.4). The condition turns out to be considerably stronger than necessary in that we need to bound only an average likelihood ratio. Following Schwartz's (1965) original approach involving test functions, Ghosal et al. (1999b) constructed the required tests using a much weaker condition on metric entropies. These authors considered the total variation distance $d_V(p_1, p_2) = \int |p_1 - p_2|$

[4] The ε-bracketing Hellinger entropy of a set is the logarithm of the number ε-brackets with respect to the Hellinger distance needed to cover the set; see van der Vaart and Wellner (1996) for details on this and the related concepts.

hs25 v.2005/04/21 Prn:19/05/2005; 15:11 F:hs25013.tex; VTEX/VJ p. 16
aid: 25013 pii: S0169-7161(05)25013-7 docsubty: REV

(which is equivalent to d_H), constructed a test directly for a point null against a small variation ball using Hoeffding's inequality, and combined the resulting tests using the condition on the metric entropy.

For a subset S of a metric space with a metric d on it, let $N(\varepsilon, S, d)$, called the ε-covering number of S with respect to the metric d, stand for the minimum number of ε-balls needed to cover S. The logarithm of $N(\varepsilon, S, d)$ is often called the ε-entropy.

Assume that we have i.i.d. observations from a density $p \in \mathcal{P}$, a space of densities. Let p_0 stand for the true density and consider the variation distance d_V on $\mathcal{P}$. Let Π be a prior on $\mathcal{P}$.

THEOREM 3. *Suppose that $p_0 \in \mathrm{KL}(\Pi)$. If given any $\varepsilon > 0$, there exist $\delta < \varepsilon/4$, $c_1, c_2 > 0$, $\beta < \varepsilon^2/8$ and $\mathcal{P}_n \subset \mathcal{P}$ such that $\Pi(\mathcal{P}_n^c) \leqslant c_1 \mathrm{e}^{-nc_2}$ and $\log N(\delta, \mathcal{P}_n, d_V) \leqslant n\beta$, then $\Pi(P\colon d_V(P, P_0) > \varepsilon | X_1, \ldots, X_n) \to 0$ a.s. $[P_0^\infty]$.*

Barron (1999) also noted that the testing condition in Schwartz's theorem is, in a sense, also necessary for posterior consistency to hold under Schwartz's condition on Kullback–Leibler support.

THEOREM 4. *Let $\mathcal{P}$ be a space of densities, $p_0 \in \mathcal{P}$ be the true density and P_0 be the probability measure corresponding to p_0. Let $p_0 \in \mathrm{KL}(\Pi)$. Then the following conditions are equivalent*:

(1) *There exists a β_0 such that $P_0\{\Pi(U^c | X_1, \ldots, X_n) > \mathrm{e}^{-n\beta_0}$ infinitely often$\} = 0$.*
(2) *There exist subsets $V_n, W_n \subset \mathcal{P}$, $c_1, c_2, \beta_1, \beta_2 > 0$ and a sequence of test functions $\Phi_n(X_1, \ldots, X_n)$ such that*
   (a) $U^c \subset V_n \cup W_n$,
   (b) $\Pi(W_n) \leqslant c_1 \mathrm{e}^{-nc_2}$,
   (c) $P_0\{\Phi_n > 0$ *infinitely often*$\} = 0$ *and* $\sup\{\mathrm{E}_p(1 - \Phi_n)\colon\ p \in V_n\} \leqslant c_2 \mathrm{e}^{-n\beta_2}$.

In a semiparametric problem, an additional Euclidean parameter is present apart from an infinite-dimensional parameter, and the Euclidean parameter is usually of interest. Diaconis and Freedman (1986a, 1986b) demonstrated that putting a prior that gives consistent posterior separately for the nonparametric part may not lead to a consistent posterior when the Euclidean parameter is incorporated in the model. The example described below appeared to be counter-intuitive when it first appeared.

EXAMPLE 2. Consider i.i.d. observations from the location model $X = \theta + \epsilon$, where $\theta \in \mathbb{R}$, $\epsilon \sim F$ which is symmetric. Put any nonsingular prior density on θ and the symmetrized Dirichlet process prior on F with a Cauchy center measure. Then there exists a symmetric distribution F_0 such that if the X observations come from F_0, then the posterior concentrates around two wrong values $\pm\gamma$ instead of the true value $\theta = 0$.

A similar phenomenon was observed by Doss (1985a, 1985b). The main problem in the above is that the posterior distribution for θ is close to the parametric posterior with a Cauchy density, and hence the posterior mode behaves like the M-estimator based on

hs25 v.2005/04/21 Prn:19/05/2005; 15:11 F:hs25013.tex; VTEX/VJ p. 17
aid: 25013 pii: S0169-7161(05)25013-7 docsubty: REV

the criterion function $m(x,\theta) = \log(1+(x-\theta)^2)$. The lack of concavity of m leads to undesirable solutions for some peculiar data generating distribution like F_0. Consistency however does obtain for the normal base measure since $m(x,\theta) = (x-\theta)^2$ is convex, or even for the Cauchy base measure if F_0 has a strongly unimodal density. Here, addition of the location parameter θ to the model destroys the delicate tail-free structure, and hence Freedman's (1963, 1965) consistency result for tail-free processes cannot be applied. Because the Dirichlet process selects only discrete distribution, it is also clear that Schwartz's (1965) condition on Kullback–Leibler support does not hold. However, as shown by Ghosal et al. (1999c), if we start with a prior on F that satisfies Schwartz's (1965) condition in the nonparametric model (that is, the case of known $\theta = 0$), then the same condition holds in the semiparametric model as well. This leads to weak consistency in the semiparametric model (without any additional testing condition) and hence consistency holds for the location parameter θ. The result extends to more general semiparametric problems. Therefore, unlike the tail-free property, Schwartz's condition on Kullback–Leibler support is very robust which is not altered by symmetrization, addition of a location parameter or formation of mixtures. Thus Schwartz's theorem is the right tool for studying consistency in semiparametric models.

Extensions of Schwartz's consistency theorem to independent, nonidentically distributed observations have been obtained by Amewou-Atisso et al. (2003) and Choudhuri et al. (2004a). The former does not use sieves and hence is useful only when weak topology is put on the infinite-dimensional part of the parameter. In semiparametric problems, this topology is usually sufficient to derive posterior consistency for the Euclidean part. However, for curve estimation problems, stronger topologies need to be considered and sieves are essential. Consistency in probability instead of that in the almost sure sense allows certain relaxations in the condition to be verified. Choudhuri et al. (2004a) considered such a formulation which is described below.

THEOREM 5. *Let $Z_{i,n}$ be independently distributed with density $p_{i,n}(\cdot;\theta)$, $i = 1,\dots,r_n$, with respect to a common σ-finite measure, where the parameter θ belongs to an abstract measurable space Θ. The densities $p_{i,n}(\cdot,\theta)$ are assumed to be jointly measurable. Let $\theta_0 \in \Theta$ and let $\overline{\Theta}_n$ and $\mathcal{U}_n$ be two subsets of Θ. Let θ have prior Π on Θ. Put $K_{i,n}(\theta_0,\theta) = \mathrm{E}_{\theta_0}(\Lambda_i(\theta_0,\theta))$ and $V_{i,n}(\theta_0,\theta) = \mathrm{var}_{\theta_0}(\Lambda_i(\theta_0,\theta))$, where $\Lambda_i(\theta_0,\theta) = \log\frac{p_{i,n}(Z_{i,n};\theta_0)}{p_{i,n}(Z_{i,n};\theta)}$.*

(A1) *Prior positivity of neighborhoods.*
Suppose that there exists a set B with $\Pi(B) > 0$ such that

(i) $\dfrac{1}{r_n^2}\displaystyle\sum_{i=1}^{r_n} V_{i,n}(\theta_0,\theta) \to 0$ *for all $\theta \in B$,*

(ii) $\displaystyle\liminf_{n\to\infty} \Pi\left(\left\{\theta \in B\colon \frac{1}{r_n}\sum_{i=1}^{r_n} K_{i,n}(\theta_0,\theta) < \varepsilon\right\}\right) > 0$ *for all $\varepsilon > 0$.*

(A2) *Existence of tests.*
Suppose that there exists test functions $\{\Phi_n\}$, $\Theta_n \subset \overline{\Theta}_n$ and constants $C_1, C_2, c_1, c_2 > 0$ such that

(i) $\mathrm{E}_{\theta_0}\Phi_n \to 0$,

hs25 v.2005/04/21 Prn:19/05/2005; 15:11 F:hs25013.tex; VTEX/VJ p. 18
aid: 25013 pii: S0169-7161(05)25013-7 docsubty: REV

(ii) $\sup_{\theta\in\mathcal{U}_n^c\cap\Theta_n} \mathrm{E}_\theta(1-\Phi_n) \leqslant C_1\,\mathrm{e}^{-c_1 r_n}$,

(iii) $\Pi(\overline{\Theta}_n \cap \Theta_n^c) \leqslant C_2\,\mathrm{e}^{-c_2 r_n}$.

Then $\Pi(\theta \in \mathcal{U}_n^c \cap \overline{\Theta}_n | Z_{1,n}, \ldots, Z_{r_n,n}) \to 0$ *in* $P_{\theta_0}^n$*-probability.*

Usually, the theorem will be applied to $\overline{\Theta}_n = \Theta$ for all n. If, however, condition (A2) could be verified only on a part of Θ which may possibly depend on n, the above formulation could be useful. However, the final conclusion should then be complemented by showing that $\Pi(\overline{\Theta}_n^c | Z_1, \ldots, Z_{r_n}) \to 0$ in $P_{\theta_0}^n$-probability by some alternative method.

The first condition (A1) asserts that certain sets, which could be thought of as neighborhoods of the true parameter θ_0, have positive prior probabilities. This condition ensures that the true value of the parameter is not excluded from the support of the prior. The second condition (A2) asserts that the hypothesis $\theta = \theta_0$ can be tested against the complement of a neighborhood for a topology of interest with a small probability of type I error and a uniformly exponentially small probability of type II error on most part of the parameter space in the sense that the prior probability of the remaining part is exponentially small.

The above theorem is also valid for a sequence of priors Π_n provided that (A1)(i) is strengthened to uniform convergence.

It should be remarked that Schwartz's condition on the Kullback–Leibler support is not necessary for posterior consistency to hold. This is clearly evident in parametric nonregular cases, where Kullback–Leibler divergence to some direction could be infinity. For instance, as in Ghosal et al. (1999a), for the model $p_\theta = \mathrm{Uniform}(0,\theta)$ density, $0 < \theta \leqslant 1$, the Kullback–Leibler numbers $\int p_1 \log(p_1/p_\theta) = \infty$. However, the posterior is consistent at $\theta = 1$ if the prior Π has 1 in its support. Modifying the model to uniform$(\theta-1, \theta+1)$, we see that the Kullback–Leibler numbers are infinite for every pair. Nevertheless, consistency for a general parametric family including such nonregular cases holds under continuity and positivity of the prior density at θ_0 provided that the general conditions of Ibragimov and Has'minskii (1981) can be verified; see Ghosal et al. (1995) for details. For infinite-dimensional models, consistency may hold without Schwartz's condition on Kullback–Leibler support by exploiting special structure of the posterior distribution as in the case of the Dirichlet or a tail-free process. For estimation of a survival distribution using a Lévy process prior, Kim and Lee (2001) concluded consistency from the explicit expressions for pointwise mean and variance and monotonicity. For densities, consistency may also be shown by using some alternative conditions. One approach is by using the so-called Le Cam's inequality: For any two disjoint subsets $U, V \subset \mathfrak{M}(\mathfrak{X})$, test function Φ, prior Π on $\mathfrak{M}(\mathfrak{X})$ and probability measure P_0 on $\mathfrak{X}$,

$$\int \Pi(V|x)\,\mathrm{d}P_0(x) \leqslant d_V(P_0, \lambda_U) + \int \Phi\,\mathrm{d}P_0 + \frac{\Pi(V)}{\Pi(U)} \int (1-\Phi)\,\mathrm{d}\lambda_V, \tag{3.6}$$

where $\lambda_U(B) = \int_U P(B)\,\mathrm{d}\Pi(P)/\Pi(U)$, the conditional expectation of $P(B)$ with respect to the prior Π restricted to the set U. Applying this inequality to V the complement of a neighborhood of P_0 and n i.i.d. observations, it may be shown that posterior

hs25 v.2005/04/21 Prn:19/05/2005; 15:11 F:hs25013.tex; VTEX/VJ p. 19
aid: 25013 pii: S0169-7161(05)25013-7 docsubty: REV

consistency in the weak sense holds provided that for any $\beta, \delta > 0$,

$$e^{n\beta}\Pi\big(P\colon d_V(P, P_0) < \delta/n\big) \to \infty. \tag{3.7}$$

Combining with appropriate testing conditions, stronger notions of consistency could be derived. The advantage of using this approach is that one need not control likelihood ratios now, and hence the result could be potentially used for undominated families as well, or at least can help reduce some positivity condition on the true density p_0. On the other hand, (3.7) is a quantitative condition on the prior unlike Schwartz's, and hence is more difficult to verify in many examples.

Because the testing condition is a condition only on a model and is more difficult to verify, there have been attempts to prove some assertion on posterior convergence using Schwartz's condition on Kullback–Leibler support only. While Theorem 4 shows that the testing condition is needed, it may be still possible to show some useful results by either weakening the concept of convergence, or even by changing the definition of the posterior distribution! Barron (1999) showed that if $p_0 \in \mathrm{KL}(\Pi)$, then

$$n^{-1}\sum_{i=1}^{n} \mathrm{E}_{p_0}\left(\log\frac{p_0(X_i)}{p(X_i|X_1,\ldots,X_{i-1})}\right) \to 0, \tag{3.8}$$

where $p(X_i|X_1,\ldots,X_{i-1})$ is the predictive density of X_i given $X_1,\ldots,X_{i-1}$. It may be noted that the predictive distribution is equal to the posterior mean of the density function. Hence in the Cesàro sense, the posterior mean density converges to the true density with respect to Kullback–Leibler neighborhoods, provided that the prior puts positive probabilities on Kullback–Leibler neighborhoods of p_0. Walker (2003), using a martingale representation of the predictive density, showed that the average predictive density converges to the true density almost surely under d_H. Walker and Hjort (2001) showed that the following pseudo-posterior distribution, defined by

$$\Pi_\alpha(p \in B|X_1,\ldots,X_n) = \frac{\int_B \prod_{i=1}^n p^\alpha(X_i)\,\mathrm{d}\Pi(p)}{\int_B \prod_{i=1}^n p^\alpha(X_i)\,\mathrm{d}\Pi(p)} \tag{3.9}$$

is consistent at any $p_0 \in \mathrm{KL}(\Pi)$, provided that $0 < \alpha < 1$.

Walker (2004) obtained another interesting result using an idea of restricting to a subset and looking at the predictive distribution (in this case, in the posterior) somewhat similar to that in Le Cam's inequality. If V is a set such that $\liminf_{n\to\infty} d_H(\lambda_{n,V}, p_0) > 0$, where $\lambda_{n,V}(B) = (\Pi(V|X_1,\ldots,X_n))^{-1}\int_V p(B)\,\mathrm{d}\Pi(p|X_1,\ldots,X_n)$, then $\Pi(V|X_1,\ldots,X_n) \to 0$ a.s. under P_0. A martingale property of the predictive distribution is utilized to prove the result. If V is the complement of a suitable weak neighborhood of p_0, then $\liminf_{n\to\infty} d_H(\lambda_{n,V}, p_0) > 0$, and hence the result provides an alternative way of proving the weak consistency result without appealing to Schwartz's theorem. Walker (2004) also considered other topologies.

The following is another result of Walker (2004) proving sufficient conditions for posterior consistency in terms of a suitable countable covering.

THEOREM 6. *Let $p_0 \in \mathrm{KL}(\Pi)$ and $V = \{p\colon d_H(p, p_0) > \varepsilon\}$. Let there exists $0 < \delta < \varepsilon$ and $V_1, V_2, \ldots$ a countable disjoint cover of V such that $d_H(p_1, p_2) < 2\delta$*

hs25 v.2005/04/21 Prn:19/05/2005; 15:11 F:hs25013.tex; VTEX/VJ p. 20
aid: 25013 pii: S0169-7161(05)25013-7 docsubty: REV

for all $p_1, p_2 \in V_j$ *and for all* $j = 1, 2, \ldots$, *and* $\sum_{j=1}^{\infty} \sqrt{\Pi(V_j)} < \infty$. *Then* $\Pi(V|X_1, \ldots, X_n) \to 0$ *a.s.* $[p_0^{\infty}]$.

While the lack of consistency is clearly undesirable, consistency itself is a very weak requirement. Given a consistency result, one would like to obtain information on the rates of convergence of the posterior distribution and see whether the obtained rate matches with the known optimal rate for point estimators. In finite-dimensional problems, it is well known that the posterior converges at a rate of $n^{-1/2}$ in the Hellinger distance; see Ibragimov and Has'minskii (1981) and Le Cam (1986).

Conditions for the rate of convergence given by Ghosal et al. (2000) and described below are quantitative refinement of conditions for consistency. A similar result, but under a much stronger condition on bracketing entropy numbers, was given by Shen and Wasserman (2001).

THEOREM 7. *Let* $\varepsilon_n \to 0$, $n\varepsilon_n^2 \to \infty$ *and suppose that there exist* $\mathcal{P}_n \subset \mathcal{P}$, *constants* $c_1, c_2, c_3, c_4 > 0$ *such that*

(i) $\log D(\varepsilon_n, \mathcal{P}_n, d) \leqslant c_1 n\varepsilon_n^2$, *where* D *stands for the packing number*;
(ii) $\Pi(\mathcal{P} \setminus \mathcal{P}_n) \leqslant c_2 \,\mathrm{e}^{-(c_3+4)n\varepsilon_n^2}$;
(iii) $\Pi(p\colon \int p_0 \log \frac{p_0}{p} < \varepsilon_n^2, \int p_0 \log^2 \frac{p_0}{p} < \varepsilon_n^2) \geqslant c_4 \,\mathrm{e}^{-c_3 n\varepsilon_n^2}$.

Then for some M, $\Pi(d(p, p_0) > M\varepsilon_n | X_1, X_2, \ldots, X_n) \to 0$.

More generally, the entropy condition can be replaced by a testing condition, though, in most applications, a test is constructed from entropy bounds. Some variations of the theorem are given by Ghosal et al. (2000), Ghosal and van der Vaart (2001) and Belitser and Ghosal (2003).

While the theorems of Ghosal et al. (2000) satisfactorily cover i.i.d. data, major extensions are needed to cover some familiar situations such as regression with a fixed design, dose response study, generalized linear models with an unknown link, Whittle estimation of a spectral density and so on. Ghosal and van der Vaart (2003a) considered the issue and showed that the basic ideas of the i.i.d. case work with suitable modifications. Let d_n^2 be the average squared Hellinger distance defined by $d_n^2(\theta_1, \theta_2) = n^{-1} \sum_{i=1}^{n} d_H^2(p_{i,\theta_1}, p_{i,\theta_2})$. Birgé (1983) showed that a test for θ_0 against $\{\theta\colon d_n(\theta, \theta_1) < d_n(\theta_0, \theta_1)/18\}$ with error probabilities at most $\exp(-nd_n^2(\theta_0, \theta_1)/2)$ may be constructed. To find the intended test for θ_0 against $\{\theta\colon d_n(\theta, \theta_0) > \varepsilon\}$, one therefore needs to cover the alternative by d_n balls of radius $\varepsilon/18$. The number of such balls is controlled by the d_n-entropy numbers. Prior concentration near θ_0 controls the denominator as in the case of i.i.d. observations. Using these ideas, Ghosal and van der Vaart (2003a) obtained the following theorem on convergence rates that is applicable to independent, nonidentically distributed observations, and applied the result to various non-i.i.d. models.

THEOREM 8. *Suppose that for a sequence* $\varepsilon_n \to 0$ *such that* $n\varepsilon_n^2$ *is bounded away from zero, some* $k > 1$, *every sufficiently large* j *and sets* $\Theta_n \subset \Theta$, *the following conditions*

hs25 v.2005/04/21 Prn:19/05/2005; 15:11 F:hs25013.tex; VTEX/VJ p. 21
aid: 25013 pii: S0169-7161(05)25013-7 docsubty: REV

are satisfied:

$$\sup_{\varepsilon>\varepsilon_n} \log N\bigl(\varepsilon/36, \{\theta \in \Theta_n\colon d_n(\theta,\theta_0) < \varepsilon\}, d_n\bigr) \leqslant n\varepsilon_n^2, \tag{3.10}$$

$$\Pi_n(\Theta \setminus \Theta_n)/\Pi_n\bigl(B_n^*(\theta_0,\varepsilon_n;k)\bigr) = \mathrm{o}\bigl(\mathrm{e}^{-2n\varepsilon_n^2}\bigr), \tag{3.11}$$

$$\frac{\Pi_n(\theta \in \Theta_n\colon j\varepsilon_n < d_n(\theta,\theta_0) \leqslant 2j\varepsilon_n)}{\Pi_n(B_n^*(\theta_0,\varepsilon_n;k))} \leqslant \mathrm{e}^{n\varepsilon_n^2 j^2/4}. \tag{3.12}$$

Then $P_{\theta_0}^{(n)}\Pi_n(\theta\colon d_n(\theta,\theta_0) \geqslant M_n\varepsilon_n|X^{(n)}) \to 0$ *for every* $M_n \to \infty$.

Ghosal and van der Vaart (2003a) also considered some dependent cases such as Markov chains, autoregressive model and signal estimation in presence of Gaussian white noise.

When one addresses the issue of optimal rate of convergence, one considers a smoothness class of the involved functions. The method of construction of the optimal prior with the help of bracketing or spline functions, as in Ghosal et al. (2000) requires the knowledge of the smoothness index. In practice, such information is not available and it is desirable to construct a prior that is adaptive. In other words, we wish to construct a prior that simultaneously achieves the optimal rate for every possible smoothness class under consideration. If only countably many models are involved, a natural and elegant method would be to consider a prior that is a mixture of the optimal priors for different smoothness classes. Belitser and Ghosal (2003) showed that the strategy works for an infinite-dimensional normal. Ghosal et al. (2003) and Huang (2004) obtained similar results for the density estimation problem.

Kleijn and van der Vaart (2002) considered the issue of misspecification, where p_0 may not lie in the support of the prior. In such a case, consistency at p_0 cannot hold, but it is widely believed that the posterior concentrates around the Kullback–Leibler projection p^* of p_0 to the model; see Berk (1966) for some results for parametric exponential families. Under suitable conditions which could be regarded as generalizations of the conditions of Theorem 7, Kleijn and van der Vaart (2002) showed that the posterior concentrates around p^* at a rate described by a certain entropy condition and concentration rate of the prior around p^*. Kleijn and van der Vaart (2002) also defined a notion of covering number for testing under misspecification that turns out to be the appropriate way of measuring the size of the model in the misspecified case. A weighted version of the Hellinger distance happens to be the proper way of measuring distance between densities that leads to a fruitful theorem on rates in the misspecified case. A useful theorem on consistency (in the sense that the posterior distribution concentrates around p^*) follows as a corollary.

When the posterior distribution converges at a certain rate, it is also important to know whether the posterior measure, after possibly a random centering and scaling, converges to a nondegenerate measure. For smooth parametric families, convergence to a normal distribution holds and is popularly known as the Bernstein–von Mises theorem; see Le Cam and Yang (2000) and van der Vaart (1998) for details. For a general parametric family which need not be smooth, a necessary and sufficient condition in terms of the

hs25 v.2005/04/21 Prn:19/05/2005; 15:11 F:hs25013.tex; VTEX/VJ p. 22
aid: 25013 pii: S0169-7161(05)25013-7 docsubty: REV

limiting likelihood ratio process for convergence of the posterior (to some nondegenerate distribution using some random centering) is given by Ghosh et al. (1994, 1995). For infinite-dimensional cases, results are relatively rare. Some partial results were obtained by Lo (1983, 1986) for Dirichlet process, Shen (2002) for certain semiparametric models, Susarla and Van Ryzin (1978) and Kim and Lee (2004) for certain survival models respectively with the Dirichlet process and Lévy process priors. However, it appears from the work of Cox (1993) and Freedman (1999) that Bernstein–von Mises theorem does not hold in most cases when the convergence rate is slower than $n^{-1/2}$. Freedman (1999) indeed showed that for the relatively simple problem of the estimation of the mean of an infinite-dimensional normal distribution with independent normal priors, the frequentist and the Bayesian distribution of L_2-norm of the difference of the Bayes estimate and the parameter differ by an amount equal to the scale of interest, and the frequentist coverage probability of a Bayesian credible set for the parameter is asymptotically zero. However, see Ghosal (2000) for a partially positive result.

4. Estimation of cumulative probability distribution

4.1. Dirichlet process prior

One of the nicest properties of the Dirichlet distribution, making it hugely popular, is its conjugacy for estimating a distribution function (equivalently, the probability law) with i.i.d. observations. Consider $X_1, \ldots, X_n$ are i.i.d. samples from an unknown cumulative distribution function (cdf) F on $\mathbb{R}^d$. Suppose F is given a Dirichlet process prior with parameters (M, G). Then the posterior distribution is again a Dirichlet process with the two parameters updated as

$$M \mapsto M + n \quad \text{and} \quad G \mapsto (MG + n\mathbb{F}_n)/(M + n), \tag{4.1}$$

where $\mathbb{F}_n$ is the empirical cdf. This may be easily shown by reducing the data to counts of sets from a partition, using the conjugacy of the finite-dimensional Dirichlet distribution for the multinomial distribution and passing to the limit with the aid of the martingale convergence theorem. Combining with (2.1), this implies that the posterior expectation and variance of $F(x)$ are given by

$$\begin{aligned} \widetilde{\mathbb{F}}_n(x) &= \mathrm{E}\big(F(x)|X_1, \ldots, X_n\big) = \frac{M}{M+n}G(x) + \frac{n}{M+n}\mathbb{F}_n(x), \\ \operatorname{var}\big(F(x)|X_1, \ldots, X_n\big) &= \frac{\widetilde{\mathbb{F}}_n(x)(1 - \widetilde{\mathbb{F}}_n(x))}{1 + M + n}. \end{aligned} \tag{4.2}$$

Therefore the posterior mean is a convex combination of the prior mean and the empirical cdf. As the sample size increases, the behavior of the posterior mean is inherited from that of the empirical probability measure. Also M could be interpreted as the strength in the prior or the "prior sample size".

The above discussion may lull us to interpret the limiting case $M \to 0$ as noninformative. Indeed, Rubin (1981) proposed $\mathrm{Dir}(n, \mathbb{F}_n)$ as the Bayesian bootstrap, which corresponds to the posterior obtained from the Dirichlet process by letting $M \to 0$.

hs25 v.2005/04/21 Prn:19/05/2005; 15:11 F:hs25013.tex; VTEX/VJ p. 23
aid: 25013 pii: S0169-7161(05)25013-7 docsubty: REV

However, some caution is needed while interpreting the case $M \to 0$ as noninformative because of the role of M in also controlling the number of ties among samples drawn from P, where P itself is drawn from the Dirichlet process. Sethuraman and Tiwari (1982) pointed out that as $M \to 0$, the Dirichlet process converges weakly to the random measure which is degenerate at some point θ distributed as G by property (ii) of convergence of Dirichlet measures mentioned in Section 2.1. Such a prior is clearly "very informative", and hence is unsuitable as a noninformative prior.

To obtain posterior consistency, note that (4.1) converges a.s. to the true cdf generating data. An important consequence of the above assertions is that the posterior distribution based on the Dirichlet process, not just the posterior mean, is consistent for the weak topology. Thus, by the weak convergence property of Dirichlet process, the posterior is consistent with respect to the weak topology. It can also be shown that, the posterior is consistent in the Kolmogorov–Smirnov distance defined as $d_{\mathrm{KS}}(F_1, F_2) = \sup_x |F_1(x) - F_2(x)|$. The space of cdf's under d_{KS} is however neither separable nor complete.

If the posterior distribution of F is given a prior that is a mixture of Dirichlet process, the posterior distribution is still a mixture of Dirichlet processes; see Theorem 3 of Antoniak (1974). However, mixtures may lead to inconsistent posterior distribution, unlike a single Dirichlet process. Nevertheless, if M_θ is bounded in θ, then posterior consistency holds.

4.2. Tail-free and Polya tree priors

Tail-free priors are extremely flexible, yet have some interesting properties. If the distribution function generating the i.i.d. data is given a tail-free prior, the posterior distribution is also tail-free. Further, as mentioned in Section 3, Freedman (1963, 1965) showed that the posterior obtained from a tail-free process prior is weakly consistent. The tail-free property helps reduce a weak neighborhood to a neighborhood involving only finitely many variables in the hierarchical representation, and hence the problem reduces to a finite-dimensional multinomial distribution, where consistency holds. Indeed Freedman's original motivation was to avoid pitfall as in Example 1.

A Polya tree prior may be used if one desires some smoothness of the random cdf. The most interesting property of a Polya tree process is its conjugacy. Conditional on the data $X_1, \ldots, X_n$, the posterior distribution is again a Polya tree with respect to the same partition and α_ε updated to $\alpha_\varepsilon^* = \alpha_\varepsilon + \sum_{i=1}^n I\{X_i \in B_\varepsilon\}$. Besides, they lead to a consistent posterior in the weak topology as Polya trees are also tail-free processes.

4.3. Right censored data

Let X be a random variable of interest that is right censored by another random variable Y. The observation is (Z, Δ), where $Z = \min(X, Y)$ and $\Delta = I(X > Y)$. Assume that X and Y are independent with corresponding cdf F and H, where both F and H are unknown. The problem is to estimate F. Susarla and Van Ryzin (1976) put a Dirichlet process prior on F. Blum and Susarla (1977) found that the posterior distribution for i.i.d. data can be written as a mixture of Dirichlet processes. Using this idea, Susarla and Van Ryzin (1978) obtained that the posterior is mean square consistent with rate

hs25 v.2005/04/21 Prn:19/05/2005; 15:11 F:hs25013.tex; VTEX/VJ p. 24
aid: 25013 pii: S0169-7161(05)25013-7 docsubty: REV

$O(n^{-1})$, almost surely consistent with rate $O(\log n/n^{1/2})$, and that the posterior distribution of $\{F(u)\colon 0 < u < T\}$, $T < \infty$, converges weakly to a Gaussian process whenever F and H are continuous and that $P(X > u)P(Y > u) > 0$. The mixture representation is however cumbersome. Ghosh and Ramamoorthi (1995) showed that the posterior distribution can also be written as a Polya tree process (with partitions dependent on the uncensored samples). They proved consistency by an elegant argument.

Doksum (1974) found that the neutral to right process for F form a conjugate family for the right censored data. Viewed as a prior on the cumulative hazard process, the prior can be identified with an independent increment process. An updating mechanism is described by Kim (1999) using a counting process approach. Beta processes, introduced by Hjort (1990), also form a conjugate family. Kim and Lee (2001) obtained sufficient conditions for posterior consistency for a Lévy processes prior, which includes Dirichlet processes and beta processes. Under certain conditions, the posterior also converges at the usual $n^{-1/2}$ rate and admits a Bernstein–von Mises theorem; see Kim and Lee (2004).

5. Density estimation

Density estimation is one of the fundamental problems of nonparametric inference because of its applicability to various problems including cluster analysis and robust estimation. A common approach to constructing priors on the space of probability densities is to use Dirichlet mixtures where the kernels are chosen depending on the sample space. The posterior distributions are analytically intractable and the MCMC techniques are different for different kernels. Other priors useful for this problem are Polya tree processes and Gaussian processes. In this section, we discuss some of the computational issues and conditions for consistency and convergence rates of the posterior distribution.

5.1. Dirichlet mixture

Consider that the density generating the data is a mixture of densities belonging to some parametric family, that is, $p_F(x) = \int \psi(x,\theta)\,\mathrm{d}F(\theta)$. Let the mixing distribution F be given a $\mathrm{Dir}(M,G)$ prior. Viewing $p_F(x)$ as a linear functional of F, the prior expectation of $p_F(x)$ is easily found to be $\int \psi(x,\theta)\,\mathrm{d}G(\theta)$. To compute the posterior expectation, the following hierarchical representation of the above prior is often convenient:

$$X_i \overset{\text{ind}}{\sim} \psi(\cdot,\theta_i), \qquad \theta_i \overset{\text{i.i.d.}}{\sim} F, \qquad F \sim \mathrm{Dir}(M,G). \tag{5.1}$$

Let $\Pi(\boldsymbol{\theta}|X_1,\ldots,X_n)$ stand for the distribution of $(\theta_1,\ldots,\theta_n)$ given $(X_1,\ldots,X_n)$. Observe that given $\boldsymbol{\theta} = (\theta_1,\ldots,\theta_n)$, the posterior distribution of F is Dirichlet with base measure $MG + n\mathbb{G}_n$, where $\mathbb{G}_n(\cdot,\boldsymbol{\theta}) = n^{-1}\sum_{i=1}^n \delta_{\theta_i}$, the empirical distribution of $(\theta_1,\ldots,\theta_n)$. Hence the posterior distribution of F may be written as a mixture of Dirichlet processes. The posterior mean of $F(\cdot)$ may be written as

$$\frac{M}{M+n}G(\cdot) + \frac{n}{M+n}\int G_n(\cdot,\boldsymbol{\theta})\Pi(\mathrm{d}\boldsymbol{\theta}|X_1,\ldots,X_n) \tag{5.2}$$

hs25 v.2005/04/21 Prn:19/05/2005; 15:11 F:hs25013.tex; VTEX/VJ p. 25
aid: 25013 pii: S0169-7161(05)25013-7 docsubty: REV

and the posterior mean of the density at x becomes

$$\frac{M}{M+n}\int \psi(x,\theta)\,\mathrm{d}G(\theta) + \frac{n}{M+n}\frac{1}{n}\sum_{i=1}^{n}\int \psi(x,\theta_i)\Pi(\mathrm{d}\boldsymbol{\theta}|X_1,\ldots,X_n). \tag{5.3}$$

The Bayes estimate is thus composed of a part attributable to the prior and a part due to observations. Ferguson (1983) remarks that the factor $n^{-1}\sum_{i=1}^{n}\int \psi(x,\theta_i)\Pi(\mathrm{d}\boldsymbol{\theta}|X_1,\ldots,X_n)$ in the second term of (5.3) can be viewed as a partially Bayesian estimate with the influence of the prior guess reduced. The evaluation of the above quantities depend on $\Pi(\mathrm{d}\boldsymbol{\theta}|X_1,\ldots,X_n)$. The joint prior for $(\theta_1,\theta_2,\ldots,\theta_n)$ is given by the generalized Polya urn scheme

$$G(\mathrm{d}\theta_1)\times\frac{(MG(\mathrm{d}\theta_2)+\delta_{\theta_1})}{M+1}\times\cdots\times\frac{\left(MG(\mathrm{d}\theta_n)+\sum_{i=1}^{n-1}\delta_{\theta_i}\right)}{M+n}. \tag{5.4}$$

Further, the likelihood given $(\theta_1,\theta_2,\ldots,\theta_n)$ is $\prod_{i=1}^{n}\psi(X_i,\theta_i)$. Hence H can be written down using the Bayes formula. Using the above equations and some algebra, Lo (1984) obtained analytical expressions of the posterior expectation of $f(x)$. However, the formula is of marginal use because the number of terms grows very fast with the sample size. Computations are thus done via MCMC techniques as in the special case of normal mixtures described in the next subsection; see the review article Escobar and West (1998) for details.

5.1.1. Mixture of normal kernels

Suppose that the unknown density of interest is supported on the entire real line. Then a natural choice of the kernel is $\phi_\sigma(x-\mu)$, the normal density with mean μ and variance σ^2. The mixture distribution F is given Dirichlet process prior with some base measure MG, while G is often given a normal/inverse-gamma distribution to achieve conjugacy. Thus, under G, $\sigma^{-2}\sim\text{Gamma}(s,\beta)$, a gamma distribution with shape parameter s and scale parameter β, and $(\mu|\sigma)\sim N(m,\sigma^2)$. Let $\theta=(\mu,\sigma)$. Then the hierarchical model is

$$X_i|\theta_i \overset{\text{ind}}{\sim} N\left(\mu_i,\sigma_i^2\right),\qquad \theta_i \overset{\text{i.i.d.}}{\sim} F,\qquad F\sim\text{Dir}(M,G). \tag{5.5}$$

Given $\boldsymbol{\theta}=(\theta_1,\ldots,\theta_n)$, the distribution of F may be updated analytically. Thus, if one can sample from the posterior distribution of $\boldsymbol{\theta}$, Monte Carlo averages may be used to find the posterior expectation of F and thus the posterior expectation of $p(x)=\int\phi_\sigma(x-\mu)\,\mathrm{d}F(x)$. Escobar (1994) and Escobar and West (1995) provided an algorithm for sampling from the posterior distribution of $\boldsymbol{\theta}$. Let $\theta_{-i}=\{\theta_1,\ldots,\theta_{i-1},\theta_{i+1},\ldots,\theta_n\}$. Then

$$(\theta_i|\theta_{-i},x_1,\ldots,x_n)\sim q_{i0}G_i(\theta_i)+\sum_{j=1,j\neq i}^{n}q_{ij}\delta_{\theta_j}(\theta_i), \tag{5.6}$$

where $G_i(\theta_i)$ is the bivariate normal/inverse-gamma distribution under which

$$\begin{aligned}\sigma_i^{-2}&\sim\text{Gamma}\left(s+1/2,\beta+(x_i-m)^2/2\right),\\ (\mu_i|\sigma_i)&\sim N\left(m+x_i,\sigma_i^2\right)\end{aligned} \tag{5.7}$$

hs25 v.2005/04/21 Prn:19/05/2005; 15:11 F:hs25013.tex; VTEX/VJ p. 26
aid: 25013 pii: S0169-7161(05)25013-7 docsubty: REV

and the weights q_{ij}'s are defined by $q_{i0} \propto M\Gamma(s+1/2)(2\beta)^s\Gamma(s)^{-1}\{2\beta+(x_i-m)^2\}^{-(s+1/2)}$ and $q_{ij} \propto \sqrt{\pi}\phi_{\sigma_i}(x_i-\mu_i)$ for $j \neq i$. Thus a Gibbs sampler algorithm is described by updating $\boldsymbol{\theta}$ componentwise through the conditional distribution in (5.6). The initial values of θ_i could be a sample from G_i.

The bandwidth parameter σ is often kept constant depending on the sample size, say σ_n. This leads to only the location mixture. In that case a Gibbs sampler algorithm is obtained by keeping σ_i fixed at σ_n in the earlier algorithm and updating only the location components μ_i.

Consistency of the posterior distribution for Dirichlet mixture of normals was studied by Ghosal et al. (1999b). Let p_0 stand for the true density.

THEOREM 9. *If $p_0 = \int \phi_\sigma(x-\mu)\,\mathrm{d}F_0(\mu,\sigma)$, where F_0 is compactly supported and in the weak support of Π, then $p_0 \in \mathrm{KL}(\Pi)$.*

If p_0 is not a mixture of normals but is compactly supported, 0 is in the support of the prior for σ, and $\lim_{\sigma\to 0}\int p_0\log(p_0/p_0*\phi_\sigma)=0$, then $p_0 \in \mathrm{KL}(\Pi)$.

If $p_0 \in \mathrm{KL}(\Pi)$, the base measure G of the underlying Dirichlet process is compactly supported and $\Pi(\sigma < t) \leqslant c_1\,\mathrm{e}^{-c_2/t}$, then the posterior is consistent at p_0 for the total variation distance d_V. If the compact support G is replaced by the condition that for every $\varepsilon > 0$, there exist a_n, σ_n with $a_n/\sigma_n < \varepsilon n$ satisfying $G[-a_n,a_n] < \mathrm{e}^{-n\beta_1}$ and $\Pi(\sigma < \sigma_n) \leqslant \mathrm{e}^{-n\beta_2}$ for $\beta_1,\beta_2 > 0$, then also consistency for d_V holds at any $p_0 \in \mathrm{KL}(\Pi)$.

The condition $p_0 \in \mathrm{KL}(\Pi)$ implies weak consistency by Schwartz's theorem. The condition for $p_0 \in \mathrm{KL}(\Pi)$ when p_0 is neither a normal mixture nor compactly supported, as given by Theorem 5 of Ghosal et al. (1999b) using estimates of Dirichlet tails, is complicated. However, the conditions holds under strong integrability conditions on p_0. The base measure for the Dirichlet could be normal and the prior on σ could be a truncated inverse gamma possibly involving additional parameters. Better sufficient condition for $p_0 \in \mathrm{KL}(\Pi)$ is given by Tokdar (2003). Consider a location-scale mixture of normal with a prior Π on the mixing measure. If p_0 is bounded, nowhere zero, $\int p_0|\log p_0| < \infty$, $\int p_0\log(p_0/\psi) < \infty$ where $\psi(x) = \inf\{p_0(t)\colon x-1 \leqslant t \leqslant x+1\}$, $\int |x|^{2+\delta}p_0(x)\,\mathrm{d}x < \infty$, and every compactly supported probability lies in $\mathrm{supp}(\Pi)$, then $p_0 \in \mathrm{KL}(\Pi)$. The moment condition can be weakened to only δ-moment if Π is Dirichlet. In particular, the case that p_0 is Cauchy could be covered.

Convergence rates of the posterior distribution were obtained by Ghosal and van der Vaart (2001, 2003b) respectively the "super smooth" and the "smooth" cases. We discuss below the case of location mixtures only, where the scale gets a separate independent prior.

THEOREM 10. *Assume that $p_0 = \phi_{\sigma_0} * F_0$, and the prior on σ has a density that is compactly supported in $(0,\infty)$ but is positive and continuous at σ_0. Suppose that F_0 has compact support and the base measure G has a continuous and positive density on an interval containing the support of F_0 and has tails $G(|z| > t) \lesssim \mathrm{e}^{-b|t|^\delta}$. Then the posterior converges at a rate $n^{-1/2}(\log n)^{\max(\frac{2}{\delta},\frac{1}{2})+\frac{1}{2}}$ with respect to d_H. The condition*

hs25 v.2005/04/21 Prn:19/05/2005; 15:11 F:hs25013.tex; VTEX/VJ p. 27
aid: 25013 pii: S0169-7161(05)25013-7 docsubty: REV

of compact support of F_0 could be replaced by that of sub-Gaussian tails if G is normal, in which case the rate is $n^{-1/2}(\log n)^{3/2}$.

If instead p_0 is compactly supported, twice continuously differentiable and $\int (p_0''/p_0)^2 p_0 < \infty$ and $\int (p_0'/p_0)^4 p_0 < \infty$, and the prior on (σ/σ_n) has a density that is compactly supported in $(0,\infty)$, where $\sigma_n \to 0$, then the posterior converges at a rate $\max((n\sigma_n)^{-1/2}(\log n), \sigma_n^2 \log n)$. In particular, the best rate $\varepsilon_n \sim n^{-2/5}(\log n)^{-4/5}$ is obtained by choosing $\sigma_n \sim n^{-1/5}(\log n)^{-2/5}$.

The proofs are the result of some delicate estimates of the number of components a discrete mixing distribution must have to approximate a general normal mixture. Some further results are given by Ghosal and van der Vaart (2003b) when p_0 does not have compact support.

5.1.2. Uniform scale mixtures

A nonincreasing density on $[0,\infty)$ may be written as a mixture of the form $\int \theta^{-1} I\{0 \leqslant x \leqslant \theta\} F(\mathrm{d}\theta)$ by a well known representation theorem of Khinchine and Shepp. This lets us put a prior on this class from that on F. Brunner and Lo (1989) considered this idea and put a Dirichlet prior for F. Coupled with a symmetrization technique as in Section 2.2.3, this leads to a reasonable prior for the error distribution. Brunner and Lo (1989) used this approach for the semiparametric location problem. The case of asymmetric error was treated by Brunner (1992) and that of semiparametric linear regression by Brunner (1995).

5.1.3. Mixtures on the half line

Dirichlet mixtures of exponential distributions may be considered as a reasonable model for a decreasing, convex density on the positive half line. More generally, mixtures of gamma densities, which may be motivated by Feller approximation procedure using a Poisson sampling scheme in the sense of Petrone and Veronese (2002), may be considered to pick up arbitrary shapes. Such a prior may be chosen to have a large weak support. Mixtures of inverse gammas may be motivated similarly by Feller approximation using a gamma sampling scheme. In general, a canonical choice of a kernel function could be made once a Feller sampling scheme appropriate for the domain could be specified. For a general kernel, weak consistency may be shown exploiting Feller approximation property as in Petrone and Veronese (2002).

Mixtures of Weibulls or lognormals are dense in the stronger sense of total variation distance provided that we let the shape parameter of the Weibull to approach infinity or that of the lognormal to approach zero. To see this, observe that these two kernels form location-scale families in the log-scale, and hence are approximate identities. Kottas and Gelfand (2001) used these mixtures for median regression, where asymmetry is an important aspect. The mixture of Weibull is very useful to model observations of censored data because its survival function has a simpler expression compared to that for the mixtures of gamma or lognormal. Ghosh and Ghosal (2003) used these mixtures to model a proportional mean structure censored data. The posterior distribution was computed using an MCMC algorithm for Dirichlet mixtures coupled with imputation of censored

hs25 v.2005/04/21 Prn:19/05/2005; 15:11 F:hs25013.tex; VTEX/VJ p. 28
aid: 25013 pii: S0169-7161(05)25013-7 docsubty: REV

data. Posterior consistency can be established by reducing the original model to a standard regression model with unknown error for which the results of Amewou-Atisso et al. (2003) apply. More specifically, consistency holds if the true baseline density is in the Kullback–Leibler support of the Dirichlet mixture prior. The last condition can be established under reasonable conditions using the ideas of Theorem 9 and its extension by Tokdar (2003).

5.1.4. *Bernstein polynomials*

On the unit interval, the family of beta distributions form a flexible two-parameter family of densities and their mixtures form a very rich class. Indeed, mixtures of beta densities with integer parameters are sufficient to approximate any distribution. For a continuous probability distribution function F on $(0, 1]$, the associated Bernstein polynomial $B(x; k, F) = \sum_{j=0}^{k} F(j/k)\binom{k}{j}x^j(1-x)^{k-j}$, which is a mixture of beta distributions, converges uniformly to F as $k \to \infty$. Using an idea of Diaconis that this approximation property may be exploited to construct priors with full topological support, Petrone (1999a, 1999b) proposed the following hierarchical prior called the Bernstein polynomial prior:

- $f(x) = \sum_{j=1}^{k} w_{j,k}\beta(x; j, k-j+1)$,
- $k \sim \rho(\cdot)$,
- $(\boldsymbol{w}_k = (w_{1,k}, \ldots, w_{k,k})|k) \sim H_k(\cdot)$, a distribution on the k-dimensional simplex.

Petrone (1999a) showed that if for all k, $\rho(k) > 0$ and $\boldsymbol{w}_k$ has full support on Δ_k, then every distribution on $(0, 1]$ is in the weak support of the Bernstein polynomial prior, and every continuous distribution is in the topological support of the prior defined by the Kolmogorov–Smirnov distance.

The posterior mean, given k, is

$$E\big(f(x)|k, x_1, \ldots, x_n\big) = \sum_{j=1}^{k} E(w_{j,k}|x_1, \ldots, x_n)\beta(x; j, k-j+1), \quad (5.8)$$

and the distribution of k is updated to $\rho(k|x_1, \ldots, x_n)$. Petrone (1999a, 1999b) discussed MCMC algorithms to compute the posterior expectations and carried out extensive simulations to show that the resulting density estimates work well.

Consistency is given by Petrone and Wasserman (2002). The corresponding results on convergence rates are obtained by Ghosal (2001).

THEOREM 11. *If p_0 is continuous density on $[0, 1]$, the base measure G has support all of $[0, 1]$ and the prior probability mass function $\rho(k)$ for k has infinite support, then $p_0 \in \mathrm{KL}(\Pi)$. If further $\rho(k) \lesssim \mathrm{e}^{-\beta k}$, then the posterior is consistent for d_H.*

If p_0 is itself a Bernstein polynomial, then the posterior converges at the rate $n^{-1/2}\log n$ with respect to d_H.

If p_0 is twice continuously differentiable on $[0, 1]$ and bounded away from zero, then the posterior converges at the rate $n^{-1/3}(\log n)^{5/6}$ with respect to d_H.

hs25 v.2005/04/21 Prn:19/05/2005; 15:11 F:hs25013.tex; VTEX/VJ p. 29
aid: 25013 pii: S0169-7161(05)25013-7 docsubty: REV

5.1.5. Random histograms

Gasparini (1996) used the Dirichlet process to put a prior on histograms of different bin width. The sample space is first partitioned into (possibly an infinite number of) intervals of length h, where h is chosen from a prior. Mass is distributed to the intervals according to a Dirichlet process, whose parameters $M = M_h$ and $G = G_h$ may depend on h. Mass assigned to any interval is equally distributed over that interval. The method corresponds to Dirichlet mixtures with a uniform kernel $\psi(x, \theta, h) = h^{-1}$, $x, \theta \in (jh, (j+1)h)$ for some j.

If $n_j(h)$ is the number of X_i's in the bin $[jh, (j+1)h)$, it is not hard to see that the posterior is of the same form as the prior with $M_h G_h$ updated to $M_h G_h + \sum_j n_j(h) I[jh, (j+1)h)$ and the prior density $\pi(h)$ of h changed to

$$\pi^*(h) = \frac{\pi(h) \prod_{j=1}^{\infty} (M_h G_h([jh, (j+1)h)))^{(n_j(h)-1)}}{M_h + n}. \tag{5.9}$$

The predictive density with no observations is given by $\int f_h(x)\pi(h)\,\mathrm{d}h$, where $f_h(x) = h^{-1} \sum_{j=-\infty}^{\infty} G_h([jh, (j+1)h)) I_{[jh,(j+1)h]}(x)$. In view of the conjugacy property, the predictive density given n observations can be easily written down. Let P_h stand for the histogram of bin-width h obtained from the probability measure P. Assume that $G_h(j)/G_h(j-1) \leqslant K_h$. If $\int x^2 p_0(x)\,\mathrm{d}x < \infty$ and $\lim_{h\to 0} \int p_0(x) \log \frac{p_{0,h}}{p_0} = 0$, then the posterior is weakly consistent at p_0. Gasparini (1996) also gave additional conditions to ensure consistency of the posterior mean of p under d_H.

5.2. Gaussian process prior

For density estimation on a bounded interval I, Leonard (1978) defined a random density on I through $f(x) = \frac{\mathrm{e}^{Z(x)}}{\int_I \mathrm{e}^{Z(t)}\,\mathrm{d}t}$, where $Z(x)$ is a Gaussian process with mean function $\mu(x)$ and covariance kernel $\sigma(x, x')$. Lenk (1988) introduces an additional parameter ξ to obtain a conjugate family. It is convenient to introduce the intermediate lognormal process $W(x) = \mathrm{e}^{Z(x)}$. Denote the distribution of W by $LN(\mu, \sigma, 0)$. For each ξ define a positive valued random process $LN(\mu, \sigma, \xi)$ on I whose Radon–Nikodym derivative with respect to $LN(\mu, \sigma, 0)$ is $(\int_I W(x, \omega)\,\mathrm{d}x)^{\xi}$. The normalization $f(x, \omega) = \frac{W(x)}{\int W(t)\,\mathrm{d}t}$ gives a random density and the distribution of this density under $LN(\mu, \sigma, \xi)$ is denoted by $LNS(\mu, \sigma, \xi)$. If $X_1, \ldots, X_n$ are i.i.d. f and $f \sim LNS(\mu, \sigma, \xi)$, then the posterior is $LNS(\mu^*, \sigma, \xi^*)$, where $\mu^*(x) = \mu(x) + \sum_{i=1}^{n} \sigma(x_i, x)$ and $\xi^* = \xi - n$.

The interpretation of the parameters are somewhat unclear. Intuitively, for a stationary covariance kernel, a higher value of $\sigma(0)$ leads to more fluctuations in $Z(x)$ and hence more noninformative. Local smoothness is controlled by $-\sigma''(0)$ – smaller value implying a smoother curve. The parameter ξ, introduced somewhat unnaturally, is the least understood. Apparently, the expression for the posterior suggests that $-\xi$ may be thought of as the "prior sample size".

5.3. Polya tree prior

A Polya tree prior satisfying $\sum_{m=1}^{\infty} a_m^{-1} < \infty$ admits densities a.s. by Kraft (1964) and hence may be considered for density estimation. The posterior expected density is given

hs25 v.2005/04/21 Prn:19/05/2005; 15:11 F:hs25013.tex; VTEX/VJ p. 30
aid: 25013 pii: S0169-7161(05)25013-7 docsubty: REV

by

$$\mathrm{E}\big(f(x)|X_1,\ldots,X_n\big)=\alpha(x)\prod_{m=1}^{\infty}\frac{2a_m+2N(B_{\epsilon(m)})}{2a_m+N(B_{\epsilon(m)})+N(B'_{\epsilon(m)})}, \tag{5.10}$$

where $N(B_{\epsilon(m)})$ stand for the number of observations falling in $B_{\epsilon(m)}$, the set in the m-level partition which contains x and $N(B'_{\epsilon(m)})$ is the number of observations falling in its sibling $B'_{\epsilon(m)}$. From Theorem 3.1 of Ghosal et al. (1999c), it follows that under the condition $\sum_{m=1}^{\infty} a_m^{-1/2}<\infty$, any p_0 with $\int p_0\log(p_0/\alpha)<\infty$ satisfies $p_0\in \mathrm{KL}(\Pi)$ and hence the weak consistency holds. Consistency under d_H has been obtained by Barron et al. (1999) under the rather strong condition that $a_m=8^m$. This high value of 8^m appears to be needed to control the roughness of the Polya trees. Using the pseudo-posterior distribution as described in Section 3, Walker and Hjort (2001) showed that the posterior mean converges in d_H solely under the condition $\sum_{m=1}^{\infty} a_m^{-1/2}<\infty$. Interestingly, they identify the posterior mean with the mean of a pseudo-posterior distribution that also comes from a Polya tree prior with a different set of parameters.

6. Regression function estimation

Regression is one of the most important and widely used tool in statistical analysis. Consider a response variable Y measured with some covariate X that may possibly be multivariate. The regression function $f(x)=\mathrm{E}(Y|X=x)$ describes the overall functional dependence of Y on X and thus becomes very useful in prediction. Spatial and geostatistical problems can also be formulated as regression problems. Classical parametric models such as linear, polynomial and exponential regression models are increasingly giving way to nonparametric regression model. Frequentist estimates of the regression functions such as the kernel estimate, spline or orthogonal series estimators have been in use for a long time and their properties have been well studied. Some nonparametric Bayesian methods have also been developed recently. The Bayesian analysis depends on the dependence structure of Y on X and are handled differently for different regression models.

6.1. Normal regression

For continuous response, a commonly used regression model is $Y_i=f(X_i)+\epsilon_i$, where ϵ_i are assumed to be i.i.d. mean zero Gaussian errors with unknown variance and be independent of X_i's. Leading nonparametric Bayesian techniques, among some others, include those based on (i) Gaussian process prior, (ii) orthogonal basis expansion, and (iii) free-knot splines.

Wahba (1978) considered a Gaussian process prior for f. The resulting Bayes estimator is found to be a smoothing spline with the appropriate choice of the covariance kernel of the Gaussian process. A commonly used prior for f is defined through the stochastic differential equation $\frac{\mathrm{d}^2 f(x)}{\mathrm{d}x^2}=\tau\frac{\mathrm{d}W(x)}{\mathrm{d}x}$, where $W(x)$ is a Wiener process. The scale parameter τ is given an inverse gamma prior while the intercept term $f(0)$ is given

hs25 v.2005/04/21 Prn:19/05/2005; 15:11 F:hs25013.tex; VTEX/VJ p. 31
aid: 25013 pii: S0169-7161(05)25013-7 docsubty: REV

an independent Gaussian prior. Ansley et al. (1993) described an extended state-space representation for computing the Bayes estimate. Barry (1986) used a similar prior for multiple covariates and provided asymptotic result for the Bayes estimator.

Another approach to putting a nonparametric prior on f is through an orthogonal basis expansion of the form $f(x) = \sum_{j=1}^{\infty} b_j \psi_j(x)$ and then putting a prior on the coefficients b_j's. Smith and Kohn (1997) considered such an approach when the infinite series is truncated at some predetermined finite stage k. Zhao (2000) considered a sieve prior putting an infinitely supported prior on k. Shen and Wasserman (2001) investigated the asymptotic properties for this sieve prior and obtained a convergence rate $n^{-q/(2q+1)}$ under some restriction on the basis function and for a Gaussian prior on the b_j's. Variable selection problem is considered in Shively et al. (1999) and Wood et al. (2002a). Wood et al. (2002b) extended this approach to spatially adaptive regression, while Smith et al. (1998) extended the idea to autocorrelated errors.

A free-knot spline approach is considered by Denison et al. (1998) and DiMatteo et al. (2001). They modeled f as a polynomial spline of fixed order (usually cubic), while putting a prior on the number of the knots, the location of the knots and the coefficients of the polynomials. Since the parameter space is canonical, computations are done through Monte Carlo averages while samples from the posterior distribution is obtained by reversible jump MCMC algorithm of Green (1995).

6.2. Binary regression

In this case, $Y|X = x \sim \text{binom}(1, f(x))$ so that $f(x) = \mathrm{P}(Y = 1|X = x) = \mathrm{E}(Y|X = x)$. Choudhuri et al. (2004b) induced a prior on $f(x)$ by using a Gaussian process $\eta(x)$ and mapping $\eta(x)$ into the unit interval as $f(x) = H(\eta(x))$ for some strictly increasing continuous chosen "link function" H. The posterior distribution of $f(x)$ is analytically intractable and the MCMC procedure depends on the choice of link function. The most commonly used link function is the probit link in which H is the standard normal cdf. In this case, an elegant Gibbs sampler algorithm is obtained by introducing some latent variables following an idea of Albert and Chib (1993).

Let $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^{\mathrm{T}}$ be the random binary observations measured along with the corresponding covariate values $\boldsymbol{X} = (X_1, \ldots, X_n)^{\mathrm{T}}$. Let $\boldsymbol{Z} = (Z_1, \ldots, Z_n)^{\mathrm{T}}$ be some unobservable latent variables such that conditional on the covariate values $\boldsymbol{X}$ and the functional parameter η, Z_i's are independent normal random variables with mean $\eta(X_i)$ and variance 1. Assume that the observations Y_i's are functions of these latent variables defined as $Y_i = I(Z_i > 0)$. Then, conditional on $(\eta, \boldsymbol{X})$, Y_i's are independent Bernoulli random variables with success probability $\Phi(\eta(X_i))$ and thus leads to the probit link model. Had we observed Z_i's, the posterior distribution of η could have been obtained analytically, which is also a Gaussian process by virtue of the conjugacy of the Gaussian observation with a Gaussian prior for the mean. However, $\boldsymbol{Z}$ is unobservable. Given the data $(\boldsymbol{Y}, \boldsymbol{X})$ and the functional parameter η, Z_i's are conditionally independent and their distributions are truncated normal with mean $\eta(X_i)$ and variance 1, where Z_i is right truncated at 0 if $Y_i = 0$, while Z_i is right truncated at 0 if $Y_i = 1$, then Z_i is taken to be positive. Now, using the conditional distributions of $(\boldsymbol{Z}|\eta, \boldsymbol{Y}, \boldsymbol{X})$ and $(\eta|\boldsymbol{Z}, \boldsymbol{Y}, \boldsymbol{X})$, a Gibbs sampler algorithm is formulated for sampling from the distribution

hs25 v.2005/04/21 Prn:19/05/2005; 15:11 F:hs25013.tex; VTEX/VJ p. 32
aid: 25013 pii: S0169-7161(05)25013-7 docsubty: REV

of $(\boldsymbol{Z}, \eta|\boldsymbol{Y}, \boldsymbol{X})$. Choudhuri et al. (2004b) also extended this Gibbs sampler algorithm to the link function that is a mixture of normal cdfs. These authors also showed that the posterior distribution is consistent under mild conditions, as stated below.

THEOREM 12. *Let the true response probability function $f_0(x)$ be continuous, $(d+1)$-times differentiable and bounded away form* 0 *and* 1*, and that the underlying Gaussian process has mean function and covariance kernel $(d+1)$-times differentiable, where d is the dimension of the covariate X. Assume that the range of X is bounded.*

If the covariate is random having a nonsingular density $q(x)$, then for any $\varepsilon > 0$, $\Pi(f\colon \int |f(x) - f_0(x)|q(x)\,\mathrm{d}x > \varepsilon | X_1, Y_1, \ldots, X_n, Y_n) \to 0$ in P_{f_0}-probability.

If the covariates are nonrandom, then for any $\varepsilon > 0$, $\Pi(f\colon n^{-1}\sum_{i=1}^{n} |f(X_i) - f_0(X_i)| > \varepsilon | Y_1, \ldots, Y_n) \to 0$ in P_{f_0}-probability.

To prove the result, conditions of Theorem 3 and Theorem 5 respectively for random and nonrandom covariates, are verified. The condition on the Kullback–Leibler support is verified by approximating the function by a finite Karhunene–Loèvè expansion and by the nonsingularity of the multivariate normal distributions. The testing condition is verified on a sieve that is given by the maximum of f and its $(d+1)$ derivatives bounded by some $M_n = \mathrm{o}(n)$. The complement of the sieve has exponentially small prior probability if M_n is not of smaller order than $n^{1/2}$.

Wood and Kohn (1998) considered the integrated Wiener process prior for the probit transformation of f. The posterior is computed via Monte Carlo averages using a data augmentation technique as above. Yau et al. (2003) extended the idea to multinomial problems. Holmes and Mallick (2003) extended the free-knot spline approach to generalized multiple regression treating binary regression as a particular case.

A completely different approach to semiparametric estimation of f is to nonparametrically estimate the link function H while using a parametric form, usually linear, for $\eta(x)$. Observe that H is a nondecreasing function with range $[0, 1]$ and this is an univariate distribution function. Gelfand and Kuo (1991), and Newton et al. (1996) used a Dirichlet process prior for H. Mallick and Gelfand (1994) modeled H as a mixture of beta cdf's with a prior probability on the mixture weights, which resulted in smoother estimates. Basu and Mukhopadhyay (2000) modeled the link function as Dirichlet scale mixture of truncated normal cdf's. Posterior consistency results for these procedures were obtained by Amewou-Atisso et al. (2003).

7. Spectral density estimation

Let $\{X_t\colon t = 1, 2, \ldots\}$ be a stationary time series with autocovariance function $\gamma(\cdot)$ and spectral density $f^*(\omega^*) = (2\pi)^{-1}\sum_{r=-\infty}^{\infty}\gamma(r)\,\mathrm{e}^{-ir\omega^*}$, $-\pi < \omega^* \leqslant \pi$. To estimate f^*, it suffices to consider the function $f(\omega) = f^*(\pi\omega)$, $0 \leqslant \omega \leqslant 1$, by the symmetry of f^*. Because the actual likelihood of f is difficult to handle, Whittle (1957, 1962) proposed a "quasi-likelihood"

$$L_n(f|X_1, \ldots, X_n) = \prod_{l=1}^{v} \frac{1}{f(\omega_l)}\,\mathrm{e}^{-I_n(\omega_l)/f(\omega_l)}, \tag{7.1}$$

hs25 v.2005/04/21 Prn:19/05/2005; 15:11 F:hs25013.tex; VTEX/VJ p. 33
aid: 25013 pii: S0169-7161(05)25013-7 docsubty: REV

where $\omega_l = 2l/n$, v is the greatest integer less than or equal to $(n-1)/2$, and $I_n(\omega) = |\sum_{t=1}^{n} X_t \, \mathrm{e}^{-it\pi\omega}|^2/(2\pi n)$ is the periodogram. A pseudo-posterior distribution may be obtained by updating the prior using this likelihood.

7.1. Bernstein polynomial prior

Normalizing f to $q = f/\tau$ with the normalizing constant $\tau = \int f$, Choudhuri et al. (2004a) induced a prior on f by first putting a Bernstein polynomial prior on q and then putting an independent prior on τ. Thus, the prior on f is described by the following hierarchical scheme:

- $f(\omega) = \tau \sum_{j=1}^{k} F((j-1)/k, j/k]\beta(\omega; j, k-j+1)$;
- $F \sim \mathrm{Dir}(M, G)$, where G has a Lebesgue density g;
- k has probability mass function $\rho(k) > 0$ for $k = 1, 2, \ldots$;
- The distribution of τ has Lebesgue density π on $(0, \infty)$;
- F, k, and τ are a priori independent.

The pseudo-posterior distribution is analytically intractable and hence is computed by an MCMC method. Using the Sethuraman representation for F as in (2.2), (f, k, τ) may be reparameterized as $(\theta_1, \theta_2, \ldots, Y_1, Y_2, \ldots, k, \tau)$. Because the infinite series in (2.2) is almost surely convergent, it may be truncated at some large L. Then one may represent F as $F = \sum_{l=1}^{L} V_l \delta_{\theta_l} + (1 - V_1 - \cdots - V_L)\delta_{\theta_0}$, where $\theta_0 \sim G$ and is independent of the other parameters. The last term is added to make F a distribution function even after the truncation. Now the problem reduces to a parametric one with finitely many parameters $(\theta_0, \theta_1, \ldots, \theta_L, Y_1, \ldots, Y_L, k, \tau)$. The functional parameter f may be written as a function of these univariate parameters as

$$f(\omega) = \tau \sum_{j=1}^{k} w_{j,k} \beta(\omega; j, k-j+1), \tag{7.2}$$

where $w_{j,k} = \sum_{l=0}^{L} V_l I\{\frac{j-1}{k} < \theta_l \leqslant \frac{j}{k}\}$ and $V_0 = 1 - V_1 - \cdots - V_L$. The posterior distribution of $(\theta_0, \theta_1, \ldots, \theta_L, Y_1, \ldots, Y_L, k, \tau)$ is proportional to

$$\left[\prod_{m=1}^{v} \frac{1}{f(2m/n)} \, \mathrm{e}^{-U_m/f(2m/n)}\right]\left[\prod_{l=1}^{L} M(1-y_l)^{M-1}\right]\left[\prod_{l=0}^{L} g(\theta_l)\right]\rho(k)\pi(\tau). \tag{7.3}$$

The discrete parameter k may be easily simulated from its posterior distribution given the other parameters. If the prior on τ is an inverse gamma distribution, then the posterior distribution of τ conditional on the other parameters is also inverse gamma. To sample from the posterior density of θ_i's or Y_i's conditional on the other parameters, Metropolis algorithm is within the Gibbs sampling step is used. The starting values of τ may be set to the sample variance divided by 2π, while the starting value of k may be set to some large integer K_0. The approximate posterior mode of θ_i's and Y_i's given the starting values of τ and k may be considered as the starting values for the respective variables.

Let f_0^* be the true spectral density. Assume that the time series satisfies the conditions

hs25 v.2005/04/21 Prn:19/05/2005; 15:11 F:hs25013.tex; VTEX/VJ p. 34
aid: 25013 pii: S0169-7161(05)25013-7 docsubty: REV

(M1) the time series is Gaussian with $\sum_{r=0}^{\infty} r^{\alpha}\gamma(r); < \infty$ for some $\alpha > 0$;
(M2) for all ω^*, $f_0^*(\omega^*) > 0$;

and the prior satisfies

(P1) for all k, $0 < \rho(k) \leqslant C\,\mathrm{e}^{-ck(\log k)^{1+\alpha'}}$ for some constants $C, c, \alpha' > 0$;
(P2) g is bounded, continuous, and bounded away from zero;
(P3) the prior on τ is degenerate at the true value $\tau_0 = \int f_0$.

Using the contiguity result of Choudhuri et al. (2004c), the following result was shown by Choudhuri et al. (2004a) under the above assumptions.

THEOREM 13. *For any $\varepsilon > 0$, $\Pi_n\{f^*\colon \|f^* - f_0^*\|_1 > \varepsilon\} \to 0$ in $P^n_{f_0^*}$-probability, where Π_n is the pseudo-posterior distribution computed using the Whittle likelihood of and $P^n_{f_0^*}$ is the actual distribution of the data $(X_1, \ldots, X_n)$.*

REMARK 1. The conclusion of Theorem 13 still holds if the degenerated prior on τ is replaced by a sequence of priors distribution that asymptotically bracket the true value, that is, the prior support of τ is in $[\tau_0 - \delta_n, \tau_0 + \delta_n]$ for some $\delta_n \to 0$. A two-stage empirical Bayes method, by using one part of the sample to consistently estimate τ and the other part to estimate q, may be considered to construct the above asymptotically bracketing prior.

7.2. Gaussian process prior

Since the spectral density is nonnegative valued function, a Gaussian process prior may be assigned to $g(\omega) = \log(f(\omega))$. Because the Whittle likelihood in (7.1) arises by assuming that $I_n(\omega_l)$'s are approximately independent exponential random variables with mean $f(\omega_l)$, one may obtain a regression model of the form $\log(I_n(\omega_l)) = g(\omega_l) + \epsilon_l$, where the additive errors ϵ_l's are approximately i.i.d. with the Gumbel distribution.

Carter and Kohn (1997) considered an integrated Wiener process prior for g. They described an elegant Gibbs sampler algorithm for sampling from the posterior distribution. Approximating the distribution of ϵ_l's as a mixture of five known normal distribution, they introduced latent variables indicating the mixture components for the corresponding errors. Given the latent variables, conditional posterior distribution of g is obtained by a data augmentation technique. Given g, the conditional posterior distribution of the latent variables are independent and samples are easily drawn from their finite support.

Gangopadhyay et al. (1998) considered the free-not spline approach to modeling g. In this case, the posterior is computed by the reversible jump algorithm of Green (1995). Liseo et al. (2001) considered a Brownian motion process as prior on g. For sampling from the posterior distribution, they considered the Karhunen–Loévé series expansion for the Brownian motion and then truncated the infinite series to a finite sum.

hs25 v.2005/04/21 Prn:19/05/2005; 15:11 F:hs25013.tex; VTEX/VJ p. 35
aid: 25013 pii: S0169-7161(05)25013-7 docsubty: REV

8. Estimation of transition density

Estimation of the transition density of a discrete-time Markov process is an important problem. Let Π be a prior on the transition densities $p(y|x)$. Then the predictive density of a future observation X_{n+1} given the data $X_1, \ldots, X_n$ equals to $\mathrm{E}(p(\cdot|X_n)|X_1, \ldots, X_n)$, which is the Bayes estimate of the transition density p at X_n. The prediction problem thus directly relates to the estimation of the transition density.

Tang and Ghosal (2003) considered a mixture of normal model

$$p(y|x) = \int \phi_\sigma\big(y - H(x;\theta)\big)\,\mathrm{d}F(\theta,\sigma), \tag{8.1}$$

where θ is possibly vector valued and $H(x;\theta)$ is a known function. Such models are analogous to the normal mixture models in the density estimation where the unknown probability density is modeled as $p(y) = \int \phi_\sigma(y-\mu)\,\mathrm{d}F(\mu,\sigma)$. A reasonable choice for the link function H in (8.1) could be of the form $\tau + \gamma\psi(\delta + \beta x)$ for some known function ψ.

As in density estimation, this mixture model may be represented as

$$X_i \sim N\big(H(X_{i-1};\theta_i), \sigma_i^2\big), \qquad (\theta_i,\sigma_i) \overset{\text{i.i.d.}}{\sim} F. \tag{8.2}$$

Here, unlike a parametric model, the unknown parameters are varying along with the index of the observation, and are actually drawn as i.i.d. samples from an unknown distribution. Hence the model is "dynamic" as opposed to a "static" parametric mixture model.

Tang and Ghosal (2003) let the mixing distribution F have a Dirichlet process prior $\mathrm{Dir}(M, G)$. As in density estimation, the hierarchical representation (8.2) helps develop Gibbs sampler algorithms for sampling from the posterior distribution. However, because of the nonstandard forms of the conditionals, special techniques, such as the "no gaps" algorithm of MacEachern and Muller (1998) need to be implemented.

To study the large sample properties of the posterior distribution, Tang and Ghosal (2003) extended Schwartz's (1965) theorem to the context of an ergodic Markov processes. For simplicity, X_0 is assumed to be fixed below, although the conclusion extends to random X_0 also.

THEOREM 14. *Let $\{X_n, n \geqslant 0\}$ be an ergodic Markov process with transition density $p \in \mathcal{P}$ and stationary distribution π. Let Π be a prior on $\mathcal{P}$. Let $p_0 \in \mathcal{P}$ and π_0 be respectively the true values of p and π. Let U_n be a sequence of subsets of $\mathcal{P}$ containing p_0.*

Suppose that there exist a sequence of tests Φ_n, based on $X_0, X_1, \ldots, X_n$ for testing the pair of hypotheses H_0: $p = p_0$ against H: $p \in U_n^c$, and subsets $V_n \subset \mathcal{P}$ such that

(i) *p_0 is in the Kullback–Leibler support of Π, that is $\Pi\{p\colon K(p_0,p) < \varepsilon\} > 0$, where*

$$K(p_0,p) = \iint \pi_0(x)p_0(y|x)\log\frac{p_0(y|x)}{p(y|x)}\,\mathrm{d}y\,\mathrm{d}x,$$

hs25 v.2005/04/21 Prn:19/05/2005; 15:11 F:hs25013.tex; VTEX/VJ p. 36
aid: 25013 pii: S0169-7161(05)25013-7 docsubty: REV

(ii) $\Phi_n \to 0$ *a.s.* $[P_{f_0}^{\infty}]$,
(iii) $\sup_{p \in U_n^c \cap V_n} \mathrm{E}_p(1-\Phi_n) \leqslant C_1 \mathrm{e}^{-n\beta_1}$ *for some constants* C_1 *and* β_1,
(iv) $\Pi(p \in V_n^c) \leqslant C_2 \mathrm{e}^{-n\beta_2}$ *for some constants* C_2 *and* β_2.

Then $\Pi(p \in U_n | X_0, X_1, \ldots, X_n) \to 1$ *a.s.* $[P_0^{\infty}]$, *where* $[P_0^{\infty}]$ *denote the distribution of the infinite sequence* $(X_0, X_1, \ldots)$.

Assume that $p_0(y|x)$ is of the form (8.1). Let F_0 denote the true mixing distribution, and π_0 denote the corresponding invariant distribution. Let the sup-L_1 distance on the space of transition probabilities be given by $d(p_1, p_2) = \sup_x \int |p_1(y|x) - p_2(y|x)| \, \mathrm{d}y$. Let H be uniformly equicontinuous in x and the support of G be compact containing the support of F_0. Tang and Ghosal (2003) showed that (i) the test $I\{\sum_{i=1}^{k} \log \frac{p_1(X_{2i}|X_{2i-1})}{p_0(X_{2i}|X_{2i-1})} > 0\}$, where $n = 2k$ or $2k+1$, for testing p_0 against a small ball around p_1, has exponentially small error probabilities, (ii) the space of transition probabilities supported by the prior is compact under the sup-L_1 distance, and (iii) the Kullback–Leibler property holds at p_0. By the compactness property, a single test can be constructed for the entire alternative having exponentially small error probabilities. It may be noted that because of the compactness of $\mathcal{P}$, it is not necessary to consider sieves. Thus by Theorem 14, the posterior distribution is consistent at p_0 with respect to the sup-L_1 distance.

The conditions assumed in the above result are somewhat stringent. For instance $H(x, \beta, \delta, \tau) = \tau + \gamma\psi(\delta + \beta x)$, then ψ is necessarily bounded, ruling out the linear link. If a suitable weaker topology is employed, Tang and Ghosal (2003) showed that consistency can be obtained under weaker conditions by extending Walker's (2004) approach to Markov processes. More specifically, the Kullback–Leibler property holds if H satisfies uniform equicontinuity on compact sets only. If now a topology is defined by the neighborhood base $\{f\colon \int |\int g_i(y) f(y|x) \, \mathrm{d}y - \int g_i(y) f_0(y|x) \, \mathrm{d}y| \nu(x) \, \mathrm{d}x < \varepsilon, i = 1, \ldots, k\}$, where ν is a probability density, then consistency holds if σ is bounded below and contains σ_0 in its support. If further σ is also bounded above and the θ is supported on a compact set, then consistency also holds in the integrated-L_1 distance integrated with respect to ν. For a linear link function $H(x, \rho, b) = \rho x + b$, $|\rho| < 1$, the compactness condition can be dropped, for instance, if the distribution of b under G is normal.

9. Concluding remarks

In this article, we have reviewed Bayesian methods for the estimation of functions of statistical interest such as the cumulative distribution function, density function, regression function, spectral density of a time series and the transition density function of a Markov process. Function estimation can be viewed as a problem of the estimation of one or more infinite-dimensional parameter arising in a statistical model. It has been argued that the Bayesian approach to function estimation, commonly known as Bayesian nonparametric estimation, can provide an important, coherent alternative to

hs25 v.2005/04/21 Prn:19/05/2005; 15:11 F:hs25013.tex; VTEX/VJ p. 37
aid: 25013 pii: S0169-7161(05)25013-7 docsubty: REV

more familiar classical approaches to function estimation. We have considered the problems of construction of appropriate prior distributions on infinite-dimensional spaces. It has been argued that, because of the lack of subjective knowledge about every detail of a distribution in an infinite-dimensional space, some default mechanism of prior specification needs to be followed. We have discussed various important priors on infinite-dimensional spaces, and their merits and demerits. While certainly not exhaustive, these priors and their various combinations provide a large catalogue of priors in a statistician's toolbox, which may be tried and tested for various curve estimation problems including, but not restricted to, the problems we discussed. Due to the vastness of the relevant literature and the rapid growth of the subject, it is impossible to even attempt to mention all the problems of Bayesian curve estimation. The material presented here is mostly a reflection of the authors' interest and familiarity. Computation of the posterior distribution is an important issue. Due to the lack of useful analytical expressions for the posterior distribution in most curve estimation problems, computation has to be done by some numerical technique, usually by the help of Markov chain Monte Carlo methods. We described computing techniques for the curve estimation problems considered in this chapter. The simultaneous development of innovative sampling techniques and computing devices has brought tremendous computing power to nonparametric Bayesians. Indeed, for many statistical problems, the computing power of a Bayesian now exceeds that of a non-Bayesian. While these positive developments are extremely encouraging, one should however be extremely cautious about naive uses of Bayesian methods for nonparametric problems to avoid pitfalls. We argued that it is important to validate the use of a particular prior by using some benchmark criterion such as posterior consistency. We discussed several techniques of proving posterior consistency and mentioned some examples of inconsistency. Sufficient conditions for posterior consistency are discussed in the problems we considered. Convergence rates of posterior distributions have also been discussed, together with the related concepts of optimality, adaptation, misspecification and Berntsein–von Mises theorem.

The popularity of Bayesian nonparametric methods is rapidly growing among practitioners as theoretical properties are increasingly better understood and the computational hurdles are being removed. Innovative Bayesian nonparametric methods for complex models arising in biomedical, geostatistical, environmental, econometric and many other applications are being proposed. Study of theoretical properties of nonparametric Bayesian beyond the traditional i.i.d. set-up has started to receive attention recently. Much more work will be needed to bridge the gap. Developing techniques of model selection, the Bayesian equivalent of hypothesis testing, as well as the study of their theoretical properties will be highly desirable.

References

Albert, J., Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *J. Amer. Statist. Assoc.* **88**, 669–679.

Amewou-Atisso, M., Ghosal, S., Ghosh, J.K., Ramamoorthi, R.V. (2003). Posterior consistency for semiparametric regression problems. *Bernoulli* **9**, 291–312.

hs25 v.2005/04/21 Prn:19/05/2005; 15:11 F:hs25013.tex; VTEX/VJ p. 38
aid: 25013 pii: S0169-7161(05)25013-7 docsubty: REV

Ansley, C.F., Kohn, R., Wong, C. (1993). Nonparametric spline regression with prior information. *Biometrika* **80**, 75–88.

Antoniak, C. (1974). Mixtures of Dirichlet processes with application to Bayesian non-parametric problems. *Ann. Statist.* **2**, 1152–1174.

Barron, A.R. (1988). The exponential convergence of posterior probabilities with implications for Bayes estimators of density functions. Unpublished manuscript.

Barron, A.R. (1999). Information-theoretic characterization of Bayes performance and the choice of priors in parametric and nonparametric problems. In: Bernardo, J.M., et al. (Eds.), *Bayesian Statistics, vol. 6*. Oxford University Press, New York, pp. 27–52.

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

Barry, D. (1986). Nonparametric Bayesian regression. *Ann. Statist.* **14**, 934–953.

Basu, S., Mukhopadhyay, S. (2000). Bayesian analysis of binary regression using symmetric and asymmetric links. *Sankhyā, Ser. B* **62**, 372–387.

Belitser, E.N., Ghosal, S. (2003). Adaptive Bayesian inference on the mean of an infinite-dimensional normal distribution. *Ann. Statist.* **31**, 536–559.

Berger, J.O., Guglielmi, A. (2001). Bayesian and conditional frequentist testing of a parametric model versus nonparametric alternatives. *J. Amer. Statist. Assoc.* **96** (453), 174–184.

Berk, R. (1966). Limiting behavior of the posterior distribution when the model is incorrect. *Ann. Math. Statist.* **37**, 51–58.

Birgé, L. (1983). Robust testing for independent non-identically distributed variables and Markov chains. In: Florens, J.P., et al. (Eds.), *Specifying Statistical Models. From Parametric to Non-Parametric. Using Bayesian or Non-Bayesian Approaches*. In: *Lecture Notes in Statistics*, vol. 16. Springer-Verlag, New York, pp. 134–162.

Blackwell, D. (1973). Discreteness of Ferguson selection. *Ann. Statist.* **1**, 356–358.

Blackwell, D., Dubins, L.E. (1962). Merging of opinions with increasing information. *Ann. Math. Statist.* **33**, 882–886.

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

Blum, J., Susarla, V. (1977). On the posterior distribution of a Dirichlet process given randomly right censored observations. *Stochastic Process. Appl.* **5**, 207–211.

Brunner, L.J. (1992). Bayesian nonparametric methods for data from a unimodal density. *Statist. Probab. Lett.* **14**, 195–199.

Brunner, L.J. (1995). Bayesian linear regression with error terms that have symmetric unimodal densities. *J. Nonparametr. Statist.* **4**, 335–348.

Brunner, L.J., Lo, A.Y. (1989). Bayes methods for a symmetric unimodal density and its mode. *Ann. Statist.* **17**, 1550–1566.

Carter, C.K., Kohn, R. (1997). Semiparametric Bayesian inference for time series with mixed spectra. *J. Roy. Statist. Soc., Ser. B* **59**, 255–268.

Choudhuri, N., Ghosal, S., Roy, A. (2004a). Bayesian estimation of the spectral density of a time series. *J. Amer. Statist. Assoc.* **99**, 1050–1059.

Choudhuri, N., Ghosal, S., Roy, A. (2004b). Bayesian nonparametric binary regression with a Gaussian process prior. Preprint.

Choudhuri, N., Ghosal, S., Roy, A. (2004c). Contiguity of the Whittle measure in a Gaussian time series. *Biometrika* **91**, 211–218.

Cifarelli, D.M., Regazzini, E. (1990). Distribution functions of means of a Dirichlet process. *Ann. Statist.* **18**, 429–442.

Cox, D.D. (1993). An analysis of Bayesian inference for nonparametric regression. *Ann. Statist.* **21**, 903–923.

Dalal, S.R. (1979). Dirichlet invariant processes and applications to nonparametric estimation of symmetric distribution functions. *Stochastic Process. Appl.* **9**, 99–107.

Denison, D.G.T., Mallick, B.K., Smith, A.F.M. (1998). Automatic Bayesian curve fitting. *J. Roy. Statist. Soc., Ser. B Stat. Methodol.* **60**, 333–350.

Diaconis, P., Freedman, D. (1986a). On the consistency of Bayes estimates (with discussion). *Ann. Statist.* **14**, 1–67.

Diaconis, P., Freedman, D. (1986b). On inconsistent Bayes estimates. *Ann. Statist.* **14**, 68–87.
DiMatteo, I., Genovese, C.R., Kass, R.E. (2001). Bayesian curve-fitting with free-knot splines. *Biometrika* **88**, 1055–1071.
Doksum, K.A. (1974). Tail free and neutral random probabilities and their posterior distributions. *Ann. Probab.* **2**, 183–201.
Doob, J.L. (1948). Application of the theory of martingales. Coll. Int. du CNRS, Paris, pp. 22–28.
Doss, H. (1985a). Bayesian nonparametric estimation of the median. I. Computation of the estimates. *Ann. Statist.* **13**, 1432–1444.
Doss, H. (1985b). Bayesian nonparametric estimation of the median. II. Asymptotic properties of the estimates. *Ann. Statist.* **13**, 1445–1464.
Doss, H., Sellke, T. (1982). The tails of probabilities chosen from a Dirichlet prior. *Ann. Statist.* **10**, 1302–1305.
Dykstra, R.L., Laud, P.W. (1981). A Bayesian nonparametric approach to reliability. *Ann. Statist.* **9**, 356–367.
Escobar, M. (1994). Estimating normal means with a Dirichlet process prior. *J. Amer. Statist. Assoc.* **89**, 268–277.
Escobar, M., West, M. (1995). Bayesian density estimation and inference using mixtures. *J. Amer. Statist. Assoc.* **90**, 577–588.
Escobar, M., West, M. (1998). Computing nonparametric hierarchical models. In: *Practical Nonparametric and Semiparametric Bayesian Statistics*. In: *Lecture Notes in Statistics*, vol. 133. Springer, New York, pp. 1–22.
Ferguson, T.S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1**, 209–230.
Ferguson, T.S. (1974). Prior distribution on the spaces of probability measures. *Ann. Statist.* **2**, 615–629.
Ferguson, T.S. (1983). Bayesian density estimation by mixtures of Normal distributions. In: Rizvi, M., Rustagi, J., Siegmund, D. (Eds.), *Recent Advances in Statistics*, pp. 287–302.
Ferguson, T.S., Phadia, E.G. (1979). Bayesian nonparametric estimation based on censored data. *Ann. Statist.* **7**, 163–186.
Freedman, D. (1963). On the asymptotic distribution of Bayes estimates in the discrete case I. *Ann. Math. Statist.* **34**, 1386–1403.
Freedman, D. (1965). On the asymptotic distribution of Bayes estimates in the discrete case II. *Ann. Math. Statist.* **36**, 454–456.
Freedman, D. (1999). On the Bernstein–von Mises theorem with infinite-dimensional parameters. *Ann. Statist.* **27**, 1119–1140.
Fristedt, B. (1967). Sample function behavior of increasing processes with stationary independent increments. *Pacific J. Math.* **21**, 21–33.
Fristedt, B., Pruitt, W.E. (1971). Lower functions for increasing random walks and subordinators. *Z. Wahsch. Verw. Gebiete* **18**, 167–182.
Gangopadhyay, A.K., Mallick, B.K., Denison, D.G.T. (1998). Estimation of spectral density of a stationary time series via an asymptotic representation of the periodogram. *J. Statist. Plann. Inference* **75**, 281–290.
Gasparini, M. (1996). Bayesian density estimation via Dirichlet density process. *J. Nonparametr. Statist.* **6**, 355–366.
Gelfand, A.E., Kuo, L. (1991). Nonparametric Bayesian bioassay including ordered polytomous response. *Biometrika* **78**, 657–666.
Ghosal, S. (2000). Asymptotic normality of posterior distributions for exponential families with many parameters. *J. Multivariate Anal.* **74**, 49–69.
Ghosal, S. (2001). Convergence rates for density estimation with Bernstein polynomials. *Ann. Statist.* **29**, 1264–1280.
Ghosal, S., van der Vaart, A.W. (2001). Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. *Ann. Statist.* **29**, 1233–1263.
Ghosal, S., van der Vaart, A.W. (2003a). Convergence rates for non-i.i.d. observations. Preprint.
Ghosal, S., van der Vaart, A.W. (2003b). Posterior convergence rates of Dirichlet mixtures of normal distributions for smooth densities. Preprint.
Ghosal, S., Ghosh, J.K., Samanta, T. (1995). On convergence of posterior distributions. *Ann. Statist.* **23**, 2145–2152.

Ghosal, S., Ghosh, J.K., Ramamoorthi, R.V. (1997). Noninformative priors via sieves and consistency. In: Panchapakesan, S., Balakrishnan, N. (Eds.), *Advances in Statistical Decision Theory and Applications*. Birkhäuser, Boston, pp. 119–132.

Ghosal, S., Ghosh, J.K., Ramamoorthi, R.V. (1999a). Consistency issues in Bayesian nonparametrics. In: Ghosh, S. (Ed.), *Asymptotics, Nonparametrics and Time Series: A Tribute to Madan Lal Puri*. Marcel Dekker, New York, pp. 639–668.

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

Ghosal, S., Ghosh, J.K., Ramamoorthi, R.V. (1999c). Consistent semiparametric Bayesian inference about a location parameter. *J. Statist. Plann. Inference* **77**, 181–193.

Ghosal, S., Ghosh, J.K., van der Vaart, A.W. (2000). Convergence rates of posterior distributions. *Ann. Statist.* **28**, 500–531.

Ghosal, S., Lember, Y., van der Vaart, A.W. (2003). On Bayesian adaptation. *Acta Appl. Math.* **79**, 165–175.

Ghosh, J.K., Ramamoorthi, R.V. (1995). Consistency of Bayesian inference for survival analysis with or without censoring. In: Koul, H. (Ed.), *Analysis of Censored Data*. In: *IMS Lecture Notes Monograph Series*, vol. 27. Inst. Math. Statist., Hayward, CA, pp. 95–103.

Ghosh, J.K., Ramamoorthi, R.V. (2003). *Bayesian Nonparametrics*. Springer-Verlag, New York.

Ghosh, S.K., Ghosal, S. (2003). Proportional mean regression models for censored data. Preprint.

Ghosh, J.K., Ghosal, S., Samanta, T. (1994). Stability and convergence of posterior in non-regular problems. In: Gupta, S.S., Berger, J.O. (Eds.), *Statistical Decision Theory and Related Topics V*. Springer-Verlag, New York, pp. 183–199.

Green, P. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**, 711–732.

Grenander, U. (1981). *Abstract Inference*. Wiley, New York.

Hanson, T., Johnson, W.O. (2002). Modeling regression error with a mixture of Polya trees. *J. Amer. Statist. Assoc.* **97**, 1020–1033.

Hjort, N.L. (1990). Nonparametric Bayes estimators based on beta processes in models for life history data. *Ann. Statist.* **18**, 1259–1294.

Hjort, N.L. (1996). Bayesian approaches to non- and semiparametric density estimation. In: Bernardo, J., et al. (Eds.), *Bayesian Statistics, vol. 5*, pp. 223–253.

Hjort, N.L. (2000). Bayesian analysis for a generalized Dirichlet process prior. Preprint.

Hjort, N.L. (2003). Topics in nonparametric Bayesian statistics (with discussion). In: Green, P.J., Hjort, N., Richardson, S. (Eds.), *Highly Structured Stochastic Systems*. Oxford University Press, pp. 455–487.

Holmes, C.C., Mallick, B.K. (2003). Generalized nonlinear modeling with multivariate free-knot regression splines. *J. Amer. Statist. Assoc.* **98**, 352–368.

Huang, T.Z. (2004). Convergence rates for posterior distributions and adaptive estimation. *Ann. Statist.* **32**, 1556–1593.

Ibragimov, I.A., Has'minskii, R.Z. (1981). *Statistical Estimation: Asymptotic Theory*. Springer-Verlag, New York.

Iswaran, H., Zarepour, M. (2002a). Exact and approximate sum representation for the Dirichlet process. *Canad. J. Statist.* **26**, 269–283.

Iswaran, H., Zarepour, M. (2002b). Dirichlet prior sieves in finite normal mixture models. *Statistica Sinica*, 269–283.

Kim, Y. (1999). Nonparametric Bayesian estimators for counting processes. *Ann. Statist.* **27**, 562–588.

Kim, Y., Lee, J. (2001). On posterior consistency of survival models. *Ann. Statist.* **29**, 666–686.

Kim, Y., Lee, J. (2004). A Bernstein–von Mises theorem in the nonparametric right-censoring model. *Ann. Statist.* **32**, 1492–1512.

Kleijn, B., van der Vaart, A.W. (2002). Misspecification in infinite-dimensional Bayesian statistics. Preprint.

Kottas, A., Gelfand, A.E. (2001). Bayesian semiparametric median regression modeling. *J. Amer. Statist. Assoc.* **96**, 1458–1468.

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

Lavine, M. (1992). Some aspects of Polya tree distributions for statistical modeling. *Ann. Statist.* **20**, 1222–1235.

Lavine, M. (1994). More aspects of Polya tree distributions for statistical modeling. *Ann. Statist.* **22**, 1161–1176.

Le Cam, L.M. (1986). *Asymptotic Methods in Statistical Decision Theory*. Springer-Verlag, New York.

Le Cam, L., Yang, G.L. (2000). *Asymptotics in Statistics*, second ed. Springer-Verlag.

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.

Leonard, T. (1978). Density estimation, stochastic processes, and prior information. *J. Roy. Statist. Soc., Ser. B* **40**, 113–146.

Liseo, B., Marinucci, D., Petrella, L. (2001). Bayesian semiparametric inference on long-range dependence. *Biometrika* **88**, 1089–1104.

Lo, A.Y. (1982). Bayesian nonparametric statistical inference for Poisson point process. *Z. Wahsch. Verw. Gebiete* **59**, 55–66.

Lo, A.Y. (1983). Weak convergence for Dirichlet processes. *Sankhyā, Ser. A* **45**, 105–111.

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

Lo, A.Y. (1986). A remark on the limiting posterior distribution of the multiparameter Dirichlet process. *Sankhyā, Ser. A* **48**, 247–249.

MacEachern, S.N., Muller, P. (1998). Estimating mixture of Dirichlet process models. *J. Comput. Graph. Statist.* **7**, 223–228.

Mallick, B.K., Gelfand, A.E. (1994). Generalized linear models with unknown link functions. *Biometrika* **81**, 237–245.

Mauldin, R.D., Sudderth, W.D., Williams, S.C. (1992). Polya trees and random distributions. *Ann. Statist.* **20**, 1203–1221.

Muliere, P., Tardella, L. (1998). Approximating distributions of functionals of Ferguson–Dirichlet priors. *Canad. J. Statist.* **30**, 269–283.

Newton, M.A., Czado, C., Chappell, R. (1996). Bayesian inference for semiparametric binary regression. *J. Amer. Statist. Assoc.* **91**, 142–153.

Nieto-Barajas, L.E., Walker, S.G. (2004). Bayesian nonparametric survival analysis via Lévy driven Markov process. *Statistica Sinica* **14**, 1127–1146.

Petrone, S. (1999a). Random Bernstein polynomials. *Scand. J. Statist.* **26**, 373–393.

Petrone, S. (1999b). Bayesian density estimation using Bernstein polynomials. *Canad. J. Statist.* **26**, 373–393.

Petrone, S., Veronese, P. (2002). Nonparametric mixture priors based on an exponential random scheme. *Statist. Methods Appl.* **11**, 1–20.

Petrone, S., Wasserman, L. (2002). Consistency of Bernstein polynomial posteriors. *J. Roy. Statist. Soc., Ser. B* **64**, 79–100.

Regazzini, E., Guglielmi, A., Di Nunno, G. (2002). Theory and numerical analysis for exact distributions of functionals of a Dirichlet process. *Ann. Statist.* **30**, 1376–1411.

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

Schwartz, L. (1965). On Bayes procedures. *Z. Wahsch. Verw. Gebiete* **4**, 10–26.

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

Sethuraman, J., Tiwari, R. (1982). Convergence of Dirichlet measures and interpretation of their parameters. In: Gupta, S.S., Berger, J.O. (Eds.), *Statistical Decision Theory and Related Topics. III, vol. 2*. Academic Press, New York, pp. 305–315.

Shen, X. (2002). Asymptotic normality of semiparametric and nonparametric posterior distributions. *J. Amer. Statist. Assoc.* **97**, 222–235.

Shen, X., Wasserman, L. (2001). Rates of convergence of posterior distributions. *Ann. Statist.* **29**, 687–714.

Shively, T.S., Kohn, R., Wood, S. (1999). Variable selection and function estimation in additive nonparametric regression using a data-based prior (with discussions). *J. Amer. Statist. Assoc.* **94**, 777–806.

Smith, M., Kohn, R. (1997). A Bayesian approach to nonparametric bivariate regression. *J. Amer. Statist. Assoc.* **92**, 1522–1535.

Smith, M., Wong, C., Kohn, R. (1998). Additive nonparametric regression with autocorrelated errors. *J. Roy. Statist. Soc., Ser. B* **60**, 311–331.

hs25 v.2005/04/21 Prn:19/05/2005; 15:11 F:hs25013.tex; VTEX/VJ p. 42
aid: 25013 pii: S0169-7161(05)25013-7 docsubty: REV

Susarla, V., Van Ryzin, J. (1976). Nonparametric Bayesian estimation of survival curves from incomplete observations. *J. Amer. Statist. Assoc.* **71**, 897–902.

Susarla, V., Van Ryzin, J. (1978). Large sample theory for a Bayesian nonparametric survival curve estimator based on censored samples. *Ann. Statist.* **6**, 755–768.

Tang, Y., Ghosal, S. (2003). Posterior consistency of Dirichlet mixtures for estimating a transition density. Preprint.

Tokdar, S.T. (2003). Posterior consistency of Dirichlet location-scale mixtures of normals in density estimation and regression. Preprint.

van der Vaart, A.W. (1998). *Asymptotic Statistics*. Cambridge University Press.

van der Vaart, A.W., Wellner, J.A. (1996). *Weak Convergence and Empirical Processes*. Springer-Verlag, New York.

Wahba, G. (1978). Improper priors, spline smoothing and the problem of guarding against model errors in regression. *J. Roy. Statist. Soc., Ser. B* **40**, 364–372.

Walker, S.G. (2003). On sufficient conditions for Bayesian consistency. *Biometrika* **90**, 482–490.

Walker, S.G. (2004). New approaches to Bayesian consistency. *Ann. Statist.* **32**, 2028–2043.

Walker, S.G., Hjort, N.L. (2001). On Bayesian consistency. *J. Roy. Statist. Soc., Ser. B* **63**, 811–821.

Walker, S.G., Muliere, P. (1997). Beta-Stacy processes and a generalization of the Polya-urn scheme. *Ann. Statist.* **25**, 1762–1780.

Wasserman, L. (1998). Asymptotic properties of nonparametric Bayesian procedures. In: Dey, D., et al. (Eds.), *Practical Nonparametric and Semiparametric Bayesian Statistics*. In: *Lecture Notes in Statistics*, vol. 133. Springer-Verlag, New York, pp. 293–304.

Whittle, P. (1957). Curve and periodogram smoothing. *J. Roy. Statist. Soc., Ser. B* **19**, 38–63.

Whittle, P. (1962). Gaussian estimation in stationary time series. *Bull. Int. Statist. Inst.* **39**, 105–129.

Wood, S., Kohn, R. (1998). A Bayesian approach to robust binary nonparametric regression. *J. Roy. Statist. Soc., Ser. B* **93**, 203–213.

Wood, S., Kohn, R., Shively, T., Jiang, W. (2002a). Model selection in spline nonparametric regression. *J. Roy. Statist. Soc., Ser. B* **64**, 119–139.

Wood, S., Jiang, W., Tanner, M. (2002b). Bayesian mixture of splines for spatially adaptive nonparametric regression. *Biometrika* **89**, 513–528.

Yau, P., Kohn, R., Wood, S. (2003). Bayesian variable selection and model averaging in high-dimensional multinomial nonparametric regression. *J. Comput. Graph. Statist.* **12**, 23–54.

Zhao, L.H. (2000). Bayesian aspects of some nonparametric problems. *Ann. Statist.* **28**, 532–552.