Bayesian Density Estimation

Jayanta K. Ghosh

Abstract. This is a brief exposition of posterior consistency issues in Bayesian nonparametrics especially in the context of Bayesian Density estimation,

1991 Mathematics Subject Classification: 62A15,62G99.
Keywords and Phrases: Dirichlet mixtures,density estimation

1 Introduction

We describe popular methods of Bayesian density estimation and explore sufficient conditions for the posterior given data to converge to a true underlying distribution P_0 as the data size increases. One of the advantages of Bayesian density estimates is that,unlike classical frequentist methods,choice of the right amount of smoothing is not such a serious problem.

Section 2 provides a general background to infinite dimensional problems of inference such as Bayesian nonparametrics, semiparametrics and density estimation. Bayesian nonparametrics has been around for about twenty five years but the other two areas,specially the last, is of more recent vintage. Section 3 indicates in broad terms why different tools are needed for these three different problems and then Section 4 focuses on our main problem of interest ,namely,positive posterior consistency results for Bayesian density estimation.

2 Background

Let $X_1, X_2, \dots, X_n$ be i.i.d. random variables with unknown common probability measure P on $(\mathbf{R}, \mathcal{B})$, where $\mathbf{R}$ is the real line and $\mathcal{B}$ the Borel $\sigma-$ field. Typically P lies in some given set of probability measures $\mathcal{P}$. In Bayesian analysis, a statistician puts a probability measure Π on $\mathcal{P}$ equipped with a suitable $\sigma-$ field $\mathcal{B}_{\mathcal{P}}$ and assumes that the unknown P is distributed over $\mathcal{P}$ according to Π and, given P, $X_1, X_2, \dots, X_n$ are i.i.d. with common distribution $\mathcal{P}$. This completely specifies the joint distribution of the random P and the random Xs. Hence, in principle one can calculate the conditional probability $\Pi(B|X_1, X_2, \dots, X_n)$ of P lying in some subset B. This is the posterior in distinction with $\Pi(B)$ which is the prior probability of B. Consistency of posterior to be defined below is a sort of partial validation of this method of analysis. We now define posterior consistency at P_0. Suppose unknown to the Bayesian statistician,$X_1, X_2, \dots, X_n$ are i.i.d. $\sim P_0$, where P_0 is a given element of $\mathcal{P}$ and not random. Suppose that $\mathcal{P}$ is also equipped with a topology and the topology and $\mathcal{B}_{\mathcal{P}}$ are compatible in the sense that the neighborhoods B of P_0 are $\mathcal{B}_{\mathcal{P}}$ measurable.

Definition: $\Pi(. \,|X_1, X_2, \ldots, X_n)$ is consistent at P_0 if for all neighborhoods B of P_0, as $n \to \infty$,

$$\Pi(B|X_1, X_2, \ldots, X_n) \to 1 \text{ a.s } P_0$$

This property depends on both Π and P_0. It would be desirable to have this property at various P_0's that seem plausible to the Bayesian who is using this posterior.

An old result of Doob shows that such a property holds for all but a $\pi-$null set of P_0's. Unfortunately, this result is too weak to settle whether consistency holds for a particular P_0. It is well known that this property holds for a wide class of priors and all P_0's if $\mathcal{P}$ is finite dimensional,e.g., when $\mathcal{P}$ is the set of all normal distributions $N(\mu, \sigma^2)$ with mean μ and variance σ^2, $-\infty < \mu < \infty, \sigma^2 > 0$. In contrast the answer is usually no when $\mathcal{P}$ is infinite dimensional as in density estimation.

There are three broad classes of infinite dimensional problems —(fully) nonparametric inference like making inference about an unknown distribution function, a semiparametric problem like estimating the point of symmetry of an unknown symmetrical distribution function, and density estimation. The set $\mathcal{P}$ is different for these three cases. In the first case,which is classical, $\mathcal{P}$ is the class of all probability measures on $(\mathbf{R}, \mathcal{B})$. In the third case and, in fact also in the second, we work instead with the set of probability measures P on $(\mathbf{R}, \mathcal{B})$ which have a density f with respect to the Lebesgue measure. In the first two problems the set $\mathcal{P}$ is equipped with the weak topology and the natural tools are the use of tail free priors or a theorem of Schwartz(1965). In the third case the natural topology is that induced by the L_1 or the Hellinger metric. The natural tool is a new theorem that makes use of the notion of metric entropy or packing numbers for the space of densities in addition to one of Schwartz's conditions.

3 Notations and other technicalities

3.1 Nonparametrics

We start with the nonparametric problem. Let $\mathcal{P}$ be the class of all probability measures on $(\mathbf{R}, \mathcal{B})$; $\mathcal{P}$ be equipped with the weak topology and $\mathcal{B}_{\mathcal{P}}$ the corresponding Borel $\sigma-$ field. Equivalently, $\mathcal{B}_{\mathcal{P}}$ is the smallest $\sigma-$ field which makes the evaluation maps $P \mapsto P(A)$ measurable for each A in $\mathcal{B}$.

The most popular prior on $(\mathcal{P}, \mathcal{B}_{\mathcal{P}})$ is the Dirichlet process due to Ferguson(1973,1974). It is specified by its finite dimensional distributions as follows. Let α be a finite non zero measure on $(\mathbf{R}, \mathcal{B})$. Let $A_1, A_2, \ldots, A_k$ form a measurable partition. Then $P(A_1), P(A_2), \ldots, P(A_k)$ have a finite dimensional Dirichlet distribution with parameters $\alpha(A_1), \alpha(A_2), \ldots, \alpha(A_k)$. If $\alpha(A_i) > 0, i = 1, 2, \ldots, k$ then this distribution has a density with respect $(k-1)$ dimensional Lebesgue measure that has the form

$$\frac{\Gamma(\mathbf{R})}{\prod_1^k \Gamma(\alpha(A_i))} \prod_1^k p_i^{\alpha(A_i)-1}, \qquad\qquad 0 < p_i, \sum_1^k p_i = 1$$

If $k = 2$, one gets the beta distribution. Integrating out p_1 one gets

$$E(P(A_i)) = \alpha(A_i)/\alpha(\mathbf{R}) = \bar{\alpha}(\mathbf{A_i}). \tag{1}$$

It can be shown that the posterior given $X_1, X_2, \ldots, X_n$ is again a Dirichlet with $\alpha + \sum_1^n \delta_{X_i}$, in place of α, where δ_{X_i} is the point mass at X_i. Using this fact and (1), one gets immediately,

$$E\left(P(A)|X_1, X_2, \ldots, X_n\right) = \frac{\alpha(\mathbf{R})}{\alpha(\mathbf{R}) + \mathbf{n}} \bar{\alpha}(A) + \frac{n}{\alpha(\mathbf{R}) + \mathbf{n}} \left(\frac{1}{n} \sum \delta_{X_i}(A)\right) \tag{2}$$

which is a convex combination of the prior guess $\bar{\alpha}(A)$ and the frequentist nonparametric maximum likelihood estimate $P_n(A) = \frac{1}{n} \sum \delta_{X_i}(A)$. The weights reflect the Bayesian's confidence in prior guess. One can elicit or choose $\bar{\alpha}(.)$ and $\alpha(\mathbf{R})$ — and hence $\alpha(.)$— from these considerations.

We denote the Dirichlet process by D_α.

Proposition. If Π is D_α and B is a weak neighborhood of true P_0, then $\Pi\left(B|X_1, X_2, \ldots, X_n\right) \to 1$ a.s. (P_0), i.e., posterior consistency holds for all P_0.

At the heart of this fact is the property of being tailfree,vide Ferguson(1974), which allows one to reduce an infinite dimensional problem to a finite dimensional problem and invoke posterior consistency for the latter. This idea as well as the introduction of Dirichlet for another infinite dimensional problem goes back to Freedman(1963).

3.2 Semiparametrics

We start with a famous example of Diaconis and Freedman(1986). Suppose we wish to make inference about θ and $P_\theta(.) = P(. - \theta)$ where θ is real and $P(.)$ is symmetric around zero. To put a prior distribution for P_θ one first chooses a P' using a D_α, symmetrizes P' to get P and independently chooses θ. Diaconis and Freedman(1986) show that the posterior for θ need not be consistent in the weak topology.

Various people have observed that semiparametrics should involve probability measures with densities but the Dirichlet assigns probability one to the set of discrete measures. However choosing priors on densities is not enough.

Ghosal, Ghosh and Ramamoorthi(1998) have pointed out that one may argue that the Diaconis–Freedman counter example occurs because of the breakdown of the tailfree property. They show that posterior consistency can be proved provided a condition used by Schwartz(1965) holds. Priors for which posterior consistency holds are exhibited in Ghosal, Ghosh and Ramamoorthi(1998).

The version of Schwartz's(1965) theorem one has to use for this purpose is given below. We now work with $\mathcal{P}$ = the set of probability measures P having a density f with respect to Lebesgue measure. For two such probability measures P_1, P_2, with densities f_1, f_2 the Kullback–Leibler number $K(P_1, P_2)$ is defined as $\int_{\mathbf{R}} f_1 \log \frac{f_1}{f_2} dx$.

$K(P_1, P_2)$ is always ≥ 0 and may be ∞. It is not a metric but measures the divergence between P_1 and P_2 with the extreme tail of the density playing an important role.

Theorem 1 *Suppose P_0 belongs to the Kullback–Leibler support of Π, i.e., for all $\delta > 0$,*

$$\Pi\{K(P_0, P) < \delta\} > 0 \tag{3}$$

Then $\Pi\,(B|X_1, X_2, \ldots, X_n) \to 1$ a.s. (P_0), for all weak neighborhoods B of P_0.

As Ghosal, Ghosh and Ramamoorthi(1998) show property(3)—unlike the tail-free property — continues to hold even with the addition of a finite dimensional parameter.

For later reference as well as completeness we record Schwartz's(1965) theorem in its original form and an extension due to Barron(1988,1998).

Theorem 2 *Let Π be a prior on $\mathcal{P}$, and $P_0 \in B$. Assume the following conditions:*

1. *$\Pi(K(P_0, P) < \delta) > 0$ for all $\delta > 0$;*
2. *There exists a uniformly consistent sequence of tests for testing $H_0 : P = P_0$ vs. $H_1 : P \in B^c$, i.e., there exists a sequence of tests $\phi_n(X_1, X_2, \ldots, X_n)$ such that as $n \to \infty$,*

$$E_{P_0}\phi_n(X_1, X_2, \ldots, X_n) \to 0 \text{ and } \inf_{P \in B^c} E_P\phi_n(X_1, X_2, \ldots, X_n) \to 1.$$

Then $\Pi(B|X_1, X_2, \ldots, X_n) \to 1$ a.s. P_0.

Theorem 3 ((Barron(1988,1998))) *Let Π be a prior on $\mathcal{P}$, and P_0 be in $\mathcal{P}$ and B be a neighborhood of P_0. Assume that $\Pi(K(P_0, P) < \delta) > 0$ for all $\epsilon > 0$. Then the following are equivalent.*

1. *There exists a β_0 such that*

$$P_0\{\Pi(B^c|X_1, X_2, \ldots, X_n) > e^{-n\beta_0} \text{ infinitely often}\} = 0;$$

2. *There exist subsets V_n, W_n of $\mathcal{P}$, positive numbers $c_1, c_2, \beta_1, \beta_2$ and a sequence of tests $\{\phi_n(X_1, X_2, \ldots, X_n)\}$ such that*
   - *(a) $B^c = V_n \cup W_n$,*
   - *(b) $\Pi(W_n) \leq C_1 e^{-n\beta_1}$,*
   - *(c) $P_0\{\phi_n(X_1, X_2, \ldots, X_n) > 0$ infinitely often$\} = 0$ and $\inf_{P \in V_n} E_P\phi_n \geq 1 - c_2 e^{-n\beta_2}$.*

4 Density estimation

4.1 Dirichlet mixture of Normals

We illustrate with what seems to be currently the most popular and successful Bayesian method, first proposed by Lo(1984) and implemented in the early nineties via Markov Chain Monte Carlo(1994) by Escobar, Mueller and West (94).

Choose a random $P' \sim D_\alpha$. Since P' is discrete, as observed before,form a convolution with a normal density $N(0,h)$. Let $P = P' * N(0,h)$.

Since the smoothness of P depends on h and one does not know how much smoothness is right, put a prior(usually, inverse gamma)on h also. This completes the specification of a prior,which is often called a Dirichlet mixture of normal. It turns out that for MCMC to be feasible one needs α also to be normal. Simulations and heuristic calculations show that one can improve the rate of convergence by adding a location and scale parameter to α and by putting a prior on these parameters also. The following discussion can handle these refinements as well as general nonnormal α. However for the normal α,one can supplement the discussion below with non trivial heuristic argument that throws light on how convergence takes place. For lack of space the heuristic argument will not be given.

4.2 Posterior consistency for general priors

The basic theorem is the following which improves on an earlier result of Barron,Schervish and Wasserman(1997).

Let $\mathcal{P}_0 \subset \mathcal{P}$. For $\delta > 0$, the L_1- metric entropy of $\mathcal{P}_0$, denoted by $J(\delta, \mathcal{P}_0)$ is $\log a(\delta)$, where $a(\delta)$ is the minimum over all k such that there exist $P_1, P_2, \cdots, P_k$ in $\mathcal{P}$ with $\mathcal{P}_\prime \subset \cup_1^k \{P : \|P - P_i\|_1 < \delta\}$.

Theorem 4 (Ghosal,Ghosh and Ramamoorthi) *Let Π be a prior on $\mathcal{P}$. If $P_0 \in \mathcal{P}$ and $\Pi(K(P_0,P) < \epsilon) > 0$ for all $\epsilon > 0$. If for each $\epsilon > 0$ there is a $\delta < \epsilon, c_1, c_2 > 0, \beta < \frac{\epsilon^2}{2}$ and also $\mathcal{P}_n$ such that*

1. $\Pi(\mathcal{P}_n^c) < C_1 e^{-n\beta_1}$ *for large n*

2. $J(\delta, \mathcal{P}_n) < n\beta$

then $\Pi(B|X_1, X_2, \ldots, X_n) \rightarrow 1$ a.s.$P_0 n$ for all L_1-neighborhoods B of P_0.

The proof of this theorem is based on the result of Barron recorded in Section 3. The first assumption is the condition assumed in Theorem 1 in Section 3 while the two remaining assumptions take care of conditions(2) and (3) of Barron's Theorem.

4.3 Application to Dirichlet mixture of normals

One has to have two sets of tools to verify the two conditions in Theorem 4. The set or sieve $\mathcal{P}_n$ for verifying the condition is: fix a δ and β as in the theorem then

$$\mathcal{P}_n = \left\{ P = P' * N(0,h); P'[-\sqrt{n}, \sqrt{n}] > 1 - \delta, h > \frac{c(\delta,\beta)}{\sqrt{n}} \right\}$$

Various sufficient conditions which entail application of Theorem 4 are given in Ghosal, Ghosh and Ramamoorthi(97. For example if P_0 is smooth unimodal with finite Shannon entropy and compact support, like the uniform on $[a, b]$ then P_0 belongs to the Kullback–Leibler support of the prior. For unbounded support the tails of P_0 and $\bar{\alpha}$ have to be compatible in a certain way.

4.4 Concluding Remarks

Theorem 4 can also be used to study posterior consistency for Gaussian process priors and Bayesian histograms (Barron(1988,1998) and Ghosh and Ramamoorthi(1998)).

One may also ask whether the Bayes estimate $E(P|X_1, X_2, \ldots, X_n)$ is consistent. It is easy to show that posterior consistency in the weak topology or the topology induced by L_1 norm implies Bayes consistency.

One may also ask questions about rates of convergence and non-informative or default priors which attain a minimax rate of convergence for the posterior or Bayes estimates. This issue is currently under investigation by Ghosal,Ghosh and van der Vaart and by Wassserman and Shen.

A final important remark. In recent work Barron(1998) shows if we focus on the cumulative Kullback-Leibler predictive loss (also called the entropy loss) an elegant consistency theory can be built up using only Kullback-Leibler support.

References

Barron, A. R. (1986). On uniformly consistent tests and Bayes consistency. Unpublished manuscript.

Barron, A. R., Schervish, M. and Wasserman, L. (1996). The consistency of posterior distributions in non parametric problems. Preprint.

Barron, A. R. (1998). Information-theoretic characterizations of Bayes Performance and the Choice of Priors in Parametric and Nonparametric Problems.*Bayesian Statistics 6*, Editors *J.M. Bernardo,J.O. Berger, A.P. Dawid and A.F.M. Smith* Oxford University press

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

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

Ferguson, T. (1974). Prior distribution on the spaces of probability measures. *Ann. Statist.* 2 615-629.

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

Ghosal, S., Ghosh, J. K. and Ramamoorthi, R. V. (1998). Consistent semiparametric estimation about a location parameter. Submitted.

Ghosal, S., Ghosh, J. K. and Ramamoorthi, R. V. (1997b). Posterior consistency of Dirichlet mixtures in density estimation. Technical report # WS-490, Vrije Universiteit, Amsterdam.

S.Ghosal, J.K. Ghosh and R.V. Ramamoorthi (1997) Consistency issues in Bayesian Nonparametrics . *Asymptotics, Nonparametrics and Time series - a tribute to M.L. Puri* Edited by S.Ghosh.Marcel Dekker

Ghosh, J. K. and Ramamoorthi R. V. (1998). *Lecture notes on Bayesian asymptotics.* Under preparation.

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

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

West, M., Muller, P. and Escobar, M. D. (1994). Hierarchical priors and mixture models, with applications in regression and density estimation. In *Aspects of uncertainty: A Tribute to D. V. Lindley.* 363–386.

Jayanta K. Ghosh
Indian Statistical Institute
203 B.T. Road
Calcutta 700 035
India