The Annals of Statistics
1984, Vol. 12, No. 1, 351–357

ON A CLASS OF BAYESIAN NONPARAMETRIC ESTIMATES: I. DENSITY ESTIMATES[1]

BY ALBERT Y. LO

University of Pittsburgh and Rutgers University

Given a positive, normalized kernel and a finite measure on an Euclidean space, we construct a random density by convoluting the kernel with the Dirichlet random probability indexed by the finite measure. The posterior distribution of the random density given a sample is classified. The Bayes estimator of the density function is given.

1. Introduction and summary. The problem of density estimation from a nonBayesian point of view has been considered by a number of authors (for a good list of references, see Tapia and Thompson, 1978). Here we consider the density estimation problem from a Bayesian point of view and obtain such estimates under a weighted squared error-loss and when random density function is generated by convoluting a known (density) kernel with a Dirichlet process (Ferguson, 1973). The next section presents the notation and the statistical description of the problem and gives the Bayes estimator $\hat{f}$ of the density function. In the same section, some computational aspects of $\hat{f}$ are also given. The third section discusses the choice of K and α. This is illustrated with a few numerical examples.

2. Prior and posterior distributions. Throughout the paper, let $\mathscr{X}$, $\mathscr{R}$ both be Borel subsets of Euclidean spaces and $\mathscr{F}$, $\mathscr{B}$ the σ-fields generated by the open sets relative to $\mathscr{X}$ and $\mathscr{R}$ respectively. Let α denote a finite measure on $(\mathscr{R}, \mathscr{B})$. Let $K(x, u)$ denote a nonnegative valued kernel defined on $(\mathscr{X} \times \mathscr{R}, \mathscr{F} \otimes \mathscr{B})$ such that for each $u \in \mathscr{R}$, $\int_{\mathscr{X}} K(x, u)\, dx = 1$ and for each $x \in \mathscr{X}$, $\int_{\mathscr{R}} K(x, u)\alpha(du) < \infty$. Let Θ be the space of distributions on $(\mathscr{R}, \mathscr{B})$ and $\mathscr{A}$ the σ-field on Θ generated by weak convergence. For each $G \in \Theta$, we define $f(x \mid G) = \int_{\mathscr{R}} K(x, u)G(du)$, $x \in \mathscr{X}$. Then $f(\cdot \mid G)$ is a density function by virtue of Fubini's theorem. Let $\mathscr{L}_K$ be the family of $f(\cdot \mid G)$'s when G varies in Θ. That $\mathscr{L}_K$ is made rich by a suitable choice of $K(\cdot, \cdot)$ is well known (see Section 3). It can also be shown that $f(x \mid G)$ is $\mathscr{A}$-measurable by first showing this when K is an indicator and then passing to the limit. Let $\mathscr{P}_\alpha$ be a Dirichlet probability on $(\Theta, \mathscr{A})$ with index measure α (Ferguson, 1973). This is denoted by $G \sim \mathscr{P}_\alpha$. Note that for each $x \in \mathscr{X}$, $\int_\Theta f(x \mid G)\mathscr{P}_\alpha(dG) = \int_{\mathscr{R}} K(x, u)(\alpha(du)/\alpha(\mathscr{R}))$. This is the marginal density of x and it is also the Bayes estimator for the no sample problem with respect to the quadratic type loss function.

We denote the posterior distribution of the parameter G given a sample X_1,

Received August 1978; revised September 1983.

[1] This research is supported in part by NSF Grant MCS-80-03629 and NSF Grant MCS-81-02523. The first draft was done at the University of California, Berkeley.

AMS 1980 *subject classifications.* Primary, 62A15; secondary, 62C10, 62G05.

Key words and phrases. Decision theory, Bayesian nonparametric method, density estimation.

$\cdots$, X_n by $\mathscr{P}^n$. The Bayes estimator of the density function $f(x \mid G) = \int K(x, u)G(du)$ under squared error loss can be obtained as an application of

LEMMA 1. *Let g be a positive valued or quasiintegrable (with respect to the joint probability of $\mathscr{P}_\alpha$ and G) function defined on $(\mathscr{R} \times \Theta, \mathscr{B} \otimes \mathscr{A})$, Then*

$$\int_\Theta \int_{\mathscr{R}} g(u, G)G(du)\,\mathscr{P}_\alpha(dG) = \int_{\mathscr{R}} \int_\Theta g(u, G)\,\mathscr{P}_{\alpha+\delta_u}(dG)\,\frac{\alpha(du)}{\alpha(\mathscr{R})} \tag{2.1}$$

where δ_u is the Dirac probability measure degenerate at u.

PROOF. Apply Theorem 1 of Section 3 of Ferguson (1973) and a limiting argument. □

We shall use Lemma 1 to evaluate the posterior distributions. To simplify our notation, we define the measure $\mu_{n,K,\alpha}$ on $(\mathscr{R}^n, \mathscr{B}^n)$, the n-folded product measure space of $(\mathscr{R}, \mathscr{B})$, by $\mu_{n,K,\alpha}(d\mathbf{u}) = [\prod_{i=1}^n K(X_i, u_i)] \prod_{i=1}^n (\alpha + \sum_{j=1}^{i-1} \delta_{u_j})(du_i)$, where $\mathbf{u} = (u_1, \cdots, u_n)$. We denote the conditional expectation given the observations $X_1, \cdots, X_n$ by $E^{\mathbf{X}}$.

THEOREM 1.

$$E^{\mathbf{X}}g(G) = \frac{\int_{\mathscr{R}^n} \int_\Theta g(G)\,\mathscr{P}_{\alpha+\sum_{i=1}^n \delta_{u_i}}(dG)\mu_{n,K,\alpha}(d\mathbf{u})}{\int_{\mathscr{R}^n} \mu_{n,K,\alpha}(d\mathbf{u})}, \quad n \geq 1 \tag{2.2}$$

for any nonnegative or integrable function g on $(\Theta, \mathscr{R})$.

PROOF. We shall evaluate the numerator of (2.2) by repeated application of Lemma 1. We first write the numerator of (2.2) as

$$\int_\Theta g(G) \prod_{i=1}^n \int_{\mathscr{R}} K(x_i, u_i)G(du_i)\,\mathscr{P}_\alpha(dG)$$

$$= \int_\Theta \int_{\mathscr{R}} K(X_1, u_1)g(G) \prod_{i=2}^n \int_{\mathscr{R}} K(X_i, u_i)G(du_i)G(du_1)\,\mathscr{P}_\alpha(dG).$$

It has been noted earlier that the integrand is measurable with respect to $\mathscr{B} \otimes \mathscr{A}$. Thus Lemma 1 is applicable. The above quantity is equal to

$$\int_{\mathscr{R}} \int_\Theta K(X_1, u_1)g(G) \prod_{i=2}^n \int_{\mathscr{R}} K(X_1, u_i)G(du_i)\,\mathscr{P}_{\alpha+\delta_{u_1}}(dG)\,\frac{\alpha(du_1)}{\alpha(\mathscr{R})}$$

$$= \int_{\mathscr{R}^n} \int_\Theta g(G)_{\alpha+\sum_{i=1}^n \delta_{u_i}}(dG) \prod_{i=1}^n K(X_i, u_i) \prod_{i=1}^n \frac{(\alpha + \sum_{j=1}^{i-1} \delta_{u_j})(du_i)}{\alpha(\mathscr{R}) + i - 1}.$$

Applying this argument a total of n times gives the numerator, and replacing $g(G)$ by 1 to obtain a similar result for the denominator completes the proof. □

The evaluation of the Bayes rules via (2.2) is rather similar for several types of functionals of G and we give the derivation of the Bayes rule of the density

function below to illustrate the idea. We need the following combinatoric result to derive the Bayes rule for $f(x \mid G)$ under weighted squared-error loss in Theorem 2 below. Let α be a finite measure on $\mathscr{R}$, m a positive integer and g_i, $i = 1, \cdots, m$ are positive or α-integrable functions. Let $\mathbf{P}$ be a partition of $1, \cdots, m$ and $N(\mathbf{P})$ be the number of cells in the partition. Thus, $\mathbf{P} = \{C_i, i = 1, \cdots, N(\mathbf{P})\}$, where C_i is the ith cell of the partition. Let e_i be the number of elements in C_i. Note that C_i and e_i, $i = 1, \cdots, N(\mathbf{P})$ depend on $\mathbf{P}$.

LEMMA 2.

$$\int_{\mathscr{R}^m} \prod_{i=1}^m g_i(u_i) \prod_{i=1}^m (\alpha + \sum_{j=1}^{i-1} \delta_{u_j})(du_i) = \sum_{\mathbf{P}} \phi(\mathbf{P}) \tag{2.3}$$

where

$$\phi(\mathbf{P}) = \prod_{i=1}^{N(\mathbf{P})} \left\{(e_i - 1)! \int_{\mathscr{R}} \prod_{\ell \in C_i} g_\ell(u)\alpha(du)\right\} \tag{2.4}$$

and the sum is over all possible partitions $\mathbf{P}$ *of* $\{1, \cdots, m\}$.

PROOF. The left hand side of (2.3) can be written as a sum of $m!$ terms. These terms, which are in fact integrals, can be grouped according to the number of distinct u_i, doubles in $\{u_1, \cdots, u_m\}$, triplets in $\{u_1, \cdots, u_m\}$, $\cdots$. Grouping these integrals according to the possible choices of u_i, it can be seen that the integrals are of the type

$$\prod_{i=1}^{N(\mathbf{P})} \int_{\mathscr{R}} \prod_{\ell \in C_i} g_\ell(u)\alpha(du). \tag{2.5}$$

There are a total of $(e_i - 1)!$ ways to obtain the above integral for $i = 1, \cdots, N(\mathbf{P})$, hence the total number of ways to obtain (2.5) is $\prod_{i=1}^{N(\mathbf{P})} (e_i - 1)!$ and thus (2.3) is obtained. □

THEOREM 2.

$$\begin{aligned} &E^{\mathbf{X}} f(x \mid G) \\ &= \frac{\alpha(\mathscr{R})}{\alpha(\mathscr{R}) + n} \left\{\int K(x, u) \frac{\alpha(du)}{\alpha(\mathscr{R})}\right\} \\ &\quad + \frac{n}{\alpha(\mathscr{R}) + n} \sum_{\mathbf{P}} W(\mathbf{P}) \sum_{i=1}^{N(\mathbf{P})} \frac{e_i}{n} \left\{\frac{\int K(x, u) \prod_{\ell \in C_i} K(X_\ell, u)\alpha(du)}{\int \prod_{\ell \in C_i} K(X_\ell, u)\alpha(du)}\right\} \end{aligned} \tag{2.6}$$

where $W(\mathbf{P}) = (\phi(\mathbf{P})/\sum_{\mathbf{P}} \phi(\mathbf{P}))$.

PROOF OF THEOREM 2. Expression (2.6) can be reduced to

$$\frac{1}{\alpha(\mathscr{R}) + n} \cdot \frac{\int_{\mathscr{R}^{n+1}} \mu_{n+1,K,\alpha}(d\mathbf{u})}{\int_{\mathscr{R}^n} \mu_{n,K,\alpha}(d\mathbf{u})}$$

by writing x as X_{n+1}. By Lemma 2, the numerator becomes $\sum_{\mathbf{P}'} \phi(\mathbf{P}')$ where $\mathbf{P}'$

is a partition of $\{1, \cdots, n+1\}$ and the denominator becomes $\sum_{\mathbf{P}} \phi(\mathbf{P})$ where $\mathbf{P}$ is a partition of $\{1, \cdots, n\}$. It remains to show that (2.6) is equal to $(1/\alpha(\mathscr{R}) + n) \cdot (\sum_{\mathbf{P}'} \phi(\mathbf{P}')/\sum_{\mathbf{P}} \phi(\mathbf{P}))$.

We shall expand $\sum_{\mathbf{P}'} \phi(\mathbf{P}')$ as a sum over $\mathbf{P}$. To do this, observe that each $\mathbf{P} = \{C_1, \cdots, C_{N(\mathbf{P})}\}$ can be identified with $\{C_0, C_1, \cdots, C_{N(\mathbf{P})}\}$ with C_0 being the empty set. In this case each $\mathbf{P}$ corresponds to $N(\mathbf{P}) + 1$ different $\mathbf{P}'$'s by assigning the element $n+1$ to each of $C_0, C_1, \cdots, C_{N(\mathbf{P})}$. Also note that if $\mathbf{P}_1$ and $\mathbf{P}_2$ are two different partitions of $\{1, \cdots, n\}$, the $N(\mathbf{P}_1) + 1$ partitions generated by $\mathbf{P}_1$ and the $N(\mathbf{P}_2) + 1$ partitions generated by $\mathbf{P}_2$ are all different. By running $\mathbf{P}$ over the class of all partitons of $\{1, \cdots, n\}$, we obtain the class of all partitions of $\{1, \cdots, n+1\}$. Using this, $\sum_{\mathbf{P}'} \phi(\mathbf{P}')$ can be written as

$$\sum_{\mathbf{P}} \Bigg\{ \int K(X_{n+1}, u)\alpha(du) \prod_{i=1}^{N(\mathbf{P})} (e_i - 1)! \int \prod_{\ell \in C_i} K(X_\ell, u)\alpha(du) + \sum_{i=1}^{N(\mathbf{P})} e_i! \int K(X_{n+1}, u) \prod_{\ell \in C_i} K(X_\ell, u)\alpha(du) \prod_{j \neq i}^{N(\mathbf{P})} (e_j - 1)! \int \prod_{\ell \in C_j} K(X_\ell, u)\alpha(du) \Bigg\}.$$

Rewriting X_{n+1} as x and simplifying, we arrive at (2.6). □

REMARK 2.1. A referee pointed out that the computational aspect of the expression (2.2), which appeared first in Lo (1978) had been investigated independently by Kuo (1980) and Ghorai and Rubin (1982). It was known that both the numerator and the denominator of (2.2) could be reduced to a sum of products of single integrals. In particular, Kuo (1980) noticed that the number of summands in this sum is equal to the Bell's exponential number and proposed a Monte Carlo method for its evaluation whereas Berry and Christensen (1979) gave approximations to cut down on computations. On the other hand, Ghorai and Rubin (1982) obtained Lemma 2 by mathematical induction.

3. The choice of K, α and some examples. The choice of K defines the model and is the difficult part. One easy case is that of the histogram model obtained by choosing $K(x, u) = \sum_j (1/\ell(\Delta_j))I_{\{x \in \Delta_j, u \in \Delta_j\}}$ in $\mathscr{L}_K$, where $\{\Delta_j\}$ is a partition of $\mathscr{X} = \mathscr{R}$ and ℓ is Lebesgue measure. In this case expression (2.6) readily reduces to the usual Bayes estimates of the cell probabilities. Other determinations of K are more difficult. Roughly speaking, $\{K(\cdot, u), u \in \mathscr{R}\}$ are the extreme points of the convex family $\mathscr{L}_K$ and hence the determination of K rests on the existence of a Choquet-type theorem representing elements of $\mathscr{L}_K$. In the following we list the available results. A broad model is also given in case no such theorem is available.

Suppose $\mathscr{L}_K$ is the family of scale mixtures of the uniform densities. A theorem of Klintchine and Shepp says $\mathscr{L}_K$ is the family of all decreasing densities supported by $[0, \infty]$ (Feller 1970, page 158). For the same reason, the family of unimodal densities with mode at zero is equal to the family of mixtures of uniform

$(-u_1, u_2)$ densities with $u_1 > 0$ and $u_2 > 0$, which further reduces to the family of symmetric and unimodal densities if $u_1 = 0$. On the other hand, a theorem of Bernstein says the family of all completely monotone densities is characterized by its exponential extreme points. Furthermore, this last class is a smooth subclass of the family of all decreasing densities. The characterization of $\mathscr{L}_K$ of scale mixtures of normal has been studied by a number of authors; see in particular Theorem 2 of Eaton (1981) which also contains some interesting results in the multivariate case.

In view of the lack of Choquet-type theorems for the models considered here, a $\mathscr{L}_K$ which is rich enough to contain all densities in its closure is quite desirable. It can be shown that in the real line case such a family consists of densities of the form

$$f(x \mid G) = \int_0^\infty \int_{-\infty}^\infty \tau K(\tau(x - u))G(du, d\tau) \tag{3.1}$$

where $K(\cdot)$ is a prescribed nondegenerated density on $(-\infty, \infty)$. Here we anticipate the problem of choosing K parallels that of choosing K for the classical kernel estimates. Details and references of the latter are available in Tapia and Thompson (1978). Nevertheless, we find that a standard normal K to be convenient in practice.

Expression (2.6) suggests that once K is chosen, the choice of α should be based on the usual prior-posterior analysis when sampling from the $K(\cdot, u)$ densities with $\alpha(du)/\alpha(\mathscr{R})$ being the prior distribution of the "parameter" $u \in \mathscr{R}$. Consequently, for computational purposes, $\alpha(du)/\alpha(\mathscr{R})$ would be a member of the conjugate family of priors when sampling from the $K(\cdot, u)$ densities (DeGroot, 1970, Chapter 9). In this case, $\phi(\mathbf{P})$ is given by (2.4) and expression (2.6) can be reduced to integral-free forms. The following examples concerning densities on the real line are made to illustrate this. Computations are carried out by an IBM 360 system via a subroutine that generates all partitions of the set $\{1, \cdots, n\}$.

A sample of size 8 (−1.47, −0.85, −0.52, −0.11, 0.058, 0.39, 0.985, 1.61) from a standard normal density is drawn and three estimates are plotted in Figure 1 using the following three models: location mixtures of normal kernels with fixed variance 1, scale mixtures of normal kernels with fixed location 0, and location and scale mixtures. The α's are the $N(0, 1)$ distribution and the gamma (1, 1) distribution in the first two models. The last model corresponds to (3.1) and hence α is supported by $[0, \infty) \times (-\infty, \infty)$ with joint distribution specified by: $\tau \sim$ gamma (1, 1) and given τ, $u \sim N(0, 1/\tau)$.

Figure 2 gives estimates based on a sample of size 8 (0.08, 0.12, 0.42, 0.61, 0.96, 1.21, 1.40, 2.30) from a standard exponential density using scale mixtures of uniform kernel and scale mixtures of exponential kernels. The corresponding α's are Pareto distributions with parameters 0.1 and 1, and the gamma (1, 1) distribution. The two irregular points on the curve obtained by using uniform kernels are caused by the two largest observations 1.40 and 2.30, whereas these do not seem to affect the smoothness of the second estimate.

In Figure 3, the two-sample data (Control: 6.8, 3.1, 5.8, 4.5, 3.3, 4.7, 4.2, 4.9;

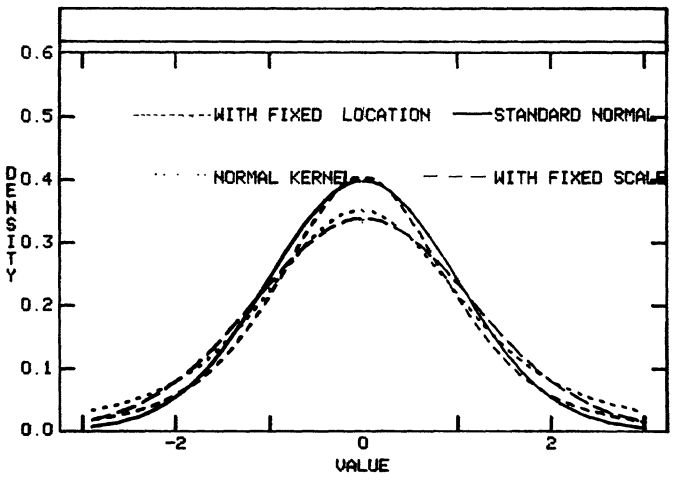

FIG. 1.

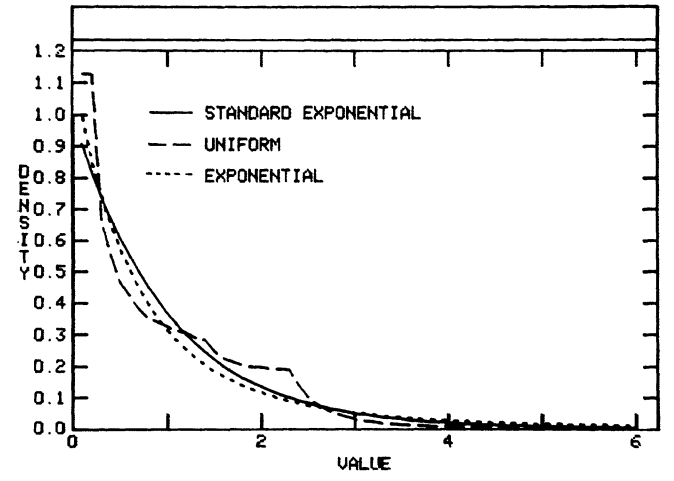

FIG. 2.

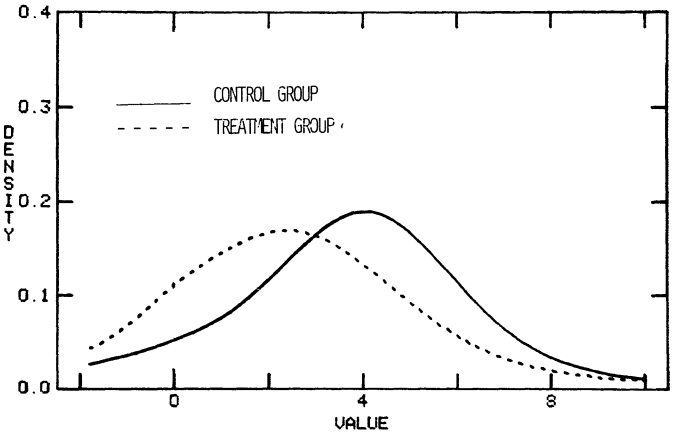

FIG. 3.

Treatment: 4.4, 2.5, 2.8, 2.1, 6.6, 0.0, 4.8, 2.3) of Lehmann (1974, page 37) are used to derive estimates for the model specified by (3.1) with joint distribution of α being identical to that of Figure 1.

4. Concluding remarks. Asymptotic theory is not available (besides the result of Doob (1949) which says that one always has consistency for almost all parameter values) except for bounded K in the convolution model (this statement applies also to Lo, 1980). In this latter case, convergence of the posteriors to a limiting distribution is achieved by $\sqrt{n}$ rescaling (Lucien Le Cam, private conversation).

The relatively simple form of expression (2.7) is promising for solving the following problems from other viewpoints. (I) The minimax rule of this model may be derived by choosing an appropriate sequence of $\alpha(du)$'s and exploiting a constant risk type argument. (II) Deflating the prior $\alpha(du)/\alpha(\mathscr{R})$ appropriately, perhaps a "heuristic" or a maximum likelihood estimate can be uncovered (the existence and uniqueness of the latter are guaranteed by a convexity argument).

Acknowledgment. The problem was suggested by David Blackwell. Looking at density estimation as a convolution operation was revealed in a conversation with Lucien Le Cam. Referees' comments helped improve the presentation. I thank them all.

REFERENCES

BERRY, D. and CHRISTENSEN, R. (1979). Empirical Bayes estimation of a binomial parameter via mixture of Dirichlet processes. *Ann. Statist.* **7** 558–568.

DE GROOT, M. (1970). *Optimal Statistical Decisions*. McGraw, New York.

DOOB, J. (1949). Application of the theory of martingales. *Coll. Int. du CNRS*. Paris, 22–28.

EATON, M. (1981). On the projection of isotropic distributions. *Ann. Statist.* **9** 391–400.

FELLER, W. (1970). *An Introduction to Probability Theory and Its Applications, Vol. II*. Wiley, New York.

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

GHORAI, J. K. and RUBIN, H. (1982). Bayes risk consistency of nonparametric Bayes density estimates. *Australian J. Statist.* **24** 51–66.

KUO, L. (1980). Computations of mixtures of Dirichlet processes. Technical Report No. 96, Department of Statistics, University of Michigan, Ann Arbor.

LEHMANN, E. (1975). *Nonparametric Statistical Methods Based on Ranks*. Holden-Day, New York.

LO, A. Y. (1978). Bayesian nonparametric density methods. Technical Report, University of California, Berkeley.

LO, A. Y. (1980). On a class of Bayesian nónparametric estimates: II rate function estimates. Technical Report, Rutgers University.

TAPIA, R. A. and THOMPSON, J. R. (1978). Nonparametric probability density estimation. Johns Hopkins University Press, Baltimore, Maryland.

DEPARTMENT OF STATISTICS
RUTGERS UNIVERSITY
BUSCH CAMPUS
NEW BRUNSWICK, NEW JERSEY 08903