BAYESIAN DENSITY ESTIMATION BY MIXTURES OF NORMAL DISTRIBUTIONS[1]

Thomas S. Ferguson

Department of Mathematics
University of California
Los Angeles, California

I. INTRODUCTION

This paper is concerned with the estimation of an arbitrary density f(x) on the real line. We model this density as a mixture of a countable number of normal distributions in the form

$$f(x) = \Sigma_1^\infty \; p_i h(x|\mu_i,\sigma_i), \tag{1}$$

where $h(x\,|\mu,\sigma)$ is the density of $N(\mu,\sigma^2)$, the normal distribution with mean μ and variance σ^2. There are a countably infinite number of parameters of the model, $(p_1,p_2,\ldots,\mu_1,\mu_2,\ldots,\sigma_1,\sigma_2,\ldots)$. Using such mixtures, any distribution on the real line can be approximated to within any preassigned accuracy in the Lévy metric, and any density on the real line can be approximated similarly in the L_1 norm. Thus the problem may be considered nonparametric.

Let $x_1,\ldots,x_n$ represent a sample of size n from f(x). Consider the problem of estimating f(x) at some fixed x or of estimating some functional of f(x) such as the mean, $\int xf(x)dx$,

[1] *This research was partially supported by the National Science Foundation under Grant MCS77-2121.*

ISBN 0-12-589320-5

using squared error loss based on $x_1,\ldots,x_n$. We consider a Bayesian approach to this problem. This entails placing a joint distribution on the parameters of the model and attempting to evaluate the posterior expectation of $f(x)$ or $\int xf(x)dx$ given the sample.

There are several advantages of such an approach. *1. Use of prior information.* It gives the statistician a formal method of combining some of his prior information with the data. *2. Consistency.* The argument of Doob (1948) shows that these estimates should be consistent for *almost all* f chosen by the prior. However, a direct consistency result seems difficult to obtain and rates of convergence look even more difficult, even though direct consistency and rates of convergence are easy to obtain for other methods of density estimation such as kernel estimation. *3. Automatic adaptation.* Asymptotic theory for kernel estimators involves problems of letting the window size tend to zero at some rate as the sample size tends to infinity. Such problems are automatically taken care of in the Bayesian framework. In particular, larger windows for more remote observations are seen to occur naturally. *4. Small sample optimality.* Classical methods have only a large sample justification and look rather ad hoc if the sample is small. However, Bayes estimates with squared error loss are generally admissible.

We are not interested in estimating the parameters of the model, $p_1,p_2,\ldots,\mu_1,\mu_2,\ldots,\sigma_1,\sigma_2,\ldots$. It would not make much sense to do so since the parameters are not identifiable. One may attempt to obtain identifiability by writing (1) in the form

$$f(x) = \int h(x\,|\,\mu,\sigma)\,dG(\mu,\sigma) \tag{2}$$

where G is the probability measure on the half-plane $\{(\mu,\sigma): \sigma>0\}$ that gives mass p_i to the point (μ_i,σ_i), $i=1,2,\ldots$. It has been shown by Teicher (1960) that if G is restricted to the class of finite probability measures, then G is identifiable, but that if G is unrestricted then G is not identifiable. The identifiability of G when G is restricted to the class of countable probability measures is still an open question.

II. THE PRIOR

We describe the prior distribution of $(p_1,p_2,\ldots,\mu_1,\mu_2,\ldots,\sigma_1,\sigma_2,\ldots)$ as follows:

(a) $(p_1,p_2,\ldots)$ and $(\mu_1,\mu_2,\ldots,\sigma_1,\sigma_2,\ldots)$ are independent.

(b) Let $q_1,q_2,\ldots$ be i.i.d. q_i having the beta distribution, Be(M,1) (i.e. the common density of the q_i is $Mq^{M-1}I_{[0,1]}(q)$), and let $p_1 = 1-q_1$, $p_2 = q_1(1-q_2),\ldots,p_j = (\Pi_{i=1}^{j-1}q_i)(1-q_j),\ldots$.

(c) (μ_1,σ_1), $(\mu_2,\sigma_2),\ldots$ are i.i.d. with common distribution the usual gamma-normal conjugate prior for the two-parameter normal distribution, namely, the precision (reciprocal of variance) $\rho_i = 1/\sigma_i^2$ has the gamma distribution $G(\alpha,2/\beta)$ (i.e. the density of ρ_i is $[1/\Gamma(\alpha)](\beta/2)^{\alpha}e^{-\rho\beta/2}\rho^{\alpha-1}I_{(0,\infty)}(\rho)$) and given ρ_i,μ_i is distributed as the normal distribution with mean μ and precision $\rho_i\tau$.

There are five parameters of the prior, $M>0$, $\alpha>0$, $\beta>0$, μ, and $\tau>0$. Note that the distribution of $(p_1,p_2,\ldots)$ depends only on M, and that the distribution of $(\mu_1,\sigma_1,\mu_2,\sigma_2,\ldots)$ depends only on α, β, μ and τ.

The prior guess at f(x), denoted by $f_0(x)$, is the expectation of f(x) under the prior distribution.

$$f_0(x) = Ef(x) = \Sigma_1^{\infty}Ep_i\ Eh(x|\mu_i,\sigma_i) = Eh(x|\mu,\sigma)$$

$$= \frac{\Gamma(\alpha+\frac{1}{2})}{\Gamma(\alpha)\Gamma(\frac{1}{2})}\sqrt{\frac{\tau}{(\tau+1)\beta}}\left(1+\frac{\tau}{(\tau+1)\beta}(x-\mu)^2\right)^{-(\alpha+\frac{1}{2})}$$

With the change of variable $y = \sqrt{\frac{2\alpha\tau}{(\tau+1)\beta}}(x-\mu)$, y has a t-distribution with 2α degrees of freedom (in the generalized sense since α need not be rational). The mean of $f_0(x)$ is μ and the variance is $(\tau+1)\beta/\ (2\tau(\alpha+1))$. Thus the prior distribution does not admit a prior guess that is not symmetric. A more general prior guess may be achieved through the use of mixtures, but the resulting formulas become more complicated and are not investigated in this paper.

Interpretation of M. There are two somewhat independent interpretations of M. The first concerns the *relative sizes of the probabilities,* p_i. A small value of M means there is a big difference in the p_i; generally, p_1 is large compared to p_2,p_2 is large compared to p_3,... etc. If M is large, there will be many small probabilities that tail off to zero slowly. As an aid to understanding this feature of M, Table I has been constructed which shows the expected values, standard deviations and the correlation of the two largest p_i.

The other interpretation of M is as *prior information.* A small M means that you don't trust your prior guess much (the estimate will be strongly influenced by the observations), and a large value of M means you do trust your prior guess (the estimate does not depend much on the observations.) In this regard, M is measured in units of sample size: M represents the number of observations for which you would be willing to trade your prior information.

Thus it appears that this prior cannot express the opinions of a persons who believes strongly that there is a big difference in the probabilities p_i. We will see below in (6) an estimate of f(x) in which the influence of M as prior information is at least partially removed.

Choice of prior parameters. The following considerations are an aid to choosing the parameters of the prior to express the opinions of the statistician.

Since $E\mu_i = \mu$, μ should be chosen as the statistician's prior guess at the center of mass. Since Var $\mu_i = E\sigma_i^2/\tau$, τ

should be chosen approximately as $E\sigma_i^2/\text{Var}\ \mu_i$. If the uncertainty in the values of the μ_i is greater than (equal to, less than) the average variance, then τ should be chosen less than (equal to, greater than) 1.

Since $E\rho_i = 2\alpha/\beta$ (or $E\sigma_i^2 = \beta/2(\alpha-1)$ for $\alpha>1$), choose $2\alpha/\beta \sim E\rho_i$, leaving one parameter, say β, to be chosen in a way that reflects how diffuse the ρ_i are thought to be. Since $\text{Var}\ \rho_i = 4\alpha/\beta^2$, choose β large if the ρ_i are expected to be close to $2\alpha/\beta$ and choose β small if the ρ_i, and hence the σ_i^2, are expected to be diffuse.

One should choose M to reflect the statistician's belief on the relative sizes of the probabilities. For this, moderate values of M, from .5 to 5 say, are appropriate; Table I will aid this choice.

TABLE I. Expectations, Standard Deviations and Correlation of the two Largest Probabilities for Various Values of M, based on 10,000 Monte Carlo Trials

M	EP_1	EP_2	$s.d.P_1$	$s.d.P_2$	Corr.
.1	.938	.058	.124	.115	-.984
.2	.881	.101	.158	.134	-.961
.5	.756	.172	.192	.137	-.890
1	.625	.210	.192	.112	-.758
2	.476	.213	.164	.080	-.487
5	.296	.170	.106	.049	-.031
10	.195	.125	.070	.032	.238
20	.122	.085	.040	.020	.383

III. CONNECTION WITH THE DIRICHLET PROCESS

The prior distribution of the parameters $(p_1,p_2,\dots,\mu_1,\mu_2,\dots,\sigma_1,\sigma_2,\dots)$ has been chosen so that the distribution function, G, of (2) is a Dirichlet process with parameter $\alpha = MG_0$, where $G_0 = \mathcal{E}G$ is the conjugate prior for (μ,σ^2) for the normal distribution given in (c) of Section II. That this is so follows from the representation of the Dirichlet process given by Sethuraman and Tiwari (1981). Their representation of the Dirichlet process, G, with parameter MG_0 is of the form

$$G = \Sigma_1^\infty p_i \delta_{\theta_i}$$

where $(p_1,p_2,\dots)$ and $(\theta_1,\theta_2,\dots)$ are independent, the distribution of the p_i as in (b) of Section II, and $\theta_1,\theta_2,\dots$ are i.i.d. G_0. (The following discussion is quite general; for the specific case introduced in Section II, θ_i represents (μ_i,σ_i), $i=1,\dots$.) This is similar to but simpler than the representation in Ferguson (1973) which describes the distribution of the p_i by their order statistics $p_{(1)} \geq p_{(2)} \geq \cdots$. The actual distribution of $p_{(1)}$ is difficult to exhibit and even $\mathcal{E}p_{(1)}$ seems difficult to obtain, as pointed out by J.F.C. Kingman (1975). However the representation of Sethuraman and Tiwari makes it easy to evaluate $\mathcal{E}p_{(1)}, \mathcal{E}p_{(2)},\dots$ etc. by Monte Carlo. The results of such a computation are found in Table I.

If $G \in \mathcal{D}(\alpha)$ with $\alpha = MG_0$, and if $X_1,\dots,X_n$ is a sample from a distribution with density $f(x) = \int h(x|\theta)dG(\theta)$, then the posterior distribution of G given $X_1,\dots,X_n$ has been found by Antoniak (1974) to be a mixture of Dirichlet processes

$$G|x_1,\dots,x_n \in \int\dots\int \mathcal{D}(\alpha+\Sigma_1^n \delta_{\theta_i})dH(\theta_1,\dots,\theta_n|x_1,\dots,x_n). \qquad (3)$$

One may consider the observations $X_1,\dots,X_n$ to be chosen by first choosing $\theta_1,\dots,\theta_n$ i.i.d. from $G(\theta)$, and then X_i from $h(x|\theta_i)$ $i=1,\dots,n$ independently. With this interpretation

$H(\theta_1,\ldots,\theta_n|x_1,\ldots,x_n)$ is the posterior distribution of $\theta_1,\ldots,\theta_n$ given $x_1,\ldots,x_n$. Using (3), the posterior expectation of $G(\theta)$ given $x_1,\ldots,x_n$ may be written as follows. Let $\hat{G}_n$ denote the empirical distribution of $\theta_1,\ldots,\theta_n$, $\hat{G}_n = \frac{1}{n}\Sigma_1^n \delta_{\theta_i}$. Then, since the expectation of $\mathcal{D}(\alpha+\Sigma_1^n \delta_{\theta_i})$ is $(MG_0+n\,\hat{G}_n)/(M+n)$,

$$E(G(\theta)|x_1,\ldots,x_n) = \frac{M}{M+n} G_0(\theta) + \frac{n}{M+n} \int\ldots\int \hat{G}_n(\theta)\,dH(\theta_1,\ldots,\theta_n| x_1,\ldots,x_n) \tag{4}$$

and

$$f_n(x) = E(f(x)|x_1,\ldots,x_n) = \frac{M}{M+n} f_0(x) + \frac{n}{M+n} \hat{f}_n(x) \tag{5}$$

where

$$\hat{f}_n(x) = \frac{1}{n} \Sigma_1^n \int\ldots\int h(x\,|\theta_i)\,dH(\theta_1,\ldots,\theta_n|x_1,\ldots,x_n). \tag{6}$$

The estimate (6) may be considered as a partially Bayesian estimate of f(x) with the influence of the prior guess at f(x) removed.

This approach to density estimation including the special case of normal-gamma shape for G_0, has been treated by Lo (1978). Lo has found the following useful representation of the function H.

$$dH(\theta_1,\ldots,\theta_n|x_1,\ldots,x_n) = \frac{(\Pi_1^n h(x_i|\theta_i))\Pi_{i=1}^n d(MG_0+\Sigma_{j=1}^{i-1}\delta_{\theta_j})(\theta_i)}{M^{(n)}\ h(x_1,\ldots,x_n)} \tag{7}$$

where

$$h(x_1,\ldots,x_n) = \int\ldots\int (\Pi_{i=1}^n h(x_i|\theta_i))\Pi_{i=1}^n d(MG_0 + \Sigma_{j=1}^{i-1}\delta_{\theta_j})(\theta_i)/M^{(n)}. \tag{8}$$

With this notation we may write (5) as

$$E(f(x)|x_1,\ldots,x_n) = h(x,x_1,\ldots,x_n)/h(x_1,\ldots,x_n). \tag{9}$$

The two extreme cases of these estimates as $M \to 0$ and $M \to \infty$ are worth noting. As $M \to 0$, it becomes more and more likely that all the θ_i are equal, resulting in the estimate

$$\lim_{M\to 0} E(f(x)|x_1,\dots,x_n) = \frac{\int h(x|\theta)\Pi_1^n h(x_i|\theta)dG_0(\theta)}{\int \Pi_1^n h(x_i|\theta)dG_0(\theta)} . \tag{10}$$

This is the parametric estimate of $f(x)$ when θ is chosen from $G_0(\theta)$ and $x_1,\dots,x_n$ are chosen i.i.d. from $h(x|\theta)$. As $M \to \infty$, the estimate (5) converges to $f_0(x)$ which is not very useful. However, the estimate (6) which is less dependent on M as a measure of prior information also converges. Since as $M \to \infty$ it becomes more and more likely that all θ_i are distinct,

$$\lim_{M\to\infty} \hat{f}_n(x) = \frac{1}{n} \Sigma_1^n f(x|x_i), \tag{11}$$

where

$$f(x|x_i) = \int h(x|\theta)h(x_i|\theta)dG_0(\theta)/\int h(x_i|\theta)dG_0(\theta)$$

is the Bayes estimate of the density $h(x|\theta)$ based on a single observation, x_i, from $h(x|\theta)$ when θ has prior distribution $G_0(\theta)$. This is a variable kernel estimate.

If the density $h(x|\theta)$ were $N(\theta,\tau^2)$ with known τ^2, and if θ were $N(\mu,\sigma^2)$ with known μ and σ^2, then $f(x|x_i)$ is

$$N\left(\frac{\tau^2\mu+\sigma^2 x_i}{\tau^2+\sigma^2} , \frac{\tau^2+2\sigma^2}{\tau^2+\sigma^2}\right).$$

This yields a variable kernel estimate with constant window size, but centered at a point between x_i and μ, as is typical of shrinkage estimates.

For the problem treated in this paper, θ represents (μ,σ^2), which has a normal-gamma prior, while $h(x|\theta)$ is $N(\mu,\sigma^2)$. In this case, $f(x|x_i)$ is the density of a $t_{2\alpha+1}$-distribution centered at

$\frac{\tau\mu+x_i}{\tau + 1}$, with scale $((\tau+2)(\beta+ \frac{\tau}{\tau+1}(x_i-\mu)^2)/((2\alpha+1)(\tau+1)))^{\frac{1}{2}}$. In addition to the shrinkage phenomenon, the window size depends on the observations with larger windows for observations x_i farther from μ. Unfortunately, one has lost, through this interchange of limits as $M \to \infty$ and $n \to \infty$, the valuable Bayesian property of not having to worry about the window size as a function of n. However, one may hope to let α and τ depend on n, $\alpha_n \to \infty$ and $\tau_n \to 0$, to obtain good asymptotic properties of the estimate (11).

IV. COMPUTATIONS

Since computation of $E(f(x)|x_1,\ldots,x_n)$ depends on computing the ratio (9), let us concentrate on the denominator, $h(x_1,\ldots,x_n)$. For this we use the analogue of a Monte Carlo technique of Kuo (1980) developed for a Bayesian approach to the empirical Bayes decision problem.

First, we expand the product measure in Lo's representation.

$$\Pi_{i=1}^{n}(MG_0+\Sigma_{j=1}^{i-1}\delta_{\theta_j})(d\theta_i)/M^{(n)}$$
$$= G_0(d\theta_1)(\tfrac{M}{M+1}G_0(d\theta_2) + \tfrac{1}{M+1}\delta_{\theta_1}(d\theta_2))\ldots \tag{12}$$
$$(\tfrac{M}{M+N-1}G_0(d\theta_n) + \tfrac{1}{M+n-1}\delta_{\theta_1}(d\theta_n)+\ldots+\tfrac{1}{M+n-1}\delta_{\theta_{n-1}}(d\theta_n)).$$

When the product is expanded, there are n! terms, but some of them are equal. For example, a term that begins $G_0(d\theta_1)\delta_{\theta_1}(d\theta_2)\delta_{\theta_2}(d\theta_3)\ldots$ is the same as a term that begins $G_0(d\theta_1)\delta_{\theta_1}(d\theta_2)\delta_{\theta_1}(d\theta_3)\ldots$; in both, $\theta_1 = \theta_2 = \theta_3$. Each term of the expansion determines a partition, $Q = \{K_1,\ldots,K_m\}$, of $\{x_1,\ldots,x_n\}$ with the property that $\theta_1 = \theta_j$ in the term if and only if x_i and x_j are in the same set $K \in Q$. Thus,

$$h(x_1,\ldots,x_n) = \sum_Q P_M(Q)Z(Q) \tag{13}$$

where $P_M(Q)$ represents the probability that partition Q is selected, and

$$Z(Q) = \prod_{K\in Q} \int \prod_{x_i\in K} h(x_i|\theta)dG_0(\theta). \tag{14}$$

The Monte Carlo technique of Kuo entails sampling partitions, Q, at random according to $P_M(Q)$, and evaluating $Z(Q)$. This is repeated N times with partitions $Q_1,\ldots,Q_N$, and the average $\overline{Z(Q)} = \Sigma_1^N Z(Q_i)/N$ is taken as the Monte Carlo estimate of $h(x_1,\ldots,x_n)$.

As the computations are being carried out, one may obtain an estimate of the standard error of the Monte Carlo estimate. namely

$$(\Sigma_1^N (Z(Q_i) - \overline{Z(Q)})^2) /\{N(N-1)\}^{\frac{1}{2}}.$$

Though useful in most situations, it should be realized that this can be very misleading as an estimate of error, since typical situations arise in which the true variance is very large due to a few values of Z(Q) that are very large and have small probabilities of appearing in the sample. Error reduction techniques are discussed in the next section.

Kuo's method of choosing Q is as follows. Take any ordering of the x_i. Start a set of the partition with x_1. For k=1,... n-1, repeat the following operations: *Let* x_{k+1} *start a new set of the partition with probability* M/(M+k); *otherwise, with probability* 1/(M+k) *each, put* x_{k+1} *into the set containing* x_i *for* i=1,...k.

To adapt this method to compute the estimate, (6), we randomly choose a partition Q in this manner, and in addition to computing the value of the denominator, Z(Q), we also compute the value of the numerator, call if Y(Q), namely,

$$Y(Q) = Z(Q) \sum_{K \in Q} \frac{|K|}{n} \frac{\int f(x|\theta) \prod_K f(x_i|\theta)\, dG_0(\theta)}{\int \prod_K f(x_i|\theta)\, dG_0(\theta)} .$$

The Monte Carlo estimate of $\hat{f}(x)$ is then

$$\Sigma Y(Q_i)/\Sigma Z(Q_i)$$

and its standard error can be estimated using the usual asymp-formula for the variance of a ratio of means,

$$\mathrm{Var}(\overline{Y}/\overline{Z}) \sim \frac{1}{n\mu_z^2} (\sigma_y^2 - 2\sigma_{yz} \frac{\mu_y}{\mu_z} + \sigma_z^2 \frac{\mu_y^2}{\mu_z^2}).$$

To test the computational method, a specific case with n = 5 was chosen so that exact expectations could be made for comparison. The observed values are taken to be $x_1 = 1.0$, $x_2 = 1.1$, $x_3 = 1.9$, $x_4 = 2.3$, and $x_5 = 2.6$. Simple values of the parameters were chosen and a small Monte Carlo of size N = 10 was carried out. The results are found in Table II. The estimates seem very good for so small a value of N, but with a larger value of n it is expected that a larger value of N is also needed.

TABLE II. Density Estimate with Prior Parameters $M = 1$, $\alpha = 1$, $\beta = 1$, $\mu = 2$, and $\tau = .5$, and with Observations 1, 1.1, 1.9, 2.3, and 2.6. Based on a Monte Carlo of size $N = 10$

x	Prior guess $f_0(x)$	Estimated $\hat{f}_n(x)$	Estimated $f_n(x)$	Estimated st. err. $\hat{f}_n(x)$	Exact $\hat{f}_n(x)$
0	.081	.055	.059	.002	.053
.5	.125	.128	.128	.005	.127
1.0	.188	.272	.258	.011	.270
1.5	.256	.434	.404	.009	.432
2.0	.289	.457	.429	.015	.467
2.5	.256	.315	.306	.009	.329
3.0	.188	.159	.163	.003	.159
3.5	.125	.069	.078	.003	.065
4.0	.081	.030	.039	.002	.027

Consider now an example that illustrates the way one goes about choosing the parameters to express prior beliefs or fit data. Taking the five data points of the previous example, suppose we expect one main peak and one smaller local maximum, the rest being quite small in comparison; then we choose M = 1 or M = 2 for definiteness. We expect the center of mass to be around 2 so we choose $\mu = 2$. (If we use data-aided choice we might choose $\mu = \bar{x} = 1.8$). We might set the standard deviation of the μ_i at $(2.3 - 1.1)/2 = .6$, and if we expect the σ_i to range from near 0 to .5 we might set $E\sigma_i^2 = .06$ giving a ratio of $\tau = .12/.36 = 1/3$, so $\tau = .1$ or $\tau = .5$ might do. We choose a small value of β, say $\beta = .5$, and solve $E\sigma_i^2 = \beta/(2\alpha-1)$ giving an α close to 5. Table III contains the exact values of the estimate (6) of the density for certain parameter values close to these, the first column being the one of choice. The density estimate is seen to have one large peak close to 2.25 and a smaller one close to 1.25.

TABLE III. Exact Values of the Estimate (6) for the Data Points 1, 1.1, 1.9, 2.3, 2.6. Parameter values M = 1, μ = 2

	α = 5		α = 5		α = 1	
	β = .5		β = .1		β = .1	
x	τ=.5	τ=.1	τ=.5	τ=.1	τ=.5	τ=.1
0	.002	.003	0	0	.022	.015
.25	.010	.013	.002	0	.042	.032
.50	.039	.063	.013	.005	.082	.082
.75	.136	.238	.085	.103	.160	.220
1.00	.337	.512	.374	.824	.282	.460
1.25	.509	.511	.663	.604	.400	.477
1.50	.488	.316	.394	.069	.452	.342
1.75	.464	.309	.307	.272	.511	.362
2.00	.607	.498	.671	.537	.619	.496
2.25	.694	.659	.831	.623	.614	.597
2.50	.475	.538	.575	.776	.418	.504
2.75	.181	.247	.074	.163	.198	.241
3.00	.046	.072	.005	.010	.085	.088
3.25	.010	.016	.001	.001	.039	.034
3.50	.002	.003	0	0	.019	.015
3.75	0	.001	0	0	.010	.007

V. IMPROVING THE MONTE CARLO

In some situations, it is possible to change the Monte Carlo sampling method to reduce the variance of the estimate. See Rubenstein (1981) for a review of such methods. In the problems treated here, there may be a partition Q with a large value of Z(Q) and a small value of $P_M(Q)$ that is rarely chosen in the Monte Carlo sampling even though it contributes significantly to the expectation. As an example, suppose there are 3 Q's with probabilities .01, .50, and .49 and values of Z(Q) 1000, 2, 1, respectively. The true mean is h = .01 (1000) + .50 (2) + .49 (1) = 11.49. With a Monte Carlo of size N = 10, there is a large variance; if the sample contains the value 1000, the estimate is greater than 100, and if it doesn't, the estimate is at most 2. If it is known which Z(Q) are expected to be large, one may change the probabilities and values; for example, we may write

$$h = .85\ (\frac{.01}{.85}\ 1000) + .10\ (\frac{.5}{.1}\ 2) + .05\ (\frac{.49}{.05}\ 1)$$
$$= .85\ (11.76) + .10\ (10) + .05\ (9.8) = 11.49\ .$$

The Monte Carlo of size N = 10, on values 11.76, 10, 9.8 with probabilities .85, .10, .05 respectively has a small variance.

The problem is to tell which Z(Q) are likely to be large and by how much. It is useful to consider two separate sources of variation in the Z(Q): the number of sets in the partition, and their within set variation.

Often the number of sets in the partition significantly affects Z(Q), the larger the number of sets the smaller the Z(Q). In the example with n = 5 observations, 1, 1.1, 1.9, 2,3, and 2.6, and paramter values M = 1, α = 1, β = 1, μ = 0, and τ = 1, Z(Q) goes down roughly as $.5^m$ where $m = |Q|$. In the more reasonable case that μ = 2, it goes down roughly as $.8^m$. This being so, we can improve the Monte Carlo as follows. *Choose the partition* *using Kuo's method but using a different value of* M *say* M'; *after* Z(Q) *has been evaluated multiply it by* $(M/M')^m M'^{(n)}/M^{(n)}$ *where* $m = |Q|$. In the cases mentioned M'/M should be chosen roughly as .5 or .8.

For a fixed number of sets in the partition, the value of Z(Q) is largest if the sets contain contiguous order statistics so that the within set variation is small. In the method of Kuo, the distribution of the assignment of the x_i to the sets of the partition is invariant under interchange of subscripts. Needed is a method of choosing "clumpy" partitions in which x_i's that are close in value have a higher probability of being in the same set. Here is one possibility involving rank-dependent grouping.

Order the x_i's, $x_1 < x_2 < \ldots < x_n$, choose a value of $t \geq 1$, and let x_1 start a set of the partition. For k=1,2,...,n-1, repeat the following operation. *Let* x_{k+1} *start a new set of the partition with probability* M/(M+k). *Otherwise, put* x_{k+1} *into the already created set* K *with probability proportional to* $\Sigma_{i\in K}\, t^i$. *The ratio of the old probability of* K *to the new probability is*

$$ratio = \frac{|K|}{k}\,\frac{\Sigma_1^k\, t^i}{\Sigma_{i\in K}\, t^i}\ .$$

Keep a running product of the ratios as you go along and multiply Z(Q) *by this product when finished.*

The old probabilities are those given by the method of Kuo which uses t = 1. One drawback of this method is that $Q_1 = \{\{1,2,3\}, \{4,5\}\}$ and $Q_2 = \{\{1,2\}, \{3,4,5\}\}$ have different probabilities even though they are symmetric. One can avoid this by randomizing with probability ½ ordering the x_i in ascending or descending order.

In some cases it may be preferable to use value-dependent grouping, in which x_{k+1} is assigned to a set K with probability proportional to some function of the values of the observations in K such as $\Sigma_{i \in K} \exp\{-t|x_i - x_{k+1}|\}$ or $|K| \exp\{-t\Sigma_{i \in K}(x_i - x_{k+1})^2/|K|\}$ for some $t \geq 0$.

It is easy to combine these two methods simultaneously changing M to M' and changing the probability of assignment to the already created sets K. Both methods have been tried on the 5-point data set used in the example with parameters similar to those mentioned. In both cases, a small reduction in the variance of the estimate (about 20%) was noted.

VI. SUMMARY

The problem of nonparametric density estimation is considered from a Bayesian viewpoint. The density is assumed to be a countable mixture of normal distributions with arbitrary means and variances, and a prior joint distribution is placed on the mixing probabilities and the means and variances. The resulting model is seen to be equivalent to Lo's model, and a method of Kuo is adapted to carry out the computations. A simple example is investigated to show the feasibility of the method.

REFERENCES

Antoniak, Charles E. (1974). "Mixtures of Dirichlet Processes with Application to Bayesian Nonparametric Problems." *Ann. Statist. 2,* 1152-1174.

Doob, J. (1948). "Application of the Theory of Martingales." *Le Calcul des Probabilités et ses Applications. Colloques Internationaux du Centre National de la Recherche Scientifique,* Paris, 23-28.

Ferguson, Thomas S. (1973). "A Bayesian Analysis of Some Nonparametric Problems." *Ann. Statist. 1,* 209-230.

Kingman, J.F.C. (1975). "Random Discrete Distributions." *J.R.S.S. Series B 37,* 1-15.

Kuo, Lynn (1980). Computations and Applications of Mixtures of Dirichlet Processes, Ph.D. thesis, UCLA.

Lo, Albert Y. (1978). On a Class of Bayesian Nonparametric Estimates: I. Density Estimates, Mimeographed preprint, Rutgers, University, New Brunswick, N.J.

Rubenstein, Reuven Y. (1981). *Simulation and the Monte Carlo Method,* John Wiley and Sons, New York.

Sethuraman J. and Tiwari, Ram C. (1982). Convergence of Dirichlet Measures and the Interpretations of their Parameter, *Statistical Decision Theory and Related Topics, III,* Editors, S.S. Gupta and J.O. Berger, Academic Press, New York.

Teicher, Henry (1960). "On the Mixture of Distributions." *Ann. Math. Statist. 31,* 55-73.