Estimating Normal Means With a Dirichlet Process Prior

Michael D. Escobar*

In this article, the Dirichlet process prior is used to provide a nonparametric Bayesian estimate of a vector of normal means. In the past there have been computational difficulties with this model. This article solves the computational difficulties by developing a "Gibbs sampler" algorithm. The estimator developed in this article is then compared to parametric empirical Bayes estimators (PEB) and nonparametric empirical Bayes estimators (NPEB) in a Monte Carlo study. The Monte Carlo study demonstrates that in some conditions the PEB is better than the NPEB and in other conditions the NPEB is better than the PEB. The Monte Carlo study also shows that the estimator developed in this article produces estimates that are about as good as the PEB when the PEB is better and produces estimates that are as good as the NPEB estimator when that method is better.

KEY WORDS: Empirical Bayes; Gibbs sampler; Importance sampling; Mixtures of Dirichlet processes; Nonparametric Bayes.

1. INTRODUCTION

Suppose that $Y_1, Y_2, \ldots, Y_n$ are observed such that the Y_i's given the X_i's are independent and have a normal distribution with mean X_i and variance 1, the X_i's given G are independent with distribution G, and G and the X_i's are all unknown. This article introduces a new way to estimate the X_i values from the observed Y_i's by using a nonparametric Bayesian estimator that uses a Dirichlet process prior. Previous attempts to use a nonparametric Bayesian estimator have been limited due to computational difficulties. The main objective of this article is to introduce a new method for calculating the nonparametric Bayesian estimator and to compare this estimator with other methods of estimating the X_i's.

Before presenting the Dirichlet process prior and the nonparametric Bayes estimator, some typical ways to estimate the vector of X_i's are discussed. The most common method is to estimate the X_i by Y_i. This method, called the *straight estimator* in this article, has many desirable properties, such as being the maximum likelihood estimate, the least squares estimate, and the minimum variance unbiased estimator. But when the number of means is greater than 3, Stein (1955) showed that the straight estimate is inadmissible under the squared error loss function.

When G is known, the posterior mean is the best estimator under squared error loss and the Bayes estimator is the posterior mean; that is,

$$E(X_i \mid Y_i) = \frac{\int X_i \phi(Y_i - X_i)\, dG(X_i)}{\int \phi(Y_i - X_i)\, dG(X_i)},$$

where ϕ is the density of the standard normal distribution function. If G is unknown, then G can be estimated by Bayesian or frequentist methods. Methods of estimating X_i by first estimating G using frequentist methods and then substituting the estimate of G into the preceding equation are called empirical Bayes methods. These methods were first introduced by Robbins (1955).

There are two classes of empirical Bayes methods: parametric and nonparametric (see Morris 1983). If G is assumed to belong to a parametric family, like the family of normal distributions, then the estimator is the James–Stein estimator or one of its relatives (see Efron and Morris 1973ab, 1975; James and Stein 1963). To obtain a nonparametric empirical Bayes estimator, one can first estimate G by nonparametric maximum likelihood methods (see, for instance, Laird 1978, 1981; Lindsay 1983).

Instead of using a frequentist estimate of G, this article presents a nonparametric Bayesian analysis using a Dirichlet process as a prior distribution on the family of distributions for G. The Dirichlet process prior (see Ferguson 1973, 1974 and references therein) is a prior distribution on the family of distribution functions that is dense in the space of distributions. Antoniak (1974) showed that if a Dirichlet process prior is used for G in this problem, then the posterior distribution of X_i is sampled from a mixture of Dirichlet processes.

In the past it has been difficult to estimate values from a mixture of Dirichlet processes. Berry and Christensen (1979) used a parametric approximation for binomial models. Kuo (1986) and Lo (1984) have independently developed similar Monte Carlo integration algorithms to estimate from a mixture of Dirichlet processes. But these algorithms do not sample values conditionally on the data, which can lead to very inefficient estimates (see Escobar 1992 for a more detailed discussion). With mixtures of Dirichlet processes, sampling from the data vector using importance sampling techniques is critical.

The methods introduced in this article are based on a Monte Carlo integration that always samples points conditional on the data. In Section 2 an algorithm to estimate the

* Michael D. Escobar is Assistant Professor, Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213-3890 and Assistant Professor, Department of Preventive Medicine and Biostatistics, University of Toronto, Ontario, M5S 1A8 Canada. Much of this research was carried out at Yale University as part of the requirements for a Ph.D. degree in statistics under the direction of John Hartigan, who gave generously of his time, energy, and ideas. The author also thanks Robert Kass and Mark Schervish for their helpful comments on an earlier draft of this manuscript and two referees and the associate editor for greatly improving the proofs of Theorems 1 and 2 and the presentation in general. This research was supported in part by National Science Foundation Grant DCR-8401636, National Cancer Institute Grant RO1-CA54852-01, and a National Research Service Award from National Institutes of Mental Health Grant MN15758. Computational support was provided in part by Biomedical Research Support Grant RR05443 awarded by the National Institutes of Mental Health Division of Research Resources.

Journal of the American Statistical Association
March 1994, Vol. 89, No. 425, Theory and Methods

values of X_i is developed that uses a Dirichlet process prior with fixed parameters as the prior distribution for G. To estimate the X_i values, samples from the posterior distribution are obtained by reiterating a Markov chain constructed from the easy-to-sample conditional distributions. We also show that the limiting distribution of the sample obtained from the reiterated Markov chain is the posterior distribution. Previously, the idea of using a reiterated Markov chain has been used in image processing by Geman and Geman (1984). Recently there has been extensive research, independent from the author's work, on calculating posterior distributions by sampling from a reiterated Markov chains by, among others, Tanner and Wong (1987), Rubin (1988), and Gelfand and Smith (1990). Gelfand and Smith (1990) extended the Gibb's sampler method of Geman and Geman (1984) to estimate general posterior distributions by using the reiterated Markov chain based on conditional distributions.

In Section 3 methods of selecting a prior distribution for the parameters of the Dirichlet process are discussed, and methods of estimating the posterior distribution for these parameters are developed. To calculate the posterior distribution of these parameters, an importance sampling algorithm is developed. The parameters of the Dirichlet process are important, because they enable the estimator from the Dirichlet process prior to behave almost like the James–Stein estimator, the nonparametric empirical Bayes estimator, or a combination of these two estimators. Choosing the correct parameter allows the Dirichlet process prior to behave like the better of these estimators for a given data set.

In Section 4 the method developed in this article is compared to the straight estimator, the Bayes estimator if the true G is known, the James–Stein estimator, and the nonparametric empirical Bayes estimator. The James–Stein estimator is a global estimator; it assumes that G is a member of a parametric family and then uses all the data to calculate the parameters. If G is a member of this parametric family, then the James–Stein estimator can be extremely efficient. But if G is not a member of this family and is instead, for example, multimodal, then the James–Stein estimator will not be very efficient; however, it will do no worse than the straight estimator. The nonparametric empirical Bayes methods (NPEB), such as the nonparametric maximum likelihood estimator, are local estimators. They do not assume that G is from any parametric family and use only the nearby data values to estimate the value of G at a given location. If the distribution G is multimodal, then the NPEB may be very efficient at finding the different modes and then shrinking the estimates to the center of the modes. But if the distribution is very disperse, which happens if G is a normal distribution with a variance of 4 or 10, then there will be few data points in the local area of the estimation. This will lead to very poor estimates. When G is a normal distribution with a variance of 10, then the NPEB does worse then the simple straight estimator. The estimate based on the Dirichlet process can act as a local or a global estimator, depending on the value of an adjustable parameter. By using the data to adjust this parameter, the Dirichlet process estimator can mimic either the James–Stein estimator or the NPEB estimator. It will do well if either the James–Stein estimator or the NPEB estimator does well, and it will avoid some of the pitfalls of these estimators.

Although the methods introduced in this article are used to calculate the posterior means of several normal distributions, these methods can be used to perform a wide range of nonparametric Bayesian analyses. Many analysis of variance, linear regression, and random-effects models can be reduced to the problem of estimating the mean of several normal populations. To study the estimation problem more clearly, the linear structure in the means and unequal variances have not been considered. Section 5 shows how the posterior expectation of a function of X can be calculated and also how to extend the algorithm to models with different error distributions. With these extensions, the methods in this article could be used to calculate a wide range of nonparametric Bayesian problems.

Section 6 presents a final discussion. The Appendix contains detailed proofs of the theorems presented in Sections 2 and 3.

2. THE ESTIMATOR

Assume that G is sampled from a Dirichlet process with parameters G_0 and A_0, where G_0 is a probability measure and A_0 is a positive real constant. The parameter G_0 is a location parameter for the Dirichlet process prior. It is the best guess at what G is believed to be and is the mean distribution for the Dirichlet process. The parameter A_0 is a measure of the strength in the belief that G is G_0. Therefore, the parameter A_0 is a type of dispersion parameter for the Dirichlet process prior. In this section it is assumed that A_0 and G_0 are fixed.

To simplify the use of the Dirichlet process prior, note that when G is integrated over its prior distribution, the sequence of X_i's follows a general Polya urn scheme; that is,

$$X_1 \sim G_0,$$

$$X_n | X_1, \ldots, X_{n-1} \begin{cases} = X_j \text{ with probability } \dfrac{1}{A_0 + n - 1} \\ \sim G_0 \text{ with probability } \dfrac{A_0}{A_0 + n - 1} . \end{cases}$$

From this it is easy to sample a sequence $X_1, \ldots, X_n, \ldots,$ given G_0 and A_0.

The closed form of the joint probability of $X_1, \ldots, X_n$ is

$$dF(X_1, \ldots, X_n) = \prod_{i=1}^{n} \frac{[A_0 G_0(dX_i) + \sum_{j=1}^{i-1} \delta(X_j, dX_i)]}{A_0 + i - 1}, \quad (1)$$

where $\delta(X, \cdot)$ is a measure defined by

$$\delta(X, B) = \begin{cases} 1 \text{ when } X \in B \\ 0 \text{ when } X \notin B. \end{cases}$$

(For more on the relationship between a generalized Polya urn scheme and the Dirichlet process prior, see Blackwell and MacQueen 1973 and Ferguson 1973).

In this article the same notation is used for a set and the indicator function for a set; that is, $\{x \le a\}$ is the indicator function for the set $\{x \le a\}$. Also, the symbols $\mathbf{X}$ and $\mathbf{Y}$ will represent the vectors $(X_1, \ldots, X_n)$ and $(Y_1, \ldots, Y_n)$.

Antoniak (1974) showed that the distribution of $\mathbf{X}$ given $\mathbf{Y}$ is a mixture of Dirichlet processes. A mixture of Dirichlet process can be difficult to use; however, Escobar (1988) has shown that the conditional distribution of X_i given all the other X_j's (with $i \neq j$) and the data $\mathbf{Y}$ is a mixture of a discrete distribution with weights on the other X_j's and a distribution that is usually close to a normal distribution. The proof of the following theorem is contained in Appendix A.

Theorem 1. The conditional distribution of X_i given X_j, $j \neq i$, and $\mathbf{Y}$ has the following closed form:

$$dF[X_i \mid X_j, i \neq j, \mathbf{Y}] = \frac{\phi(Y_i - X_i)A_0G_0(dX_i) + \sum_{\substack{j=1\\ j\neq i}}^{n} \phi(Y_i - X_j)\delta(X_j, dX_i)}{A(Y_i) + \sum_{\substack{j=1\\ j\neq i}}^{n} \phi(Y_i - X_j)}, \tag{2}$$

where ϕ is the standard normal density function and $A(Y)$ is defined as $A(Y) = A_0 \int \phi(Y - X)G_0(dX)$.

The conditional distribution defined in equation (2) can be sampled according to the following rule:

$$X_i \mid X_j, i \neq j, \mathbf{Y} \begin{cases} = X_j \quad \text{with probability} \quad \dfrac{\phi(Y_i - X_j)}{A(Y_i) + \sum_{\substack{l=1\\ l\neq i}}^{n} \phi(Y_i - X_l)} \\[2ex] \sim h(X_i \mid Y_i) \quad \text{with probability} \quad \dfrac{A(Y_i)}{A(Y_i) + \sum_{\substack{l=1\\ l\neq i}}^{n} \phi(Y_i - X_l)}, \end{cases} \tag{3}$$

where h is a density function from which to sample X_i and is defined as $h(X_i \mid Y_i) = [A_0/A(Y_i)]\phi(Y_i - X)g_0(X)$, and where g_0 is the density function or the probability function corresponding to the probability measure G_0.

Note that the function $h(X_i \mid Y_i)$ is the posterior density of X_i given the data Y_i if G_0 is the prior distribution of X_i. The function $A(Y_i)$ is the marginal distribution of Y_i when G_0 is the prior distribution of X_i. In the procedure described in equation (3), the weights are proportional to $\phi(Y_i - X_j)$ and $A(Y_i)$, because $A(Y_i)$ is the marginal distribution of Y_i when the prior of X_i is G_0 and $\phi(Y_i - X_j)$ is the marginal distribution of Y_i when the prior of X_i is a point mass on X_j. Of course when the prior for X_i is a point mass on X_j, then the posterior distribution of X_i given Y_i is a point mass on X_j.

If G_0 is constant in a large area around Y, then $h(X \mid Y)$ is approximately the normal density. So if G_0 is the uniform distribution, with support that covers the range of the data with a margin of 2 or 3 on each side of the range, then $h(X \mid Y)$ is just the normal distribution with mean Y and variance 1.

The distribution defined in (2) or (3) can be used to construct a Markov chain that converges in the limit to the posterior distribution $F(\mathbf{X} \mid \mathbf{Y})$. Start by setting $(X_1, \ldots, X_n) = (x_1^{(0)}, \ldots, x_n^{(0)})$. Usually, $x_i^{(0)} = Y_i$. Define the first step of a continuous-state Markov chain by the following:

$$\begin{array}{ll} \text{Sample } x_1^{(1)} \text{ from } X_1 \mid X_2 = x_2^{(0)}, & X_3 = x_3^{(0)}, \ldots, X_n = x_n^{(0)}, Y_1, \ldots, Y_n. \\ \text{Sample } x_2^{(1)} \text{ from } X_2 \mid X_1 = x_1^{(1)}, & X_3 = x_3^{(0)}, \ldots, X_n = x_n^{(0)}, Y_1, \ldots, Y_n. \\ \vdots & \\ \text{Sample } x_n^{(1)} \text{ from } X_n \mid X_1 = x_1^{(1)}, & X_2 = x_2^{(1)}, \ldots, X_{n-1} = x_{n-1}^{(1)}, Y_1, \ldots, Y_n. \end{array}$$

These n samples are considered to be one step in the Markov chain. In a similar way $X_i^{(m)} = x_i^{(m)}$ can be sampled given $x_1^{(m-1)}, \ldots, x_n^{(m-1)}$. Define the random vector $X^{(m)}$ as the vector produced after the mth step of the Markov chain. The proof of the following theorem is in the Appendix.

Theorem 2. The distribution $F(\mathbf{X} \mid \mathbf{Y})$ is the stationary distribution of the Markov chain and $X^{(m)}$ converges in distribution to the stationary distribution, $F(\mathbf{X} \mid \mathbf{Y})$, regardless of the initial values of the Markov chain.

Note that the proof in the Appendix primarily uses theorem 2 of Feller (1971, p. 271). Schervish and Carlin (1992) presented convergence results similar to Feller's theorem under slightly different conditions and also discussed rates of convergence for similar types of Markov chains. But due to the unusual dominating measure of the posterior distribution, it is not clear how one could demonstrate that the necessary conditions could be satisfied to apply the theorems of Schervish and Carlin (1992).

To estimate X_i, for some large m draw samples of $\mathbf{X}_{(l)} \equiv X_{(l)}^{(m)}$, for $l = 1, \ldots, L$ and estimate $E[X_i \mid \mathbf{Y}]$ by

$$\hat{X}_i = \frac{1}{L} \sum_{l=1}^{L} E[X_{i(l)} \mid \mathbf{Y}, X_{j(l)}, j \neq i], \tag{4}$$

where $E[X_{i(l)} \mid \mathbf{Y}, X_{j(l)}, j \neq i]$ is calculated from the distribution defined in Equation (2).

The method of summing the conditional expectation of X_i instead of just summing the sampled X_i's was first suggested by Gelfand and Smith (1990), who called this method Rao–Blackwellization.

3. POSSIBLE PRIORS FOR THE DIRICHLET PROCESS

3.1 General Approach for Selecting the Parameter Priors

In the preceding section we assumed that A_0 and G_0, the parameters for the Dirichlet process prior, are fixed. In practice it is difficult to select appropriate values for these parameters. Instead, a prior distribution is placed on these values and a posterior distribution is calculated. Because these parameter values have an important influence on the estimation, a broad range of possible values for the parameters is chosen.

For computational simplicity, the set of values for the parameters is finite. For example, in the simulations in the next section the support for the prior for A_0 contains four positive numbers and the support for the prior on G_0 contains four distinct distributions. In this case there are 16 possible values for the pair (A_0, G_0). In the simulation the prior distribution for (A_0, G_0) puts equal weight on all the pairs.

The posterior distribution is obtained by calculating the likelihood of each (A_0, G_0) pair for the data $\mathbf{Y}$. Here it is important to use importance sampling methods, because the calculations involve integrations that have almost all the weight in a small neighborhood near the data vector, $\mathbf{Y}$. When we wish to sample $\mathbf{X}_{(l)}$ from the Markov chain, we first sample (A_0, G_0) from its posterior distribution and then use this sampled (A_0, G_0) to generate the vector $\mathbf{X}_{(l)}$ using the method described in the previous section. Choices for the prior distributions on A_0 and G_0 and the calculation of their joint posterior distribution are discussed next.

3.2 Prior Distribution for A_0

When defining a Dirichlet process prior, A_0 represents the weight of our belief that G is the distribution G_0. Although this may be hard to quantify, in this section it is shown that A_0 is related to how "clumpy" the data are. Clumpy data occur when the observations are concentrated into a few clusters. If the observed data are very sparse and not very clumpy, then we will see in the Monte Carlo study of the next section that nonparametric maximum likelihood methods may not work very well. But if the data are very clumpy, with modes that are spread out, then standard parametric empirical Bayes methods do not work very well and nonparametric empirical Bayes methods work quite well. The choice of A_0 will determine whether the estimate from the Dirichlet process prior behaves like the nonparametric empirical Bayes estimator or like the parametric empirical Bayes methods.

The value A_0 is related to the number of different X's. Define $C(A_0, n)$ as the expected number of different X's. Then (see Antoniak 1974, p. 1161),

$$C(A_0, n) = E(\text{number of different } X\text{'s}) = \sum_{i=1}^{n} \frac{A_0}{A_0 + i - 1}.$$

It is easy to show that the preceding implies that $\max(1, A_0 \ln[(A_0 + n)/A_0]) \le C(A_0, n) \le 1 + A_0 \ln([A_0 + n - 1]/A_0)$. Table 1 gives the expected number of clusters, $C(A_0, n)$, where n equals 16, 20, 50, 100, 200, and 1,000 and where A_0 equals different powers of n. This table shows that for A_0 equal to or less than n^{-1}, one expects only about one cluster; for A_0 equal to or greater than n^2, one expects almost n different clusters; for A_0 about 1, one expects only a few clusters; and for A_0 equal to n one expects about $\frac{2}{3} n$ clusters. [Actually, when A_0 equals n, $C(A_0, n) \approx n \ln(2)$.]

Table 1. Expected Number of Clusters for Different Values of A_0 and n Where n is the Sample Size and A_0 is the Precision Parameter for the Dirichlet Process

	n					
A_0	16	20	50	100	200	1,000
$n^{3.0}$	15.97	19.98	49.99	100.00	200.00	1,000.00
$n^{2.5}$	15.88	19.89	49.93	99.95	199.96	999.98
$n^{2.0}$	15.55	19.54	49.52	99.51	199.50	999.50
$n^{1.5}$	14.38	18.14	46.83	95.36	193.28	984.93
$n^{1.0}$	11.34	14.12	34.91	69.57	138.88	693.40
$n^{0.5}$	6.86	8.03	15.22	24.44	38.90	110.69
$n^{0.0}$	3.38	3.60	4.50	5.19	5.88	7.49
$n^{-0.5}$	1.75	1.72	1.60	1.50	1.41	1.24
$n^{-1.0}$	1.20	1.17	1.09	1.05	1.03	1.01
$n^{-1.5}$	1.05	1.04	1.01	1.01	1.00	1.00
$n^{-2.0}$	1.01	1.01	1.00	1.00	1.00	1.00

When there are only a few clusters, the estimate of the normal means from the Dirichlet process prior will be similar to the nonparametric empirical Bayes estimator. When there are almost n different clusters, the estimator from the Dirichlet process prior will be similar to the parametric empirical Bayes estimator. The parameter A_0 adjusts the estimator presented in this article to behave like either a parametric estimator, which uses the data in a global manner, or a nonparametric estimator, which uses the data in a local manner.

In choosing a prior on A_0 based on the expected number of clusters, I developed the following prior. Let the values of A_0 be between n^{-1} and n^2, because these values would result in almost the extreme values for the expected number of clusters; that is, one cluster or n clusters. Because of the logarithmic relationship between the expected number of clusters and A_0, it seems reasonable to pick a prior for A_0 so that $\log_n(A_0)$ is evenly spread between -1 and 2. For the simulations study I used a discrete prior on A_0 which puts a mass of 1/4 on each of the atoms $\{n^{-1}, n^0, n^1, n^2\}$. The largest and smallest atoms, n^{-1} and n^2, correspond to a belief that the expected number of clusters is nearly the minimum or maximum number of possible clusters. The middle atoms, n^0 and n^1, correspond to a belief that one expects either a few clusters or many clusters. Simple modifications to the author's prior (e.g., using a prior that has more atoms, resulting in a finer grid to cover the prior space, or using a prior that uses different weights reflecting one's own prior belief on the number of different clusters) could easily be incorporated. By using the updating mechanism presented in Section 3.4 to calculate the posterior distribution of A_0, the data are then used to modify the algorithm to produce estimates similar to either the nonparametric or parametric empirical Bayes estimates.

3.3 Prior Distribution for G_0

The parameter G_0 is the prior guess of the shape of G. The parameter G_0 is used only in the algorithm in Section 2 to calculate $A(Y)$, $h(X_i \mid Y_i)$, and $E[X_i \mid X_j, j \neq i, \mathbf{Y}]$. Thus G_0 could be chosen to facilitate the calculation of these values in the algorithm. For computational purposes, G_0 could be a conjugate prior, a uniform prior, or some mixture of these two types of distributions. Three different types of priors one could use would be a subjective prior, an empirical Bayes prior, and a noninformative prior.

The role of G_0 for the Dirichlet process priors is similar to the role that the median or mean play in the typical prior distribution; it is the location parameter. It is our best guess of where the true value is. Therefore, if there are prior subjective beliefs, prior expert opinions, or theoretical considerations that G belongs to a small, finite set of possible distributions, then the prior distribution of G_0 should have support on this set. If the set of distributions is not finite, then a finite subset of "typical" distributions that belong to this set could be chosen that represent the larger set. Because the algorithm will average over the posterior distribution of G_0, a natural smoothing occurs. Also, this prior is a third-stage prior in a hierarchical Bayesian structure. Therefore, picking a finite subset will approximate the results that would have been obtained from the infinite set of possible G_0 values.

If we decided to use the data to help determine the set of possible G_0's, then we are actually using an empirical Bayes method. When estimating normal means, it is common to assume that G is a normal distribution with unknown mean and variance. We could let G_0 be a normal distribution and use the data to estimate the mean and variance, as is done in the James–Stein estimator. If we wanted to incorporate the possibility of multimodal G's, then we could let G_0 be a mixture of one to, say, four normals and could use standard methods to estimate from the data the means, variances, and mixing parameters for the different fixed number of mixed normals. (See, for example, Titterington, Smith, and Makov 1985 for methods for calculating these mixtures.)

An alternative prior for G_0 is to use a noninformative or improper prior; that is, to let G_0 have constant weight on the real line. Of course improper priors can cause mathematical problems. But the effect of such an improper prior in our problem could be approximated by using a uniform prior with support much larger than the range of the data. To do this I must calculate the minimum and maximum of the data, which means that I am technically using an empirical Bayes approach. Of course this could be avoided by consulting experts in the field where the data is collected and having them state some extreme maximum and minimum values of the data. One advantage of the truncated prior is that it is very similar to the improper prior but avoids the computational and mathematical difficulties.

For the simulation study in the next section, the noninformative prior for G_0 was used. Define G_r as the uniform distribution on the interval $[r_1, r_2]$ with $r_1 = \{\min(\mathbf{Y}) - r\}$ and $r_2 = \{\max(\mathbf{Y}) + r\}$. The set of G_0's used in the simulation study is $G_0 \in \{G_r \mid r = 0, 1, 2, 3\}$.

3.4 Calculating the posterior distribution of (A_0, G_0)

First, the parameters (A_0, G_0) are limited to a finite number of values, S. Label the S different values $(A^{(s)}, G^{(s)})$, with $s = 1, 2, \ldots, S$. Put prior weights on the set of values $\{(A^{(s)}, G^{(s)})\}$, which we signify by $P[(A_0, G_0) = (A^{(s)}, G^{(s)})]$. In the simulation in the next section, S equals 16 and the prior distribution puts equal weight on all values of $\{(A^{(s)}, G^{(s)})\}$.

By Bayes's theorem, the posterior prior of $(A^{(s)}, G^{(s)})$ given the data $\mathbf{Y}$ is

$$P\{(A_0, G_0) = (A^{(s)}, G^{(s)}) \mid \mathbf{Y}\} = \frac{f[\mathbf{Y} \mid (A^{(s)}, G^{(s)})] P[(A_0, G_0) = (A^{(s)}, G^{(s)})]}{\sum_{t=1}^{S} f[\mathbf{Y} \mid (A^{(t)}, G^{(t)})] P[(A_0, G_0) = (A^{(t)}, G^{(t)})]}, \quad (5)$$

where

$$f[\mathbf{Y} \mid (A^{(s)}, G^{(s)})] = \int \prod_{i=1}^{n} \phi(Y_i - X_i)\, dF[\mathbf{X} \mid (A_0, G_0) = (A^{(s)}, G^{(s)})] \quad (6)$$

and $dF[X \mid (A_0, G_0) = (A^{(t)}, G^{(t)})]$ is defined in equation (1) with the parameters (A_0, G_0) set to the values $(A^{(t)}, G^{(t)})$.

The hard part in calculating Equation (5) is calculating Equation (6). Equation (6) is evaluated via Monte Carlo. It is very important to sample values X_i near Y_i; therefore, an importance sampling method is used.

Sample a vector $\mathbf{X}^m$ with the following rule: let

$$\left.\begin{aligned} X_i^m &\sim N(Y_i, 1) \\ Z_i^m &= 1 \end{aligned}\right\} \text{ with probability } \frac{A^{(s)}}{A^{(s)} + \sum_{l=1}^{i-1} \phi(Y_i - X_l^m)}$$

and for $j = 1, \ldots, i-1$

$$\left.\begin{aligned} X_i^m &= X_j^m \\ Z_i^m &= 0 \end{aligned}\right\} \text{ with probability } \frac{\phi(Y_i - X_j^m)\{j < i\}}{A^{(s)} + \sum_{l=1}^{i-1} \phi(Y_i - X_l^m)}. \quad (7)$$

Table 2. Estimated Bayes Risk of the Means When the Means Have a Normal Distribution with Variance σ^2

	σ^2				σ^2			
Method	.01	1	4	10	.01	1	4	10
		N = 8				*N* = 32		
B	10_6	49_5	84_4	93_3	0_3	53_3	79_2	91_1
JS	40_5	68_5	89_3	98_3	10_4	57_2	80_2	92_1
ML	29_4	74_4	111_5	123_8	9_3	65_3	96_3	114_3
Dir	31_4	64_4	90_4	103_4	7_3	63_2	90_2	96_1
		N = 16				*N* = 50		
B	-4_4	52_4	78_3	88_2	2_3	53_2	82_2	92_1
JS	13_3	65_5	81_2	93_2	8_3	55_2	82_2	92_1
ML	13_8	73_5	103_4	121_4	6_2	63_2	96_5	110_2
Dir	13_3	65_4	87_2	98_2	6_3	60_2	91_2	97_1

NOTE: All units are in hundredths. Subscripts are standard errors. The Bayes risk of the straight estimate is 100 units. Method symbols: B = Bayes estimator; JS = James–Stein estimator; ML = nonparametric maximum likelihood estimator; Dir = Dirichlet process estimator.

For M samples of $\mathbf{X}^m$, with $g^{(s)}$ as the density function or the probability function of $G^{(s)}$, we can estimate $f[\mathbf{Y}\mid(A^{(s)}, G^{(s)})]$ by $\hat{f}[\mathbf{Y}\mid(A^{(s)}, G^{(s)})]$ defined as

$$\hat{f}[\mathbf{Y}\mid(A^{(s)},G^{(s)})]=\frac{1}{M}\sum_{m=1}^{M}\prod_{i=1}^{n}\{Z_i^m\cdot g^{(s)}(X_i^m)+(1-Z_i^m)\}\cdot\left\{\frac{A^{(s)}+\sum_{l=1}^{i-1}\phi(Y_i-X_l^m)}{A^{(s)}+i-1}\right\}. \qquad (8)$$

The next theorem states that the preceding estimate is consistent. The proof of this theorem is contained in Appendix C.

Theorem 3. As $M \to \infty$, $\hat{f}[\mathbf{Y}\mid(A^{(s)}, G^{(s)})] \overset{\text{a.s.}}{\to} f[\mathbf{Y}\mid(A^{(s)}, G^{(s)})]$.

4. THE MONTE CARLO STUDY

A Monte Carlo study was done to compare the Dirichlet process prior estimator with the James–Stein estimator (a parametric empirical Bayes estimator), the nonparametric maximum likelihood estimator (a nonparametric empirical Bayes estimator), the straight estimator, and the Bayes estimate with known G. The Bayes estimator with known G is the best one could possibly do. Because the straight estimator that estimates X_i by Y_i is so simple, one might consider the straight estimator to be the worst estimator that one might be willing to tolerate. The overall study design is described in Section 4.1; some of the Fortran programs used in the study have been provided by Escobar (1988). The results of the Monte Carlo study are contained in Tables 2 and 3; these results are discussed in the Section 4.2.

4.1 Study Design

There are 50 sets of observations, each of size n and distribution G. For each set of observations, first a value of X_i is generated from a distribution G, and then Y_i is the sum of X_i and a generated standard normal. The value of n, which is the number of observations, is either 8, 16, 32, or 50. The unknown distribution G is either normal or symmetric Bernoulli. The normal distributions have a mean 0 and variance either .01, 1, 4, or 10. The symmetric Bernoulli samples have values of σ or $-\sigma$ with equal probability. The parameter σ has values 0, .5, 1, 2, or 5. Note that when σ is greater than 1, the unconditional distribution of Y_i is bimodal.

When calculating the estimator that uses a Dirichlet process prior, the posterior distribution is approximated by reiterating

Table 3. Estimated Bayes Risk of the Means When the Means are Equal to Either σ or $-\sigma$ With Equal Probability

	σ^2					σ^2				
Method	0	.5	1	2	5	0	.5	1	2	5
			N = 8					*N* = 32		
B	6_9	10_8	48_5	28_6	9_6	0_4	21_3	48_3	26_4	-3_3
JS	41_9	46_5	65_4	88_3	100_2	12_4	30_3	60_3	81_2	98_1
ML	29_4	44_4	78_5	68_5	45_4	10_3	32_3	64_3	40_3	12_2
Dir	28_4	41_4	63_4	73_4	53_3	8_3	30_3	65_3	62_2	6_3
			N = 16					*N* = 50		
B	1_6	31_4	53_4	31_5	3_5	-2_3	20_2	43_2	26_3	-2_3
JS	19_5	45_5	70_4	86_2	98_1	5_3	24_2	53_2	82_2	97_1
ML	15_4	44_4	81_4	57_4	28_3	4_2	26_2	53_2	37_2	9_2
Dir	19_3	44_3	72_3	68_2	28_3	3_3	23_2	54_2	52_2	4_3

NOTE: All units are in hundredths. Subscripts are standard errors. The Bayes risk of the straight estimate is 100 units. Method symbols: B = Bayes estimator; JS = James–Stein estimator; ML = nonparametric maximum likelihood estimator; Dir = Dirichlet process estimator.

the Markov chain 16 times. This approximated posterior distribution is sampled 100 times; that is, when using Equation (4), the value of m is 16 and the value of L is 100. When calculating the posterior distribution of the pair (A_0, G_0), the likelihood $f[\mathbf{Y} \mid (A^{(s)}, G^{(s)})]$ is calculated using Equation (8) by drawing M samples, where M is equal to $\min\{n^2, 1024\}$.

Normal distributions are generated by first generating a uniform distribution using algorithm AS183 (Wichmann and Hill 1982) and then calculating the inverse normal distribution function using algorithm AS111 (Beasley and Springer 1977). The symmetric Bernoulli distributions are generated using the uniform distribution generated from algorithm AS183.

The Bayes risk is calculated by first getting the difference between the errors of the test estimator and the error of the straight estimator. This difference is then squared, averaged, and subtracted from 1 to obtain an estimate of the Bayes risk under a squared error lost function. This differencing procedure improves the estimate of the Bayes risk when the Bayes risk is near 1 for a given estimator and distribution G.

4.2 Simulation Results

The results are in Tables 2 and 3. There are times when the Dirichlet process estimate is better than either the James–Stein estimator or the NPEB estimator. Even when the Dirichlet process is not the best, it is usually very close to the best of the other two. Also, each of the other two estimators have weak spots. The NPEB is worse than the simple straight estimator, sometimes 20% worse, especially in the case where G is a normal distribution with large variance. The James–Stein estimator is not very good for bimodal data. It assumes that G is close to a normal distribution and will not offer much of an improvement when G deviates significantly from a normal distribution. As the modes get further apart, the James–Stein estimator collapses to the straight estimator. The posterior estimation of (A_0, G_0) allows for the Dirichlet process estimator to act like either a parametric or nonparametric empirical Bayes method.

5. EXTENSIONS

The algorithm presented in this article could be extended to allow the use of the Dirichlet process prior in more general settings. So far, methods for estimating posterior means with a simple normal error model have been discussed; however, these methods can easily be extended to nonnormal models and to estimate other posterior expectations.

If one wished to calculate the posterior expectation of a function of X, say $\psi(X)$, then this function could be estimated by modifying Equation (4) to

$$\hat{E}[\psi(X_i) \mid \mathbf{Y}] = \frac{1}{L} \sum_{l=1}^{L} E[\psi(X_{i(l)}) \mid X_{j(l)}, j \neq i, \mathbf{Y}]$$

if $\psi(X)$ is a function of just one component of the vector X_i, or the following formula could be used:

$$\hat{E}[\psi(\mathbf{X}) \mid \mathbf{Y}] = \frac{1}{L} \sum_{l=1}^{L} \psi(\mathbf{X}_{(l)})$$

if $\psi(X)$ was a finite function of the whole $\mathbf{X}$ vector.

The method in this article can be extended to nonnormal models with relative ease. Assume that Y_i, given X_i and a fixed parameter θ_i, has the likelihood function $\lambda(Y_i \mid X_i, \theta_i)$ rather than the function $\phi(Y_i - X_i)$ in the algorithm. Replace $\phi(Y_i - X_j)$ by $\lambda(Y_i \mid X_j, \theta_i)$ everywhere in this article. Also, in the sampling procedure (7), replace $A^{(s)}$ by $A_i^{(s^*)}$, and instead of sampling from $N(Y_i, 1)$, sample from the distribution with density $\lambda^*(X \mid Y_i, \theta_i)$, where $A_i^{(s^*)} = A^{(s)} \int \lambda(Y_i \mid X, \theta)\, dX$ and

$$\lambda^*(X \mid Y_i, \theta_i) = \frac{A^{(s)}}{A_i^{(s^*)}} \lambda(Y_i \mid X, \theta_i).$$

Given these substitutions for the nonnormal case, Theorems 1 and 3 are obviously still true. In Theorem 2 the proof presented in the Appendix still applies if $\lambda(Y_i \mid X_i, \theta_i)$ is a bounded function of X_i for fixed values of Y_i and θ_i.

6. CONCLUSION

A method to calculate the posterior distribution for parameters of the Dirichlet process has been presented, and a technique to obtain Bayesian estimates from this method for the means of a normal distribution has also been shown. These estimates have been compared to empirical Bayes estimates.

The Dirichlet process estimator blends together the advantages of both the James–Stein estimator and the NPEB estimator. For fixed (A_0, G_0), the Dirichlet process provides a nonparametric empirical Bayes method that is able to estimate the local features of the unknown G distribution, similar to the NPEB estimator. By putting a prior distribution on (A_0, G_0) and then calculating the posterior distribution for (A_0, G_0), a global parameter is estimated that protects the Dirichlet process from very sparse data.

Besides using this method to estimate posterior means, the algorithm in this article could be used to perform "nonparametric" Bayesian estimates. The use of Ferguson's nonparametric prior has been limited due to its computational difficulties; even previous Monte Carlo algorithms share this problem, because they have not used importance sampling techniques. Now nonparametric Bayesian analyses can be applied to a wide range of statistical procedures.

APPENDIX: PROOFS

A.1 Proof of Theorem 1

For some constant C, by the Bayes theory for Dirichlet processes (Ferguson 1973, p. 217), $dF[X_i \mid X_j, j \neq i] = C[A_0 G_0(dX_i) + \sum_{j \neq i} \delta(X_j, dX_i)]$. Now Bayes's theorem gives

$$dF[X_i \mid X_j, j \neq i, Y_i] = \frac{\phi(Y_i - X_i)\, dF[X_i \mid X_j, j \neq i]}{\int \phi(Y_i - X_i)\, dF[X_i \mid X_j, j \neq i]}.$$

For $j \neq i$, Y_j is conditionally independent of X_i given the X_j's; therefore, $dF[X_i \mid X_j, j \neq i, \mathbf{Y}] = dF[X_i \mid X_j, j \neq i, Y_i]$, which proves the theorem.

A.2 Proof of Theorem 2

This theorem is an application of results from the ergodic theory of Markov chains. Let Ω be the support of G_0 and let $\Gamma \subset \Omega$; then the Markov chain $\{\mathbf{X}^m\}$ has the transition probability given by the stochastic kernel K, where K is defined as

$$K(\mathbf{x}, \Gamma) = \int \cdots \int \{\mathbf{z} \in \Gamma\} P_{X_n|X_1 \cdots X_{n-1}\mathbf{Y}}[dz_n | z_1, z_2, \ldots, z_{n-1}, \mathbf{y}]$$
$$\cdots P_{X_2|X_1X_3 \cdots X_n\mathbf{Y}}[dz_2 | z_1, x_3, \ldots, x_n, \mathbf{y}]$$
$$\cdot P_{X_1|X_2 \cdots X_n\mathbf{Y}}[dz_1 | x_2, \ldots, x_n, \mathbf{y}], \qquad \text{(A.1)}$$

where

$$P_{X_i|X_1 \cdots X_{i-1}X_{i+1} \cdots X_n\mathbf{Y}}[dz_i | z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_n, \mathbf{y}]$$
$$= dF(X_i = z_i | X_1 = z_1, \ldots, X_{i-1} = z_{i-1}, X_{i+1}$$
$$= \mathbf{z}_{i+1}, \ldots, X_n = z_n, \mathbf{Y} = \mathbf{y}).$$

Also, the transition probability, $K^m(\mathbf{x}, \Gamma)$, of m steps in the Markov chain is defined recursively from (A.1) and the following equation: $K^m(\mathbf{x}, \Gamma) = \int K^{m-1}(\mathbf{x}, d\mathbf{z})K(\mathbf{z}, \Gamma)$. To prove the theorem we use the following theorem and four definitions from Feller (1971, pp. 207, 271-272).

Definition 1 (Feller). A measure α is strictly positive in Ω if $\alpha(I) > 0$ for each open interval $I \subset \Omega$. The kernel K is strictly positive if $K(\mathbf{x}, I) > 0$ for each open interval I in Ω.

Given $\mathbf{x}^0$ with initial distribution γ_0, the distributions of $\mathbf{X}^{(1)}$, $\mathbf{X}^{(2)}, \ldots$ are defined recursively as $\gamma_m\{\Gamma\} = \int \gamma_{m-1}(d\mathbf{x})K(\mathbf{X}, \Gamma)$. If γ_0 is defined as an atom at $\mathbf{x}^{(0)}$, then $\gamma_m\{\Gamma\} = K^m(\mathbf{x}^{(0)}, \Gamma)$.

Definition 2 (Feller). The distribution γ_0 is a stationary distribution for K if $\gamma_m = \gamma_0$ for all m; that is, if $\gamma_0(\Gamma) = \int \gamma_0(d\mathbf{x})K(\mathbf{x}, \Gamma)$.

Definition 3 (Feller). The kernel is ergodic if there exists a strictly positive probability distribution α such that $\gamma_m \rightarrow \alpha$ independently of the initial probability distribution γ_0. That is, $K^m(\mathbf{x}, I) \rightarrow \alpha(I) > 0$ for each interval of continuity of α, regardless of the initial value of $\mathbf{x}$.

Given a function u that is bounded and continuous in the underlying interval Ω, define $u = u_0$ and, by induction, u_m as

$$u_m(\mathbf{x}) = \int K(\mathbf{x}, d\mathbf{z})u_{m-1}(\mathbf{z}). \qquad \text{(A.2)}$$

Definition 4 (Feller). The kernel K is regular if the family of transforms u_m is equicontinuous whenever u_0 is bounded and uniformly continuous in Ω.

Theorem (Feller). A strictly positive regular kernel K is ergodic if and only if it possesses a strictly positive stationary probability distribution α.

To prove Theorem 2 we show that K is a strictly positive regular kernel and $P[\mathbf{X}|\mathbf{Y}]$ is a stationary distribution of K, and then apply the preceding theorem from Feller. The kernel K and distribution $P[\mathbf{X}|\mathbf{Y}]$ are obviously both strictly positive on the support of G_0. In Part A we show that $P[\mathbf{X}|\mathbf{Y}]$ is a stationary distribution of K; in Part B we show that K is a regular kernel.

Part A. The probability $P[\mathbf{X}|\mathbf{Y}]$ is a stationary probability for K.

Proof of A. We need to show that

$$P[\{\mathbf{X} \in \Gamma\}|\mathbf{Y}] = \int P[d\mathbf{x}|\mathbf{Y}]K(\mathbf{x}, \Gamma). \qquad \text{(A.3)}$$

To simplify the notation, we prove only the $n = 3$ case. The extension to the general case is straightforward. To prove (A.3), the following basic properties are needed:

$$\int \cdots \int \psi_{23}(x_2, x_3)P_{X_1X_2X_3}[d(x_1, x_2, x_3)]$$
$$= \int \cdots \int \psi_{23}(x_2, x_3)P_{X_2X_3}[d(x_2, x_3)], \qquad \text{(A.4)}$$

where ψ_{23} is any measurable function and is not a function of x_1, and

$$\int \psi(x_1, x_2, x_3)P_{X_1|X_2X_3}[dx_1 | x_2, x_3]P_{X_2X_3}[d(x_2, x_3)]$$
$$= \int \psi(x_1, x_2, x_3)P_{X_1X_2X_3}[d(x_1, x_2, x_3)], \qquad \text{(A.5)}$$

where ψ is a measurable function. In the following series of equations, let everything be conditional on $\mathbf{Y}$. To prove the results it is necessary to show that $\int K(\mathbf{x}, \Gamma)P(d\mathbf{x}) = P(\mathbf{X} \in \Gamma)$. By repeated applications of (A.4) and (A.5),

$$\int K(\mathbf{x}, \Gamma)P(d\mathbf{x}) = \int K[(x_1, x_2, x_3), \Gamma]P[d(x_1, x_2, x_3)]$$
$$= \int \{(z_1, z_2, z_3) \in \Gamma\}K[(x_1, x_2, x_3), d(z_1, z_2, z_3)]P_{X_1X_2X_3}[d(x_1, x_2, x_3)]$$
$$= \int \{(z_1, z_2, z_3) \in \Gamma\}P_{X_3|X_1X_2}[dz_3 | z_1, z_2]P_{X_2|X_1X_3}[dz_2 | z_1, x_3]P_{X_1|X_2X_3}[dz_1 | x_2, x_3]P_{X_1X_2X_3}[d(x_1, x_2, x_3)]$$
$$= \int \{(z_1, z_2, z_3) \in \Gamma\}P_{X_3|X_1X_2}[dz_3 | z_1, z_2]P_{X_2|X_1X_3}[dz_2 | z_1, x_3]P_{X_1|X_2X_3}[dz_1 | x_2, x_3]P_{X_2X_3}[d(x_2, x_3)]$$
$$= \int \{(z_1, z_2, z_3) \in \Gamma\}P_{X_3|X_1X_2}[dz_3 | z_1, z_2]P_{X_2|X_1X_3}[dz_2 | z_1, x_3]P_{X_1X_2X_3}[d(z_1, x_2, x_3)]$$
$$= \int \{(z_1, z_2, z_3) \in \Gamma\}P_{X_3|X_1X_2}[dz_3 | z_1, z_2]P_{X_2|X_1X_3}[dz_2 | z_1, x_3]P_{X_1X_3}[d(z_1, x_3)]$$
$$= \int \{(z_1, z_2, z_3) \in \Gamma\}P_{X_3|X_1X_2}[dz_3 | z_1, z_2]P_{X_1X_2X_3}[d(z_1, z_2, x_3)]$$
$$= \int \{(z_1, z_2, z_3) \in \Gamma\}P_{X_3|X_1X_2}[dz_3 | z_1, z_2]P_{X_1X_2}[d(z_1, z_2)]$$
$$= \int \{(z_1, z_2, z_3) \in \Gamma\}P_{X_1X_2X_3}[d(z_1, z_2, z_3)]$$
$$= P[\mathbf{X} \in \Gamma].$$

Therefore, Equation (A.3) is true and, in turn, Part A is true.

Part B. The stochastic kernel K is regular.

Proof of B. We have to show that for all $u_0(\mathbf{x})$, a bounded uniformly continuous function, the family of functions $\{u_m\}$ defined by Equation (A.2) is equicontinuous.

Let the symbol $\|\mathbf{x}\|_\infty$ be the sup norm for a vector $\mathbf{x}$ defined as $\|\mathbf{x}\|_\infty = \max_i\{|x_i|\}$. (Note: Use of the sup norm instead of the usual euclidian norm is not restrictive, because for any vector $\mathbf{x} \in \mathfrak{R}^n$, $\|\mathbf{x}\|_2/\sqrt{n} \leq \|\mathbf{x}\|_\infty \leq \|\mathbf{x}\|_2$, where the euclidian norm is defined as $\|\mathbf{x}\|_2 = (\sum_{i=1}^n x_i^2)^{1/2}$. Therefore, the arguments that follow could be rewritten using the usual euclidian norm, but the sup norm is more straightforward for this proof.)

To simplify the notation, we work only with the case $n = 2$. The extension to the general case is more complex, but it follows the same basic argument.

Using Equations (A.2), (A.1), and (2), u_m is equal to the following:

$$u_m(x_1, x_2) = \int u_{m-1}(z_1, z_2) K[(x_1, x_2), d(z_1, z_2)]$$

$$= \left\{ \left[\frac{u_{m-1}(x_2, x_2)\phi(Y_2 - x_2) + \int u_{m-1}(x_2, z_2)\phi(Y_2 - z_2) A_0 G_0(dz_2)}{A(Y_2) + \phi(Y_2 - x_2)} \right] \phi(Y_1 - x_2) \right.$$

$$\left. + \int \frac{[u_{m-1}(z_1, z_1)\phi(Y_2 - z_1) + \int u_{m-1}(z_1, z_2)\phi(Y_2 - z_2) A_0 G_0(dz_2)]}{A(Y_2) + \phi(Y_2 - z_1)} \phi(Y_1 - z_1) A_0 G_0(dz_1) \right\} \cdot [A(Y_1) + \phi(Y_1 - x_2)]^{-1}.$$

To simplify the right side of this equation, define the following functions:

$$p_i(x) = \frac{\phi(Y_i - x)}{A(Y_i) + \phi(Y_i - x)},$$

$$B_m(x) = \int u_m(x, z) \frac{\phi(Y_2 - z)}{A(Y_2)} A_0 G_0(dz),$$

and

$$C_m = \int \{u_m(z, z) p_2(z) + B_m(z)[1 - p_2(z)]\} \times \frac{\phi(Y_1 - z)}{A(Y_1)} A_0 G_0(dz).$$

Therefore, $u_m(x_1, x_2)$ reduces to

$$u_m(x_1, x_2) = \{p_2(x_2) u_{m-1}(x_2, x_2) + [1 - p_2(x_2)] B_{m-1}(x_2)\} \times p_1(x_2) + [1 - p_1(x_2)] C_{m-1}.$$

Because the function $\phi(Y_i - x)$ is greater than 0 and bounded by $1/\sqrt{(2\pi)}$, and because $\underline{Y_i}$ is fixed in this theorem, $p_i(x)$ is bounded by $1/\{1 + A(Y_i)\sqrt{2\pi}\}$, which is strictly less than 1. The function $p_i(x)$ is also uniformly continuous, because it is bounded and goes to 0 as x goes to positive or negative infinity.

By the hypothesis, u_0 is bounded in absolute value; therefore, there exists a number M such that $\sup_{\mathbf{x}} |u_0(\mathbf{x})| \leq M < \infty$. By definition, u_1 is an average of u_0; therefore, u_1 is bounded by M. Repeating this argument for all m, we find that for all m, $\sup_{\mathbf{x}} |u_m(\mathbf{x})| \leq M < \infty$. Also, for each B_m and C_m, these functions are averages of the function u_m; therefore, for all m, these functions are also bounded in absolute value by M.

Also, if u_m is absolutely continuous, then so is B_m absolutely continuous with the same δ's and ε's. That is, if there exist a δ and an ε such that $|u_m(x_1', x_2') - u_m(x_1'', x_2'')| \leq \varepsilon$ whenever $\|(x_1', x_2') - (x_1'', x_2'')\|_\infty \leq \delta$, then for the same δ and ε we have $|B_m(x_2') - B_m(x_2'')| \leq \varepsilon$ whenever $|x_2' - x_2''| \leq \delta$. This is true because if $\|(x_1', x_2') - (x_1'', x_2'')\|_\infty \leq \delta$ implies that $|u_m(x_2', x_2') - u_m(x_1'', x_2'')| \leq \varepsilon$, then for all $|x_2' - x_2''| \leq \delta$ we have

$$|B_m(x_2') - B_m(x_2'')| \leq \int |u_m(x_2', z) - u_m(x_2'', z)| \times \frac{A_0 \phi(Y_2 - z) G_0(dz)}{A(Y_2)}$$

$$\leq \int \varepsilon \frac{A_0 \phi(Y - z) G_0(dz)}{A(Y_2)} = \varepsilon.$$

The rest of the proof of Part B follows from natural induction. First, because u_0 is uniformly continuous, there exist a δ_1 and an ε such that for all (x_1', x_2') and (x_1'', x_2'') such that $\|(x_1', x_2') - (x_1'', x_2'')\|_\infty \leq \delta_1$, $|u_0(x_1', x_2') - u_0(x_1'', x_2'')| \leq \varepsilon$.

Also, because $p_2(\cdot)$ and $p_1(\cdot)$ are both uniformly continuous and do not depend on m, there exist a δ_2 such that for all i ($i = 1, 2$) and for all x_2' and x_2'', where $|x_2' - x_2''| \leq \delta_2$,

$$|p_i(x_2') - p_i(x_2'')| \leq \frac{\varepsilon\sqrt{2\pi} A(Y_1)}{4M[1 + \sqrt{2\pi} A(Y_1)]}.$$

Now assume that for all (x_1', x_2') and (x_1'', x_2'') such that $\|(x_1', x_2') - (x_1'', x_2'')\|_\infty \leq \min(\delta_1, \delta_2)$, $|u_{m-1}(x_1', x_2') - u_{m-1}(x_1'', x_2'')| \leq \varepsilon$. By the preceding argument, $|B_{m-1}(x_2') - B_{m-1}(x_2'')| \leq \varepsilon$, and we have

$$|u_m(x_1', x_2') - u_m(x_1'', x_2'')|$$

$$\leq p_1(x_2')\{|p_2(x_2') - p_2(x_2'')| \cdot [|u_{m-1}(x_2', x_2')| + |B_{m-1}(x_2')|]$$

$$+ p_2(x_2'') \cdot |u_{m-1}(x_2', x_2') - u_{m-1}(x_2'', x_2'')|$$

$$+ [1 - p_2(x_2'')] \cdot |B_{m-1}(x_2') - B_{m-1}(x_2'')|\}$$

$$+ |p_1(x_2') - p_1(x_2'')| \cdot |p_2(x_2'') u_{m-1}(x_2'', x_2'')$$

$$+ [1 - p_2(x_2'')] \cdot B_{m-1}(x_2'') - C_{m-1}|$$

$$\leq [1 + \sqrt{2\pi} A(Y_1)]^{-1}[|p_2(x_2') - p_2(x_2'')| \cdot 2M + \varepsilon]$$

$$+ |p_1(x_2') - p_1(x_2'')| \cdot 2M$$

$$\leq 2M \cdot |p_2(x_2') - p_2(x_2'')| + 2M \cdot |p_1(x_2') - p_1(x_2'')|$$

$$+ \varepsilon[1 + \sqrt{2\pi} A(Y_1)]^{-1} \leq \varepsilon.$$

Therefore, because u_0 is bounded and uniformly continuous, the family of functions $\{u_m\}$ is equicontinuous by natural induction. Thus Part B is true, and the theorem is proven.

A.3 Proof of Theorem 3

If the expected value of $\hat{f}$ is equal to f in the preceding equation, then the theorem is true by the strong law of large numbers. So the expected value of $\hat{f}$ is

$$\begin{aligned}
E\{\hat{f}[\mathbf{Y}\,|(A^{(s)}, G^{(s)})]\} &= \int \prod_{i=1}^{n} \{Z_i \cdot g^{(s)}(X_i) + (1 - Z_i)\} \cdot \left\{\frac{A^{(s)} + \sum_{l=1}^{i-1} \phi(Y_i - X_l)}{A^{(s)} + i - 1}\right\} \\
&\quad \times \prod_{i=1}^{n} \left\{\frac{A^{(s)} \cdot \delta(1, dZ_i) \cdot \phi(Y_i - X_j) dX_i + \delta(0, dZ_i) \cdot \sum_{j=1}^{i-1} [\phi(Y_i - X_j)\delta(X_j, dX_i)]}{A^{(s)} + \sum_{j=1}^{i-1} \phi(Y_i - X_j)}\right\} \\
&= \int \prod_{i=1}^{n} \frac{A^{(s)} g^{(s)}(X_i) \phi(Y_i - X_i)\, dX_i + \sum_{j=1}^{i-1} \phi(Y_i - X_i)\delta(X_j, dX_i)}{A^{(s)} + i - 1} \\
&= \int \prod_{i=1}^{n} \phi(Y_i - X_i) \frac{A^{(s)} G^{(s)}(dX_i) + \sum_{j=1}^{i-1} \delta(X_j, dX_i)}{A^{(s)} + i - 1} \\
&= f[\mathbf{Y}\,|(A^{(s)}, G^{(s)})].
\end{aligned}$$

The second equality is obtained by integrating over Z_i and collecting terms.

[Received March 1991. Revised January 1993.]

REFERENCES

Antoniak, C. E. (1974), "Mixtures of Dirichlet Processes With Applications to Bayesian Nonparametric Problems," *The Annals of Statistics,* 2, 1152–1174.

Beasley, J. D., and Springer, S. G. (1977), "The Percentage Points of the Normal Distribution," *Applied Statistics,* 26, 118–121.

Berry, D. A., and Christensen, R. (1979), "Empirical Bayes Estimation of a Binomial Parameter via Mixtures of a Dirichlet Process," *The Annals of Statistics,* 7, 558–568.

Blackwell, D., and MacQueen, J. B. (1973), "Ferguson Distribution via Polya Urn Schemes," *The Annals of Statistics,* 1, 353–355.

Efron, B., and Morris, C. (1973a), "Stein's Estimation Rule and its Competitors—an Empirical Bayes Approach," *Journal of the American Statistical Association,* 68, 117–130.

——— (1973b), "Combining Possibly Related Estimation Problems," *Journal of the Royal Statistical Society,* Ser. B, 35, 379–421.

——— (1975), "Data Analysis Using Stein's Estimator and Its Generalizations," *Journal of the American Statistical Association,* 70, 311–319.

Escobar, M. D. (1988), "Estimating the Means of Several Normal Populations by Nonparametric Estimation of the Distribution of the Means," unpublished Ph.D. dissertation, Yale University, Dept. of Statistics.

——— (1992), Invited comment on "Bayesian Analysis of Mixtures: Some Results on Exact Estimability and Identification," by J.-P. Florens, M. Mouchart, and J.-M. Rolin, in *Bayesian Statistics 4,* eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, Oxford, U.K.: Oxford University Press, pp. 142–144.

Feller, W. (1971), *An Introduction to Probability Theory and Its Applications,* (Vol. 2, 2nd ed.), New York: John Wiley.

Ferguson, T. S. (1973), "A Bayesian Analysis of Some Nonparametric Problems," *The Annals of Statistics,* 1, 209–230.

——— (1974), "Prior Distributions on Space of Probability Measures," *The Annals of Statistics,* 2, 615–629.

Gelfand, A. E., and Smith, A. F. M. (1990), "Sampling-Based Approaches to Calculating Marginal Densities," *Journal of the American Statistical Association,* 85, 398–409.

Geman, S., and Geman, D. (1984), "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 6, 721–741.

James, W., and Stein, C. (1961), "Estimating With Quadratic Loss," in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 1), pp. 361–379.

Kuo, L. (1986), "Computations of Mixtures of Dirichlet Processes," *SIAM Journal of Science and Statistical Computing,* 7, 60–71.

Laird, N. M. (1978), "Nonparametric Maximum Likelihood Estimation of a Mixing Distribution," *Journal of the American Statistical Association,* 73, 805–811.

——— (1981), "Empirical Bayes Estimates Using Nonparametric Maximum Likelihood Estimate for the Prior," *Journal of Statistical Computing and Simulation,* 15, 211–220.

Lindsay, B. G. (1983), "The Geometry of Mixing Likelihoods: A General Theory," *The Annals of Statistics,* 11, 86–94.

Lo, A. Y. (1984), "On a Class of Bayesian Nonparametric Estimates: I. Density Estimates," *The Annals of Statistics,* 12, 351–357.

Morris, C. (1983), "Parametric Empirical Bayes Inference: Theory and Applications," *Journal of the American Statistical Association,* 78, 47–65.

Robbins, H. (1955), "An Empirical Bayes Approach to Statistics," in *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 1), pp. 157–164.

Rubin, D. (1988), "Using the SIR Algorithm to Simulate Posterior Distributions," in *Bayesian Statistics 3,* eds. J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith, Oxford, U.K.: Oxford University Press, pp. 395–404.

Schervish, M. J., and Carlin, B. P. (1992), "On the Convergence of Successive Substitution Sampling," *Journal of Computational and Graphical Statistics,* 1, 111–127.

Stein, C. (1955), "Inadmissibility of the Usual Estimators for the Mean of a Multivariate Normal Distribution," in *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 1), pp. 197–206.

Tanner, M. A., and Wong, W. H. (1987), "The Calculation of Posterior Distributions by Data Augmentation" (with discussion), *Journal of the American Statistical Association,* 82, 528–550.

Titterington, D. M., Smith, A. F. M., and Makov, U. E. (1985), *Statistical Analysis of Finite Mixture Distributions,* New York: John Wiley.

Wichmann, B. A., and Hill, I. D. (1982), "An Efficient and Portable Pseudo-Random Number Generator," *Applied Statistics,* 31, 188–190.