# BAYESIAN DENSITY ESTIMATION VIA DIRICHLET DENSITY PROCESSES

by

Mauro Gasparini
Purdue University

Technical Report #93-23

Department of Statistics
Purdue University

May 1993
Revised December 1994

△

# Bayesian Density Estimation via Dirichlet Density Processes

Mauro Gasparini[1]

*Purdue University*

**Abstract**

For the purpose of nonparametric density estimation, a prior distribution is constructed on the space of stepwise constant density functions, not necessarily of bounded support. In particular, the sequence of heights is conditionally distributed *a priori* as a Dirichlet process on the integers, given a bidimensional mixing parameter with location and scale components which are, in turn, assigned an arbitrary marginal prior.

Proper Bayesian estimates of the density are obtained. They are not histograms, but they share common features with the histogram and other kernel based estimators. They also incorporate prior information, like a prior guess for the density or bounds for its support, which may be particularly appealing for small sample situations, where usual density estimation methods are not satisfactory.

The estimates are computable by simple numerical methods, as opposed to other nonparametric Bayesian density estimators proposed in the literature, which display significant computational problems. Moreover a simple way to generate density functions from the posterior process allows simulation of Bayesian intervals for functionals of the density, in particular for the density itself, evaluated at selected points.

---

[1]Research partially supported by the National Science Foundation, Grant DMS-9303556.

[0]*Key words and phrases.* Nonparametric Density Estimation, Dirichlet priors, Histogram.

# 1 Introduction

Consider a statistical model in which random variables $X_1, \ldots, X_n$ are independent and identically distributed with an unknown density function $f$. The focus is on estimating $f$, or a functional $\phi(f)$ of it, in the presence of a sample $X_1 = x_1, \ldots, X_n = x_n$. The uncertainty about $f$ may be such that any finite dimensional parametrization of the density is regarded as insufficient. On the other hand, some information may be available *a priori* about its location, shape, smoothness or other characteristics. The researcher may decide to quantify this prior information by assigning a prior to a nonparametric neighborhood of some prior guess $f_o$. Then the problem arises of specifying a prior distribution for $f$, lying in an infinite dimensional space of functions. In other words, we seek a model for Bayesian nonparametric density estimation.

Several proposals have already been made to exloit the computational advantages of Dirichlet process priors in the density estimation setup (for recent discussions see Hjort [8] and Hartigan [6]). These prior processes, introduced by Ferguson [2], have proven very useful for many other nonparametric Bayes problems. The computational advantages of Dirichlet-type priors rest on the conjugate character of the Dirichlet distribution to multinomial sampling. To understand how they could be used in the context of density estimation, consider first the following very simple histogram-like prior: partition a finite interval of the real line into subintervals of equal length and imagine a random density $f$, stepwise constant over those intervals. Now suppose the heights of the density are given a joint Dirichlet prior distribution and are appropriately scaled. Next, obtain a sample $x_1, \ldots, x_n$ from $f$. Then the multinomial likelihood $\prod_{i=1}^n f(x_i)$ trivially produces a posterior on the heights of $f$ which is again a scaled Dirichlet. Leonard [10] considers models of this kind, but regards them as unsuitable to density estimation because of the lack of local prior correlation between the ordinates of $f$. It is also obvious that the finite dimension of such a rigid specification of $f$ does not allow many density functions to be in the support of the prior.

It is the purpose of this paper to consider a generalization of the simple model above, refined in such a way that one can still take advantage of the computational ease, but also obtain enough flexibility in the prior to correct for the aforementioned problems. In particular, to give a summary of the proposed method in simple terms, take the interval described above to be the whole real line and generalize the Dirichlet distribution to an infinite dimensional distribution (a process on the integers). Such a generalization is naturally given by an appropriate Dirichlet process. Next, take the bin width $\lambda$ to be a random positive quantity, thus allowing densities not necessarily stepwise constant to enter the support of the prior. Finally, also take a location parameter $\mu$ to be random.

From a methodological viewpoint, such a prior represents a specification which the researcher may not literally believe, but which allows for the approximate representation of a great deal of prior information, for example, centering around a prior guess.

In a related paper Lo [11] defines a prior random density $f$ as

$$f(z) = \int K(z,u)G(du),$$

where $K(z,u)$ is a fixed density kernel and $G$ is a Dirichlet process. Ferguson [3] details an example with a normal kernel. The explicit form obtained for the posterior expectation of $f(z)$ requires summing over all partitions of the sample size, so that exact computations are prohibitive even for small data sets. More recent works by West, Müller and Escobar ([14], v. bibliography) explore the computational usefulness of the Gibbs sampler for this kind of models. The prior considered by all these authors is kernel-based, as is the one proposed in this paper, but location and weight of the kernels are controlled by the same Dirichlet index. In this paper, instead, they vary according to an arbitrary marginal distribution. On the other hand, only a simple rectangular kernel is considered.

Approximate posterior expectations of the density are obtained by simple numerical methods. Computations are possible for any sample size, up to a certain degree of numerical accuracy. Furthermore, posterior distributions, and in particular Bayesian intervals, for any functional $\phi(f)$ of the density can be constructed by simulation.

## 2 Prior and Posterior Processes.

In the following Lemma a formal construction of the random density described in Section 1 is given.

**Lemma 1** *Let $\alpha_k(\mu,\lambda), k \in \mathcal{N} := 0, \pm 1, \pm 2, \ldots$ be a double sequence of nonnegative measurable functions defined for $0 \leq \mu < \lambda$, in such a way that*

$$\alpha(\mu,\lambda) := \sum_{k\in\mathcal{N}} \alpha_k(\mu,\lambda) < \infty \quad \forall \mu,\lambda.$$

*Let the stochastic double sequence $(M, \Lambda, G_0, G_{\pm 1}, G_{\pm 2}, \ldots)$ be defined on some probability space and take values in $\mathrm{R}^+ \times \mathrm{R}^+ \times [0,1] \times [0,1] \times \ldots$, in such a way that the distribution of $(M,\Lambda)$ has Lebesgue density $\pi(\mu,\lambda)$ and, for any integer $N$, the vector*

$$(G_0, G_{\pm 1}, \ldots, G_{\pm N}, 1 - \sum_{|k|>N} G_k)$$

*has, conditionally on any given $(\mu,\lambda)$, a Dirichlet distribution with index*

$$(\alpha_0(\mu,\lambda), \alpha_{\pm 1}(\mu,\lambda), \ldots, \alpha_{\pm N}(\mu,\lambda), \sum_{|k|>N} \alpha_k(\mu,\lambda)).$$

*Define a random step function $f$ by*

$$f(z) := \sum_{k\in\mathcal{N}} \frac{G_k}{\Lambda} \{M + (k-1)\Lambda < z \leq M + k\Lambda\} = \frac{G_{c(M,\Lambda)}}{\Lambda}, \tag{1}$$

*where $\{A\}$ is the indicator of event $A$ and $c(\mu,\lambda) := \min\{n \in \mathcal{N} : n \geq (z-\mu)/\lambda\}$. Then $f$ is almost surely a probability density function.*

**Proof.** It is easily seen that the process $(M, \Lambda, G_0, G_{\pm 1}, G_{\pm 2}, \ldots)$ with the desired properties exists. The distributions of $(M, \Lambda, G_0, G_{\pm 1}, \ldots, G_{\pm N})$ are consistent, so that existence follows from Kolmogorov's extension theorem. The rest of the proof is straightforward. ∎

A random function $f$ as in Lemma 1 is said to be a *Dirichlet Density Process* with indices $\pi(\mu,\lambda)$ and $\{\alpha_k(\mu,\lambda), k = 0, \pm 1, \pm 2, \ldots\}$, for $0 \leq \mu < \lambda < \infty$.

Notice that the assumption that $(M, \Lambda)$ has a Lebesgue density is not essential in Lemma 1, although it helps giving $f$ desirable properties. In particular, the joint density of $(M, \Lambda, G_0, G_{\pm 1}, \ldots, G_{\pm N})$ evaluated at $(\mu, \lambda, g_0, g_{\pm 1}, \ldots, g_{\pm N})$ is

$$\pi(\mu,\lambda)\frac{\Gamma(\alpha(\mu,\lambda))(\prod_{k=-N}^{N} g_k^{\alpha_k(\mu,\lambda)-1})(1-\sum_{j=-N}^{N} g_j)^{(\sum_{|i|>N}\alpha_i(\mu,\lambda))-1}}{(\prod_{k=-N}^{N}\Gamma(\alpha_k(\mu,\lambda)))\Gamma(\sum_{|i|>N}\alpha_i(\mu,\lambda))} \tag{2}$$

for each $N = 1, 2, \ldots$ and for $g_k \geq 0, \sum_k g_k \leq 1, k = 0, \pm 1, \ldots, \pm N$.

The construction in Lemma 1 implies that the infinite double sequence $\{G_k; k = 0, \pm 1, \pm 2, \ldots\}$ is, formally, a mixture of Dirichlet processes on the integers (Antoniak [1]), that is, a mixture of infinite dimensional Dirichlet distributions.

Now suppose that a Dirichlet Density Process is used as a prior in a nonparametric Bayesian approach to density estimation, when a random sample from $f$ is available.

**Theorem 1** *Let $X_1, \ldots, X_n$ be, conditionally on $f$, i.i.d. random variables with density $f$ and let $f$ be a Dirichlet Density Process as in Lemma 1. Then, conditionally on $X_1 = x_1, \ldots, X_n = x_n$, $f$ is a Dirichlet Density Process with indices*

$$\pi^*(\mu,\lambda|x_1,\ldots,x_n) = \frac{\pi(\mu,\lambda)\prod_{k\in\mathcal{N}}\alpha_k(\mu,\lambda)_{[n_k(\mu,\lambda)]}/(\lambda^n\alpha(\mu,\lambda)_{[n]})}{\int_{0\leq\mu'<\lambda'<\infty}\pi(\mu',\lambda')\prod_{k\in\mathcal{N}}\alpha_k(\mu',\lambda')_{[n_k(\mu',\lambda')]}/(\lambda'^n\alpha(\mu',\lambda')_{[n]})d\mu' d\lambda'}, \tag{3}$$

*for $0 \leq \mu < \lambda < \infty$ and*

$$\{\alpha_k(\mu,\lambda) + n_k(\mu,\lambda), k = 0, \pm 1, \pm 2, \ldots\},$$

*where*

$$n_k(\mu,\lambda) := \#\{i \leq n : \mu + (k-1)\lambda < x_i \leq \mu + k\lambda\}$$

*and $x_{[0]} := 1$, and $x_{[k]} := x(x+1)\ldots(x+k-1)$, for any $x \in \mathrm{R}$ and any positive integer $k$, is the ascending factorial.*

**Proof.** By the properties of marginals and moments of Dirichlet random vectors (see Wilks [15]) we have the likelihood

$$l(x_1,\ldots,x_n|\mu,\lambda,g_0,g_{\pm 1},\ldots,g_{\pm N}) =$$

$$
\begin{aligned}
&= \mathrm{E}(l(x_1,\ldots,x_n|M,\Lambda,G_0,G_{\pm1},G_{\pm2},\ldots)|\mu,\lambda,g_0,g_{\pm1},\ldots,g_{\pm N}) \\
&= \frac{(\prod_{k=-N}^{N} g_k^{n_k(\mu,\lambda)})}{\lambda^n}\mathrm{E}(\prod_{|k|>N} G_k^{n_k(\mu,\lambda)}|\mu,\lambda,g_0,g_{\pm1},\ldots,g_{\pm N}) \\
&= \frac{(\prod_{k=-N}^{N} g_k^{n_k(\mu,\lambda)})}{\lambda^n}(1-\sum_{i=1}^{N} g_i)^{\sum_{|j|>N} n_j(\mu,\lambda)} \qquad (4) \\
&\qquad \times \frac{\prod_{|l|>N} \alpha_l(\mu,\lambda)_{[n_l(\mu,\lambda)]}}{(\sum_{|k|>N}\alpha_k(\mu,\lambda))_{[\sum_{|j|>N} n_j(\mu,\lambda)]}}.
\end{aligned}
$$

The second product in (4) is finite, and actually void if $N$ is large enough.

The posterior finite dimensional distributions of the stochastic double sequence $(M,\Lambda,G_0,G_{\pm1},G_{\pm2},\ldots)$ thus have Lebesgue densities proportional to the product of (2) and (4). The posterior structure as in the statement of the Theorem can then be easily recognized. ∎

This useful selfreproductive property is analogous to the property described in Theorem 1 of Ferguson [2].

# 3 Bayes estimation and choice of the prior

Now consider the Bayesian estimation of $f$ in the above setup. There is no loss of generality in solving first the simpler *no data* problem, since the posterior process has the same structure as the prior. The expectation of $f(z)$ (the Bayes estimate under squared error loss) conditionally on a given $(\mu,\lambda)$ has a simple form, but the integration over $(\mu,\lambda)$ is to be done with numerical methods.

With the convention $1/0=\infty$, a *no data* estimator of $f(z)$ is the prior expectation

$$
\mathrm{E}(f(z)) = \mathrm{E}(\mathrm{E}(\frac{G_{c(M,\Lambda)}}{\Lambda}|M,\Lambda)) = \int\int_{0\le\mu\le\lambda<\infty} \frac{\alpha_{c(\mu,\lambda)}(\mu,\lambda)}{\alpha(\mu,\lambda)\lambda}\pi(\mu,\lambda)d\mu d\lambda. \qquad (5)
$$

We therefore obtain the following

**Corollary 1** *The posterior mean of $f(z)$ is given by*

$$
\tilde{f}_n(z) \;:=\; \mathrm{E}(f(z)|x_1,\ldots,x_n) = \qquad (6)
$$

$$\int\int_{0\leq\mu\leq\lambda<\infty} \frac{\alpha_{c(\mu,\lambda)}(\mu,\lambda)+n_{c(\mu,\lambda)}(\mu,\lambda)}{(\alpha(\mu,\lambda)+n)\lambda}\pi^*(\mu,\lambda|x_1,\ldots,x_n)d\mu d\lambda.$$

Now suppose that a known density $f_o$, derivative of a distribution function $F_o$, is considered a suitable prior guess for the density. The goal is to approximate this prior guess with a Dirichlet Density prior process centered around $f_o$. For this purpose, a natural choice for any given $(\mu,\lambda)$, is the following:

$$\frac{\alpha_k(\mu,\lambda)}{\alpha(\mu,\lambda)} = F_o(\mu+k\lambda)-F_o(\mu+(k-1)\lambda), \qquad k\in\mathcal{N}. \tag{7}$$

For example, if the prior guess $f_o$ is a standard normal, we are not putting any prior mass on it or on any other normal density, but we are giving probability 1 to a set of step functions which approximate it. The prior guess is well centered around $f_o$, although the analytical expression for it, formula (5), is not that transparent. For example, in the normal example of the next Section the highest relative difference between the prior guess and the prior mean occurs at $z=0$ and is about 3%.

By substituting (7) in (6) we have

$$\begin{aligned}\tilde{f}_n(z) \;=\; \int\int_{0\leq\mu\leq\lambda<\infty} &[\frac{\alpha(\mu,\lambda)}{\alpha(\mu,\lambda)+n}F_o(I_{\mu,\lambda}(z)) \\ &+(1-\frac{\alpha(\mu,\lambda)}{\alpha(\mu,\lambda)+n})F_n(I_{\mu,\lambda}(z))]\frac{1}{\lambda}\pi^*(\mu,\lambda|x_1,\ldots,x_n)d\mu d\lambda\end{aligned} \tag{8}$$

where $I_{\mu,\lambda}(z) := (\mu+(c(\mu,\lambda)-1)\lambda, \mu+c(\mu,\lambda)\lambda]$ is an interval containing $z$, $F_n$ is the empirical distribution function and the symbols $F_o$ and $F_n$ are used to denote both the distribution function and the corresponding measure. The usual representation of a Bayes estimate as a weighted average of prior and empirical is contained in the integrand of equation (8); one further averaging operation takes place when the integration over $\mu,\lambda$ is performed.

Note that estimator (8) is very similar an average shifted (Scott [12]), or WARPed (see Härdle [7]) histogram, the main difference being in the presence of a prior guess.

The limiting case with $\alpha(\mu,\lambda)\equiv 0$ and $(\mu,\lambda)$ degenerate at a specified value $(\mu_o,\lambda_o)$ reduces to a histogram with bin width $\lambda_o$.

As $n \to \infty$ - the other, more important, limiting case - the conjecture is that the posterior on $(\mu, \lambda)$ will degenerate to a unit point mass on the point $(0, 0)$, allowing for consistency of the corresponding estimates. Pointwise consistency has been proved in Gasparini [4] for the special case of $\mu$ identically equal to 0.

# 4 Toward computations

The estimation method presented in this paper is fairly intuitive but intrinsically numerical. The presence of the counts $n_k(\mu, \lambda)$ in (3) makes it impossible to have natural conjugate priors for $\pi(\mu, \lambda)$ and $\pi^*(\mu, \lambda | x_1, \ldots, x_n)$ is quite an irregular function, whose behaviour is deeply connected to the distribution of the spacings $X_i - X_j$, $i, j \leq n$. Large jumps of $\pi^*(\mu, \lambda | x_1, \ldots, x_n)$ are due, roughly speaking, to the way the pair $(\mu, \lambda)$ fits in the interstices between the spacings.

Abandoned the search for conjugate priors, the prior density $\pi(\mu, \lambda)$ is then to be approximated numerically by a discrete probability mass function over a grid of $(\mu_i, \lambda_i)$ values, $i = 1, 2, \ldots, I$ such that $0 \leq \mu_i \leq \lambda_i \leq \lambda_{max}$, where $I$ is a large number and $\lambda_{max}$ an arbitrary upper bound for the support of $\lambda$. From a practical point of view, $\lambda_{max}$ can be of the order of magnitude of the seminterquartile range of the prior guess, say. If the prior guess is not too far off, $\lambda_{max}$ is then a large fraction of the sample range and, since $\lambda$ plays a role similar to the bandwidth in kernel density estimation, few objections can be raised to that. Typically, only few of the $(\mu_i, \lambda_i)$ points carry significant posterior mass.

# 5 Frequentist Monte Carlo simulations

This section contains a comparison of the frequentist performance of the Bayes estimator $\tilde{f}_n$ to an optimal kernel density estimator. The comparison is aimed at highlighting some of the properties of $\tilde{f}_n$, rather than comparing the efficiency of the two methods, since,

loosely speaking, if the true density $f$ is smooth, a kernel estimate is perhaps the best known method over repeated sampling in the absence of prior information, but it can be outperformed, trivially, by a Bayes estimate centered around a prior guess which happens to be close to the true underlying $f$. The following simulations can only give an idea of what amount of correct prior information is necessary for the Bayes estimate to outperform the kernel estimate in a regular (normal) case. Given the affinity between $\tilde{f}_n$ and WARPed histograms, it would probably be more sensible to compare these two procedures, rather than $\tilde{f}_n$ and a kernel estimate. On the other hand, since WARPed histograms are a bridge between the histogram and the kernel estimate, they improve their performance as they converge to the latter and it is therefore impossible to find a WARPed histogram to use as a term of comparison.

1000 samples of $n = 100$ points each were pseudo-randomly generated from a normal with mean 0 and variance 1. The Bayes estimate (8) was calculated for each sample using the default prior described in the previous section: a grid of $25 \times 35$ $(\mu, \lambda)$-points were chosen over the triangle $0 \leq \mu \leq \lambda \leq 1$, and assigned an improper prior proportional to $1/\lambda$. The correct prior guess $f_o = \text{Normal}(0, 1)$ and a function $\alpha(\mu, \lambda)$ identically equal to constant values $A = 10$, $A = 50$ and $A = 100$ were used.

For the kernel density estimator, a normal kernel was chosen, since considerations about higher asymptotic efficiency of other kernels with compact support seemed irrelevant in this case, the sample size being relatively small (for density estimation standards) and the data being themselves normally generated.

Both $L_1$ and $L_2$ optimality criteria are used to evaluate the performance of a density estimate (Bayesian or kernel) $f^*$, namely the Mean Integrated Absolute Error

$$\text{MIAE}(f^*) = \text{E}(\int |f^*(z) - f(z)|dz) \tag{9}$$

and the Mean Integrated Squared Error

$$\text{MISE}(f^*) = \text{E}(\int (f^*(z) - f(z))^2 dz). \tag{10}$$

To obtain an approximation to the optimal bandwidths, we computed Monte Carlo values of MIAE and MISE over those 1000 samples for different values of the bandwith $h$. We then selected $h_1$, minimizing MIAE, and $h_2$, minimizing MISE. The integrals in formulae (9) and (10) were approximated as sums over 81 $z$-points, equally spaced between -4 and 4. For $n = 100$, the MIAE minimizer turned out to be $h_1 = .426$, up to the third decimal point. The MISE minimizer was instead $h_2 = .447$, to be compared to the asymptotic approximation $.422 = 1.06 * (100)^{-.2}$ appearing, for example, in Silverman ([12], page 45). Notice that $h_1 < h_2$, that is, minimizing $L_1$ distance requires less smoothing than minimizing $L_2$ distance, but the two values are not so different. This was already noticed by Hall and Wand ([5], page 73).

Table 1 shows the Monte Carlo averages, the fifth, the fiftieth and the ninetieth percentiles of the sampling distributions of the three different $\tilde{f}_{100}(z)$ and of two kernel estimates - corresponding to bandwidths $h_1$ and $h_2$ - for few selected values of $z$. Monte Carlo approximations to MIAE and MISE of the estimators are shown in the last two columns. As it can be seen from a comparison with the true density at the top of the table, the long run averages of the Bayes estimates are substantially closer to the true values than the kernel estimates, but their sampling variabilities are much larger, accounting for much larger MISE and MIAE if $A = 10$ or $A = 50$. For $A = 100$, the MIAE of the Bayes estimate finally becomes smaller than the MIAE of the kernel estimate. Notice once again that we are using the correct prior guess, so that as $A \to \infty$ MIAE and MISE tend to 0. *Table1 here*

The weight $A = 100$, for which the MIAE of the Bayes estimator is smaller that the MIAE of the kernel estimator, may, at first, seem quite large. However, it should be noticed that the hyperparameter $A$ in a Dirichlet Density Process has not quite the same interpretation as the prior mass in a straight Dirichlet process prior. The reason is that, for too small an $A$, too much weight is given to the "maximum likelihood" estimator of the density, which is, basically, an unreasonable combination of spikes at the observations. Keeping the weight of the prior guess, $A$, the same order of magnitude as the sample

size prevents, in this example, unwanted spiky looking Bayesian estimates for most of the samples [2]. Nevertheless for some samples, the magnitude of $\lambda$, the hyperparameter $A$ and the spacings interact in such a way that some undesirable spikeness still appears. Figure 1 contains the Bayes estimates of $f$ given the first and the second of the 1000 samples used in the simulations described above. The first is affected by the problem but the second is not, although the only difference between the two is the observed sample. *Figure1 her*

It is not clear how to avoid this phenomenon, since "adjusting" the prior based on the data to avoid spiky looking estimates - for example by starting the prior support of $\lambda$ at a value large enough - involves a circular reasoning contrary to the Bayesian dictate. Such a corrective action has not been applied sample by sample in the simulations above, giving some more explanation of the high MISE and MIAE of the Bayes estimates. Another possible solution would be to use a Bayes estimate based on a loss function that penalizes for roughness, instead of the square error loss implicit in the use of the posterior mean, although this would mean introducing a complicated structure in the problem, too far removed from the simpleminded Bayesian approximation ideas underlying the use of Dirichlet Density Processes.

The issue of how to avoid too rough Bayes estimates parallels the choice of bandwidth in kernel estimation but it is a little less crucial, since a posterior spread over a multitude of possible values of $\lambda$ is less committing than the choice of a single bandwidth.

# 6 Bayesian Monte Carlo simulations

The simple structure of Dirichlet Density Processes has the advantage that it is fairly easy to simulate from the posterior process. With repeated simulations from the posterior comes the possibility of constructing Monte Carlo posterior distributions and Monte

[2]From this point of view, the hyperparameter $A$ is more similar to the hyperparameters of a Polya tree prior, as in Lavine [9], where big values of the hyperparameters are used to for the purpose of density estimation.

Carlo Bayesian intervals for any functional $\phi(f)$ of the density $f$. The simple step character of the posterior realizations of $f$ makes the computation of $\phi(f)$ straightforward in many cases.

A method for simulating $100c\%$ Bayesian intervals for a functional $\phi(f)$ of interest from a posterior process, given one sample, goes through the following steps:

1. Generate $(\mu', \lambda')$ from $\pi^*(\mu, \lambda | x_1, \ldots, x_n)$.

2. Fix $N$ large enough that $n_k(\mu', \lambda') = 0$ and also $\alpha_k(\mu', \lambda')$ are numerically negligible (compared to $\alpha(\mu, \lambda) + n$) for all $|k| > N$.

3. Generate $g'_{-N}, \ldots, g'_{N-1}$ from a Dirichlet distribution with parameter

$$\alpha_k(\mu', \lambda') + n_k(\mu', \lambda'), k = -N, \ldots, N$$

4. Compute $\phi(f')$, where

$$f'(z) = \sum_{k=-N}^{N} \frac{g'_k}{\lambda'} \{\mu' + (k-1)\lambda' < z \leq \mu' + k\lambda'\} \tag{11}$$

for $\mu' \leq z \leq \mu' + N\lambda'$, 0 otherwise.

5. Repeat steps 1 through 4 a large number $R$ of times.

6. Report the $(1-c)/2$ and $(1+c)/2$ quantiles of the $\phi$ sample of size $R$ obtained.

For step 1, a discretization of the posterior as the one used in the previous section will do. For step 3, the usual construction of Dirichlet vectors, as independent gamma variates divided by their sum (see Wilks [15]), is computationally efficient.

Bayesian intervals have been simulated for two particularly interesting functionals, namely $f(0)$ and the mean $m = \int x f(x) dx$, given the second of the 1000 samples used in the previous Section. In this case, $R = 5000$ and the functionals are easily calculated, for a particular Monte Carlo realization $f'$ from the posterior, using the formulae

$$\begin{aligned} f'(0) &= g'_0/\lambda \\ m' &= \int x f'(x) dx = \mu' + \lambda' \sum k g'_k - \lambda'/2. \end{aligned}$$

The Monte Carlo experiment produced the following 90% Bayesian intervals:

$$\begin{aligned} 0.2774 \leq \quad & f(0) \quad \leq 0.4528 \\ -0.0841 \leq \quad & m \quad \leq 0.1559. \end{aligned}$$

*Figure2 her*

Notice that the construction of frequentist confidence intervals for densities in the presence of a relatively small sample size ($n = 100$) is quite problematic, whereas the Montecarlo Bayes procedure is straightforward.

A Montecarlo estimate of the whole posterior distribution of $f(0)$ and $m$ is provided by a histogram of the 5000 Montecarlo replicates and is shown in Figure 2.

**Acknowledgement:** The author thanks Professor Michael Woodroofe for his help and advice.

# References


[1 ] C. Antoniak (1974). Mixtures of Dirichlet processes with application to Bayesian nonparametric problems. *Ann.Statist.* **2** 1152-1174.

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

[3 ] T.S. Ferguson (1983). Bayesian density estimation by mixtures of normal distributions. In *Recent advances in Statistics* (H.M. Rizvi, J.S. Rustagi and D. Siegmund eds.) Academic Press, New York, 287-302.

[4 ] M. Gasparini (1992). Bayes Nonparametrics for biased sampling and density estimation. Unpublished Ph.d. Thesis, The University of Michigan, Ann Arbor.

[5 ] P. Hall and M. P. Wand (1988). Minimizing $L_1$ Distance in Nonparametric Density Estimation. *J. Multivariate Anal.* **26** 59-88.

[6 ] J. A. Hartigan (1994). Bayesian Histograms. Invited paper at the *Fifth International Meeting on Bayesian Statistics*, Alicante, Spain, 5-9 June, 1994.

[7 ] W. Härdle (1991). *Smoothing Techniques with Implementation in* S. Springer-Verlag.

[8 ] N.L. Hjort (1994). Bayesian approaches to non- and semiparametric density estimation. Invited paper at the *Fifth International Meeting on Bayesian Statistics*, Alicante, Spain, 5-9 June, 1994.

[9 ] M. Lavine (1992). Some aspects of Polya tree distributions for statistical modeling. *Ann. Statist.* **20** 1222-1235.

[10 ] T. Leonard (1973). A Bayesian method for histograms. *Biometrika* **60** 297-308.

[11 ] A. Y. Lo (1984). On a class of Bayesian nonparametric estimates: I.Density estimates. *Ann. Statist.* **12** 351-357.

[12 ] D. W. Scott (1985). Average shifted histograms: effective nonparametric density estimators in several dimensions. *Ann. Statist.* **13** 1024-1040.

[13 ] B. W. Silverman (1986). *Density Estimation for Statistics and Data Analysis.* Chapman and Hall.

[14 ] M. West, P. Müller and M. D. Escobar (1994). Hierarchical priors and mixture models, with application in regression and density estimation. In *Aspects of uncertainty: a tribute to D.V. Lindley*, Smith, A.F.M. and Freeman, P. eds. Wiley, New York.

[15 ] S. S. Wilks (1962). *Mathematical Statistics* Wiley, New York.

Captions for tables and figures.

- Table 1: Simulation results from 1000 pesudonormal samples of size 100
- Figure 1: $\tilde{f}_{100}(z)$ = posterior mean of $f$
- Figure 2: Montecarlo posterior distributions of $f(0)$ and $m$

BAYES AND KERNEL ESTIMATES OF A FEW POINTS OF A NORMAL DENSITY

| z | -3.0 | -2.0 | -1.0 | -0.1 | 0.0 | 0.1 | 1.0 | 2.0 | 3.0 | MISE (x1000) | MIAE (x1000) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| true f(z) | 0.0044 | 0.0540 | 0.2420 | 0.3970 | 0.3989 | 0.3970 | 0.2420 | 0.0540 | 0.0044 | | |
| Bayes estimates | | | | | | | | | | | |
| A=10 | | | | | | | | | | | |
| average | 0.0045 | 0.0571 | 0.2408 | 0.3889 | 0.3910 | 0.3898 | 0.2399 | 0.0588 | 0.0043 | 10.96 | 186.15 |
| 5th pctl | 0.0004 | 0.0141 | 0.1502 | 0.3056 | 0.3055 | 0.3104 | 0.1563 | 0.0166 | 0.0004 | | |
| 50th pctl | 0.0012 | 0.0544 | 0.2391 | 0.3897 | 0.3917 | 0.3892 | 0.2368 | 0.0563 | 0.0012 | | |
| 95th pctl | 0.0192 | 0.1059 | 0.3353 | 0.4788 | 0.4753 | 0.4746 | 0.3346 | 0.1115 | 0.0189 | | |
| A=50 | | | | | | | | | | | |
| average | 0.0048 | 0.0564 | 0.2421 | 0.3919 | 0.3931 | 0.3914 | 0.2405 | 0.0578 | 0.0048 | 7.82 | 152.91 |
| 5th pctl | 0.0016 | 0.0237 | 0.1680 | 0.3213 | 0.3177 | 0.3196 | 0.1697 | 0.0255 | 0.0016 | | |
| 50th pctl | 0.0025 | 0.0536 | 0.2403 | 0.3885 | 0.3907 | 0.3900 | 0.2386 | 0.0551 | 0.0024 | | |
| 95th pctl | 0.0154 | 0.0960 | 0.3203 | 0.4731 | 0.4752 | 0.4730 | 0.3208 | 0.0974 | 0.0154 | | |
| A=100 | | | | | | | | | | | |
| average | 0.0048 | 0.0563 | 0.2429 | 0.3921 | 0.3942 | 0.3914 | 0.2406 | 0.0571 | 0.0048 | 6.24 | 132.74 |
| 5th pctl | 0.0024 | 0.0306 | 0.1801 | 0.3250 | 0.3231 | 0.3244 | 0.1797 | 0.0314 | 0.0023 | | |
| 50th pctl | 0.0030 | 0.0531 | 0.2406 | 0.3881 | 0.3923 | 0.3886 | 0.2384 | 0.0547 | 0.0030 | | |
| 95th pctl | 0.0140 | 0.0918 | 0.3122 | 0.4670 | 0.4704 | 0.4644 | 0.3116 | 0.0915 | 0.0146 | | |
| Kernel estimates | | | | | | | | | | | |
| optimal L1 lambda = .426 | | | | | | | | | | | |
| average | 0.0082 | 0.0676 | 0.2406 | 0.3654 | 0.3671 | 0.3657 | 0.2402 | 0.0676 | 0.0080 | 5.532 | 138.74 |
| 5th pctl | 0.0005 | 0.0376 | 0.1889 | 0.3102 | 0.3109 | 0.3087 | 0.1886 | 0.0386 | 0.0007 | | |
| 50th pctl | 0.0070 | 0.0664 | 0.2416 | 0.3658 | 0.3687 | 0.3669 | 0.2405 | 0.0665 | 0.0068 | | |
| 95th pctl | 0.0207 | 0.0997 | 0.2915 | 0.4201 | 0.4203 | 0.4204 | 0.2961 | 0.1013 | 0.0197 | | |
| optimal L2 lambda = .447 | | | | | | | | | | | |
| average | 0.0086 | 0.0688 | 0.2403 | 0.3626 | 0.3643 | 0.3629 | 0.2399 | 0.0688 | 0.0085 | 5.508 | 139.17 |
| 5th pctl | 0.0007 | 0.0398 | 0.1905 | 0.3100 | 0.3106 | 0.3085 | 0.1910 | 0.0406 | 0.0009 | | |
| 50th pctl | 0.0075 | 0.0677 | 0.2414 | 0.3633 | 0.3655 | 0.3641 | 0.2402 | 0.0678 | 0.0074 | | |
| 95th pctl | 0.0211 | 0.1004 | 0.2894 | 0.4145 | 0.4150 | 0.4151 | 0.2928 | 0.1013 | 0.0198 | | |

Table 1: (author: Mauro Gasparini) Simulation results from 1000 pseudonormal samples of size 100

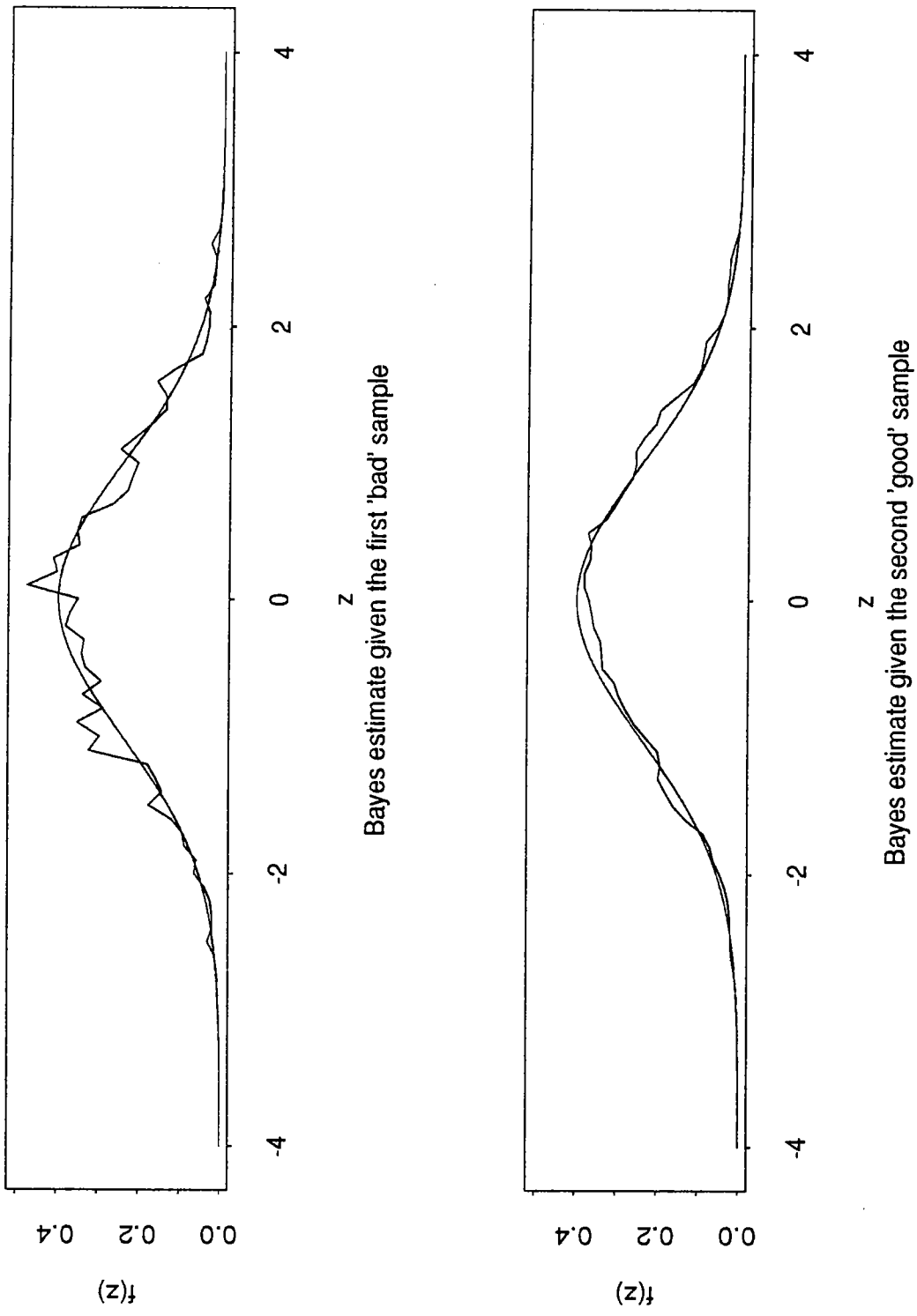


Figure 1: (author: Mauro Gasparini) $\tilde{f}_{100}(z)$ = posterior mean of $f$

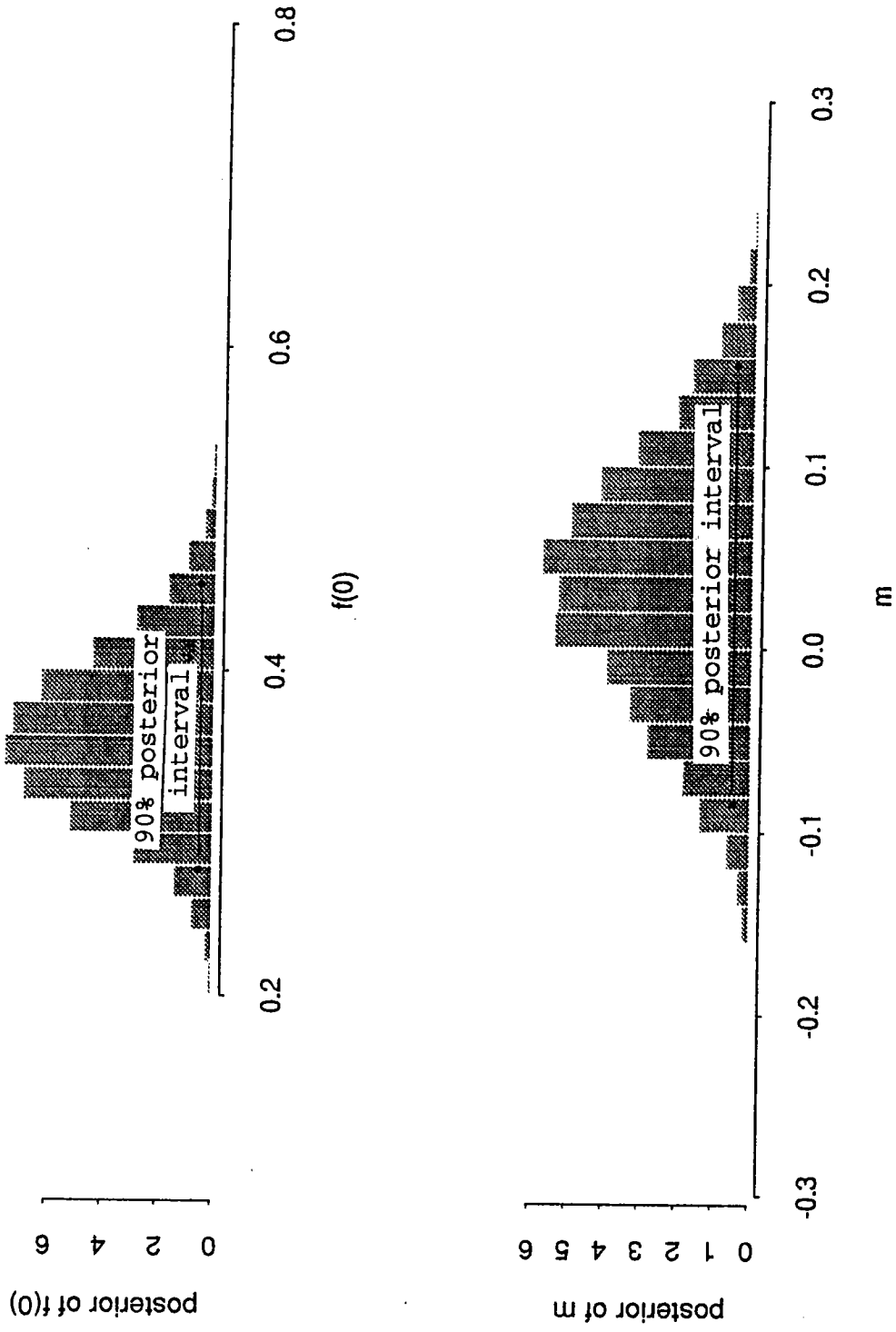


Figure 2: (author: Mauro Gasparini) Montecarlo posterior distributions of $f(0)$ and $m$