Sankhyā : The Indian Journal of Statistics
San Antonio Conference: Selected articles
2002, Volume 64, Series A, Pt.2, pp 306-322

ON A NONPARAMETRIC RECURSIVE ESTIMATOR OF THE MIXING DISTRIBUTION

By MICHAEL A. NEWTON
University of Wisconsin-Madison, USA

SUMMARY. Routinely in statistical applications hierarchical models arise in which unobserved random effects contribute to heterogeneity amongst sampling units. An easily computable, smooth nonparametric estimate of the underlying mixing distribution can be derived as an approximate nonparametric Bayes estimate under a Dirichlet process prior. I discuss the recursive estimation algorithm, its consistency properties, and its application in several examples, including its use as a model diagnostic in the analysis of DNA microarray gene expression data.

1. Nonparametric Mixture Models

A useful stochastic model to account for heterogeneity amongst experimental units is the nonparametric mixture model. From an unknown distribution F there arise independent and identically distributed but unobserved draws $\phi_1, \phi_2, \ldots, \phi_n$ in a set Φ. Given these ϕ_i's, conditionally independent random variables $x_1, x_2, \ldots, x_n$ are observed, where the law of x_i given ϕ_i is described by a known sampling density $p(x|\phi)$ with respect to some dominating measure μ on the sample space. This framework covers many examples, and I consider two for illustration:

1. $\phi_i \in \Phi = [0, 1]$ and x_i is binomially distributed with success probability ϕ_i and sample size m. I study an application of this model to a well-studied experiment on thumbtack tossing in which $m = 9$, $n = 320$ and x_i counts the number of times the ith tack lands with its point facing up (Beckett and Diaconis, 1994).

Paper received October 2000; revised November 2001.
AMS (2000) subject classification. Primary 65C60; secondary 62G07.
Keywords and phrases. Random effects distribution, hierarchical modeling, stochastic approximation algorithm, Dirichlet process.

2. $\phi_i \in \Phi = (-\infty, \infty)$, and $x_i = (u_i, v_i)$ is such that the components u_i and v_i are conditionally identically distributed and independent from a Gamma distribution with a known shape a and scale $\exp(\phi_i)$. This hierarchical model arises in a study of high-throughput gene expression measurements (Newton, *et al.* 2001). The data (u_i, v_i) measure the expression of gene i from cells under two different conditions. The numerical example has $n = 13,027$ which is the number of genes on a particular oligonucleotide microarray for the mouse genome.

The general problem is to estimate the mixing distribution F from data $x_1, \ldots, x_n$.

Professor C.R. Rao himself has considered finite mixture models to enable statistical classification (Rao, 1948). I consider here the nonparametric case which dates to Robbins (1950). Lindsay's (1995) monograph provides a clear discussion of advances on this nonparametric problem, many of which have centered on the nonparametric maximum likelihood estimator (NPMLE): i.e. the distribution that maximizes the loglikelihood

$$l(F) = \sum_{i=1}^{n} \log \left\{ \int p(x_i|\phi)\, dF(\phi) \right\}. \tag{1}$$

To obtain the NPMLE is to maximize a concave function over a convex domain; various algorithms are known for evaluating the solution, and fundamental theoretical results provide some characterization of this object (e.g., the NPMLE is discrete with support size at most the number of distinct values in the sample $x_1, \ldots, x_n$.).

Nonparametric Bayesian estimation of F has been considered since at least the work of Antoniak (1974), but the evident computational challenges have been met only recently with the advent of Markov chain Monte Carlo and related solutions (e.g., Liu, 1996; Dey, Mueller, and Sinha, 1999). In the present paper I consider the approximate Bayes estimator proposed recently in Newton and Zhang (1999) which is evaluated using a computationally efficient recursive algorithm. I derive some first order asymptotic theory for the estimator, and evaluate it in the two examples outlined above. I also show that if the recursion is applied to a large with-replacement Monte Carlo sample from the observed data, then it provides a smooth approximation to the NPMLE.

2. **Predictive Recursion**

I suppose that the unknown mixing distribution F has a density function $f(\phi)$ with respect to some dominating measure $\tilde{\mu}$ for arguments $\phi \in \Phi$. The recursive estimation algorithm can be defined in some generality, though theoretical results are restricted at present to finite Φ and counting measure $\tilde{\mu}$. Estimation starts with a prior guess – a density $f_0(\phi)$ – and a user-supplied weight sequence $w_1, w_2, \ldots, w_n$ for $w_i \in [0, 1]$. These weights are invoked purely for computational reasons; they have no particular interpretation in terms of the model. A sequence of density estimates is formed from the simple recursion

$$f_i(\phi) = (1 - w_i) f_{i-1}(\phi) + w_i \frac{p(x_i|\phi) f_{i-1}(\phi)}{c(x_i, f_{i-1})} \tag{2}$$

where $c(x, f) = \int p(x|\phi) f(\phi)\, d\tilde{\mu}(\phi)$. That is, the updated estimate f_i is a mixture of the current estimate f_{i-1} and *new information* contained in datum x_i in the form of a posterior density of ϕ_i using f_{i-1} as the prior. Three points are noteworthy.

1. *Exactness for* $n = 1$: A Bayesian who summarizes prior uncertainty about $F = \int f$ in a Dirichlet process centered at $F_0 = \int f_0$ and with mass parameter $(1 - w_1)/w_1$ obtains (2) exactly as his/her estimate of f on the basis of a single observation x_1. More specifically, the Bayes estimate of $F = \int f$ under squared-error loss is $\int f_1$. Taken further, a Bayesian who approximates uncertainty in F after $x_1, \ldots, x_{i-1}$ with a single Dirichlet process centered at $F_{i-1} = \int f_{i-1}$ is obliged to use (2) to update his/her opinion. These claims follow from the well-known Polya sequence characterization of the Dirichlet process, and they were developed in some detail in Newton, Quintana, and Zhang (1998).

2. *Order Dependence*: Unless there is no information loss in producing x_i from ϕ_i, the recursive estimate f_n obtained through (2) depends on the order by which the observations $x_1, \ldots, x_n$ are processed, and thus it is not the Bayes estimate. That the observations are exchangeable but the estimator is not symmetric is a deficiency, and several solutions are possible, as I describe in the numerical examples (§4).

3. *Evaluation*: In spite of its status as an approximation, the recursion (2) is very simple to implement numerically. In the one and two-dimensional examples studied to date, I keep track of f_i on a grid or

lattice of ϕ values, and update on that finite set. Some sort of numerical integration is required to evaluate the normalizing constant $c(x_i, f_{i-1})$ at each step.

3. First Order Sampling Theory

Using some Markov chain theory I establish the existence of a limiting distribution for the recursion (2) in the special case that Φ is finite. Again, the set up has independent and identically distributed observations $x_1, x_2, \ldots$ from a mixture model with unknown mixing distribution $f(\phi)$ on Φ and known sampling component $p(x|\phi)$. Starting with a prior guess $f_0(\phi)$ on Φ and a weight sequence $w_1, w_2, \ldots$ in $[0, 1]$, form the recursion

$$f_n(\phi) = (1 - w_n) f_{n-1}(\phi) + w_n \frac{p(x_n|\phi) f_{n-1}(\phi)}{c(x_n, f_{n-1})}.$$

LEMMA 1 *If $\sum_n w_n$ diverges, then surely there exists a probability vector $f_\infty = \{f_\infty(\phi) : \phi \in \Phi\}$ such that $f_n \to f_\infty$ as $n \to \infty$.*

To know the specific nature of this limiting distribution f_∞ requires further investigation, but it is interesting that a limit exists for every sequence of data, rather than for almost every sequence, as one might expect. The method of proof is similar to one used in Newton and Zhang (1999) for a related model.

PROOF OF LEMMA 1: Consider any single realization of data $x_1, x_2, \ldots$. I proceed by constructing an artificial Markov chain $z_0, z_1, z_2, \ldots$ in Φ where $z_0 \sim f_0$ and for $n \geq 1$,

$$z_n = \begin{cases} z_{n-1} & \text{with probability } w_n \\ y_n & \text{with probability } 1 - w_n \end{cases}$$

where $y_n \in \Phi$ are independent (given the x sequence) and

$$\Pr(y_n = \phi) = \frac{f_{n-1}(\phi) p(x_n|\phi)}{c(x_n, f_{n-1})}.$$

One may readily compute the matrix P_n holding transition probabilities from z_{n-1} to z_n. Noting that the off-diagonal entries of P_n are constant within columns, I calculate the δ-coefficient

$$\begin{aligned} \delta(P_n) &= \frac{1}{2} \sup_{i,j \in \Phi} \sum_{k \in \Phi} |P_n(i,k) - P_n(j,k)| \\ &= 1 - w_n. \end{aligned}$$

The δ-coefficient characterizes properties of the chain and I use it extensively. See Isaacson and Madsen (1976), Chapter V.

Convergence of the artificial Markov chain to a stationary distribution may be assessed by considering the $m-$step transition matrix $P_n P_{n+1} \ldots P_{n+m-1}$ for which the δ-coefficient $d_{m,n}$ is necessarily bounded by the product of coefficients from the individual steps:

$$d_{m,n} \leq \prod_{i=0}^{m-1} \delta(P_{n+i}) = \prod_{i=0}^{m-1} (1 - w_{n+i}).$$

Simply,

$$\log(d_{m,n}) \leq \sum_{i=0}^{m-1} \log(1 - w_{n+i}) \leq - \sum_{i=0}^{m-1} w_{n+i}.$$

The assumption taken that $\sum_n w_n$ diverges implies that for any n, $d_{m,n} \to 0$ as $m \to \infty$. In Markov chain terminology, the chain z_n is thus weakly ergodic and finiteness of Φ further implies that it is strongly ergodic. But strong ergodicity is equivalent to the existence of a stationary distribution, f_∞ on Φ to which the $n-$step marginal distributions of the chain converge. A calculation confirms that the $n-$step marginal of z_n is precisely f_n, thus motivating the particular construction and completing the proof. □

The limiting distribution f_∞ has yet to be characterized, and may depend on several factors. For example, if one takes $w_n = 1$ for all n, then f_1 is the parametric posterior distribution of a single ϕ given prior f_0, and likewise f_n is the parametric posterior for a single common ϕ given prior f_0. One expects this posterior distribution to converge to a point mass measure as $n \to \infty$. However, the data-generating mixture distribution f may support multiple values ϕ, so f_∞ would be off the mark.

The sequence $\{f_n\}$ depends in subtle ways on the initial guess f_0. For example if $f_0(\phi) = 0$ for some ϕ then $f_n(\phi) = 0$ for all n.

To see the form of f_n more clearly, note that by repeated substitution into the basic recursion (2) one obtains

$$f_n(\phi) = v_{n,0} f_0(\phi) + \sum_{i=1}^{n} v_{n,i} \frac{f_{i-1}(\phi) p(x_i|\phi)}{c(x_i, f_{i-1})} \tag{3}$$

where $\{v_{n,i}\}$ are other weights formed from the primary weights $\{w_n\}$. Specifically,

$$v_{n,i} = \begin{cases} \prod_{j=1}^{n}(1 - w_j) & \text{if } i = 0 \\ w_i \prod_{j=i+1}^{n}(1 - w_j) & \text{if } 0 < i < n \\ w_n & \text{if } i = n. \end{cases} \tag{4}$$

These non-negative weights satisfy $\sum_{i=0}^{n} v_{n,i} = 1$.

One gets further insight by rewriting (3) in terms of the limiting distribution f_∞:

$$f_n(\phi) = \sum_{i=1}^{n} v_{n,i} \frac{f_\infty(\phi) p(x_i|\phi)}{c(x_i, f_\infty)} + E_n \tag{5}$$

where the error term E_n is

$$E_n = v_{n,0} f_0(\phi) + \sum_{i=1}^{n} v_{n,i} d_i \left(\frac{a_i}{b_i} - \frac{a}{\tilde{b}_i} \right)$$

and $a_i = f_{i-1}(\phi)$, $a = f_\infty(\phi)$, $b_i = c(f_{i-1}, x_i)$, $\tilde{b}_i = c(f_\infty, x_i)$, and $d_i = p(x_i|\phi)$. For clarity I restrict the representation (5) to parameter values ϕ at which $f_\infty(\phi) > 0$. (If you like, one can define $E_n = f_n(\phi)$ in the case $f_\infty(\phi) = 0$.) The following describes sufficient conditions for the error term to vanish.

LEMMA 2 *If, in addition to the assumptions of Lemma 1,*

$$0 < \underline{p} := \inf_{x,\phi} p(x|\phi) \leq \sup_{x,\phi} p(x|\phi) := \overline{p} < \infty \tag{6}$$

and $\max_{0 \leq i \leq n} v_{n,i} \to 0$ *as* $n \to \infty$*, then, surely,* $E_n \to 0$ *as* $n \to \infty$*.*

This result seems to severely restrict the choice of sampling model, but practically it is not a restriction because already the parameter space Φ is finite and so the boundaries are readily avoided.

PROOF OF LEMMA 2: Fix a realization of data $x_1, x_2, \ldots$ and thus fix a sequence $\{f_n\}$ converging to a limiting distribution f_∞ by Lemma 1. Ignore the initial term $v_{n,0} f_0(\phi)$ since it vanishes. Rewrite E_n so that the *ith* term involves the common denominator $b_i \tilde{b}_i$, and then invoke the triangle inequality:

$$\begin{aligned} |E_n| &= \left| \sum_{i=1}^{n} v_{n,i} \frac{d_i (a_i \tilde{b}_i - a_i b_i - a b_i + a b_i)}{b_i \tilde{b}_i} \right| \\ &\leq \sum_{i=1}^{n} v_{n,i} \frac{a_i d_i |b_i - \tilde{b}_i|}{b_i \tilde{b}_i} + \sum_{i=1}^{n} v_{n,i} \frac{d_i |a - a_i|}{\tilde{b}_i} \;=\; E_{n,1} + E_{n,2}. \end{aligned}$$

Let us consider first $E_{n,2}$. By the regularity condition on the sampling model $p(x|\phi)$,

$$E_{n,2} \leq \left(\overline{p}/\underline{p}\right) \sum_{i=1}^{n} v_{n,i}|a - a_i|.$$

Convergence of $E_{n,2}$ to 0 follows because it is a weighted Cesàro average, and $|a - a_n| \to 0$. More specifically, I split up the sum into a tail where $|a - a_i|$ is small and an initial part in which $\max_{0\leq i\leq n} v_{n,i} \to 0$ forces terms to be small for large n. Convergence of $E_{n,1}$ to 0 follows similarly. Again by model assumptions

$$E_{n,1} \leq \left(\overline{p}/\underline{p}^2\right) \sum_{i=1}^{n} v_{n,i}|\tilde{b}_i - b_i|.$$

Further,

$$|\tilde{b}_i - b_i| = \left|\sum_{\phi\in\Phi} p(x_i|\phi)[f_\infty(\phi) - f_{i-1}(\phi)]\right| \leq \overline{p} \sum_{\phi\in\Phi} |f_\infty(\phi) - f_{i-1}(\phi)| := K_i$$

where K_n is converging to 0 by convergence of f_n, and $E_{n,1}$ converges by the same Cesàro argument as above. □

In Lemma 2, conditions are placed on the weights $v_{n,i}$ rather than directly on the user-supplied weights w_n which give rise to the $v_{n,i}$ through (4).

LEMMA 3 *With primary weights w_n of the form $w_n = c/n^\alpha$ where $0 < \alpha < 1$ and $\alpha \leq c < 1$, one has $\max_{0\leq i\leq n} v_{n,i} = w_n$ and hence this maximum converges to 0. If $\alpha = 1$, the maximum also converges to zero, but in this case $\max_{0\leq i\leq n} v_{n,i} = v_{n,0} = \prod_i (1 - w_i)$.*

PROOF OF LEMMA 3: Use concavity of the function $h(u) = u^\alpha$ to bound the ratio $v_{n,i+1}/v_{n,i}$. □

I have established a representation (5) for the recursive approximation f_n as an empirical weighted average of posterior-like terms

$$f_\infty(\phi)p(x_i|\phi)/c(x_i, f_\infty)$$

and an error term E_n that vanishes under some conditions. The next step would seem to be to invoke a law of large numbers to show that the average

converges to its expectation under sampling of the x_i's. There is some difficulty with this, however, because I have not shown that the limiting vector $f_\infty(\phi)$ is non-random. If it is, then there is no trouble in getting almost sure convergence. But the subtle dependence created by the recursion seems to invalidate the use of a zero-one law, for example, to conclude that f_∞ is non-random. A proof of non-randomness of f_∞ has eluded me so far, but there are reasons to expect this will hold. Numerical work, for example, indicates that for large n the estimator f_n is insensitive to the particular $\{x_i\}$ realization. Another argument would be to look at the difference between the recursion f_n from initial distribution f_0 and a parallel recursion g_n using the same data and model but starting instead at a different distribution g_0. I hope to find conditions which ensure that $|f_n - g_n|$ is decreasing, and hence that the limit f_∞ is independent of any initial segment of the data. These arguments remain to be worked out.

Assuming that f_∞ is almost surely constant, the law of large numbers applied to (5) will imply that at ϕ such that $f_\infty(\phi) > 0$,

$$f_\infty(\phi) = \int \frac{f_\infty(\phi) p(x|\phi)}{c(x, f_\infty)} \, c(x, f) \, d\mu(x) \tag{7}$$

where recall that f is the true mixing distribution governing the samples. In other words, the recursive estimator f_n converges to a solution of an *asymptotic self consistency equation*. If the model happens to ensure a unique solution to (7), then certainly the true distribution f is this solution, and thus f_n is consistent.

The user-supplied weight sequence w_n affects the estimator f_n, and, though minimal conditions are established in Lemmas 1-3, I cannot say what weight sequence is optimal in terms of efficiency of estimation or rate of convergence. In particular examples I can measure the approximation error numerically and I find that weights on the order $n^{-1/2}$ work quite well. Weights which drop faster seem to produce overly smooth final estimates, and weights which drop much more slowly yield more variable solutions. In so far as the recursion is a type of stochastic approximation scheme there is some guidance from existing theory, but a fuller theoretical connection remains to be established. For some related algorithms the weight sequence must drop faster than $1/\log(n)$, for example (Kushner and Yin, 1997, page 110). Also, the potential to attain an optimal convergence rate by averaging iterates is intriguing and deserves further attention (Polyak and Juditsky, 1992), but I do not consider these issues here.

4. Numerical Examples

4.1. A Binomial Mixture. Experimental data from Beckett and Diaconis (1994) on thumbtack tossing helps to illustrate the methodology. There are $n = 320$ experimental units. Each unit corresponds to a thumbtack which is tossed $m = 9$ times to produce data x_i, the number of times out of m that tack i lands with its point facing up. The present framework assumes that the tack-specific success probability ϕ_i arises from an unknown population f on the unit interval. The summary frequencies $n(x)$ in this data set are:

x	0	1	2	3	4	5	6	7	8	9
$n(x)$	0	3	13	18	48	47	67	54	51	19

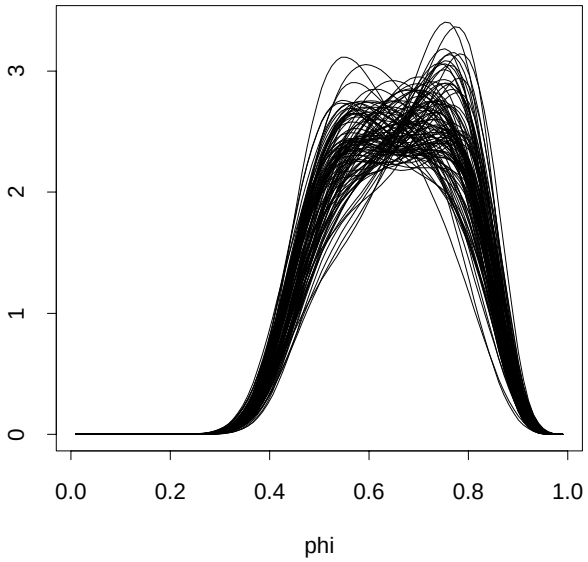

Figure 1. Recursive estimates of the mixing distribution, thumbtack example: Each curve is the resulting f_n after processing the thumbtack data in a random order. Shown are results from 100 different orders.

Fig. 1 shows some numerical results from running the recursive algorithm in this binomial mixture example. Calculations are done on a grid of 100 ϕ values in the unit interval. The initial guess f_0 is taken to be the uniform distribution. The weight sequence $\{w_n\}$ has $w_n = [(4/3)(1/3 + n)]^{-1/2}$, but others gave similar results. Calculations were repeated for 100 random orderings of the $n = 320$ data points, and the resulting curves f_n are plotted in Fig. 1. There is a high amount of variation, and thus a high degree of dependence on processing order, in this case. However, all estimates place

the bulk of their mass in the same place. Note also that two modes are apparent. One way to cope with the variation is simply to average the density estimates pointwise over the orderings. The result, not shown, has two primary modes, and is similar to the best available approximations to the Bayes estimate shown in Liu (1996). I have noticed that here and in other examples there is a tendency of the recursive estimate to be somewhat smoother than the actual Bayes estimate (Newton, Quintana, and Zhang, 1998).

In §4.3 I consider a different way to handle variation over the processing order and I observe that for large n the order dependence vanishes.

4.2. DNA Microarrays The nonparametric mixture methodology is used in the present example as a diagnostic procedure to check the adequacy of a certain parametric model. Newton *et al.* (2001) presents a parametric mixture model for high-throughput gene expression measurements obtained from DNA microarrays. Several microarray technologies are in wide use; basically, all these devices enable the simultaneous measurement of the abundance of molecular transcripts copied from each of n genes in the cells under study. The ith spot on a microarray yields measurements interpreted as estimates of the expression of the ith gene in certain cells. Our numerical example is from an experiment using Affymetrix oligonucleotide arrays performed by Sam Nadler and colleagues at the University of Wisconsin on the comparison of gene expression between two types of mouse cells. There are $n = 13,027$ genes and the quantitative AVE-DIFF scores from two arrays are shown in Fig. 2. These are part of a larger study (Nadler *et al.*, 2000). They are presented here to demonstrate that the proposed recursive algorithm can provide a model diagnostic when using parametric mixture models. (I note as an aside that extensive current research is underway to develop statistical methods for microarray data. See, for example, `http://www.stat.Berkeley.EDU/users/terry/zarray/Html/`.)

The parametric models described in Newton *et al.* (2001) are used to compare gene expression between two conditions. Fig. 2 is a scatterplot comparing expression data u_i under one condition to expression data v_i under a second condition, plotted on a logarithmic scale. Of interest is to know which genes i are significantly differentially expressed between the two conditions, or, more simply, which points are unusually far from the diagonal line $u = v$. Contour lines in Fig. 2 are based on a fitted parametric model and are used to infer significantly differentially expressed genes. They are obtained as follows. Under a null hypothesis H_0, the gene i measurements $x_i = (u_i, v_i)$ arise from a common Gamma distribution with shape parameter a and scale

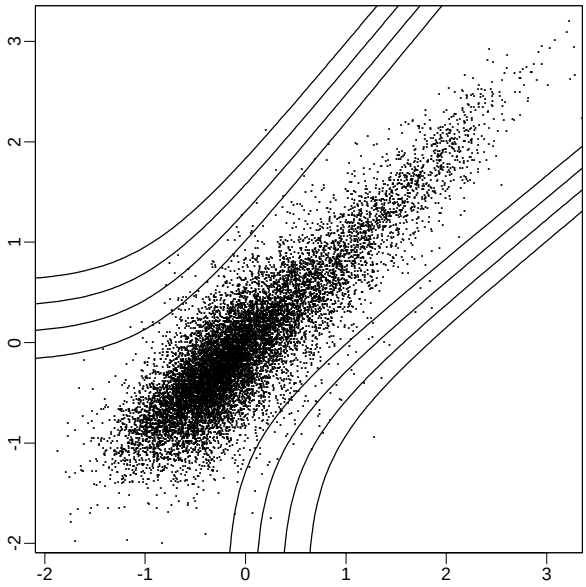

Figure 2. DNA Microarray, Parametric Odds: Scatterplot shows $n = 13,027$ pairs $x_i = (u_i, v_i)$ of gene expression in two mouse tissues. The scale is logarithmic, base 10. Contours are at odds of $1:1$, $10:1$, $100:1$, and $1000:1$ favouring H_A over H_0 as you move from the diagonal line $u = v$. An estimated fraction $p = 0.0325$ of the genes satisfy H_A.

parameter $\theta_i = \exp(\phi_i)$. On this hypothesis, there is no real differential expression and differences between u_i and v_i are merely measurement noise. This Gamma distribution has density:

$$p_0(x|\theta) = p(u|\theta)p(v|\theta) \propto (uv)^{a-1} \exp\{-\theta(u+v)\}. \tag{8}$$

Thus, for example, both u_i and v_i have expected value a/θ_i. Under an alternative hypothesis H_A, there is real differential expression between the two conditions, and this is modeled by allowing u_i and v_i to have distinct sampling distributions so that jointly:

$$p_A(x|\theta, \theta^*) = p(u|\theta)p(v|\theta^*) \propto (uv)^{a-1} \exp\{-\theta u - \theta^* v\}.$$

Interestingly, whether H_0 is true or H_A is true is a question relative only to spot i, so the scientific question of interest amounts to $n = 13,027$ different hypothesis tests. To handle this problem, a hierarchical model is formed in which the scale parameters θ_i (and θ_i^* under H_A) themselves arise from some mixing distribution F. The parametric Gamma model is used for F in the calculations of Newton *et al.* (2001). By integrating the latent scale

parameters and estimating other fixed parameters, one obtains marginal densities for x both under H_0 and under H_A. The contour lines in Fig. 2 indicate level sets of the posterior odds of H_A to H_0:

$$\text{odds}(u, v) = \left(\frac{p}{1-p}\right)\left(\frac{p_A(u,v)}{p_0(u,v)}\right)$$

and p is the estimated proportion of genes for which H_A is true. The specific forms of $p_A(u, v)$ and $p_0(u, v)$ are compound Gamma densities. (See Newton *et al.* (2001) for more details.) The odds of real change (i.e. H_A) increase as you move from the diagonal. An important feature of the resulting calculation is that genes for which the overall expression is low are treated more conservatively than highly expressed genes.

There is great interest to know if the conclusions drawn by the above hierarchical model are sensitive to the parametric form of the mixing distribution F. To assess this I used the recursive algorithm (2) to estimate F nonparametrically. Since most of the genes probably satisfy the null hypothesis, and because it is simpler to cast the mixture problem in this case, I ran the recursion through the n observations $x_i = (u_i, v_i)$ using the sampling model $p_0(x|\theta)$ in (8). I used the parametric maximum likelihood estimated shape parameter $a = 4.04$ so that the sampling-model densities are considered fixed and known. Furthermore, I found that numerical stability was obtained by running the recursion in the parameterization $\phi = \log(\theta)$. For the initial guess $f_0(\phi)$ I used the (log-transformed) Gamma mixing distribution which had been estimated by maximum likelihood in the parametric model. This had shape parameter 0.78 and scale parameter 0.10. The recursive computations were done on a grid of 350 ϕ values spanning the central 0.998 mass of the parametric mixing distribution, an interval from about -6 to +4. The recursion was run out five times using independent random orderings of the n genes and using a weight sequence proportional to the inverse square root. I observed very little variation across orderings.

Panel A in Fig. 3 compares the average of these five recursive estimates (wiggly solid curve f_n) to the estimated parametric mixing distribution (dotted line). The central tendency and spread of the nonparametric estimate matches the parametric estimate very closely, but the nonparametric estimate has much more local structure. Panels B through D in Fig. 3 characterize the nonparametric fit in terms of the implied mixed distributions. Panel C, for example, shows contours of the inferred joint density

$$p_0(u, v) = \int f_n(\phi) p_0(u, v|\phi)\, d\tilde{\mu}(\phi)$$

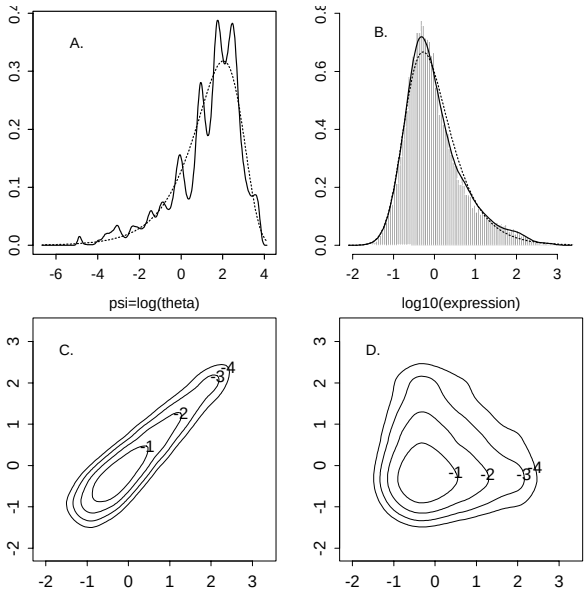

Figure 3. DNA Microarray, Recursive Estimation: (A) parametric and nonparametric estimation of the mixing distribution, (B) marginal distribution of both u_i and v_i, (C) joint distribution of $x_i = (u_i, v_i)$ under H_0, (D) joint distribution of x_i under H_A.

where $p_0(u, v|\phi)$ has the Gamma form in (8) and again $\phi = \log(\theta)$. This joint density has more fine structure than the compound Gamma distribution implied by the parametric model. Panel D shows the same integral, but against $p_A(u, v|\phi, \phi^*)$. It may be helpful to observe that genes satisfying H_0 should present data $x_i = (u_i, v_i)$ according to the density in Panel C and those satisfying H_A should present data according to that in Panel D. Panel B compares the marginal, one-dimensional fit of the parametric (dashed) and nonparametric (solid) model to a histogram of the observed measurements. Clearly the nonparametric fit respects the empirical marginal better than the parametric fit.

Finally, Fig. 4 shows the inference (odds function) that would occur if I use the nonparametric mixing distribution instead of the parametric mixing distribution. Up to the ratio $p/(1-p)$, these contours are simply at level sets of the ratio of Panel D to Panel C from Fig. 3. The main message is that conclusions about which genes are significantly differentially expressed are not affected greatly by the use of the parametric model instead of the more flexible nonparametric model in this case (compare Fig. 4 to Fig. 2). This provides some support, therefore, to the use of the simpler parametric form.

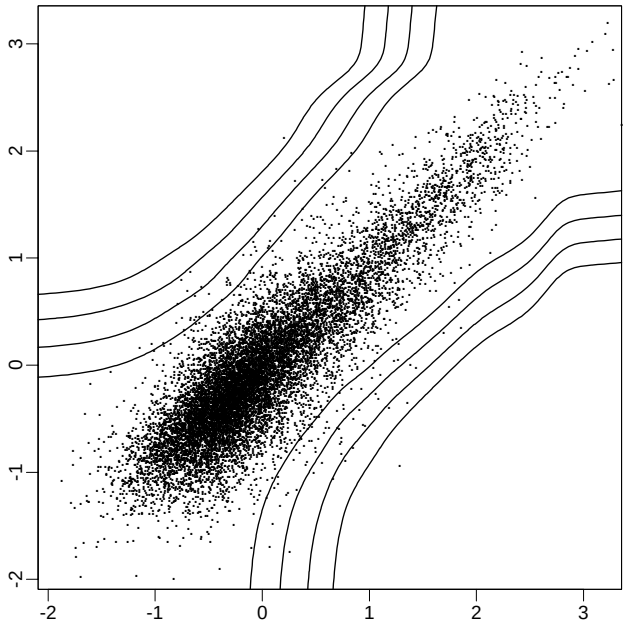

Figure 4. DNA Microarray, Nonparametric Odds: Same as Fig. 2 but using nonparametric estimate of mixing distribution.

Note that a problem of this magnitude ($n = 13,027$) is a numerical challenge to any nonparametric mixing procedure and the recursive algorithm produces an estimate with very little trouble.

Calculations for this example could be extended in various ways. I have assumed that the shape parameter of the Gamma observation component is known as estimated from the fully parametric model. Typical experiments involve some replication and in that case there will be enough information in the data to simultaneously identify this shape parameter and the nonparametrically determined mixing distribution. I have also fixed the mixing proportion p at its value estimated from the fully parametric model, and I have run the recursion on the null hypothesis. To generalize this requires some care because on the alternative hypothesis one needs to specify a two-dimensional mixing distribution f. Essentially the true mixing distribution f is itself a finite mixture of components from each of the hypotheses. The recursion may be applied in this larger context and can lead to simultaneous estimates of the mixing distribution under each hypothesis and the mixing proportion p. I am testing the methodology and hope to report on it soon.

4.3. *NPMLEs and the Binomial Mixture Revisited.* That the recursive

estimator f_n depends on the order in which observations x_i are processed is an unappealing feature. In theory and practice this dependence subsides for large n, but something must be done about it for any fixed n. In the examples presented so far I averaged the estimate obtained from several random orders. Numerically this is very efficient, but here and in other examples I note some oversmoothing relative to the more elusive Bayes estimate. (Of course the recursion is orders of magnitude simpler to compute than the Bayes estimate.)

An alternative procedure was suggested in Newton and Zhang (1999). The idea is to treat the observed sample $\{x_i\}$ itself as a population, just as one does in bootstrap sampling. One runs the recursion from an initial distribution, as before, but it is applied to an arbitrarily large sample, say of size $N >> n$, from the empirical population. In other words, one takes a large sample of size N with replacement from the observed sample of size n and processes that. As far as I can tell, the resulting estimator is independent of the particular post-sample realization. Fig. 5 summarizes such a calculation for the thumbtack example. I took $N = 1000n$ using inverse square root weights. The final estimate is in Panel A. Panel B compares the empirical distribution of x values (solid) to the fitted marginal distribution from the recursive estimate (dashed), and we see that the procedure provides a very good model fit.

It is interesting to reflect on the calculations from §3 as they relate to the Monte Carlo procedure summarized in Fig. 5. I showed that as the sample size increases the recursive estimate should converge to a solution of the asymptotic self consistency equations (7). But because the population under study is empirically derived from the observed data, these equations are nothing more than the sample self consistency equations known well in mixture modeling. Canceling $f_\infty(\phi)$ from both sides and rearranging things slightly, the equations become

$$0 = \frac{1}{n} \sum_{x=0}^{m} \left(\frac{p(x|\phi)}{c(x, f_\infty)} - 1 \right) n(x) \tag{9}$$

where $n = 320$ and $n(x)$ is the empirical frequency of x in the observed sample. It is known (e.g. Lindsay, 1995) that these equations characterize the nonparametric maximum likelihood estimator. In fact the right hand side of (9) is the directional derivative of the nonparametric loglikelihood (1). Panel C in Fig. 5 evaluates the right hand side of (9) for the estimator f_N in Panel A. One sees that this Monte Carlo sample of size $N = 1000n$ produces a final estimate that solves equations (9) with a reasonably high degree of

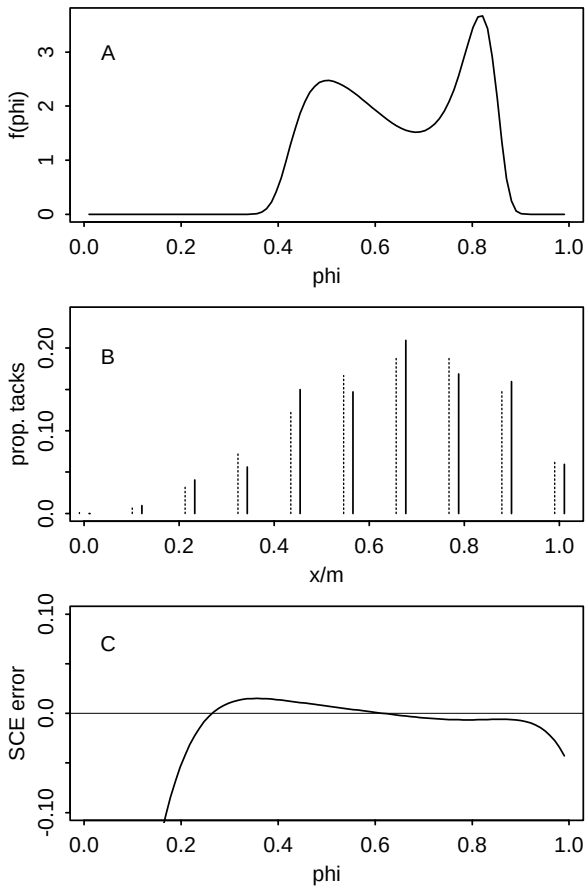

Figure 5. Binomial Mixture Revisited: (A) f_N by Monte Carlo sampling, (B) sample relative frequencies (solid) and marginal model fit (dashed), (C) discrepancy in self consistency equations.

precision. In other words, this procedure provides a numerically efficient and structurally simple way to approximate the NPMLE.

Acknowledgements. In earlier work Yunlei Zhang conceived the Markov chain convergence argument which is critical in §3. Brian Yandell, Sam Nadler, and Alan Attie gave me access to the DNA microarray data. Also Peter Hoff, Christina Kendziorski, Tom Kurtz, Fernando Quintana, and two anonymous referees have given me helpful comments on this work, which has been supported in part by grant R01 64364 from the National Cancer Institute. S-language code to implement the recursive algorithm is available at my web site `http://www.stat.wisc.edu/~newton/`.

References

Antoniak, C.E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Ann. Statist.*, **2**, 1152-1174.

Beckett, L. and Diaconis, P. (1994). Spectral analysis for discrete longitudinal data. *Adv. Math.*, **103**, 107-128.

Dey, D., Muller, P. and Sinha, D. (1999). *Practical Nonparametric and Semiparametric Bayesian Statistics.* Springer, New York.

Isaacson, D.L. and Madsen, R.W. (1976). *Markov Chains: Theory and Applications.* Wiley, New York.

Kushner, H.J. and Yin, G.G. (1997). *Stochastic Approximation Algorithms and Applications.* Springer, New York.

Lindsay, B. (1995). *Mixture models: Theory, geometry and applications.* Institute of Mathematical Statistics, Hayward, CA.

Liu, J.S. (1996). Nonparametric hierarchical Bayes via sequential imputations, *Ann. Statist.*, **24**, 911-930.

Nadler, S.T., Stoehr, J.P., Schueler, K.M., Tanimoto, G., Yandell, B.S. and Attie, A.D. (2000). The expression of adipogenic genes is decreased in obesity and diabetes mellitus. *Proc. Natl. Acad. Sci.* USA **97**, 11371-11376.

Newton, M.A., Kendziorski, C.M., Richmond, C.S., Blattner, F.R. and Tsui, K.W. (2001). On differential variability of expression ratios: Improving statistical inference about gene expression changes from microarray data. *J. Computational Biol.*, **8**, 37-52.

Newton, M.A., Quintana, F.A. and Zhang, Y. (1998). Nonparametric Bayes methods using predictive updating. In *Practical Nonparametric and Semiparametric Bayesian Statistics*, Eds. D. Dey, P. Muller and D. Sinha, 45-61. Springer, New York.

Newton, M.A. and Zhang, Y. (1999). A recursive algorithm for nonparametric analysis with missing data. *Biometrika*, **86**, 15-26.

Polyak, B.T. and Juditsky, A.B. (1992). Acceleration of stochastic approximation by averaging. *SIAM J. Control Optim.*, **30**, 838-855.

Rao, C.R. (1948). The utilization of multiple measurements in problems of biological classification (with discussion). *J. Roy. Statist. Soc. Ser. B*, **10**, 159-193.

Robbins, H. (1950). A generalization of the method of maximum likelihood: Estimating a mixing distribution (abstract). *Ann. Math. Statist.*, **21**, 314-315.

Michael A. Newton

Department of Statistics

University of Wisconsin–Madison

1210 West Dayton St.

Madison, WI 53706-1685, USA

E-mail: newton@stat.wisc.edu