ON A NONPARAMETRIC RECURSIVE ESTIMATOR
OF THE MIXING DISTRIBUTION

by

Michael A. Newton
Department of Statistics and
Department of Biostatistics and Medical Informatics
University of Wisconsin-Madison
On a nonparametric recursive estimator of the mixing distribution

Michael A. Newton
Department of Statistics, University of Wisconsin–Madison,
1210 West Dayton St., Madison WI 53706-1685, U.S.A., and
Department of Biostatistics and Medical Informatics, University of Wisconsin–Madison,
600 Highland Ave., Madison WI 53792, U.S.A., e-mail: newton@stat.wisc.edu

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.

This manuscript contains work presented at the conference Statistics: Reflections on the past and visions for the future in honor of Professor C.R. Rao on the occasion of his 80th birthday, held at the University of Texas, San Antonio, March 2000. The manuscript was prepared in August 2000, and is simultaneously Technical Report #1021 of the Department of Statistics at the University of Wisconsin-Madison.
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 $\Phi$. Given these $\phi_i$'s, conditionally independent random variables $x_1, x_2, \ldots, x_n$ are observed, where the law of $x_i$ given $\phi_i$ is described by a known sampling density $p(x|\phi)$ with respect to some dominating measure $\mu$ on the sample space. This framework covers many examples, and we shall consider two for illustration:

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

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 $\alpha$ and scale $\exp(\phi_i)$. This hierarchical model arises in a study of high-throughput gene expression measurement (Newton, et al. 2000). The data $(u_i, v_i)$ measure the expression of gene $i$ from cells under two different conditions. The numerical example has $n = 13027$ 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). We 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

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

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). In the present paper we consider the approximate Bayes estimator proposed recently in Newton and Zhang (1999) which is evaluated using a computationally efficient
recursive algorithm. We derive some first order asymptotic theory for the estimator, and evaluate it in the two examples outlined above.

2 Predictive Recursion

We suppose that the unknown mixing distribution $F$ has a density function $f(\phi)$ with respect to some dominating measure $\mu$ for arguments $\phi \in \Phi$. The recursive estimation algorithm can be defined in some generality, though our theoretical results are restricted at present to finite $\Phi$ and counting measure $\mu$. 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 $w_i \in [0, 1]$. A sequence of density estimates is formed from the simple recursion

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

where $c(x, f) = \int p(x|\phi)f(\phi) d\mu(\phi)$. 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 $\phi_i$ 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 $F_0 = \int f_0$ obtains (2) exactly as his/her estimate of $f$ on the basis of a single observation $x_1$. 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. This claim follows from the Poly a sequence characterization of the Dirichlet process, as shown in Newton, Quintana, and Zhang (1998).

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

3. **Evaluation**: Numerically it is very simple to implement the recursion (2). In the one and two-dimensional examples studied to date, we keep track of $f_i$ on a grid or lattice of $\phi$ 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 we establish the existence of a limiting distribution for the recursion (2) in the special case that $\Phi$ 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 \(\Phi\) and known sampling component \(p(x|\phi)\). Starting with a prior guess \(f_0(\phi)\) on \(\Phi\) and a weight sequence \(w_1, w_2, \ldots\) in \([0, 1]\), we 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 \(f_\infty\) 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 different model.

**Proof of Lemma 1:** Consider any single realization of data \(x_1, x_2, \ldots\). We proceed by constructing an artificial Markov chain \(z_0, z_1, z_2, \ldots\) in \(\Phi\) 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, we can calculate the \(\delta\)-coefficient

\[
\delta(P_n) = \frac{1}{2} \sup_{i,j \in \Phi} \sum_{k \in \Phi} |P_n(i, k) - P_n(j, k)|
\]

\[
= 1 - w_n.
\]

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

Convergence of our artificial Markov chain to a stationary distribution may be assessed by considering the \(m\)-step transition matrix \(P_n P_{n+1} \cdots P_{n+m-1}\) for which the \(\delta\)-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 $\Phi$ further implies that it is strongly ergodic. But strong ergodicity is equivalent to the existence of a stationary distribution, $f_\infty$, on $\Phi$ 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 $f_\infty$ has yet to be characterized, and may depend on several factors. For example, if we take $w_n = 1$ for all $n$, then $f_1$ is the parametric posterior distribution of a single $\phi$ given prior $f_0$, and likewise $f_n$ is the parametric posterior for a single common $\phi$ 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 $\phi$, so $f_\infty$ 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 $\phi$ 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) we obtain

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

(3)

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

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

(4)

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

We get further insight by rewriting (3) in terms of the limiting distribution $f_\infty$:

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

(5)

where the error term $E_n$ is

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

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 we may restrict the representation (5) to parameter values $\phi$ at which $f_\infty(\phi) > 0$. (If you like, we 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, we have

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

(6)

and $\max_{0 \leq i \leq n} v_{n,i} \to 0$ as $n \to \infty$, then, surely, $E_n \to 0$ as $n \to \infty$. 5
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 \( f_\infty \) by Lemma 1. Ignore the initial term \( v_{n,0} f_0(\phi) \) since it vanishes. Rewrite \( E_n \) so that the \( i \)th term involves the common denominator \( b_i b_i \), and then invoke the triangle inequality:

\[
|E_n| = \left| \sum_{i=1}^{n} v_{n,i} d_i (a_i b_i - a_i b_i - a_i + a_i) \right| \\
\leq \sum_{i=1}^{n} v_{n,i} d_i |b_i - b_i| \frac{b_i b_i}{b_i b_i} + \sum_{i=1}^{n} v_{n,i} d_i |a_i| \frac{b_i b_i}{b_i}
\]

\[
= E_{n,1} + E_{n,2}.
\]

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

\[
E_{n,2} \leq (\bar{p}/p^2) \sum_{i=1}^{n} v_{n,i} |a_i - a_i|.
\]

Convergence of \( E_{n,2} \) to 0 follows because it is a weighted Cesàro average, and \( |a_i - a_i| \to 0 \). More specifically, we split up the sum into a tail where \( |a_i - a_i| \) 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

\[
E_{n,1} \leq (\bar{p}/p^2) \sum_{i=1}^{n} v_{n,i} |\bar{b}_i - b_i|.
\]

Further,

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

where \( K_i \) 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 \( \alpha \leq c < 1 \), we have \( \max_{0 \leq i \leq n} v_{n,i} = w_n \) 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 \( v_{n,i+1}/v_{n,i} \). #

We 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 we 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 $f_{\infty}$ is non-random. A proof of non-randomness of $f_{\infty}$ has eluded us so far, but there are reasons to expect this will hold. Our numerical work, for example, indicates that for large $n$ the estimator $f_n$ is insensitive to the particular $\{x_i\}$ realization. Another argument is 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 $g_0$. We hope to find conditions which ensure that $|f_n - g_n|$ is decreasing, and hence that the limit $f_{\infty}$ is independent of any initial segment of the data. These arguments remain to be worked out.

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

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

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 we would have consistency of $f_n$.

4 Numerical Examples

4.1 A Binomial Mixture

Experimental data from Beckett and Diaconis (1994) on thumbtack tossing helps to illustrate the methodology. We have $n = 320$ 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 $\phi_i$ arises from an unknown population $f$ on the unit interval. The summary frequencies $n(x)$ in this data set are:

<table>
<thead>
<tr>
<th>$x$</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n(x)$</td>
<td>0</td>
<td>3</td>
<td>13</td>
<td>18</td>
<td>48</td>
<td>47</td>
<td>67</td>
<td>54</td>
<td>51</td>
<td>19</td>
</tr>
</tbody>
</table>

Fig. 1 shows some numerical results from running the recursive algorithm in this binomial mixture example. Calculations are done on a grid of 100 $\phi$ 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. We see 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. We also note 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). We 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 we consider a different way to handle variation over the processing order and we 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. (2000) presents a parametric mixture model for high-throughput gene expression measurements obtained from DNA microarrays. The $i$th spot on a microarray yields measurements interpreted as estimates of the expression of the $i$th gene in certain cells. Our numerical example is from an experiment done by Sam Nadler at the University of Wisconsin on the expression of $n = 13027$ genes in mouse tissues as measured by oligonucleotide arrays. The parametric models described in Newton et al. (2000) are used to compare gene expression under 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 parametric model fit 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 $\alpha$ and scale parameter $\theta_i = \exp(\phi_i)$. If so, there is no real differential expression and differences between $u_i$ and $v_i$ are merely measurement noise. This Gamma distribution has density:

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

Thus, for example, both $u_i$ and $v_i$ have expected value $\alpha/\theta_i$. Under an alternative hypothesis $H_A$, there is a 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:

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

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 $n = 13027$ different hypothesis tests. To handle this problem, a hierarchical model is
formed in which the scale parameters \( \theta_i \) (and \( \theta^*_i \) 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., (2000). 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 \( H_0 \):

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

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. (2000) 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 we 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, we run the recursion through the \( n \) observations \( x_i = (u_i, v_i) \) using the sampling model \( p_0(x; \theta) \) in (8). We used the parametric maximum likelihood estimated scale parameter \( \alpha = 4.04 \) so that the sampling-model densities are considered fixed and known. Furthermore, we found that numerical stability was obtained by running the recursion in the parameterization \( \phi = \log(\theta) \). For the initial guess \( f_0(\phi) \) we 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 \( \phi \) 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 inverse square root. We 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

\[
p_0(u, v) = \int f_n(\phi)p_0(u, v|\phi) \, d\phi
\]

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) 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 we 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 is not affected greatly by the use of the parametric model instead of the more flexible nonparametric model (compare Fig. 4 to Fig. 2). This provides some support, therefore, to the use of the simpler parametric form.

We note that a problem of this magnitude \( (n = 13027) \) is a numerical challenge to any nonparametric mixing procedure and the recursive algorithm produces an estimate with very little trouble.

### 4.3 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 we see that the for large \( n \) this dependence subsides, but something must be done about it for any fixed \( n \). In the examples presented so far we averaged the estimate obtained from several random orders. Numerically this is very efficient, but here and in other examples we 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. We run the recursion from an initial distribution, as before, but it is applied to an arbitrarily large sample, say of size \( N \), from the empirical population. In other words, we take a large sample of size \( N \) with replacement from the observed sample of size \( n \) and we process that. As far as we can tell, the resulting estimator is independent of the particular post-sample realization. Fig. 5 summarizes such a calculation for the thumback example. We 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. We 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

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

(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. We see that this Monte Carlo sample of size \( N = 1000n \) produces a final estimate that solves equations (9) with a reasonably high degree of 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 and Sam Nadler gave me access to the DNA microarray data. Also Peter Hoff, Christina Kendziorski, Tom Kurtz, and Fernando Quintana have given me helpful comments on this work, which has been supported in part by grant R01 144-JL89 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


Figure 1: Recursive estimates of mixing distribution, thumbtack example: Each curve is the result $f_n$ after processing the thumbtack data in a random order. Shown are results from 100 different orders.
Figure 2: DNA Microarray, Parametric Odds: Scatterplot shows $n = 13027$ 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 $1000:1$ favoring $H_A$ over $H_0$ as you move from the diagonal line $u = v$. An estimated fraction $p = 0.0325$ of the genes satisfy $H_A$. 
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 $H_A$. 
Figure 4: DNA Microarray, Nonparametric Odds: Same as Fig. 2 but using nonparametric estimate of mixing distribution.
Figure 5: Binomial Mixture Revisited: (A) $f_X$ by Monte Carlo sampling, (B) sample relative frequencies (solid) and marginal model fit (dashed), (C) discrepancy in self consistency equations.