Statistics and Computing (1998) **8**, 177–188

Mixture structure analysis using the Akaike Information Criterion and the bootstrap

JEFFREY L. SOLKA[1]*, EDWARD J. WEGMAN[2], CAREY E. PRIEBE[3], WENDY L. POSTON[1] and GEORGE W. ROGERS[1]

[1] *Dahlgren Division of the Naval Surface Warfare Center, Systems Research and Technology Department, Advanced Computation Technology Group, Code B10, Dahlgren VA 22448-5100, USA*
[2] *Center for Computational Statistics, George Mason University, Fairfax, VA 22030-4444, USA*
[3] *Department of Mathematical Sciences, The Johns Hopkins University, Baltimore, MD 21218, USA*

Received September 1997 and accepted January 1998

Given i.i.d. observations $x_1, x_2, x_3, \ldots, x_n$ drawn from a mixture of normal terms, one is often interested in determining the number of terms in the mixture and their defining parameters. Although the problem of determining the number of terms is intractable under the most general assumptions, there is hope of elucidating the mixture structure given appropriate caveats on the underlying mixture. This paper examines a new approach to this problem based on the use of Akaike Information Criterion (AIC) based pruning of data driven mixture models which are obtained from resampled data sets. Results of the application of this procedure to artificially generated data sets and a real world data set are provided.

Keywords: AIC, bootstrap, cluster analysis, mixture models

1. Introduction

Given $X = \{\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n\}$ where each $\vec{x}_i$ is d-dimensional and i.i.d. according to an unknown density $f_0(\vec{x})$, one is often interested in estimating $f_0(\vec{x})$. This problem occurs in such areas as exploratory data analysis, classification and regression. There are a variety of approaches to the multivariate density estimation problem (Scott, 1992). An often used parametric approach is that of finite mixture models (Everitt and Hand, 1981) in combination with the expectation maximization (EM) method of Dempster *et al.* (1977). One difficulty with this tactic is that one needs some idea as to the appropriate number of terms in the mixture model as well as the approximate parameter values. Given this information, the EM algorithm is guaranteed to converge to at least a local maxima in the likelihood surface when properly constrained.

Some of the previous non-parametric approaches include histograms, frequency polygons, adaptive histograms, average shifted histograms, kernel estimators and k nearest neighbour estimators. The reader is referred to the recent work of Scott (1992) for a good discussion of these density estimation techniques. These approaches are beneficial in that they possess desirable asymptotic consistency properties, and robustness with regard to non-normality. They are at a disadvantage as compared to the mixture model approach when it is suspected that the unknown true density is a mixture of a number of normal terms and one would like to estimate the posteriori probability of underlying term membership for an unlabelled observation.

This type of problem exists in the areas of medical diagnosis and image processing. In medical diagnosis the term membership may play an important role in identification of the underlying mechanism of disease or the identification of appropriate tissue type (Carmen and Merickel, 1990). In the general problem of image analysis the term membership may pertain to region type.

* To whom correspondence should be addressed.
E-mail: jsolka@nswc.navy.mil

A recently developed density estimation technique that circumvents some of the problems of the above techniques is the adaptive mixtures density estimator (AMDE) of Priebe and Marchette (1993) and Priebe (1994). This procedure is a blend of the finite mixtures and kernel estimator approaches. It is essentially a mixtures approach that allows for the creation of new terms in a data-driven manner. We have successfully applied this technique, in combination with fractal-based features, to the detection of man-made objects in land images (Solka *et al.*, 1992) and aerial (Priebe *et al.*, 1993) images, the general problem of texture classification (Solka *et al.*, 1993) and the measurement of breast parenchymal tissue density (Priebe *et al.*, 1994). The adaptive mixtures estimator is asymptotically consistent like the kernel estimator, but it has the added benefit of creating additional terms at a rate which is considerably less than the $O(n)$ creation rate associated with the kernel estimator.

One drawback to the adaptive mixtures estimator is that while there is asymptotic L_1 convergence for the procedure, this convergence is achieved through the creation of an asymptotically infinite number of terms. Thus the procedure will result in an overly complex model given enough data. Maximum likelihood estimation solutions derived from a reordered starting point yields robustness as a density, but with potentially different underlying term structure. Investigation of reordering properties is an interesting but involved subject which will be the focus of future analyses. Preliminary investigations can be found in Solka (1995).

In this work we are interested in a model whose complexity more closely matches that of the unknown distribution. The approach is to use the adaptive mixtures procedure as a starting point to generate a mixture model with (potentially many) extra degrees of freedom or parameters and to prune this model to a much smaller mixture model. This pruning of terms which is based on the use of the Akaike Information Criterion (AIC) is performed to obtain a model that matches the underlying distribution not only in a functional sense but also with regard to model complexity. Subsequent sections will detail how well our pruning-based procedure meets these goals.

The issues addressed herein are analogous to some of the standard issues of cluster analysis. There have been many different approaches to the clustering problem. The reader is referred to the paper of Milligan and Copper (1985) for a comparison of the relative effectiveness of over 30 different clustering procedures. These include Bayesian approaches (Binder, 1978), information theoretic approaches (Wallace and Boulton, 1968) and mixture-based approaches (McLachlan, 1987). Given the close connection between our approach and the mixture-based approaches, we feel it prudent to steer the reader to several good sources on mixture analysis. There have been many works written which provide a good summary of the recent attempts in the literature to apply mixture models to cluster analysis. These include such standard texts as Everitt and Hand (1981), McLachlan and Basford (1988) and Titterington *et al.* (1985).

The AIC was originally developed as a tool to choose between two statistical estimators of differing complexities (Akaike, 1974). The AIC is written as a function of the likelihood $\mathscr{L}$ and number of free parameters $\mathscr{M}$ in a model as follows

$$\mathrm{AIC}(\hat{f}) = -2\ln(\mathscr{L}) + 2\mathscr{M}.$$

In Akaike's original paper the AIC was applied to time series analysis, but subsequent work has applied the technique to ISODATA-based (Carman and Merickel, 1990) and general clustering (Bozdogan and Sclove, 1984), and finite mixture analysis (Liang *et al.*, 1992).

We have developed a new approach to finite mixture determination which employs AIC-based pruning of AMDE estimates. This approach differs from the work of Liang in two ways. Liang chooses to make an initial guess as to the appropriate complexity of the data's finite mixture model and then adjusts the number of terms in the model up and down by adding and removing a term until no further improvement is possible. Our approach begins with an overdetermined data-driven model that is produced by the AMDE procedure and then uses AIC in combination with the expectation maximization technique to prune superfluous terms from the model. So whereas Liang's approach adds and subtracts terms to the model our approach just removes terms. The second difference is that where Liang's approach is to produce a single best solution to the finite mixtures question, our approach produces a distribution of model complexities from which estimates of the appropriate model complexity can be made.

Section 2 develops the methodology for term pruning in the case of finite mixture models obtained from finite sample application of the adaptive mixture procedure. Section 3 presents results indicating that we can improve upon an overdetermined mixture model and in some cases determine the true model complexity. Section 4 concludes with a discussion of the relevance of these results.

2. Approach

Our approach combines elements of non-parametric density estimation, parametric density estimation, and information-based pruning. The non-parametric AMDE is used as the starting point of our procedure. We begin our discussions with an overview of AMDE.

2.1. *Adaptive mixture density estimation*

Given an unknown distribution $f_0(\vec{x})$ we seek to model the distribution using $\hat{f}(\vec{x})$ defined by

$$\hat{f}(\vec{x};\hat{\Psi}) = \sum_{i=1}^{g} \hat{\pi}_i K(\vec{x};\hat{\theta}_i), \tag{1}$$

where K is some fixed density parameterized by $\hat{\theta}_i$, and $\hat{\Psi} = (\hat{\pi}_1, \hat{\theta}_1, \hat{\pi}_2, \hat{\theta}_2, \ldots, \hat{\pi}_g, \hat{\theta}_g)$ and g is the number of terms in the mixture. The $\hat{\pi}_i$ are referred to as the mixing proportions. (We can assume for much of what follows that K is taken to be the normal distribution, in which case $\hat{\theta}_i$ becomes $\{\hat{\mu}_i, \hat{\Sigma}_i\}$.) In the simplest case the mixture is assumed to have a single term and the parameters that need to be estimated are the mean and covariance of the distribution.

The basic stochastic approach to parameter estimation is to update the estimate $\hat{\Psi}$ of the true parameters Ψ_0 recursively based on the latest estimate $\hat{\Psi}_n$ and the newest data point $\vec{x}_{n+1}$. That is,

$$\hat{\Psi}_{n+1} = \hat{\Psi}_n + \Phi_n(\vec{x}_{n+1};\hat{\Psi}_n) \tag{2}$$

for some update function Φ_n. The specific form of the update equation that we use is the one suggested by Titterington (1984). If we let $I_c(\Psi)$ be the complete Fisher information matrix, then the version of the recursive update formula we will use is

$$\hat{\Psi}_{n+1} = \hat{\Psi}_n + \left(nI_c(\hat{\Psi}_n)\right)^{-1}\left(\frac{\partial}{\partial\hat{\Psi}}\right)\log\left(\hat{f}(\vec{x}_{n+1};\hat{\Psi}_n)\right) \tag{3}$$

where the derivatives represent the vector of partial derivatives with respect to the terms of $\hat{\Psi}$.

In the case of mixtures of multivariate normals, we may write the recursive update equations as

$$\hat{\tau}_{n+1}^{(i)} = \frac{\pi_n^{(i)}\hat{f}^{(i)}(\vec{x}_{n+1};\hat{\theta}_n)}{\sum_{t=1}^{g}\pi_n^{(t)}\hat{f}^{(t)}(\vec{x}_{n+1};\hat{\theta}_n)} \tag{4}$$

$$\hat{\pi}_{n+1}^{(i)} = \hat{\pi}_n^{(i)} + \frac{1}{n}\left(\hat{\tau}_{n+1}^{(i)} - \hat{\pi}_n^{(i)}\right) \tag{5}$$

$$\hat{\mu}_{n+1}^{(i)} = \hat{\mu}_n^{(i)} + \frac{\hat{\tau}_{n+1}^{(i)}}{n\hat{\pi}_n^{(i)}}\left(x_{n+1} - \hat{\mu}_n^{(i)}\right), \tag{6}$$

and

$$\hat{\Sigma}_{n+1}^{(i)} = \hat{\Sigma}_n^{(i)} + \frac{\hat{\tau}_{n+1}^{(i)}}{n\hat{\pi}_n^{(i)}}\left[\left(\vec{x}_{n+1} - \hat{\mu}_n^{(i)}\right)\left(\vec{x}_{n+1} - \hat{\mu}_n^{(i)}\right)^T - \hat{\Sigma}_n^{(i)}\right]. \tag{7}$$

This is where $\hat{\tau}_{n+1}^{(i)}$ is the estimated posteriori probability of $\vec{x}_{n+1}$ belonging to the ith term of the mixture, $\hat{\pi}_{n+1}^{(i)}$ is the estimated mixing coefficient, $\hat{\mu}_{n+1}^{(i)}$ is the d-dimensional estimated mean, $\hat{\Sigma}_{n+1}^{(i)}$ is the $d \times d$ estimated covariance matrix of the ith term, and $\hat{f}^{(t)}(x)$ is the functional form of the tth term. We will use a superscripted (i) to denote the ith term in the case of the recursive update equations and a subscripted i in the case of the iterative equations to follow.

The adaptive mixtures density estimation (AMDE) stochastic approximation approach is to update $\hat{\Psi}$ recursively, the estimate of the true parameters Ψ_0, while at the same time providing the capability to expand the extent of the parameter space $\hat{\Psi}$ if dictated by the underlying complexity of the data. We note that in the AMDE case our parameter space $\hat{\Psi}$ is given by $\hat{\Psi} = (\hat{\pi}_1, \hat{\theta}_1, \hat{\pi}_2, \hat{\theta}_2, \ldots)$. The procedure

$$\hat{\Psi}_{n+1} = \hat{\Psi}_n + A \cdot U_n(\vec{x}_{n+1};\hat{\Psi}_n) + B \cdot C_n(\vec{x}_{t+1};\hat{\Psi}_t, t), \tag{8}$$

is used to update the density recursively where $A = [1 - P_n(\vec{x}_{n+1};\hat{\Psi}_n)]$ and $B = P_n(\vec{x}_{n+1};\vec{\Psi}_n)$. P_n represents a possibly stochastic create decision and takes on values 0 or 1. U_n updates the current parameters using Equations 4–7 while C_n adds a new term to the model. As is implicit in the equation, the decision to add a new term is a function of the current data point, our current estimation of the parameters and time. The time dependence is important in those cases for which we wish to anneal the probability of creation as a function of training time. The models produced by the AMDE procedure are good functional estimates, but are typically overdetermined with regard to the number of terms.

The exact nature of the creation process is as follows. The Mahalanobis distance from the new observation x_t to each of the terms in the model is computed using

$$MHD(i) = \left(x_t - \hat{\mu}^{(i)}\right)^T \hat{\Sigma}^{-1(i)}\left(x_t - \hat{\mu}^{(i)}\right).$$

If $MHD(i) > \tau_c$ (a 'create' threshold) for every term then a new term is created at $\vec{\mu}^{(\text{new})} = x_t$, with a covariance given by $\Sigma^{(\text{new})} = \mathscr{I}(\Sigma^{(i)})$ and a mixing coefficient of $\hat{\pi}^{(\text{new})} = 1/n$ assuming x_t is the nth data point. $\mathscr{I}(.)$ is a weighted average based on posterior probability. We also note that the mixing coefficients of the remaining terms are all rescaled by $(n-1)/n$

2.2. *Approaches to AIC-based pruning of AMDE-generated mixture models*

Previous work in the literature has examined the application of the AIC to the determination of the number of terms in a finite mixture (Liang *et al.*, 1992). The AIC/n estimates -2 times the expected value of the log likelihood of the estimated model (Akaike, 1972)

$$\frac{\text{AIC}}{n} = -2E\left[\int f_0 \log\hat{f}\right]. \tag{9}$$

AIC is defined in terms of likelihood, $\mathscr{L}$, and the number of free parameters in the model, $\mathscr{M}$, as

$$\text{AIC}(\hat{f}) = -2\ln(\mathscr{L}) + 2\mathscr{M} = -2\ln[\hat{f}(\vec{x})] + 2\mathscr{M}. \tag{10}$$

One uses the AIC to choose between models of differing complexities by selecting the model with the minimum

AIC. This choice is equivalent to maximizing the mean likelihood of the model.

Using the idea as a starting point we have developed a procedure that uses a single or set of adaptive mixture density estimates and produces a pruned model with a lower complexity. This procedure uses AIC to evaluate the appropriateness of lower complexity models that have been subjected to the iterative EM method. In the iterative EM method the update equation takes the form

$$\hat{\Psi}_{n+1} = \hat{\Psi}_n + \Phi(\vec{X}; \hat{\Psi}_n), \tag{11}$$

where Φ is the update function and $\vec{X}$ is the set of observations. In the case of mixtures of multivariate normals we may write the iterative update equations as

$$\hat{\tau}_{ij} = \frac{\hat{\pi}_i \hat{f}_i(\vec{x}_j; \hat{\theta})}{\sum_{t=1}^{g} \hat{\pi}_t \hat{f}_t(\vec{x}_j; \hat{\theta})}, \tag{12}$$

$$\hat{\pi}_i = \sum_{j=1}^{n} \frac{\hat{\tau}_{ij}}{n}, \tag{13}$$

$$\hat{\mu}_i = \frac{\sum_{j=1}^{n} \hat{\tau}_{ij} x_j}{n\hat{\pi}_i}, \tag{14}$$

and

$$\hat{\Sigma}_i = \sum_{j=1}^{n} \frac{\hat{\tau}_{ij}(x_j - \hat{\mu}_i)(x_j - \hat{\mu}_i)^T}{n\hat{\pi}_i}. \tag{15}$$

This is where $\hat{\tau}_{ij}$ is the estimated posterior probability that x_j belongs to term i, $\hat{\pi}_i$ is the estimated mixing coefficient, $\hat{\mu}_i$ is the d-dimensional estimated mean vector, and $\hat{\Sigma}_i$ is the $d \times d$ estimated covariance matrix for the ith term. So the EM algorithm is essentially a two step process. In the first step, the expectation step, the posterior probabilities of membership for each point and term are estimated using Equation 12. In the second step, the maximization step, the maximum likelihood estimates of the mixing parameters are computed based on this posterior estimate. These are computed in Equations 13 through 15.

The steps in our pruning procedure are as follows:

Step 1. Obtain $\hat{f}_g$ an initial adaptive mixtures approximation to f_0 containing g terms.
Step 2. Compute the AIC of each of the $g - 1$ term models after application of the EM method of Equations 12–15 to each of the models.
Step 3. If AIC $(\hat{f}_{g-1})$ < AIC $(\hat{f}_g)$ for one of the $g - 1$ term models then the pruning process is repeated using this model.
Step 4. Repeat this process of pruning and expectation maximization until no further improvement is possible.

It is important to point out that at each pruning step the remaining terms $\hat{\pi}_i$ are updated based on their Mahalanobis distance to the pruned term prior to updating with the EM method. In addition we note that at Step 3 of the procedure we make use of the model with the minimum AIC.

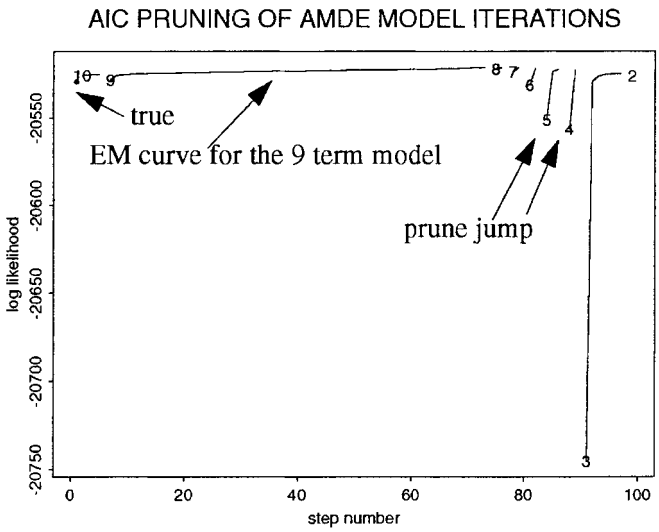

Fig. 1. *Pruning curves for the reduction of a 10 term model to a 2 term model*

Figure 1 illustrates the pruning process. The log likelihood for the true model, the original ten-term model, and the pruned and subsequently expectation maximized models are plotted. The first plotted point on the graph is plotted using a '*' and it represents the log likelihood of the underlying distribution. Then at the start of each likelihood maximization curve the number of terms is plotted. Each of these curves is separated from the next by a gap of one unit along the x-axis. The x-axis really serves as a convenient means to keep track of the number of iterations of the EM algorithm for each level of model complexity. In this case the process was able to reduce a ten-term model of the mixture $0.5N(-2, 1) + 0.5N(2, 1)$ to the appropriate two-term model. This case will be discussed in Section 3.

3. Results

In this section we present results which detail the performance of our density estimation procedure on a group of test cases. The first set of test cases consists of a suite of test densities chosen to illustrate the performance of the procedure on a variety of density types. In the second part of the results section we present simulation results that detail the performance of the estimator on a two-term mixture model as we vary the number of observations and the separation of the terms.

First we will detail results obtained from testing the pruning procedure on data sets drawn from two different bimodal two-term distributions, one four-mode four-term distribution, a standard unimodal normal distribution and the Buffalo snowfall data (Parzen, 1979) (see Fig. 2). In

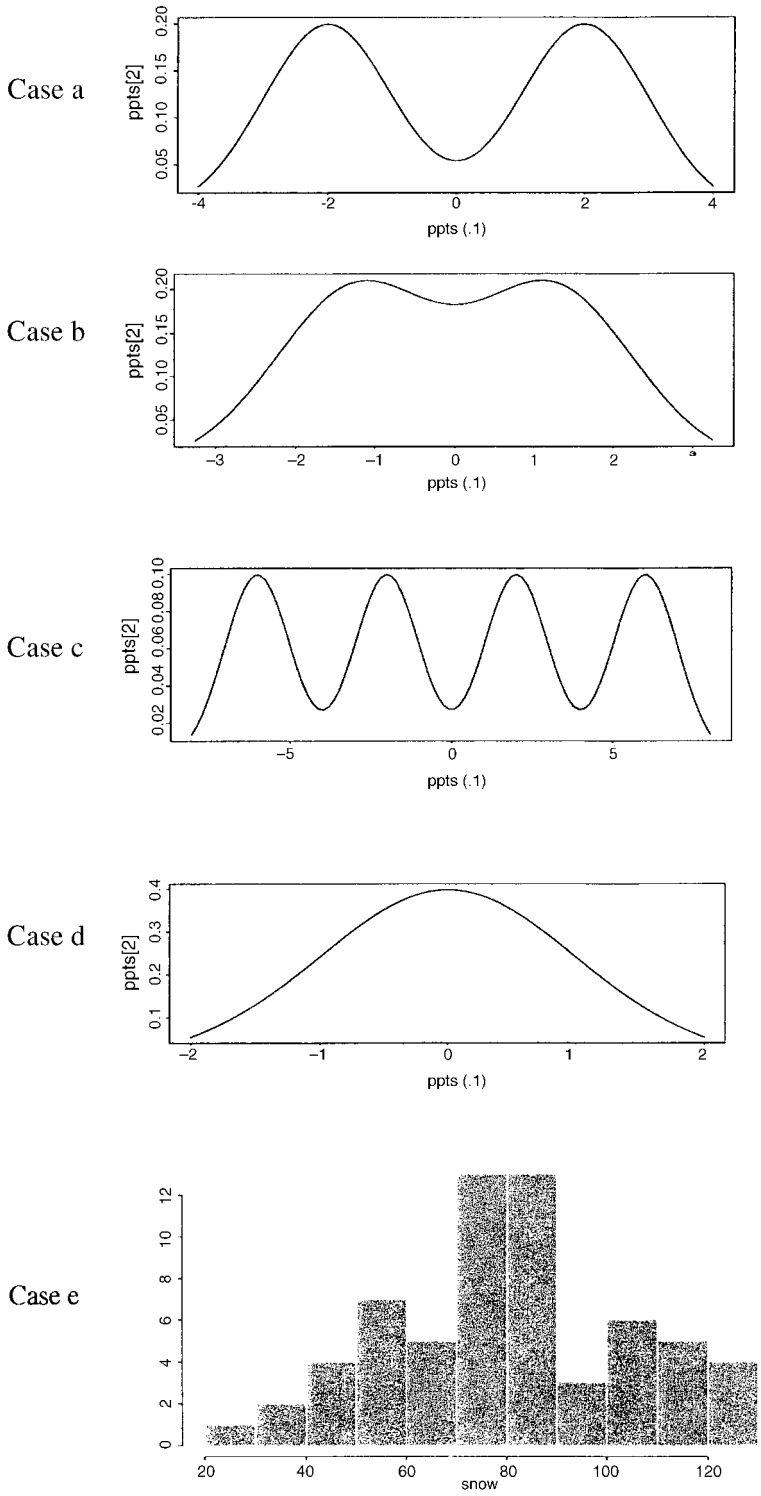

Fig. 2. *Test cases: (a) $\alpha(x) = 0.5 * N(-2, 1) + 0.5 * N(2, 1)$; (b) $\alpha(x) = 0.5 * N(-1.25, 1) + 0.5 * N(1.25, 1)$; (c) $\alpha(x) = 0.25 * N(-6, 1) + 0.25 * N(-2, 1) + 0.25 * N(2, 1) + 0.25 * N(6, 1)$; (d) $\alpha(x) = N(0, 1)$; (e) the Buffalo Snowfall data*

each simulated data case, 10 000 points were drawn from each distribution. The snowfall data consisted of 63 points.

One hundred bootstrap resamples were extracted from each of the data sets using their empirical distributions (Efron and Tibshirani, 1993). One hundred was chosen as a compromise between the suggested value of 25 for the estimate of standard error and that of 200 for the estimate of significance (Efron and Tibshirani, 1993). These resamples

are used in a way that is slightly different from the standard procedure. In standard bootstrapping one uses the resamples to estimate the standard error of a statistic whose standard error is not available in closed form. Our goal in bootstrapping is the production of a distribution on the number of terms in the models after AIC-based pruning. This distribution can then be used to estimate the number of terms in the true model. Given that the standard AMDE is highly order-dependent, the use of this reordering is essential in that it gives our estimator a chance to solve for the appropriate model complexity.

We wish to standardize the number of terms the AMDE produces under reorderings. This number obviously changes slightly under reorderings. Ten was chosen to limit the initial complexity level. This level is significantly overdetermined for our test cases and yields sufficient complexity to represent a rich class of densities (Marron and Wand, 1992). Each of these models was then subjected to the AIC-based pruning process. This process provides a model complexity distribution based on the data set.

In Figs 3a and b we present adaptive mixtures solutions for two of the resamplings of the data set drawn from $f_0(x) = 0.5 * N(-2, 1) + 0.5 * N(2, 1)$. We have included dF space plots along with the standard functional representation of the distributions. dF space plots are an effective way to display the terms in a mixture. Each term, $\pi_i N(\mu_i, \sigma_i^2)$, is plotted as a circle whose radius is proportional to π_i and whose centre is given by (μ_i, σ_i^2). The reader

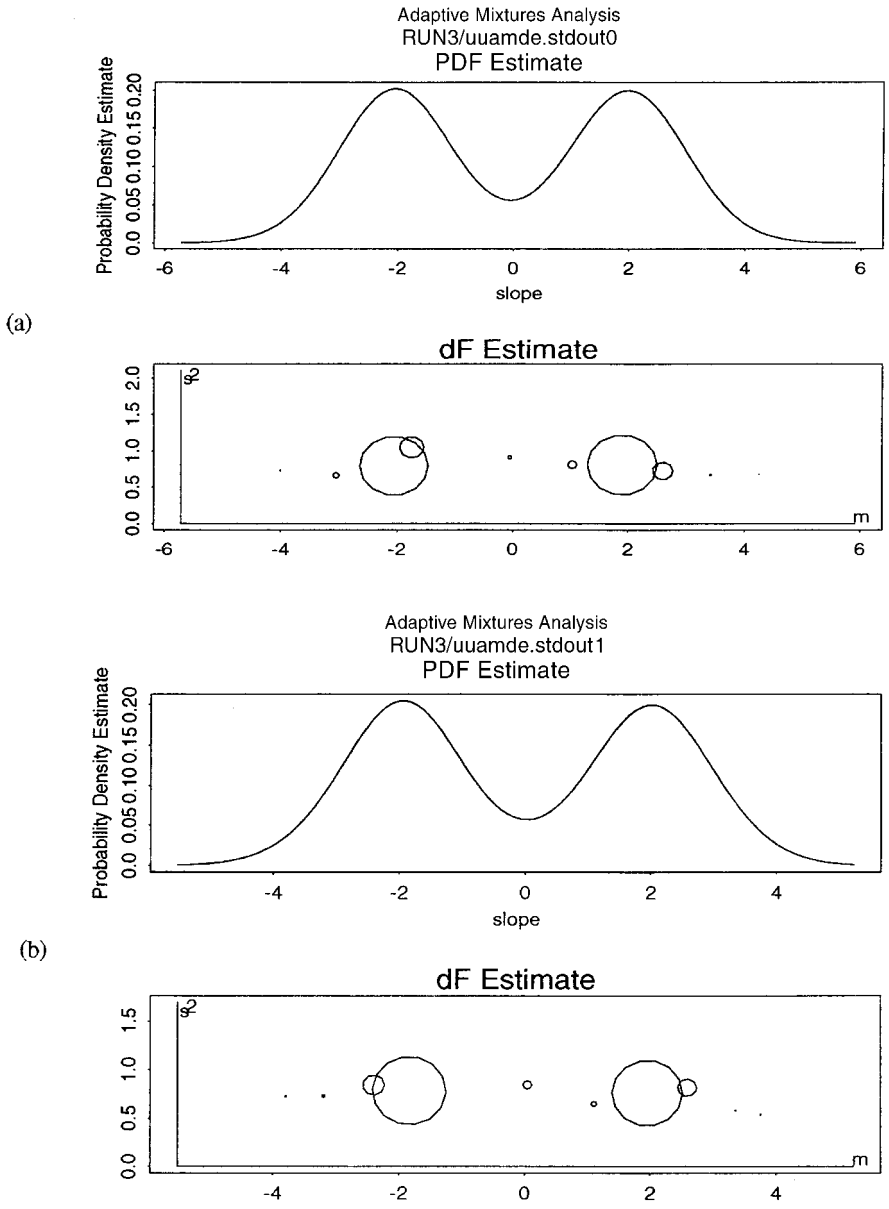

Fig. 3. *Adaptive mixtures estimates for two of the resamplings of the data set drawn from case a*

is referred to our recent paper on the visualization of mixture models for a more complete discussion of the issues associated with various methods of mixture visualization (Solka *et al.*, 1995). Where it is hard to discern the distributional structure from a standard function plot, it is quite easy in a dF space plot. We notice that the terms in each of the two solutions are markedly different. This phenomenon illustrates the multiplicity of mixtures models that lead to the same functional estimate. We also notice that there are more than the 'theoretical' number of terms needed. Each of the models is made up of ten terms. The occurrence of a matching number of terms in each model is the result of our initial constraint on the model complexity. It is important to note that although the terms are different in each solution, the location and number of modes are not, and that there are terms that are superfluous to the minimal representation of the distribution.

Table 1. *Number of terms for each case*

Case/No. of terms	1	2	3	4	5	6	7
a – Separated bimodal		**53**	16	22	5	4	
b – Bimodal		**19**	20	26	22	11	7
c – Quadmodal				**74**	13	12	1
d – Standard normal	**10**	7	20	23	24	15	1
e – Buffalo snowfall		2	15	35	29	14	1

Table 1 illustrates the results of the pruning process. For each of the five distributions we have listed the number of terms in the final pruned models for each of the 100 re-

Table 2. *Average L_1 error for each test case for each model complexity*

Case/No. of terms	1	2	3	4	5	6	7
a – Separated bimodal		**0.017**	0.025	0.023	0.025	0.043	
		(0.006)	(0.008)	(0.006)	(0.005)	(0.006)	
b – Bimodal		**0.021**	0.027	0.030	0.036	0.034	0.038
		(0.009)	(0.010)	(0.008)	(0.013)	(0.010)	(0.030)
c – Quadmodal				**0.046**	0.050	0.057	0.043
				(0.008)	(0.011)	(0.011)	–
d – Standard normal	**0.012**	0.016	0.025	0.029	0.031	0.031	0.0236
	(0.006)	(0.006)	(0.007)	(0.007)	(0.007)	(0.007)	–
e – Buffalo snowfall							

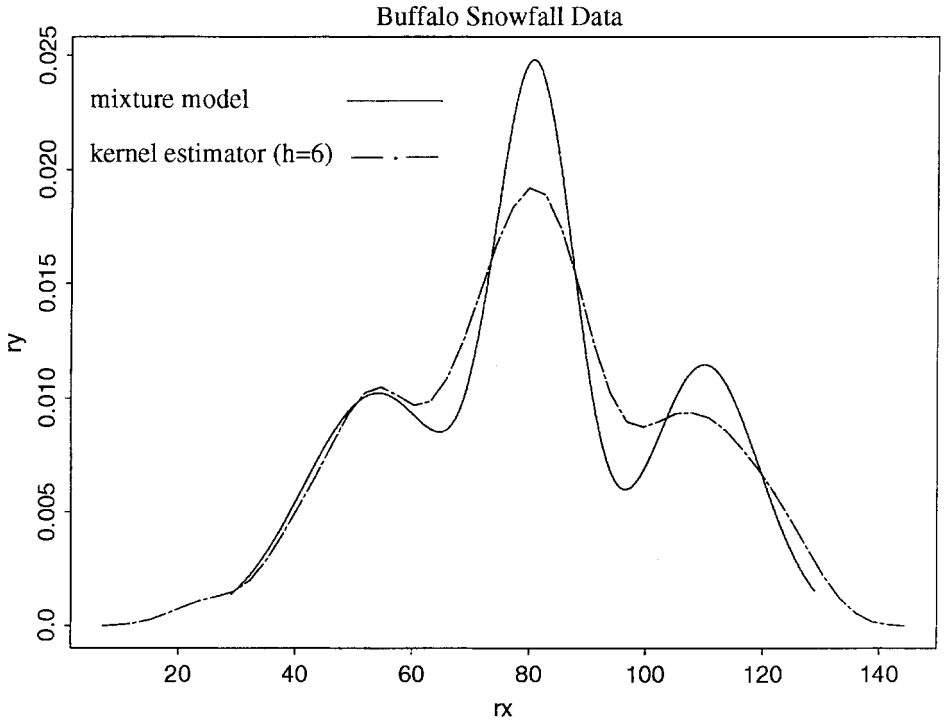

Fig. 4. *Comparison of the pruned 3-term model (solid line) which has been expectation maximized against kernel estimates of the original Buffalo snowfall data with bandwidths of 6 and 4 (dashed line)*

samples. In case a, the procedure converged to the correct solution in 53 of 100 times, 19 of 100 times in case b, 74 of 100 times in case c, and 10 of 100 times in case d. The procedure converged to a 3-term solution in 15 of the 100 times in the case of the snowfall data. The appropriate solution for the case of the snowfall data will be the subject of later discussions.

We may estimate the model complexity through the use of statistical measures on this distribution. For example one could choose the minimal order statistic as the measure of the number of terms in the minimal complexity mixture model that characterizes the data. This choice has the advantage that it represents the lowest complexity model obtainable from the procedure. Alternatively one could use the expected value of the distribution. This choice indicates the average complexity of the mixture models that represent the data set.

Table 2 presents the average L_1 error between the true mixture model and the pruned model for each of the first four cases for each of the obtained model complexities. No L_1 results were provided for the snowfall data since the true underlying model is unknown. It is encouraging to note that in all cases except the quadmodal, the minimum average error occurred at the appropriate level of model complexity. In the case of the quadmodal, a lower L_1 error was obtained for a single 7-term model. This particular case was an outlier.

The number of modes in the snowfall data has been the topic of continued debate throughout the history of density estimation. Arguments have been made in favour of tri-

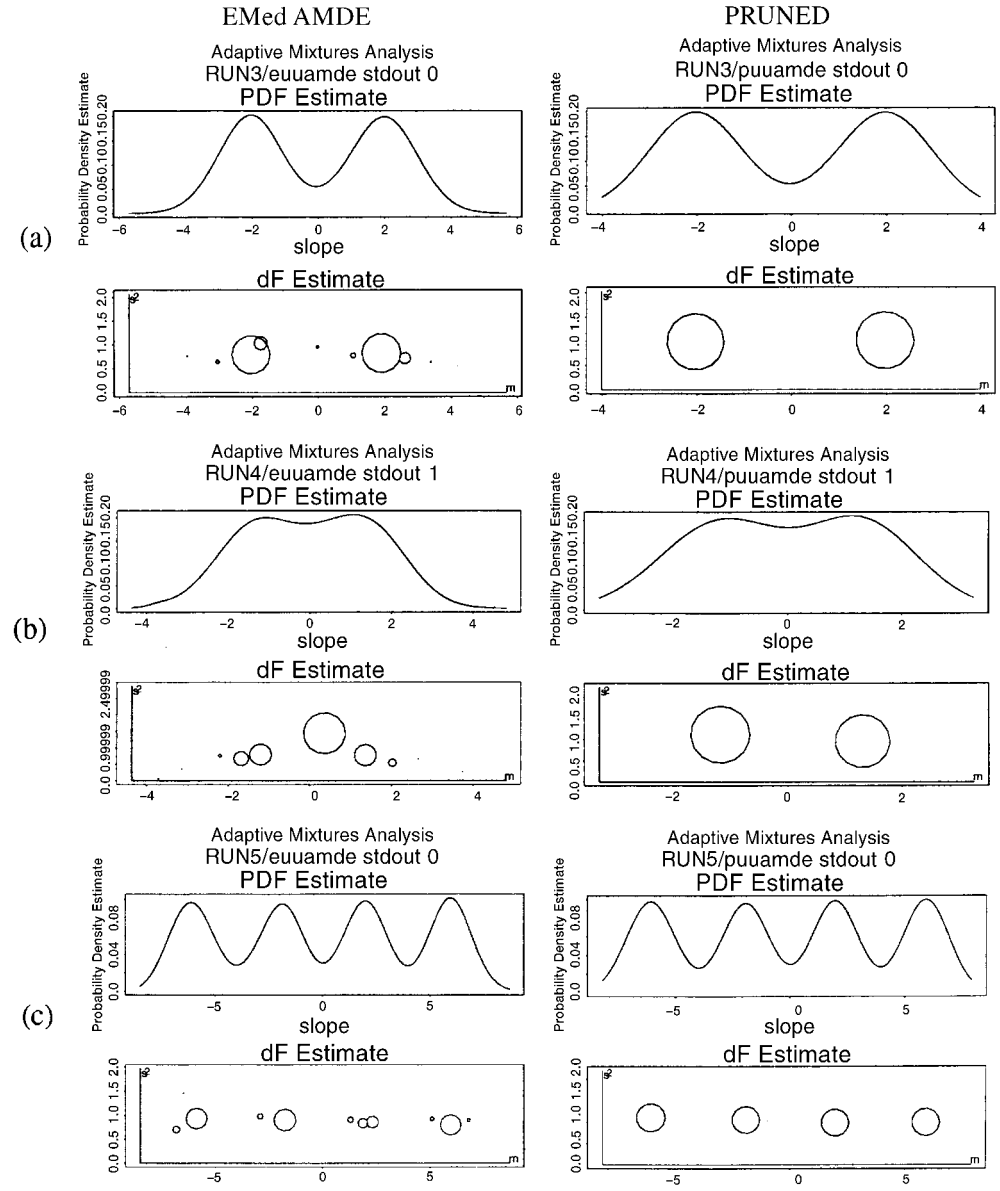

Fig. 5. *Expectation maximized adaptive mixtures estimates along with the output of the pruning process for the first three cases*

modal (Scott, 1992) and unimodal structure (Scott, 1994). In Fig. 4 we compare the output of the pruning process for one of the 3-term models to a standard kernel estimator with a bandwidth of 6. The bandwidth of 6 was chosen as an appropriate setting to illustrate the trimodality of the data (Silverman, 1986). The 3-term model has been expectation maximized against the original data in order to make a fair comparison between the two. We note that the two models are very similar in character, specifically with regard to the mode placement.

The output of the pruning procedure is discussed next. In Figs 5a, b, c and 6a and b, we present an expectation maximized adaptive mixture solution along with the output of pruning this solution. We notice that the number of terms in the solution has been reduced from ten to the appropriate number in each case. We also notice that the terms left from the process are in approximately the correct location and have about the right mixing coefficients and variances.

Finally we turn our attention to the simulation results which study the performance of the procedure on a two-term mixture model as we vary the separation of the terms and the number of points in the sample. The densities used for the study consisted of $f_0(x) = 0.3N(0,1) + 0.7N(\Delta, 2)$ where Δ varies from 1 to 4 in steps of 1 and the sample sizes assume the values of 100, 500 and 1000. One hundred data sets were generated for each possible test case and the performance of the algorithm was studied using 100 resamples per test case. The create threshold used in the adaptive mixtures step of the procedure was held constant at 4. This implies that a new term is created whenever the Mahalanobis square distance between a new observation and each data point exceeds 4. For each resample an adaptive mixtures model was built using these parameters and then we attempted to remove terms from this model using our procedure described above. Therefore, each resample produced an original and pruned mixture model.

There are numerous ways in which we may present the simulation results. One measure of complexity is the expected number of terms in the original and pruned models. This metric was used extensively in the discussions of our previous test cases. An alternative measure of the effectiveness of our procedure is the mode (most frequently occurring) complexity for the 100 original and 100 pruned models for each of the 100 data sets comprising a given test case. In Fig. 7 we present histograms of the modal distribution for the original and pruned models as measured on the 100 data sets for all of the test cases. Cross hatches that rise from left to right represent the original models while those that fall from left to right represent the pruned models. The plots are laid out so that a given column represents the test results for a particular sample size and each row represents the results at a particular separation Δ. When the bars from the original and the pruned cases occupy identical bins the bar corresponding to the pruned results has been appended to the bar corresponding to the original results.

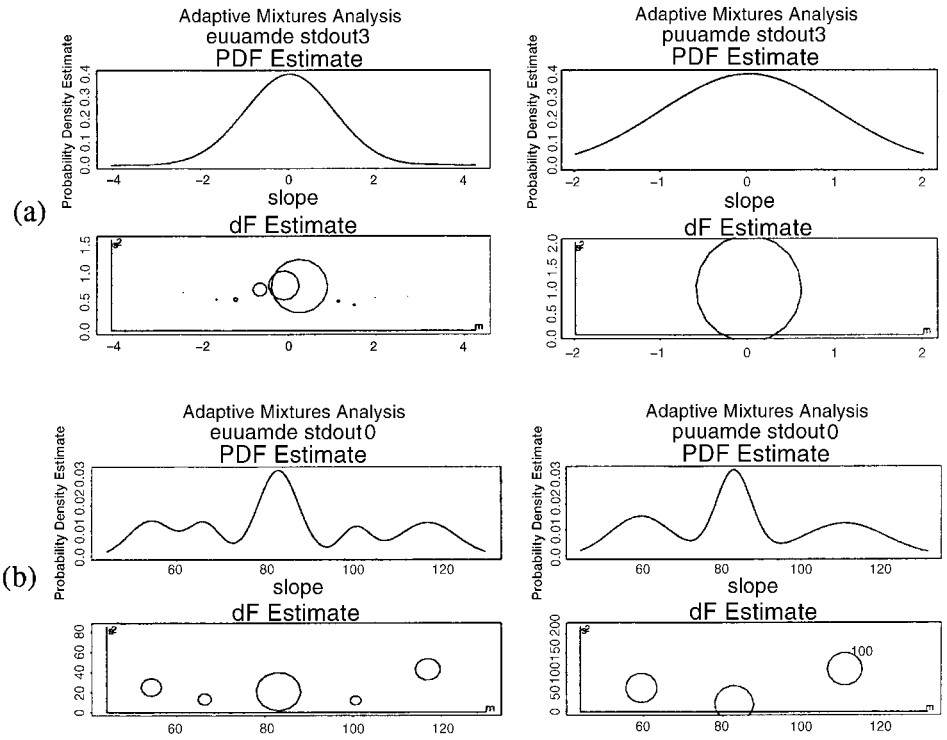

Fig. 6. *Expectation maximized adaptive mixtures estimates along with the output of the pruning process for the last two cases*

There are several things to notice about the figure. First we point out that in general the histograms for the pruned models lie to the left of those for the original ones. There is some overlap in several of the cases but the distributions represent a clear decrease in the complexity of the models in the pruned cases. Next we notice that there is an apparent increase in the number of terms in the original models as the number of data points increases. This trend is in agreement with our understanding of the inner workings of the adaptive mixtures procedure. Given the nature of our creation process the number of terms generally increases as the number of observations increases. We also notice that for a fixed sample size, the number of terms in the original model decreases as the separation between the terms, Δ, decreases. This too agrees with our intuitive understanding of the creation process. Next we turn our attention to the analysis of trends in the pruned models. In general the pruned models follow the same trends as the original models. Specifically the number of terms in the pruned models grows as the sample size increases and

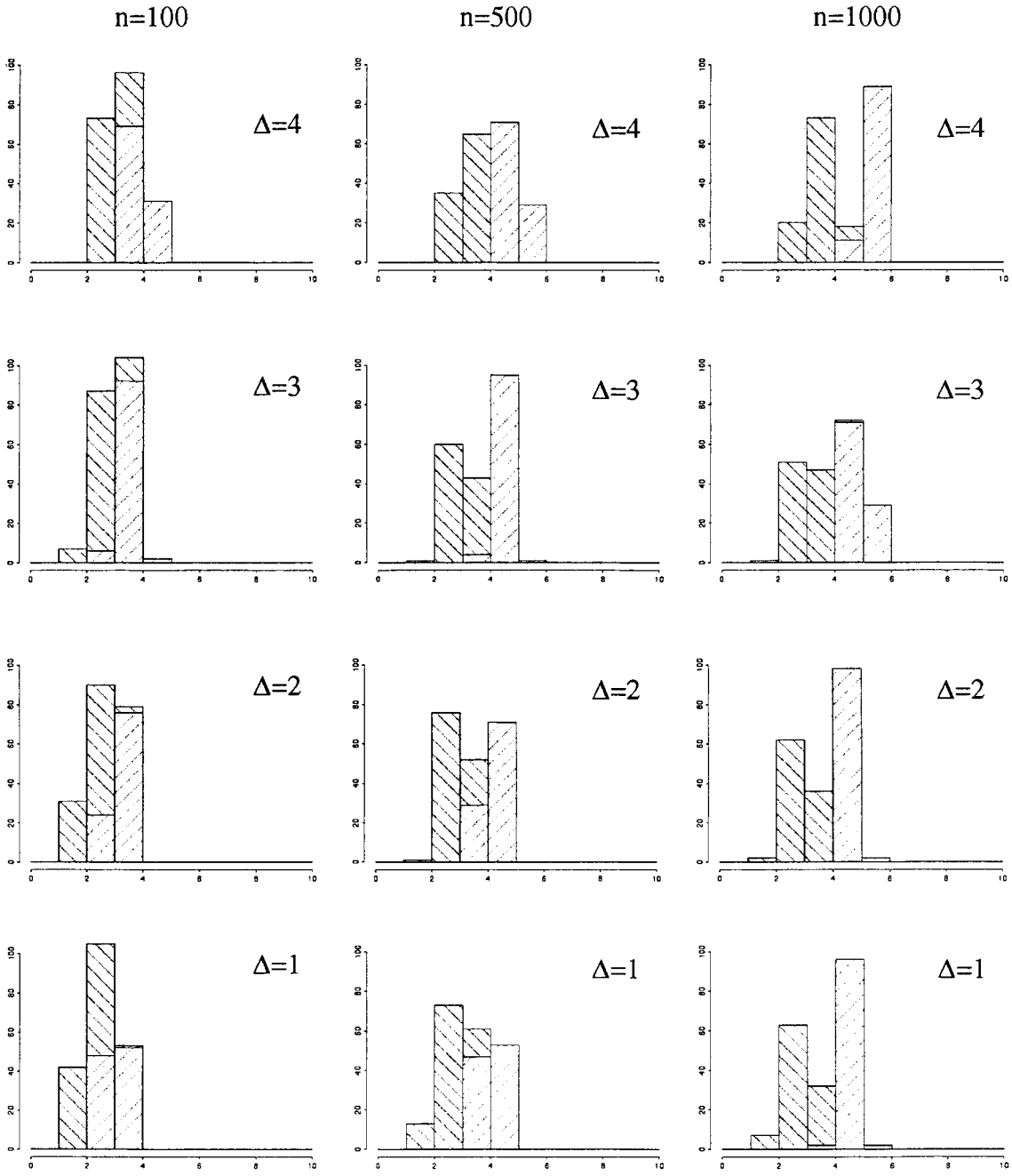

Fig. 7. *Histogram plots for original and pruned models complexity*

shrinks as the separation between the terms decreases. So the pruned model's complexity seems to be driven by the complexity in the original model.

4. Conclusions

The AMDE procedure provides a data-driven method for obtaining a good mixture model density estimate. The convergence properties of the procedure tend to guarantee that the model will be of higher complexity than the true density if the latter is a finite mixture. The exceptions to this occur when the sample size is small enough that too few terms are created by the AMDE. The AMDE does provide a useful mixture-model estimate of an unknown density; however it is not designed to nor does it yield a useful estimate of the underlying complexity in terms of the number of components present. The AIC provides a convenient tool for evaluating appropriate model complexity. It serves as a good 'rule of thumb' in choosing between models. Under appropriate conditions it has the capability to help reveal the underlying mixture which generates a data set.

In many cases we have reason to believe that the unknown density is a mixture model but of unknown complexity. In these cases we are often interested in the structure of the underlying mixture mode. It is in this case that AIC-based pruning can be used to find not only an 'optimal' model but also a distribution of pruned models which provides some knowledge about the true density. In those cases where the true model is not a mixture model we believe that the pruning process is still capable of producing more parsimonious solutions.

In this paper we have presented a new technique to help determine the unknown structure of a mixture model. This technique uses a set of adaptive mixtures solutions that have been subject to AIC-based pruning to help determine the minimum complexity mixture model that best characterizes the data. The goal of this technique is the production of a more parsimonious mixture model of an unknown distribution.

This approach embodies the spirit in which the AIC should be used, in that one is comparing two maximum likelihood solutions. There is a penalty with regard to computational complexity that occurs in the production of expectation maximized models at each step of the pruning process. However, the pruning procedure is highly parallel in nature and we could expect substantial speed-ups on a parallel machine.

Acknowledgements

The authors (JLS, WLP, GWR) would like to acknowledge the support of the NSWCDD Independent Research Program through the Office of Naval Research. The work of EJW was supported by the Army Research Office under contract number DAAH04-94-G-0267, and by the Office of Naval Research under contract number N00014-92-J-1303. In addition the work of CEP was supported through the Office of Naval Research under contract number N00014-95-1-0777. Finally, we would like to thank the reviewers for their many helpful comments.

References

Akaike, H. (1974) A new look at the statistical model identification . *IEEE Transactions on Automatic Control*, **19**, 716–23.

Binder, D. A. (1978) Bayesian cluster analysis. *Biometrika*, **65**(1), 31–8.

Bozdogan, H. and Sclove, S. L. (1984) Multi-sample cluster analysis using Akaike's information criterion. *Annals of the Institute of Statistics and Mathematics*, **36**, 163–80.

Carmen, C. S. and Merickel, M. (1990) Supervising isodata with an information theoretic stopping rule. *Pattern Recognition*, **23**, 185–97.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977) Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, **39**, 1–38.

Efron, B. and Tibshirani, R. (1993) *An Introduction to the Bootstrap*, London: Chapman and Hall.

Everitt, B. S. and Hand, D. J. (1981) *Finite Mixture Distributions*, London: Chapman and Hall.

Liang, Z., Jaszczak, R. J. and Coleman, R. E. (1992) Parameter estimation of finite mixtures using the EM algorithm and information criteria with applications to medical image processing. *IEEE Transactions on Nuclear Science*, **39**(4), 1126–33.

Marron, J. S. and Wand, M. P. (1992) Exact mean integrated squared error. *Annals of Statistics*, **20**(2), 712–36.

McLachlan, G. J. (1987) On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. *Applied Statistics*, **36**(3), 318–24.

McLachlan, G. J. and Basford, K. E. (1988) *Mixture Models*, New York: Marcel Dekker.

Milligan, G. W. and Cooper M. C. (1985) An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, **50**(1), 159–79.

Parzen, E. (1979) Nonparametric statistical data modeling. *Journal of the American Statistical Association*, **74**, 105–31.

Priebe, C. E. (1994) Adaptive mixtures. *Journal of the American Statistical Association*, **89**, 796–806.

Priebe, C. E. and Marchette, D. J. (1993) Adaptive mixture density estimation. *Pattern Recognition*, **26**(5), 771–85.

Priebe, C. E., Solka, J. L. and Rogers, G. W. (1993) Discriminant analysis in aerial images using fractal based features. In F. A. Sadjadi (ed.) *Adaptive and Learning Systems II, Proc. SPIE 1962*, pp. 196–208.

Priebe, C. E., Solka, J. L., Lorey, R. A., Rogers, G. W., Poston, W. L., Kallergi, M., Qian, W., Clarke, L. P. and Clark, R. A. (1994) The application of fractal analysis to mammographic tissue classification. *Cancer Letters*, **77**, 183–89.

Scott, D. W. (1985b) Frequency polygons. *Journal of the American Statistical Association*, **80**, 348–54

Scott, D. W. (1985b) Average shifted histograms: effective non-parametric density estimation in several dimensions. *Annals of Statistics*, **13**, 1024–40.

Scott, D. W. (1992) *Multivariate Density Estimation*, New York: John Wiley.

Scott, D. W. (1994) Multivariate Density Estimation, *Short Course Interface 1994*.

Silverman, B. W. (1986) *Density Estimation for Statistics and Data Analysis*. New York: Chapman and Hall.

Solka, J. L. (1995) Matching Model Information Content to Data Information, PhD Dissertation, George Mason University, Fairfax, Virginia.

Solka, J. L., Priebe, C. E. and Rogers, G. W. (1992) An initial assessment of discriminant surface complexity for power law features. *Simulation*, **58**(5), 311–18.

Solka, J. L., Priebe, C. E. and Rogers, G. W. (1993) A probabilistic approach to fractal based texture discrimination. In F. A. Sadjadi (ed.) *Adaptive and Learning Systems II*, *Proc. SPIE 1962*, pp. 209–18.

Solka, J. L., Priebe, C. E., Rogers, G. W., Poston, W. L. and Lorey, R. A. (1994) Maximum likelihood density estimation with term creation and annihilation. In *Computationally Intensive Statistical Methods*, *Proceedings of the 26th Symposium on the Interface*, pp. 222–25.

Solka, J. L., Poston, W. L. and Wegman, E. J. (1995) A visualization technique for studying the iterative estimation of mixture densities. *Journal of Computational and Graphical Statistics*, **4**(3), 180–97.

Sturges, H. A. (1926) The choice of a class interval. *Journal of the American Statistical Association*, **21**, 65–6.

Titterington, D. M. (1984) Recursive parameter estimation using incomplete data. *Journal of the Royal Statistical Society, Series B*, **46**, 257–67.

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

Wallace, C. S. and Boulton D. M. (1968) An information measure for classification. *Computer Journal*, **11**, 185–94.

Wegman, E. J. (1970) Maximum likelihood estimation of a unimodal density function. *Annals of Statistics*, **41**, 457–71.