BIOMETRICS 56, 483–486
June 2000

Mixture Models, Robustness, and the Weighted Likelihood Methodology

Marianthi Markatou

Department of Statistics, Columbia University, 615 Mathematics Building, New York, New York 10027, U.S.A.
email: markat@stat.columbia.edu

SUMMARY. Problems associated with the analysis of data from a mixture of distributions include the presence of outliers in the sample, the fact that a component may not be well represented in the data, and the problem of biases that occur when the model is slightly misspecified. We study the performance of weighted likelihood in this context. The method produces estimates with low bias and mean squared error, and it is useful in that it unearths data substructures in the form of multiple roots. This in turn indicates multiple potential mixture model fits due to the presence of more components than originally specified in the model. To compute the weighted likelihood estimates, we use as starting values the method of moment estimates computed on bootstrap subsamples drawn from the data. We address a number of important practical issues involving bootstrap sample size selection, the role of starting values, and the behavior of the roots. The algorithm used to compute the weighted likelihood estimates is competitive with EM, and it is similar to EM when the components are not well separated. Moreover, we propose a new statistical stopping rule for the termination of the algorithm. An example and a small simulation study illustrate the above points.

KEY WORDS: Mixture models; Robustness; Roots; Weighted likelihood.

1. Introduction

Mixtures of distributions are used extensively to model a wide variety of important practical situations. Lindsay (1995, Chapter 1) gives a detailed account of problems that can be reformulated as mixture model problems. Moreover, McLachlan and Basford (1988) and Titterington, Smith, and Makov (1985) presented references to hundreds of examples of applications in diverse fields.

One area of extensive application of finite mixture models is the analysis of fisheries data. Fisheries scientists are often interested in modeling a discrete number of fish age classes by a mixture of normal components. The presence of outliers in the sample can have a large effect on the parameter estimates and the fit. Another problem with the analysis of such data is the fact that older age classes may be represented by very few individuals in the sample. In this case, the question of how many components are present in the mixture is relevant.

A realistic goal in this situation is to obtain estimates of the parameters describing the major components of the mixture that are unaffected by a small proportion of data from components not well reflected in the sample. We propose to achieve this goal by using the weighted likelihood (WLEE) methodology (Markatou, Basu, and Lindsay, 1998). The method also provides a heuristic for identification of the potential number of components.

Another problem with maximum likelihood analysis of mixtures is that biases can occur when there is model misspecification (Gray, 1994). Weighted likelihood performs well, producing estimates with low bias.

In Section 2, we describe the WLEE methodology as it applies to the finite mixture model. The important problem of bandwidth selection is discussed. In Section 3, we discuss the algorithmic issues involved and propose a new statistical stopping rule that is based on the weighted likelihood score. In Section 4, we apply the method to a data set on scallop length measurements and report briefly simulation results.

2. Weighted Likelihood for Mixture Models

Let $X_1, X_2, \ldots, X_n$ be a random sample of completely unclassified observations from a finite mixture distribution with g components. The probability density function of an observation X is $m(x;\phi) = \Sigma_{i=1}^{g} p_i f(x;\theta_i)$, where $f(x;\theta_i)$ is the probability density or mass function of the ith subpopulation and θ_i is the parameter that describes the specific attributes of the ith component. The vector $\phi = (p_1, p_2, \ldots, p_g, \theta_1, \ldots, \theta_g)^{\mathrm{T}}$ of all unknown parameters belongs to a parameter space Ω subject to $\Sigma_{i=1}^{g} p_i = 1, p_i \geq 0, i = 1, 2, \ldots, g$, and θ_i belongs to Θ.

The weighted likelihood estimating equations are

$$\sum_{j=1}^{n} w(\delta(x_j))[\nabla_\phi \ln m(x_j;\phi)] = 0,$$

where $\delta(x_j)$ is the Pearson residual evaluated at x_j and $w(\cdot)$ is a weight function that downweights observations that have

large residuals. If $m(x;\phi)$ is a discrete probability model, $\delta(t) = (d(t) - m(t;\phi))/m(t;\phi)$ (Lindsay, 1994), where $d(t)$ is the proportion of observations in the sample with value t. If the model is continuous, $\delta(x_j) = \{f^*(x_j)/m^*(x_j;\phi)\} - 1$, where $f^*(\cdot)$ is a kernel density estimator and $m^*(\,.\,;\phi)$ is the mixture model smoothed with the same kernel used to obtain the density estimator (Markatou et al., 1998). The range of δ is the interval $[-1,\infty)$. A value of the residual close to zero indicates agreement between the data and the hypothesized model in the neighborhood of the observation x_i. On the other hand, a large value of $\delta(x_i)$ indicates a discrepancy between the data and the model in the form of an excess of observations relative to the model prediction; such an observation is called an outlier. A value of it near -1 indicates a dearth of observations relative to the model prediction and is called an inlier.

The weight functions $w(\cdot)$ are unimodal in that they decline smoothly as the residual δ departs from zero toward -1 or $+\infty$ and take the maximal value of one when the residual δ is zero. Observations consistent with the model receive a weight of approximately one; inconsistent observations receive a weight of approximately zero. The weight function we use is $w(\delta) = 1 - (\delta/(\delta+2))^2$, which corresponds to a symmetric chi-square distance. Markatou et al. (1998) discuss extensively the construction and motivation of the weight functions that can be used.

The weighted likelihood equations for θ_i are

$$\sum_{j=1}^{n}\sum_{i=1}^{g} w(\delta(x_j))\tau_{ij}u(x_j;\theta_i) = 0, \tag{1}$$

where $u(x_j;\theta_i) = \nabla_{\theta_i} \ln f(x_j;\theta_i)$, $\tau_{ij} = p_i f(x_j;\theta_i)/m(x_j;\phi)$.

For the vector $p = (p_1, p_2, \ldots, p_g)^{\mathrm{T}}$, we obtain

$$p_i = \sum_{j=1}^{n} \frac{w(\delta(x_j))}{\sum_{j=1}^{n} w(\delta(x_j))}\tau_{ij}. \tag{2}$$

The solutions of the system (1), (2) are the WLEE estimators of the parameter vector $\phi = (p_1, p_2, \ldots, p_g, \theta_1, \theta_2, \ldots, \theta_g)^{\mathrm{T}}$.

Assume the model is a mixture of $\mathrm{N}(\mu_i, \sigma^2)$ densities. Then a natural kernel for constructing δ is the normal density with variance h^2. The variance h^2 serves as the bandwidth parameter and determines the robustness properties of the WLEE estimator. We propose to select $h^2 = c\hat{\sigma}^2$, where $\hat{\sigma}^2$ is the estimated model variance and $c > 0$. Notice that the variance of the mixing distribution does not enter in the determination of h^2. Similarly, when the components of the mixture model are $\mathrm{N}(\mu_i, \sigma_i^2)$ densities, we propose to select $h^2 = c \cdot (\Sigma_{i=1}^{g} \hat{p}_i\hat{\sigma}_i^2)$, where $\hat{p}_i$ and $\hat{\sigma}_i^2$ are estimates of the proportion and variance, respectively, of the ith subpopulation, $c > 0$. To complete the selection of h^2, we need a rule for selecting c. In what follows, we will justify these proposals and provide a way to select c.

Let $\tilde{w}^* = n^{-1}\,\Sigma\, w(\delta(x_j)$ be the average of the weights at true value ϕ. For a unimodal weight function, under suitable regularity conditions, the asymptotic mean of $n(1-\tilde{w}^*)$ is

$$\wedge = -\frac{A_2}{2}\left\{\int \frac{\int k^2(x;t,h)dM_\phi(x)}{m_\phi^*(t)} dM_\phi(t) - 1\right\}, \tag{3}$$

where $A_2 = w''(0) < 0$. This expression is not elementary to compute, so if the components are well separated, we obtain the approximation

$$\wedge \cong -\frac{A_2}{2}\left[\sum_{i=1}^{g} p_i \int \frac{\int k^2(x;t,h)dF(x;\theta_i)}{f^*(t;\theta_i)} dF(t;\theta_i) - 1\right]. \tag{4}$$

Because of the normality of model and kernel, when all components have the same σ^2, the bracketed term in the mean downweighting factor is $(\sigma^2+h^2)^{3/2}/(3\sigma^2+h^2)^{1/2}h^2 - 1$. If the separation is poor, $M_\phi(x)$ will be similar to a normal with variance $\sigma^2+\sigma_Q^2$, where σ_Q^2 is the variance of the mixing distribution, giving smaller approximate mean downweighting with σ_Q^2 small. Since the mean downweighting depends on the parameter σ^2, if h^2 is held fixed, the robustness properties will vary with the true value of σ^2. Therefore, selecting $h^2 = c\sigma^2$ makes these equations parameter free and c can be obtained by solving the equation $-(A_2/2)\{(1+c)^{3/2}/c(3+c)^{1/2} - 1\} = \gamma_0$, where $\gamma_0 > 0$ is the number of observations to be downweighted on average, selected by the user. In the multivariate case, we suggest selecting the smoothing matrix as $H = c\hat{\Sigma}$, where $\hat{\Sigma}$ is the estimated covariance and the constant c is obtained by solving the equation

$$-(A_2/2)\{[(1+c)^{3/2}/c(3+c)^{1/2}]^m - 1\} = \gamma_0,$$

$m = \dim(\phi)$. This equation assumes that $\Sigma = \mathrm{diag}(\sigma_i^2)$. In the general case, equation (3) is used.

When the model is a finite mixture of normal components with parameters μ_i and σ_i^2, the bracketed term in the mean downweighting factor (4) is

$$\sum_{i=1}^{g} p_i\left(\sigma_i^2+h^2\right)^{3/2} / \left(3\sigma_i^2+h^2\right)^{1/2} h^2 - 1.$$

A first-order Taylor expansion of the function $(\sigma_i^2+h^2)^{3/2}/h^2(3\sigma_i^2+h^2)^{1/2}$ with respect to h^2 in the neighborhood of zero produces $h^2 = c(\Sigma_{i=1}^{g}\hat{p}_i\hat{\sigma}_i^2)$, where $\hat{p}_i, \hat{\sigma}_i^2$ are estimates of the proportion and variance, respectively, of the ith subpopulation and c is selected by solving the equation $-(A_2/2)\{(0.57735/c - 0.2302)\} = \gamma_0$. Plotting this function and the one obtained when all components have equal variance as a function of c, we see that the two graphs are the same for small values of c. In the multivariate case, we select $H = c\,\Sigma_{i=1}^{g}\hat{p}_i\hat{\Sigma}_i$.

3. Algorithmic Issues

To solve (1) and (2), we generate starting values by obtaining B bootstrap subsamples of size m from the data and compute the method of moment estimates (MME) using these subsamples (Lindsay, 1989; Lindsay and Basak, 1993; Furman and Lindsay, 1994). It seems that $B = 50$ bootstrap values suffice to identify important data substructures. But if a component is not well represented in the sample, the number of bootstrap values needs to be increased considerably for corresponding root identification. The motivation for this search is that different small subsets may represent regions of the sample space fit by different models. McLachlan and Peel (1998b) also use random starting points in their MIXFIT algorithm.

The bootstrap sample size is important for obtaining positive MME of variance. This depends on the model under investigation. We empirically found that, in a two-component

univariate normal mixture model, a bootstrap subsample of size equal to the number of parameters to be estimated is sufficient to produce reasonable estimates. This has a statistical interpretation of allocating one observation per parameter. On the other hand, for a mixture of two fairly well separated Poisson(θ_i), a sample size of five produces, in most bootstrap samples, a positive MME of variance and feasible estimates of the θ_i's. When the mixture is that of three Poissons, a reasonable bootstrap sample size is 15.

To terminate the weighted likelihood algorithm, we use a new statistical stopping rule that is constructed as follows. Denote the elements of the weighted score vector by $S_\ell, \ell = 1, \ldots, m = \dim(\phi)$, and let $S = (S_1, \ldots, S_m)^{\mathrm{T}}$. For any ϕ_ℓ^0, we could consider testing H_0: $\phi = \phi^0$ based on the statistic

$$n\left(\frac{1}{n}\sum_{j=1}^{n} S^0(x_j)\right)^{\mathrm{T}}\left(\frac{1}{n}\sum_{j=1}^{n} S^0(x_j)S^{0\,\mathrm{T}}(x_j)\right)^{-1} \times \left(\frac{1}{n}\sum_{j=1}^{n} S^0(x_j)\right). \tag{5}$$

This statistic is zero at $\phi = \hat{\phi}$, as $\Sigma S(x_j; \hat{\phi}) = 0$. The target now is to make sure that the value of ϕ at the jth step of the iteration is well within the confidence set with confidence coefficient $100(1-\alpha)\%$. We evaluate (5) at each iteration and terminate the algorithm when it becomes less than or equal to the α-quantile of a χ^2_m. Note that this idea can be used to evaluate the different stopping criteria proposed in the literature in connection with the computation of the maximum likelihood estimate. This can be done by evaluating a statistic analogous to (5) at each $\hat{\phi}$ produced and comparing the values with the appropriate χ^2_m quantile. The smaller the value of the statistic, the closer the estimate is to the true value of ϕ.

4. Example and Simulation

The data consist of length measurements of 222 scallops caught in a 79-m^2 area during a dredge survey in Mercury Bay, New Zealand, and can be found in Jorgensen (1990). There is a fairly well-defined component of smaller animals containing 170 observations, or 76.75% of the sample, followed by a spread-out tail of larger scallops. The data contain little information regarding the decomposition of the larger animals into components. Heterogeneity within a cohort cannot be explained as a result of growth of scallops because it is extremely small. Thus, if we take the scallops with lengths between 62 and 82 mm to constitute one component, the remaining scallops must constitute at least two components. Jorgensen (1990) argues that 35 of the larger measurements, or 15.766% of the sample, comprise one component and the remaining 17 observations, or 7.659% of the sample, belong to a third component. This last component is not well represented in the sample.

We start by fitting a two-component normal model with common σ^2. As starting values, we use both MME calculated on the entire sample and MME computed only on bootstrap samples. Let $\phi = (p, \mu_1, \mu_2, \sigma)$ be the parameter vector with μ_1 corresponding to the mean of the largest component. The MLE of ϕ is $(0.799, 72.243, 100.163, 6.295)$. The WLEE($c$ = 0.3) with $\alpha = 0.05$ for the stopping rule is (0.804, 71.643, 95.287, 4.856), estimating accurately the parameters of the two well represented in the sample components. To compute the MLE, we used the implementation of the EM algorithm given in McLachlan and Basford (1988, pp. 213–216).

Table 1

MME, MLE, and WLEE estimates and their mean squared errors. The sampling model is $0.5t_3^{(r)} + 0.5\mathrm{N}(8,1)$ *and the fitted model is* $p\mathrm{N}(\mu_1, \sigma^2) + (1-p)\mathrm{N}(\mu_2, \sigma^2)$. *The notation* $t_3^{(r)}$ *means that the sample is rescaled to have variance one.*

Estimates	$\hat{p}$	$\hat{\mu}_1$	$\hat{\mu}_2$
MME	0.5045	7.9621	−0.0266
	(0.0027)	(0.2254)	(0.3140)
MLE	0.5073	7.9497	−0.1594
	(0.0025)	(0.1998)	(0.2577)
WLE(0.050)	0.5027	7.9735	0.0416
	(0.0021)	(0.1709)	(0.0977)
WLE(0.010)	0.4924	7.9704	0.0307
	(0.0024)	(0.1498)	(0.0668)
WLE(0.001)	0.4629	7.9600	0.0243
	(0.0047)	(0.1599)	(0.0550)

When bootstrap root search was performed with 100 starting values, a root corresponding to the third component was identified only twice. We then fitted a third component. The MLE of ϕ is (0.764, 0.163, 71.557, 92.830, 110.514, 4.540) and the WLEE is (0.768, 0.156, 71.533, 92.265, 108.554, 4.38). An examination of the weights reveals that every observation has a weight of approximately one except observation 222. This is the scallop with the largest length of 126 mm and receives a weight of 0.06858. This observation was identified as the most influential by Jorgensen (1990).

To compute the WLEE estimators, we used a chi-square weight and a normal kernel for both the example and the simulation. The aim of the simulation is to assess performance under model misspecification. Samples from $0.5t_3 + 0.5\mathrm{N}(8,1)$ were generated in Splus. The fitted model was a mixture of two normals with parameters μ_i and σ^2. The data from the t_3 distribution was scaled by $3^{1/2}$ so that the variance is one. The programs were written in FORTRAN77. All calculations were carried out on a DEC5000/50 station.

Table 1 presents the MME, MLE, and WLEE using MME as starting values and, in parentheses, their associated mean squared errors. When bootstrap was implemented, only one root was identified. From all three estimates, the one exhibiting the highest bias is the MLE. However, the estimate with the smaller mean squared error is the WLEE. The mean squared error becomes increasingly smaller as the value of c decreases. This is especially true for $\hat{\mu}_2$. Further experimentation with mixtures of two t-distributions with different degrees of freedom verifies the above behavior.

Acknowledgements

This work was supported by an RGK Foundation grant and an NSF grant to M. Markatou and by an NSF grant to Ingram Olkin. The author thanks Prof B. G. Lindsay for stimulating discussions and the editor and those involved in the editorial process for helpful suggestions that improved the manuscript.

Résumé

Les problèmes associés à l'analyse des données d'un mélange de distributions incluent la présence de données aberrantes dans l'échantillon, le fait qu'un composant peut ne pas être bien représenté dans les données et le problème des biais qui apparaissent lorsque le modèle est légèrement mal spécifié. Nous étudions les performances de la vraisemblance pondérée dans ce contexte. La méthode donne des estimateurs avec un biais et une erreur quadratique moyenne faibles, et elle est utile car elle permet de faire apparaître des sous structures dans les données sous la forme de racines multiples. Cela indique alors la possibilité d'un modèle multiple qui s'ajusterait mieux à cause de la présence d'un nombre de composants plus important que celui spécifié au départ. Pour calculer les estimations par la vraisemblance pondérée nous utilisons comme valeurs initiales les estimations obtenues par la méthode des moments sur des sous-échantillons bootstrap tirés des données. Nous présentons des résultats importants pour la pratique concernant le choix de la taille des échantillons bootstrap, le rôle des valeurs initiales et le traitement des racines. L'algorithme utilisé pour calculer les estimations par la vraisemblance pondérée est compétitive avec l'algorithme EM et il est semblable à lui lorsque les composants ne sont pas bien séparés. De plus, nous proposons une nouvelle règle statistique d'arrêt pour terminer l'algorithme. Un exemple et une petite étude par simulation illustrent les points précédents.

References

Furman, D. W. and Lindsay, B. G. (1994). Measuring the relative effectiveness of moment estimators as starting values in maximizing likelihoods. *Computational Statistics and Data Analysis* **17,** 493–507.

Gray, G. (1994). Bias in misspecified mixtures. *Biometrics* **50,** 457–470.

Jorgensen, M. A. (1990). Influence-based diagnostics for finite mixture models. *Biometrics* **46,** 1047–1058.

Lindsay, B. G. (1989). Moment matrices: Applications in mixtures. *Annals of Statistics* **17,** 722–740.

Lindsay, B. G. (1994). Efficiency versus robustness: The case of minimum Hellinger distance and related methods. *Annals of Statistics* **22,** 1018–1114.

Lindsay, B. G. (1995). *Mixture Models: Theory, Geometry and Applications,* Volume 5, *NSF-CBMS Regional Conference Series in Probability and Statistics.* Hayward, California: Institute of Mathematical Studies.

Lindsay, B. G. and Basak, P. (1993). Multivariate normal mixtures: A fast consistent method of moments. *Journal of the American Statistical Association* **88,** 468–476.

Markatou, M., Basu, A., and Lindsay, B. G. (1998). Weighted likelihood estimating equations with a bootstrap root search. *Journal of the American Statistical Association* **93,** 740–750.

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

McLachlan, G. J. and Peel, D. (1998a). Robust cluster analysis via mixtures of multivariate t-distributions. In *Lecture Notes in Computer Science,* Volume 1451, A. Amin, D. Doris, P. Pudil, and H. Freeman (eds), 658–666. Berlin: Springer-Verlag.

McLachlan, G. J. and Peel, D. (1998b). MIXFIT: An algorithm for the automatic fitting and testing of normal mixture models. In *Proceedings of the 14th International Conference on Pattern Recognition,* Volume I, 553–557. Los Alamitos, California: IEEE Computer Society.

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

Received August 1998. *Revised August* 1999.
Accepted August 1999.