SIAM REVIEW
Vol. 26, No. 2, April 1984

002

MIXTURE DENSITIES, MAXIMUM LIKELIHOOD AND THE EM ALGORITHM*

RICHARD A. REDNER† AND HOMER F. WALKER‡

Abstract. The problem of estimating the parameters which determine a mixture density has been the subject of a large, diverse body of literature spanning nearly ninety years. During the last two decades, the method of maximum likelihood has become the most widely followed approach to this problem, thanks primarily to the advent of high speed electronic computers. Here, we first offer a brief survey of the literature directed toward this problem and review maximum-likelihood estimation for it. We then turn to the subject of ultimate interest, which is a particular iterative procedure for numerically approximating maximum-likelihood estimates for mixture density problems. This procedure, known as the EM algorithm, is a specialization to the mixture density context of a general algorithm of the same name used to approximate maximum-likelihood estimates for incomplete data problems. We discuss the formulation and theoretical and practical properties of the EM algorithm for mixture densities, focussing in particular on mixtures of densities from exponential families.

Key words. mixture densities, maximum likelihood, EM algorithm, exponential families, incomplete data

1. Introduction. Of interest here is a parametric family of *finite mixture densities,* i.e., a family of probability density functions of the form

$$p(x \mid \Phi) = \sum_{i=1}^{m} \alpha_i p_i(x \mid \phi_i), \qquad x = (x_1, \cdots, x_n)^T \in R^n, \tag{1.1}$$

where each α_i is nonnegative and $\sum_{i=1}^{m} \alpha_i = 1$, and where each p_i is itself a density function parametrized by $\phi_i \in \Omega_i \subseteq R^{n_i}$. We denote $\Phi = (\alpha_1, \cdots, \alpha_m, \phi_1, \cdots, \phi_m)$ and set

$$\Omega = \left\{(\alpha_1, \cdots, \alpha_m, \phi_1, \cdots, \phi_m): \sum_{i=1}^{m} \alpha_i = 1 \text{ and } \alpha_i \geqq 0, \phi_i \in \Omega_i \text{ for } i = 1, \cdots, m\right\}.$$

The more general case of a possibly infinite mixture density, expressible as

$$\int_{\Lambda} p(x \mid \Phi(\lambda)) d\alpha(\lambda), \tag{1.2}$$

is not considered here, even though much of the following is applicable with few modifications to such a density. For general references dealing with infinite mixture densities and related densities not considered here, see the survey of Blischke [13]. Also, it is understood that in determining probabilities, probability density functions are integrated with respect to a measure on R^n which is either Lebesgue measure, counting measure on some finite or countably infinite subset of R^n, or a combination of the two. In the following, it is usually obvious from the context which measure on R^n is appropriate for a particular probability density function, and so measures on R^n are not specified unless there is a possibility of confusion. It is further understood that the topology on Ω is the natural product topology induced by the topology on the real numbers. At times when it is convenient to determine this topology by a norm, we will regard elements of Ω as $(m + \sum_{i=1}^{m} n_i)$-vectors and consider norms defined on such vectors.

Finite mixture densities arise naturally—and can naturally be interpreted—as densities associated with a statistical population which is a mixture of m component populations with associated component densities $\{p_i\}_{i=1,\cdots,m}$ and mixing proportions

*Received by the editors April 6, 1982, and in revised form August 5, 1983.

†Division of Mathematical Sciences, University of Tulsa, Tulsa, Oklahoma 74104.

‡Department of Mathematics, University of Houston, Houston, Texas 77004. The work of this author was supported by the U.S. Department of Energy under grant DE-AS05-76ER05046.

$\{\alpha_i\}_{i=1,\ldots,m}$. Such densities appear as fundamental models in areas of applied statistics such as statistical pattern recognition, classification and clustering. (As examples of general references in the broad literature on these subjects, we mention Duda and Hart [46], Fukunaga [51], Hartigan [68], Van Ryzin [148], and Young and Calvert [159]. For some specific applications, see, for example, the Special Issue on Remote Sensing of the *Communications in Statistics* [33]). In addition, finite mixture densities often are of interest in life testing and acceptance testing (cf. Cox [35], Hald [65], Mendenhall and Hader [101], and other authors referred to by Blischke [13]). Finally, many scientific investigations involving statistical modeling require by their very nature the consideration of mixture populations and their associated mixture densities. The example of Hosmer [75] below is simple but typical. For references to other examples in fishery studies, genetics, medicine, chemistry, psychology and other fields, see Blischke [13], Everitt and Hand [47], Hosmer [74] and Titterington [145].

Example 1.1. According to the International Halibut Commission of Seattle, Washington, the length distribution of halibut of a given age is closely approximated by a mixture of two normal distributions corresponding to the length distributions of the male and female subpopulations. Thus the length distribution is modeled by a mixture density of the form

$$p(x \mid \Phi) = \alpha_1 p_1(x \mid \phi_1) + \alpha_2 p_2(x \mid \phi_2), \qquad x \in R^1, \tag{1.3}$$

where, for $i = 1, 2$,

$$p_i(x \mid \phi_i) = \frac{1}{\sqrt{2\pi}\sigma_i} e^{-(x-\mu_i)^2/2\sigma_i^2}, \qquad \phi_i = (\mu_i, \sigma_i^2)^T \in R^2, \tag{1.4}$$

and $\Phi = (\alpha_1, \alpha_2, \phi_1, \phi_2)$. Suppose that one would like to estimate Φ on the basis of some sample of length measurements of halibut of a given age. If one had a large sample of measurements which were labeled according to sex, then it would be an easy and straightforward matter to obtain a satisfactory estimate of Φ. Unfortunately, it is reported in [75] that the sex of halibut cannot be easily (i.e., cheaply) determined by humans; therefore, as a practical matter, it is likely to be necessary to estimate Φ from a sample in which the majority of members are not labeled according to sex.

Regarding p in (1.1) as modeling a mixture population, we say that a sample observation on the mixture is *labeled* if its component population of origin is known with certainty; otherwise, we say that it is *unlabeled*. The example above illustrates the central problem with which we are concerned here, namely that of estimating Φ in (1.1) using a sample in which some or all of the observations are unlabeled. This problem is referred to in the following as the *mixture density estimation problem*. (For simplicity, we do not consider here the problem of estimating not only Φ but also the number m of component populations in the mixture.) A variety of cases of this problem and several approaches to its solution have been the subject of or at least touched on by a large, diverse set of papers spanning nearly ninety years. We begin by offering in the next section a cohesive but very sketchy review of those papers of which we are aware which have as their main thrust some aspect of this problem and its solution. It is hoped that this survey will provide both some perspective in which to view the remainder of this paper and a starting point for those who wish to explore the literature associated with this problem in greater depth.

Following the review in the next section, we discuss at some length the method of maximum-likelihood for the mixture density estimation problem. In rough general terms, a maximum-likelihood estimate of a parameter which determines a density function is a choice of the parameter which maximizes the induced density function (called in this

context the *likelihood function*) of a given sample of observations. Maximum-likelihood estimation has been the approach to the mixture density estimation problem most widely considered in the literature since the use of high speed electronic computers became widespread in the 1960's. In §3, the maximum-likelihood estimates of interest here are defined precisely, and both their important theoretical properties and aspects of their practical behavior are summarized.

The remainder of the paper is devoted to the subject of ultimate interest here, which is a particular iterative procedure for numerically approximating maximum-likelihood estimates of the parameters in mixture densities. This procedure is a specialization to the mixture density estimation problem of a general method for approximating maximum-likelihood estimates in an incomplete data context which was formalized by Dempster, Laird and Rubin [39] and termed by them the EM *algorithm* (E for "expectation" and M for "maximization"). The EM algorithm for the mixture density estimation problem has been studied by many authors over the last two decades. In fact, there have been a number of independent derivations of the algorithm from at least two quite distinct points of view. It has been found in most instances to have the advantages of reliable global convergence,[1] low cost per iteration, economy of storage and ease of programming, as well as a certain heuristic appeal. On the other hand, it can also exhibit hopelessly slow convergence in some seemingly innocuous applications. All in all, it is undeniably of considerable current interest, and it seems likely to play an important role in the mixture density estimation problem for some time to come.

We feel that the point of view toward the EM algorithm for mixture densities advanced in [39] greatly facilitates both the formulation of a general procedure for prescribing the algorithm and the understanding of the important theoretical properties of the algorithm. Our objectives in the following are to present this point of view in detail in the mixture density context, to unify and extend the diverse results in the literature concerning the derivation and theoretical properties of the EM algorithm, and to review and add to what is known about its practical behavior.

In §4, we interpret the mixture density estimation problem as an incomplete data problem, formulate the general EM algorithm for mixture densities from this point of view, and discuss the general properties of the algorithm. In §5, the focus is narrowed to mixtures of densities from the exponential family, and we summarize and augment the results of investigations of the EM algorithm for such mixtures which have appeared in the literature. Finally, in §6, we discuss the performance of the algorithm in practice through qualitative comparisons with other algorithms and numerical studies in simple but important cases.

2. A review of the literature. The following is a very brief survey of papers which are primarily directed toward some part of the mixture density estimation problem. No attempt has been made to include papers which are strictly concerned with applications of estimation procedures and results developed elsewhere. Little has been said about papers which are of mainly historical interest or peripheral to the subjects of major interest in the sequel. For additional references relating to mixture densities as well as more detailed summaries of the contents of many of the papers touched on below, we refer the reader to the recently published monograph by Everitt and Hand [47] and the recent survey by Titterington [145]. As a convenience, this survey has been divided somewhat arbitrarily

[1]Throughout this paper, we use "global convergence" in the sense of the optimization community, i.e., to mean convergence to a *local* maximizer from *almost any starting point* (cf. Dennis and Schnabel [42, p. 5]).

by topics into four subsections. Not surprisingly, many papers are cited in more than one subsection.

2.1. The method of moments. The first published investigation relating to the mixture density estimation problem appears to be that of Pearson [109]. In that paper, as in Example 1.1, the problem considered is the estimation of the parameters in a mixture of two univariate normal densities. The sample from which the estimates are obtained is assumed to be independent and to consist entirely of unlabeled observations on the mixture. (Since this is the sort of sample dealt with in the vast majority of work on the problem at hand, it is understood in this review that all samples are of this type unless otherwise indicated.) The approach suggested by Pearson for solving the problem is known as the method of moments. The method of moments consists generally of equating some set of sample moments to their expected values and thereby obtaining a system of (generally nonlinear) equations for the parameters in the mixture density. To estimate the five independent parameters in a mixture of two univariate normal densities according to the procedure of [109], one begins with equations determined by the first five moments and, after considerable algebraic manipulation, ultimately arrives at expressions for estimates which depend on a suitably chosen root of a single ninth-degree polynomial.

From the time of the appearance of Pearson's paper until the use of computers became widespread in the 1960's, only fairly simple mixture density estimation problems were studied, and the method of moments was usually the method of choice for their solution. During this period, most energy devoted to mixture problems was directed toward mixtures of normal densities, especially toward Pearson's case of two univariate normal densities. Indeed, most work on normal mixtures during this period was intended either to simplify the job of obtaining Pearson's estimates or to offer more accessible estimates in restricted cases. In this connection, we cite the papers of Charlier [25] (who referred to the implementation of Pearson's method as "an heroic task"), Pearson and Lee [111], Charlier and Wicksell [26], Burrau [19], Strömgren [132], Gottschalk [55], Sittig [129], Weichselberger [152], Preston [116], Cohen [32], Dick and Bowden [43] and Gridgeman [57]. More recently, several investigators have studied the statistical properties of moment estimates for mixtures of two univariate normal densities and have compared their performance with that of other estimation methods. See Robertson and Fryer [125], Fryer and Robertson [50], Tan and Chang [138] and Quandt and Ramsey [118].

Some work has been done extending Pearson's method of moments to more general mixtures of normal densities and to mixtures of other continuous densities. Pollard [115] obtained moment estimates for a mixture of three univariate normal densities, and moment estimates for mixtures of multivariate normal densities were studied by Cooper [34], Day [37] and John [81]. These authors all made simplifying assumptions about the mixtures under consideration in order to reduce the complexity of the moment estimation problem. Gumbel [58] and Rider [123] derived moment estimates for the means in a mixture of two exponential densities under the assumption that the mixing proportions are known. Later, Rider offered moment estimates for mixtures of Weibull distributions in [124]. Moment estimates for mixtures of two gamma densities were treated by John [82]. Tallis and Light [136] explored the relative advantages of fractional moments in the case of mixtures of two exponential densities. Also, Kabir [83] introduced a generalized method of moments and applied it to this case.

Moment estimates for a variety of simple mixtures of discrete densities were derived more or less in parallel with moment estimates for mixtures of normal and other continuous densities. In particular, moment estimates for mixtures of binomial and

Poisson densities were investigated by Pearson [110], Muench [102], [103], Schilling [127], Gumbel [58], Arley and Buch [4], Rider [124], Blischke [12], [14] and Cohen [30]. Kabir [83] also applied his generalized method of moments to a mixture of two binomial densities. For additional information on moment estimation and many other topics of interest for mixtures of discrete densities, see the extensive survey of Blischke [13].

Before leaving the method of moments, we mention the important problem of estimating the proportions alone in a mixture density under the assumption that the component densities, or at least some useful statistics associated with them, are known. Most general mixture density estimation procedures can be brought to bear on this problem, and the manner of applying these general procedures to this problem is usually independent of the particular forms of the densities in the mixture. In addition to the general estimation procedures, a number of special procedures have been developed for this problem; these are discussed in §2.3. of this review. The method of moments has the attractive property for this problem that the moment equations are linear in the mixture proportions. Moment estimates of proportions were discussed by Odell and Basu [104] and Tubbs and Coberly [147].

2.2. The method of maximum likelihood. With the arrival of increasingly powerful computers and increasingly sophisticated numerical methods during the 1960's, investigators began to turn from the method of moments to the method of maximum likelihood as the most widely preferred approach to mixture density estimation problems. To reiterate the working definition given in the introduction, we say that a maximum-likelihood estimate associated with a sample of observations is a choice of parameters which maximizes the probability density function of the sample, called in this context the *likelihood function.* In the next section, we define precisely the maximum-likelihood estimates of interest here and comment on their properties. In this subsection, we offer a very brief tour of the literature addressing maximum-likelihood estimation for mixture densities. Of course, more is said in the sequel about most of the work mentioned below.

Actually, maximum-likelihood estimates and their associated efficiency were often the subject of wishful thinking prior to the advent of computers, and some work was done then toward obtaining maximum-likelihood estimates for simple mixtures. Specifically, we cite a curious paper of Baker [5] and work by Rao [119] and Mendenhall and Hader [101]. In both [119] and [101], iterative procedures were successfully used to obtain approximate solutions of nonlinear equations satisfied by maximum-likelihood estimates. In [119], a system of four equations in four unknown parameters was solved by the method of scoring; in [101], a single equation in one unknown was solved by Newton's method. Both of these methods are described in §6. Despite these successes, the problem of obtaining maximum-likelihood estimates was generally considered during this period to be completely intractable for computational reasons.

As computers became available to ease the burden of computation, maximum-likelihood estimation was proposed and studied in turn for a variety of increasingly complex mixture densities. As before, mixtures of normal densities were the subject of considerable attention. Hasselblad [70] treated maximum-likelihood estimation for mixtures of any number of univariate normal densities; his major results were later obtained independently by Behboodian [9]. Mixtures of two multivariate normal densities with a common unknown covariance matrix were addressed by Day [37] and John [81]. The general case of a mixture of any number of multivariate normal densities was considered by Wolfe [153], and additional work on this case was done by Duda and Hart [46] and Peters and Walker [113]. Redner [121] and Hathaway [72] proposed,

respectively, penalty terms on the likelihood function and constraints on the variables in order to deal with singularities. Tan and Chang [138] compared the moment and maximum-likelihood estimates for a mixture of two univariate normal densities with common variance by computing the asymptotic variances of the estimates. Hosmer [74] reported on a Monte Carlo study of maximum-likelihood estimates for a mixture of two univariate normal densities when the component densities are not well separated and the sample size is small.

Several interesting variations on the usual estimation problem for mixtures of normal densities have been addressed in the literature. Hosmer [75] compared the maximum-likelihood estimates for a mixture of two univariate normal densities obtained from three different types of samples, the first of which is the usual type consisting of only unlabeled observations and the second two of which consist of both labeled and unlabeled observations and are distinguished by whether or not the labeled observations contain information about the mixing proportions. Such partially labeled samples were also considered by Tan and Chang [137] and Dick and Bowden [43]. (We elaborate on the nature of these samples and how they might arise in §3.) The treatment of John [81] is unusual in that the number of observations from each component in the mixture is estimated along with the usual component parameters. Finally, a number of authors have investigated maximum-likelihood estimates for a "switching regression" model which is a certain type of estimation problem for mixtures of normal densities; see the papers of Quandt [117], Hosmer [76], Kiefer [88] and the comments by Hartley [69], Hosmer [77], Kiefer [89] and Kumar, Nicklin and Paulson [91] on the paper of Quandt and Ramsey [118]. A generalization of the model considered by these authors was touched on by Dennis [40].

Maximum-likelihood estimation has also been studied for a variety of unusual and general mixture density problems, some of which include but are not restricted to the usual normal mixture problem. Cohen [31] considered an unusual but simple mixture of two discrete densities, one of which has support at a single point; he focused in particular on the case in which the other density is a negative binomial density. Hasselblad [71] generalized his earlier results in [70] to include mixtures of any number of univariate densities from exponential families. He included a short study comparing maximum-likelihood estimates with the moment estimates of Blischke [14] for a mixture of two binomial distributions. Baum, Petrie, Soules and Weiss [8] addressed a mixture estimation problem which is both unusual and in one respect more general than the problems considered in the sequel. In their problem, the a priori probabilities of sample observations coming from the various component populations in the mixture are not independent from one observation to the next (that is, they are not simply the proportions of the component populations in the mixture), but rather are specified to follow a Markov chain. Their results are specifically applied to mixtures of univariate normal, gamma, binomial and Poisson densities, and to mixtures of general strictly log concave density functions which are identical except for unknown location and scale parameters. John [82] treated maximum-likelihood estimation for a mixture of two gamma distributions in a manner similar to that of [81]. Gupta and Miyawaki [59] considered uniform mixtures. Peters and Coberly [112] and Peters and Walker [114] treated maximum-likelihood estimates of proportions and subsets of proportions for essentially arbitrary mixture densities. Maximum-likelihood estimates were included by Tubbs and Coberly [147] in their study of the sensitivity of various proportion estimators. Other maximum-likelihood estimation problems which are closely related to those considered here are the latent structure problems touched on by Wolfe [153] (see also Lazarsfeld and Henry [93]) and the problems concerning frequency tables derived by indirect observation addressed by

Haberman [62], [63], [64]. Finally, although infinite mixture densities of the general form (1.2) are specifically excluded from consideration here, we mention a very interesting result of Laird [92] to the effect that, under various assumptions, the maximum-likelihood estimate of a possibly infinite mixture density is actually a finite mixture density. A special case of this result was shown earlier by Simar [128] in a broad study of maximum-likelihood estimation for possibly infinite mixtures of Poisson densities.

2.3. Other methods. In addition to the method of moments and the method of maximum-likelihood, a variety of other methods have been proposed for estimating parameters in mixture densities. Some of these methods are general purpose methods. Others are (or were at the time of their derivation) intended for mixture problems the forms of which make (or made) them either ill suited for the application of more widely used methods or particularly well suited for the application of special purpose methods.

For mixtures of any number of univariate normal densities, Harding [67] and Cassie [20] suggested graphical procedures employing probability paper as an alternative to moment estimates, which were at that time practically unobtainable in all but the simplest cases. Later, Bhattacharya [11] prescribed other graphical methods as a particularly simple way of resolving a mixture density into normal components. These graphical procedures work best on mixture populations which are well separated in the sense that each component has an associated region in which the presence of the other components can be ignored. More recently, Fowlkes [49] introduced additional procedures which are intended to be more sensitive to the presence of mixtures than other methods. In the bivariate case, Tarter and Silvers [140] introduced a graphical procedure for decomposing mixtures of any number of normal components.

Also for general mixtures of univariate normal densities, Doetsch [45] exhibited a linear operator which reduces the variances of the component densities without changing their proportions or means and used this operator in a procedure which determines the component densities one at a time. For applications and extensions of this technique to other densities, see Medgyessy [100] (and the review by Mallows [98]), Gregor [56] and Stanat [130]. In [126], Sammon considered a mixture density consisting of an unknown number of component densities which are identical except for translation by unknown location parameters; he derived techniques based on convolution for estimating both the number of components in the mixture and the location parameters.

A number of specialized procedures have been developed for application to the problem of estimating the proportions in a mixture under the assumption that something about the component densities is known. Choi and Bulgren [29] proposed an estimate determined by a least-squares criterion in the spirit of the minimum-distance method of Wolfowitz [154]. A variant of the method of [29] for which smaller bias and mean-square error were reported was offered by Macdonald [96]. A method termed the "confusion matrix method" was given by Odell and Chhikara [105] (see also the review of Odell and Basu [104]). In this method, an estimate is obtained by subdividing R^n into disjoint regions $R_1, \cdots, R_m$ and then solving the equation $P\hat{\alpha} = e$, in which $\hat{\alpha}$ is the estimated vector of proportions, e is a vector whose ith component is the fraction of observations falling in R_i, and the "confusion matrix" P has ijth entry

$$\int_{R_i} p_j(x \mid \phi_j)\,dx.$$

The confusion matrix method is a special case of a method of Macdonald [97], whose formulation of the problem as a least-squares problem allows for a singular or rectangular

confusion matrix. Related methods have been considered recently by Hall [66] and Titterington [146]. Earlier, special cases of estimates of this type were considered by Boes [15], [16]. Guseman and Walton [60], [61] employed certain pattern recognition notions and techniques to obtain numerically tractable confusion matrix proportion estimates for mixtures of multivariate normal densities. James [80] studied several simple confusion matrix proportion estimates for a mixture of two univariate normal densities. Ganesalingam and McLachlan [52] compared the performance of confusion matrix proportion estimates with maximum-likelihood proportion estimates for a mixture of two multivariate normal densities. Another approach to proportion estimation was taken by Anderson [3], who proposed a fairly general method based on direct parametrization and estimation of likelihood ratios. Finally, Walker [151] considered a mixture of two essentially arbitrary multivariate densities and, assuming only that the means of the component densities are known, suggested a simple procedure using linear maps which yields unbiased proportion estimates.

A stochastic approximation algorithm for estimating the parameters in a mixture of any number of univariate normal densities was offered by Young and Coraluppi [160]. In such an algorithm, one determines a sequence of recursively updated estimates from a sequence of observations of indeterminate length considered on a one-at-a-time or few-at-a-time basis. Such an algorithm is likely to be appealing when a sample of desired size is either unavailable in toto at any one point in time or unwieldy because of its size. Stochastic approximation of mixture proportions alone was considered by Kazakos [86].

Quandt and Ramsey [118] derived a procedure called the moment generating function method and applied it to the problem of estimating the parameters in a mixture of two univariate normal densities and in a switching regression model. In brief, a moment generating function estimate is a choice of parameters which minimizes a certain sum of squares of differences between the theoretical and sample moment generating functions. In a comment by Kiefer [89], it is pointed out that the moment generating function method can be regarded as a natural generalization of the method of moments. Kiefer [89] further offers an appealing heuristic explanation of the apparent superiority of moment generating function estimates over moment estimates reported by Quandt and Ramsey [118]. In a comment by Hosmer [77], evidence is presented that moment generating function estimates may in fact perform better than maximum-likelihood estimates in the small-sample case. However, Kumar et al. [91] attribute difficulties to the moment generating function method, and they suggest another method based on the characteristic function.

Minimum chi-square estimation is a general method of estimation which has been touched on by a number of authors in connection with the mixture density estimation problem, but which has not become the subject of much consideration in depth in this context. In minimum chi-square estimation, one subdivides R^n into cells $R_1, \cdots, R_k$ and seeks a choice of parameters which minimizes

$$\chi^2(\Phi) = \sum_{j=1}^{k} \frac{\{N_j - E_j(\Phi)\}^2}{E_j(\Phi)}$$

or some similar criterion function. In this expression, N_j and $E_j(\Phi)$ are, respectively, the observed and expected numbers of observations in R_j for $j = 1, \cdots, k$. For mixtures of normal densities, minimum chi-square estimates were mentioned by Hasselblad [70], Cohen [32], Day [37] and Fryer and Robertson [50]. Minimum chi-square estimates of proportions were reviewed by Odell and Basu [104] and included in the sensitivity study of Tubbs and Coberly [147]. Macdonald [97] remarked that his weighted least-squares

approach to proportion estimation suggested a convenient iterative method for computing minimum chi-square estimates.

As a final note, we mention three methods which have been proposed for general mixture density estimation problems. Choi [28] discussed the extension to general mixture density estimation problems of the least-squares method of Choi and Bulgren [29] for estimating proportions. Deely and Kruse [38] suggested an estimation procedure which is in spirit like that of Choi and Bulgren [29] and Choi [28], except that a sup-norm distance is used in place of the square integral norm. Deely and Kruse argued that their procedure is computationally feasible, but no concrete examples or computation results are given in [38]. Yakowitz [156], [157] outlined a very general "algorithm" for constructing consistent estimates of the parameters in mixture densities which are *identifiable* in the sense described in §2.5. of this review. The sense in which his "algorithm" is really an algorithm in the usually understood meaning of the word is discussed in [157].

2.4. The EM algorithm. At several points in the review, above, we have alluded to computational difficulties associated with obtaining maximum-likelihood estimates. For mixture density problems, these difficulties arise because of the complex dependence of the likelihood function on the parameters to be estimated. The customary way of finding a maximum-likelihood estimate is first to determine a system of equations called the likelihood equations which are satisfied by the maximum-likelihood estimate, and then to attempt to find the maximum-likelihood estimate by solving these likelihood equations. The likelihood equations are usually found by differentiating the logarithm of the likelihood function, setting the derivatives equal to zero, and perhaps performing some additional algebraic manipulations. For mixture density problems, the likelihood equations are almost certain to be nonlinear and beyond hope of solution by analytic means. Consequently, one must resort to seeking an approximate solution via some iterative procedure.

There are, of course, many general iterative procedures which are suitable for finding an approximate solution of the likelihood equations and which have been honed to a high degree of sophistication within the optimization community. We have in mind here principally Newton's method, various quasi-Newton methods which are variants of it, and conjugate gradient methods. In fact, the method of scoring, which was mentioned above in connection with the work of Rao [119] and which we describe in detail in the sequel, falls into the category of Newton-like methods and is one such method which is specifically formulated for solving likelihood equations.

Our main interest here, however, is in a special iterative method which is unrelated to the above methods and which has been applied to a wide variety of mixture problems over the last fifteen or so years. Following the terminology of Dempster, Laird and Rubin [39], we call this method the EM algorithm (E for "expectation" and M for "maximization"). As we mentioned in the introduction, it has been found in most instances to have the advantage of reliable global convergence, low cost per iteration, economy of storage and ease of programming, as well as a certain heuristic appeal; unfortunately, its convergence can be maddeningly slow in simple problems which are often encountered in practice.

In the mixture density context, the EM algorithm has been derived and studied from at least two distinct viewpoints by a number of authors, many of them working independently. Hasselblad [70] obtained the EM algorithm for an arbitrary finite mixture of univariate normal densities and made empirical observations about its behavior. In an extension of [70], he further prescribed the algorithm for essentially arbitrary finite mixtures of univariate densities from exponential families in [71]. The

EM algorithm of [70] for univariate normal mixtures was given again by Behboodian [9], while Day [37] and Wolfe [153] formulated it for, respectively, mixtures of two multivariate normal densities with common covariance matrix and arbitrary finite mixtures of multivariate normal densities. All of these authors apparently obtained the EM algorithm independently, although Wolfe [153] referred to Hasselblad [70]. They all derived the algorithm by setting the partial derivatives of the log-likelihood function equal to zero and, after some algebraic manipulation, obtained equations which suggest the algorithm.

Following these early derivations, the EM algorithm was applied by Tan and Chang [137] to a mixture problem in genetics and used by Hosmer [74] in the Monte Carlo study of maximum likelihood estimates referred to earlier. Duda and Hart [46] cited the EM algorithm for mixtures of multivariate normal densities and commented on its behavior in practice. Hosmer [75] extended the EM algorithm for mixtures of two univariate normal densities to include the partially labeled samples described briefly above. Hartley [69] prescribed the EM algorithm for a "switching regression" model. Peters and Walker [113] offered a local convergence analysis of the EM algorithm for mixtures of multivariate normal densities and suggested modifications of the algorithm to accelerate convergence. Peters and Coberly [112] studied the EM algorithm for approximating maximum-likelihood estimates of the proportions in an essentially arbitrary mixture density and gave a local convergence analysis of the algorithm. Peters and Walker [114] extended the results of [112] to include subsets of mixture proportions and a local convergence analysis along the lines of [113].

All of the above investigators regarded the EM algorithm as arising naturally from the particular forms taken by the partial derivatives of the log-likelihood function. A quite different point of view toward the algorithm was put forth by Dempster, Laird and Rubin [39]. They interpreted the mixture density estimation problem as an estimation problem involving incomplete data by regarding an unlabeled observation on the mixture as "missing" a label indicating its component population of origin. In doing so, they not only related the mixture density problem to a broader class of statistical problems but also showed that the EM algorithm for mixture density problems is really a specialization of a more general algorithm (also called the EM algorithm in [39]) for approximating maximum-likelihood estimates from incomplete data. As one sees in the sequel, this more general EM algorithm is defined in such a way that it has certain desirable theoretical properties by its very definition. Earlier, the EM algorithm was defined independently in a very similar manner by Baum et al. [8] for very general mixture density estimation problems, by Sundberg [135] for incomplete data problems involving exponential families (and specifically including mixture problems), and by Haberman [62], [63], [64] for mixture-related problems involving frequency tables derived by indirect observation. Haberman also refers in [64] to versions of his algorithm developed by Ceppellini, Siniscalco and Smith [21], Chen [27] and Goodman [54]. In addition, an interpretation of mixture problems as incomplete data problems was given in the brief discussion of mixtures by Orchard and Woodbury [106]. The desirable theoretical properties automatically enjoyed by the EM algorithm suggest in turn the good global convergence behavior of the algorithm which has been observed in practice by many investigators. Theorems which essentially confirm this suggested behavior have been recently obtained by Redner [121], Vardi [149], Boyles [18] and Wu [155] and are outlined in the sequel.

2.5. Identifiability and information. To complete this review, we touch on two topics which have to do with the general well-posedness of estimation problems rather than with any particular method of estimation. The first topic, identifiability, addresses the

theoretical question of whether it is possible to uniquely estimate a parameter from a sample, however large. The second topic, information, relates to the practical matter of how good one can reasonably hope for an estimate to be. A thorough survey of these topics is far beyond the scope of this review; we try to cover, below, those aspects of them which have a specific bearing on the sequel.

In general, a parametric family of probability density functions is said to be *identifiable* if distinct parameter values determine distinct members of the family. For families of mixture densities, this general definition requires a special interpretation. For the purposes of this paper, let us first say that a mixture density $p(x \mid \Phi)$ of the form (1.1) is *economically represented* if, for each pair of integers i and j between 1 and m, one has that $p_i(x \mid \phi_i) = p_j(x \mid \phi_j)$ for almost all $x \in R^n$ (relative to the underlying measure on R^n appropriate for $p(x \mid \Phi)$) only if either $i = j$ or one of α_i and α_j is zero. Then it suffices to say that a family of mixture densities of the form (1.1) is identifiable for $\Phi \in \Omega$ if for each pair $\Phi' = (\alpha_1', \cdots, \alpha_m', \phi_1', \cdots, \phi_m')$ and $\Phi'' = (\alpha_1'', \cdots, \alpha_m'', \phi_1'', \cdots, \phi_m'')$ in Ω determining economically represented densities $p(x \mid \Phi')$ and $p(x \mid \Phi'')$, one has that $p(x \mid \Phi') = p(x \mid \Phi'')$ for almost all $x \in R^n$ only if there is a permutation π of $(1, \cdots, m)$ such that $\alpha_i' = \alpha_{\pi(i)}''$ and, if $\alpha_i' \neq 0$, $\phi_i' = \phi_{\pi(i)}''$ for $i = 1, \cdots, m$. For a more general definition suitable for possibly infinite mixture densities of the form (1.2), see, for example, Yakowitz and Spragins [158].

It is tacitly assumed here that all families of mixture densities under consideration are identifiable. One can easily determine the identifiability of specific mixture densities using, for example, the identifiability characterization theorem of Yakowitz and Spragins [158]. For more on identifiability of mixture densities, the reader is referred to the papers of Teicher [141], [142], [143], [144], Barndorff-Nielsen [6], Yakowitz and Spragins [158], and Yakowitz [156], [157], and to the book by Maritz [99].

The Fisher information matrix is given by

$$I(\Phi) = \int_{R^n} [\nabla_\Phi \log p(x \mid \Phi)][\nabla_\Phi \log p(x \mid \Phi)]^T p(x \mid \Phi) d\mu, \tag{2.5.1}$$

provided that $p(x \mid \Phi)$ is such that this expression is well defined. (In writing ∇_Φ, we suppose that one can conveniently redefine Φ as a vector $\Phi = (\xi_1, \cdots, \xi_v)^T$ of unconstrained scalar parameters, and we take $\nabla_\Phi = (\partial/\partial\xi_1, \cdots, \partial/\partial\xi_v)^T$. Also, in (2.5.1), μ denotes the underlying measure on R^n appropriate for $p(x \mid \Phi)$.) The Fisher information matrix has general significance concerning the distribution of unbiased and asymptotically unbiased estimates. For the present purposes, the importance of the Fisher information matrix lies in its role in determining the asymptotic distribution of maximum-likelihood estimates (see Theorem 3.1 below).

A number of authors have considered the Fisher information matrix for finite mixture densities in a variety of contexts. We mention in particular several investigations in which the Fisher information matrix is of central interest. (There have been others in which the Fisher information matrix or some approximation of it has played a significant but less prominent role; see those of Mendenhall and Hader [101], Hasselblad [70], [71], Day [37], Wolfe [153], Dick and Bowden [43], Hosmer [74], James [80] and Ganesalingam and McLachlan [52].) Hill [73] exploited simple approximations obtained in limiting cases from a general power series expansion to investigate the Fisher information for estimating the proportion in a mixture of two normal or exponential densities. Behboodian [10] offered methods for computing the Fisher information matrix for the proportion, means and variances in a mixture of two univariate normal densities; he also provided four-place tables from which approximate information matrices for a variety of parameter values can be easily obtained. In their comparison of moment and maximum-

likelihood estimates, Tan and Chang [138] numerically evaluated the diagonal elements of the inverse of the Fisher information matrix at a variety of parameter values for a mixture of two univariate normal densities with a common variance. Using the Fisher information matrix, Chang [23] investigated the effects of adding a second variable on the asymptotic distribution of the maximum-likelihood estimates of the proportion and parameters associated with the first variable in a mixture of two normal densities. Later, Chang [24] extended the methods of [23] to include mixtures of two normal densities on variables of arbitrary dimension. For a mixture of two univariate normal densities, Hosmer and Dick [78] considered Fisher information matrices determined by a number of sample types. They compared the asymptotic relative efficiencies of estimates from totally unlabeled samples, estimates from two types of partially labeled samples, and estimates from two types of completely labeled samples.

3. Maximum likelihood. In this section, maximum-likelihood estimates for mixture densities are defined precisely, and their important properties are discussed. It is assumed that a parametric family of mixture densities of the form (1.1) is specified and that a particular $\Phi^* = (\alpha_1^*, \cdots, \alpha_m^*, \phi_1^*, \cdots, \phi_m^*) \in \Omega$ is the "true" parameter value to be estimated. As before, it is both natural and convenient to regard $p(x \mid \Phi)$ in (1.1) as modeling a statistical population which is a mixture of m component populations with associated component densities $\{p_i\}_{i=1,\cdots,m}$ and mixing proportions $\{\alpha_i\}_{i=1,\cdots,m}$.

In order to suggest to the reader the variety of samples which might arise in mixture problems, as well as to provide a framework within which to discuss samples of interest in the sequel, we introduce samples of observations in R^n of four distinct types. All of the mixture density estimation problems which we have encountered in the literature involve samples which are expressible as one or a stochastically independent union of samples of these types, although the imaginative reader can probably think of samples for mixture problems which cannot be so represented. The four types of samples and the notation which we associate with them are given as follows:

Type 1. Suppose that $\{x_k\}_{k=1,\cdots,N}$ is an independent sample of N unlabeled observations on the mixture, i.e., a set of N observations on independent, identically distributed random variables with density $p(x \mid \Phi^*)$. Then $S_1 = \{x_k\}_{k=1,\cdots,N}$ is a sample of Type 1.

Type 2. Suppose that $J_1, \cdots, J_m$ are arbitrary nonnegative integers and that, for $i = 1, \cdots, m$, $\{y_{ik}\}_{k=1,\cdots,J_i}$ is an independent sample of observations on the ith component population, i.e., a set of J_i observations on independent, identically distributed random variables with density $p_i(x \mid \phi_i^*)$. Then $S_2 = \bigcup_{i=1}^m \{y_{ik}\}_{k=1,\cdots,J_i}$ is a sample of Type 2.

Type 3. Suppose that an independent sample of K unlabeled observations is drawn on the mixture, that these observations are subsequently labeled, and that, for $i = 1, \cdots, m$, a set $\{z_{ik}\}_{k=1,\cdots,K_i}$ of them is associated with the ith component population with $K = \sum_{i=1}^m K_i$. Then $S_3 = \bigcup_{i=1}^m \{z_{ik}\}_{k=1,\cdots,K_i}$ is a sample of Type 3.

Type 4. Suppose that an independent sample of M unlabeled observations is drawn on the mixture, that the observations in the sample which fall in some set $E \subseteq R^n$ are subsequently labeled, and that, for $i = 1, \cdots, m$, a set $\{w_{ik}\}_{k=1,\cdots,M_i}$ of them is thereby associated with the ith component population while a set $\{w_{0k}\}_{k=1,\cdots,M_0}$ remains unlabeled. Then $S_4 = \bigcup_{i=0}^m \{w_{ik}\}_{k=1,\cdots,M_i}$ is a sample of Type 4.

A totally unlabeled sample S_1 of Type 1 is the sort of sample considered in almost all of the literature on mixture densities. Throughout most of the sequel, it is assumed as a convenience that samples under consideration are of this type. The major qualitative difference between completely labeled samples S_2 and S_3 of Types 2 and 3, respectively, is that the numbers K_i contain information about the mixing proportions while the numbers J_i do not. Thus if estimation of proportions is of interest, then a sample S_2 is useful only as

a subset of a larger sample which includes samples of other types. For mixtures of two univariate densities, Hosmer [75] considered samples of the forms S_1, $S_1 \cup S_2$, and $S_1 \cup S_3$. Previously, Tan and Chang [137] considered a problem involving an application of mixtures in explaining genetic variation which is almost identical to that of [75] in which the sample is of the form $S_1 \cup S_2$. Also Dick and Bowden [43] used a sample of the form $S_1 \cup S_2$ in which $m = 2$ and $J_2 = 0$. Hosmer and Dick [78] evaluated the Fisher information matrix for a variety of samples of Types 1, 2, and 3, and their unions.

A sample S_4 of Type 4 is likely to be associated with a mixture problem involving censored sampling. While the numbers M_i contain information about the mixing proportions, as do the numbers K_i of a sample S_3 of Type 3, they also contain information about the parameters of the component densities while the numbers K_i do not. An interesting and informative example of how a sample of Type 4 might arise is the following, which is in the area of life testing and is outlined by Mendenhall and Hader [101].

Example 3.1. In life testing, one is interested in testing "products" (systems, devices, etc.), recording failure times or causes, and hopefully thereby being better able to understand and improve the performance of the product. It often happens that products of a particular type fail as a result of two or more distinct causes. (An example of Acheson and McElwee [1] is quoted in [101] in which the causes of electronic tube failure are divided into gaseous defects, mechanical defects and normal deterioration of the cathode.) It is therefore natural to regard collections of such products as mixture populations, the component populations of which correspond to the distinct causes of failure. The first objective of life testing in such cases is likely to be estimation of the proportions and other statistical parameters associated with the failure component populations.

Because of restrictions on time available for testing, life testing experiments must often be concluded after a predetermined length of time has elapsed or after a predetermined number of product units have failed, resulting in censored sampling. If the causes of failure of the failed products are determined in the course of such an experiment, then the (labeled) failed products together with those (unlabeled) products which did not fail constitute a sample of Type 4.

The *likelihood function* of a sample of observations is the probability density function of the random sample evaluated at the observations at hand. When maximum-likelihood estimates are of interest, it is usually convenient to deal with the logarithm of the likelihood function, called the *log-likelihood function,* rather than with the likelihood function itself. The following are the log-likelihood functions L_1, L_2, L_3 and L_4 of samples S_1, S_2, S_3 and S_4 of Types 1, 2, 3, and 4 respectively:

$$L_1(\Phi) = \sum_{k=1}^{N} \log p(x_k \mid \Phi), \tag{3.1}$$

$$L_2(\Phi) = \sum_{i=1}^{m} \sum_{k=1}^{J_i} \log p_i(y_{ik} \mid \phi_i), \tag{3.2}$$

$$L_3(\Phi) = \sum_{i=1}^{m} \sum_{k=1}^{K_i} \log [\alpha_i p_i(z_{ik} \mid \phi_i)] + \log \frac{K!}{K_1! \cdots K_m!}, \tag{3.3}$$

$$L_4(\Phi) = \sum_{k=1}^{M_0} \log p(w_{0k} \mid \Phi) + \sum_{i=1}^{m} \sum_{k=1}^{M_i} \log [\alpha_i p_i(w_{ik} \mid \phi_i)] + \log \frac{M!}{M_0! \cdots M_m!}. \tag{3.4}$$

Note that if a sample of observations is a union of independent samples of the types considered here, then the log-likelihood function of the sample is just the corresponding sum of log-likelihood functions defined above for the samples in the union.

If S is a sample of observations of the sort under consideration, then by a *maximum-likelihood estimate* of Φ^*, we mean any choice of Φ in Ω at which the log-likelihood function of S, denoted by $L(\Phi)$, attains its largest local maximum in Ω. In defining a maximum-likelihood estimate in this way, we have taken into account two practical difficulties associated with maximum-likelihood estimation for mixture densities.

The first difficulty is that one cannot always in good conscience take Ω to be a set in which the log-likelihood function is bounded above, and so there are not always points in Ω at which L attains a global maximum over Ω. Perhaps the most notorious mixture problem for which L is not bounded above in Ω is that in which p is a mixture of normal densities and $S = S_1$, a sample of Type 1. It is easily seen in this case that if one of the mixture means coincides with a sample observation and if the corresponding variance tends to zero (or if the corresponding covariance matrix tends in certain ways to a singular matrix in the multivariate case), then the log-likelihood function increases without bound. This was perhaps first observed by Kiefer and Wolfowitz [87], who offered an example involving a mixture of two univariate normal densities to show that classically defined maximum-likelihood estimates, i.e., global maximizers of the likelihood function, need not exist. They also considered certain "modified" and "neighborhood" maximum-likelihood estimates in [87] and pointed out that in their example the former cannot exist and the latter are not consistent. For the normal mixture problem, an advantage of including labeled observations in a sample is that, with probability one, this difficulty does not occur if the sample includes more than n labeled observations from each component population. This was observed in the univariate case by Hosmer [75]. Other methods of circumventing this difficulty in the normal mixture case include the imposition of constraints on the variables (Hathaway [72]) or penalty terms on the log-likelihood function (Redner [121]).

The second difficulty is that mixture problems are very often such that the log-likelihood function attains its largest local maximum at several different choices of Φ. Indeed, if p_i and p_j are of the same parametric family for some i and j and if $S = S_1$, a sample of Type 1, then the value of $L(\Phi)$ will not change if the component pairs (α_i, ϕ_i) and (α_j, ϕ_j) are interchanged in Φ, i.e., if in effect there is "label switching" of the ith and jth component populations. The results reviewed below show that whether or not such "label switching" is a cause for concern depends on whether estimates of the particular component density parameters are of interest, or whether only an approximation of the mixture density is desired. We remark that this "label switching" difficulty can certainly occur in mixtures which are identifiable (see §2.5).

It should be mentioned that there is a problem relating to this second difficulty which can be quite difficult to overcome in practice. The problem is that the log-likelihood function can (and often does) have local maxima which are not *largest* local maxima. Thus one must in practice decide whether to accept a given local maximum as the largest or to go on to search for others. At the present time, there is little that can be done about this problem, other than to find in one way or another as many local maxima as is feasible and to compare them to find the largest. There is currently no adequately efficient and reliable way of systematically determining all local optima of a general function, although work has been done toward this end and some algorithms have been developed which have proved useful for some problems. For more information, see Dixon and Szegö [44].

In the remainder of this section, our interest is in the important general qualitative properties of maximum-likelihood estimates of mixture density parameters. For convenience, we restrict the discussion to the case which is most often addressed in the

literature, namely, that in which the sample S at hand is a sample S_1 of Type 1. We also assume that each component density p_i is differentiable with respect to ϕ_i and make the nonessential assumption that the parameters ϕ_i are unconstrained in Ω_i and mutually independent variables. It is not difficult to modify the discussion below to obtain similar statements which are appropriate for other mixture density estimation problems of interest. For a discussion of the properties of maximum-likelihood estimates of constrained variables, see the papers of Aitchison and Silvey [2] and Hathaway [72].

The traditional general approach to determining a maximum-likelihood estimate is first to arrive at a system of *likelihood equations* satisfied by the maximum-likelihood estimate and then to try to obtain a maximum-likelihood estimate by solving the likelihood equations. Basically, the likelihood equations are found by considering the partial derivatives of the log-likelihood function with respect to the components of Φ. If $\hat{\Phi} = (\hat{\alpha}_1, \cdots, \hat{\alpha}_m, \hat{\phi}_1, \cdots, \hat{\phi}_m)$ is a maximum-likelihood estimate, then one has the likelihood equations

$$\nabla_{\phi_i} L(\hat{\Phi}) = 0, \qquad i = 1, \cdots, m, \tag{3.5}$$

determined by the unconstrained parameters ϕ_i, $i = 1, \cdots, m$. (Our convention is that "∇" with a variable appearing as a subscript indicates the gradient of first partial derivatives with respect to the components of the variable.)

To obtain likelihood equations determined by the proportions, which are constrained to be nonegative and to sum to one, we follow Peters and Walker [114]. Setting $\hat{\alpha} = (\hat{\alpha}_1, \cdots, \hat{\alpha}_m)^T$, one sees that

$$0 \geqq \nabla_\alpha L(\hat{\Phi})^T(\alpha - \hat{\alpha}) \tag{3.6}$$

for all $\alpha = (\alpha_1, \cdots, \alpha_m)^T$ such that $\sum_{i=1}^m \alpha_i = 1$ and $\alpha_i \geqq 0$, $i = 1, \cdots, m$. Now (3.6) holds for all α satisfying the given constraints if and only if

$$0 \geqq \nabla_\alpha L(\hat{\Phi})^T(e_i - \hat{\alpha}), \qquad i = 1, \cdots, m,$$

with equality for those values of i for which $\hat{\alpha}_i > 0$. (Here, e_i is the vector the ith component of which is one and the other components of which are zero.) It follows that (3.6) is equivalent to

$$1 \geqq \frac{1}{N} \sum_{k=1}^{N} \frac{p_i(x_k \mid \hat{\phi}_i)}{p(x_k \mid \hat{\Phi})}, \qquad i = 1, \cdots, m, \tag{3.7}$$

with equality for those values of i for which $\hat{\alpha}_i > 0$. Finally, multiplying each side of (3.7) by $\hat{\alpha}_i$ for $i = 1, \cdots, m$ yields likelihood equations in the convenient form

$$\hat{\alpha}_i = \frac{1}{N} \sum_{k=1}^{N} \frac{\hat{\alpha}_i p_i(x_k \mid \hat{\phi}_i)}{p(x_k \mid \hat{\Phi})}, \qquad i = 1, \cdots, m. \tag{3.8}$$

We remark that it is easily seen by considering the matrix of second partial derivatives of L with respect to $\alpha_1, \cdots, \alpha_m$ that L is a concave function of $\alpha = (\alpha_1, \cdots, \alpha_m)^T$ for *any* fixed set of values $\hat{\phi}_i \in \Omega_i$, $i = 1, \cdots, m$. Thus, for any fixed $\hat{\phi}_i$, $i = 1, \cdots, m$, (3.6) and, hence, (3.7) are sufficient as well as necessary for $\hat{\alpha}$ to maximize L over the set of all α satisfying the given constraints. On the other hand, the likelihood equations (3.8) are necessary but not sufficient conditions for $\hat{\alpha}$ to maximize L for fixed $\hat{\phi}_i$, $i = 1, \cdots, m$. Indeed, $\hat{\alpha} = e_i$ satisfies (3.8) for $i = 1, \cdots, m$. In fact, it follows from the concavity of L that there is a solution of (3.8) in each (closed) face of the simplex of points α satisfying the given constraints. In spite of perhaps suffering from a

surplus of solutions, the likelihood equations (3.8) nevertheless have a useful form which takes on additional significance later in the context of the EM algorithm.

The equations (3.5) and (3.8) together constitute a full set of likelihood equations which are necessary but not sufficient conditions for a maximum likelihood estimate. Of course, some irrelevant solutions of the likelihood equations can be avoided in practice by using one of a number of procedures for obtaining a numerical solution of them (among which is the EM algorithm) which in all but the most unfortunate circumstances will yield a local maximizer of the log-likelihood function (or a singularity near which it grows without bound) rather than some stationary point which is a local minimizer or a saddle point. Still, it is natural to ask at this point the extent to which solving the likelihood equations can be expected to produce a maximum-likelihood estimate and the extent to which a maximum-likelihood estimate can be expected to be a good approximation of Φ^*.

Two general theorems are offered below which give a fair summary of the results in the literature most pertinent to the question put forth above. As a convenience, we assume that $\alpha_i^* > 0$ for $i = 1, \cdots, m$. For the purposes of the theorems and the discussion following them, this justifies writing, say, $\alpha_m = 1 - \sum_{i=1}^{m-1} \alpha_i$ and considering the redefined, locally unconstrained variable $\Phi = (\alpha_1, \cdots, \alpha_{m-1}, \phi_1, \cdots, \phi_m)$ in the modified set

$$\Omega = \left\{(\alpha_1, \cdots, \alpha_{m-1}, \phi_1, \cdots, \phi_m): \sum_{i=1}^{m-1} \alpha_i < 1 \text{ and } \alpha_i > 0, \phi_i \in \Omega_i \text{ for } i = 1, \cdots, m\right\}.$$

The likelihood equations (3.5) and (3.8) can now be written in the general unconstrained form

$$\nabla_\Phi L(\hat{\Phi}) = 0, \tag{3.9}$$

which facilitates our presenting the theorems as general results which are not restricted to the mixture problem at hand or, for that matter, to mixture problems at all. In our discussion of the theorems, all statements regarding measure and integration are made with respect to the underlying measure on R^n appropriate for $p(x \mid \Phi)$, which we denote by μ.

The first theorem states roughly that, under reasonable assumptions, there is a unique strongly consistent solution of the likelihood equations (3.9), and this solution at least locally maximizes the log-likelihood function and is asymptotically normally distributed. (*Consistent* in the usual sense means converging with probability approaching 1 to the true parameters as the sample size approaches infinity; *strongly consistent* means having the same limit with probability 1.) This theorem is a compendium of results generalizing the initial work of Cramér [36] concerning existence, consistency and asymptotic normality of the maximum-likelihood estimate of a single scalar parameter. The conditions below, on which the theorem rests, were essentially given by Chanda [22] as multidimensional generalizations of those of Cramér. With them, Chanda claimed that there exists a unique solution of the likelihood equations which is consistent in the usual sense (this fact was correctly proved by Tarone and Gruenhage [139]) and established its asymptotic normal behavior. (See also the summary in Kiefer [88], the discussion in Zacks [161], Sundberg's [134] treatment in the case of incomplete data from an exponential family, and the related material for constrained maximum-likelihood estimates in Aitchison and Silvey [2] and Hathaway [72].) Using these same conditions, Peters and Walker [113, Appendix A] showed that there is a unique strongly consistent

solution of the likelihood equations, and that it at least locally maximizes the log-likelihood function.

In stating the following conditions and in the discussion after the theorem, it is convenient to adopt temporarily the notation $\Phi = (\xi_1, \cdots, \xi_v)$, where $v = (m - 1 + \sum_{i=1}^{m} n_i)$ and $\xi_i \in R^1$ for $i = 1, \cdots, v$. Also, we remark that because the results of the theorem below implied by these conditions are strictly local in nature, there is no loss of generality in restricting Ω to be any neighborhood of Φ^* if such a restriction is necessary for the first condition to be met.

Condition 1. For all $\Phi \in \Omega$, for almost all $x \in R^n$ and for $i, j, k = 1, \cdots, v$, the partial derivatives $\partial p/\partial \xi_i$, $\partial^2 p/\partial \xi_i \partial \xi_j$ and $\partial^3 p/\partial \xi_i \partial \xi_j \partial \xi_k$ exist and satisfy

$$\left| \frac{\partial p(x \mid \Phi)}{\partial \xi_i} \right| \leqq f_i(x), \quad \left| \frac{\partial^2 p(x \mid \Phi)}{\partial \xi_i \partial \xi_j} \right| \leqq f_{ij}(x), \quad \left| \frac{\partial^3 \log p(x \mid \Phi)}{\partial \xi_i \partial \xi_j \partial \xi_k} \right| \leqq f_{ijk}(x),$$

where f_i and f_{ij} are integrable and f_{ijk} satisfies

$$\int_{R^n} f_{ijk}(x) p(x \mid \Phi^*) d\mu < \infty.$$

Condition 2. The Fisher information matrix $I(\Phi)$ given by (2.5.1) is well defined and positive definite at Φ^*.

THEOREM 3.1. *If Conditions* 1 *and* 2 *are satisfied and any sufficiently small neighborhood of* Φ^* *in* Ω *is given, then with probability* 1, *there is for sufficiently large* N *a unique solution* Φ^N *of the likelihood equations* (3.9) *in that neighborhood, and this solution locally maximizes the log-likelihood function. Furthermore,* $\sqrt{N}(\Phi^N - \Phi^*)$ *is asymptotically normally distributed with mean zero and covariance matrix* $I(\Phi^*)^{-1}$.

The second theorem is directed toward two questions left unresolved by the theorem above regarding Φ^N, the unique strongly consistent solution of the likelihood equations. The first question is whether Φ^N is really a maximum-likelihood estimate, i.e., a point at which the log-likelihood function attains its *largest* local maximum. The second is whether, even if the answer to the first question is "yes," there are maximum-likelihood estimates other than Φ^N which lead to limiting densities other than $p(x \mid \Phi^*)$. Given our assumption of identifiability of the family of mixture densities $p(x \mid \Phi)$, $\Phi \in \Omega$, one easily sees that the theorem below implies that if Ω' is any compact subset of Ω which contains Φ^* in its interior, then with probability 1, Φ^N is a maximum-likelihood estimate in Ω' for sufficiently large N. Furthermore, every other maximum-likelihood estimate in Ω' is obtained from Φ^N by the "label switching" described earlier and, hence, leads to the same limiting density $p(x \mid \Phi^*)$. Accordingly, we usually assume in the sequel that Conditions 1 through 4 are satisfied and refer to Φ^N as the unique strongly consistent maximum-likelihood estimate. The theorem is a slightly restricted version of a general result of Redner [122] which extends earlier work by Wald [150] on the consistency of maximum-likelihood estimates. It should be remarked that the result of [122] rests on somewhat weaker assumptions than those made here and is specifically aimed at families of distributions which are not identifiable.

For $\Phi \in \Omega$ and sufficiently small $r > 0$, let $N_r(\Phi)$ denote the closed ball of radius r about Φ in Ω and define

$$p(x \mid \Phi, r) = \sup_{\Phi' \in N_r(\Phi)} p(x \mid \Phi')$$

and

$$p^*(x \mid \Phi, r) = \max\{1, p(x \mid \Phi, r)\}.$$

Condition 3. For each $\Phi \in \Omega$ and sufficiently small $r > 0$,

$$\int_{R^n} \log p^*(x \mid \Phi^*, r) p(x \mid \Phi^*)\, d\mu < \infty.$$

Condition 4.

$$\int_{R^n} \log p(x \mid \Phi^*) p(x \mid \Phi^*)\, d\mu < \infty.$$

THEOREM 3.2. *Let Ω' be any compact subset of Ω which contains Φ^* in its interior, and set*

$$C = \{\Phi \in \Omega' : p(x \mid \Phi) = p(x \mid \Phi^*) \text{ almost everywhere}\}.$$

If Conditions 3 *and* 4 *are satisfied and D is any closed subset of Ω' not intersecting C, then with probability* 1,

$$\limsup_{N \to \infty} \sup_{\Phi \in D} \frac{\prod_{k=1}^{N} p(x_k \mid \Phi)}{\prod_{k=1}^{N} p(x_k \mid \Phi^*)} = 0.$$

From a theoretical point of view, Theorems 3.1 and 3.2 are adequate for mixture density estimation problems in providing assurance of the existence of strongly consistent maximum-likelihood estimates, characterizing them as solutions of the likelihood equations and prescribing their asymptotic behavior. In practice, however, one must still contend with certain potential mathematical, statistical and even numerical difficulties associated with maximum-likelihood estimates. Some possible mathematical problems have been suggested above: The log-likelihood function may have many local and global maxima and perhaps even singularities; furthermore, the likelihood equations are likely to have solutions which are not local maxima of the log-likelihood function. According to Theorem 3.1, the statistical soundness (as measured by bias and variance) of the strongly consistent maximum-likelihood estimate is determined, at least for large samples, by the Fisher information matrix $I(\Phi^*)$. As it happens, $I(\Phi^*)$ also plays a role in determining the numerical well-posedness of the problem of approximating the strongly consistent maximum-likelihood estimate for large samples.

To show how $I(\Phi^*)$ enters into the problem of numerically approximating Φ^N for large samples, we recall that the *condition* of a problem refers to the nonuniformity of response of its solution to perturbations in the data associated with the problem. This is to say that a problem is *ill-conditioned* if its solution is very sensitive to some perturbations in the data while being relatively insensitive to others. Another way of looking at it is to say that a problem is ill-conditioned if some relatively widely differing approximate solutions give about the same fit to the data, which may be very good. For maximum-likelihood estimation, then, one can associate ill-conditioning with the log-likelihood function's having a "narrow ridge" below which the maximum-likelihood estimate lies, a phenomenon often observed in the mixture density case. If a problem is ill-conditioned, then the accuracy to which its solution can be computed is likely to be limited, sometimes sharply so. Indeed, in some contexts one can demonstrate this rigorously through backward error analysis in which it is shown that the computed solution of a problem is the exact solution of a "nearby" problem with perturbed data.

For an optimization problem, the condition is customarily measured by the *condition number* of the Hessian matrix of the function to be optimized, evaluated at the solution. Denoting the Hessian by $H(\Phi)$ and assuming that it is invertible, we recall that its condition number is defined to be $\|H(\Phi)\| \cdot \|H(\Phi)^{-1}\|$, where $\|\cdot\|$ is a matrix norm of interest. (For more on the condition number of a matrix, see, for example, Stewart [131].) For the log-likelihood function at hand, the Hessian is given by

$$H(\Phi) = \sum_{k=1}^{N} \nabla_\Phi \nabla_\Phi^T \log p(x_k \mid \Phi), \tag{3.10}$$

where $\nabla_\Phi \nabla_\Phi^T = (\partial^2/\partial\xi_i \partial\xi_j)$. If Conditions 1 and 2 above are satisfied, then it follows from the strong law of large numbers (see Loéve [94]) that, with probability 1,

$$\lim_{N\to\infty} \frac{1}{N} H(\Phi^N) = -I(\Phi^*). \tag{3.11}$$

Since $\frac{1}{N}H(\Phi^N)$ has the same condition number as $H(\Phi^N)$, (3.11) demonstrates that the condition number of $I(\Phi^*)$ reflects the limits of accuracy which one can expect in computed approximations of Φ^N for large N.

To illustrate the potential severity of the statistical and numerical problems associated with maximum-likelihood estimates, we augment the material on the Fisher information matrix in the literature cited in §2.5 with Table 3.3 below, which lists approximate values of the condition number and the diagonal elements of the inverse of $I(\Phi^*)$ for a mixture of two univariate normal densities (see (1.3) and (1.4)) at a variety of choices of Φ^*. To prepare this table, we took $\Phi = (\alpha_1, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$ and numerically evaluated $I(\Phi^*)$, its condition number, and its inverse for selected values of Φ^* using IMSL Library routines DCADRE, EIGRS, and LINV2P on a CDC7600.[2] The choices of Φ^* were obtained by taking $\alpha_1^* = .3$ and $\sigma_1^{2*} = \sigma_2^{2*} = 1$ and varying the mean separation $\mu_1^* - \mu_2^*$. In the table, the condition number of $I(\Phi^*)$ is denoted by κ, and the first through fifth diagonal elements of $I(\Phi^*)^{-1}$ are denoted by $I^{-1}(\alpha_1)$, $I^{-1}(\mu_1)$, $I^{-1}(\mu_2)$, $I^{-1}(\sigma_1^2)$ and $I^{-1}(\sigma_2^2)$, respectively.

Table 3.3 reinforces one's intuitive understanding that, for mixture density estimation problems, maximum-likelihood estimates are more appealing, from both statistical and numerical standpoints, if the component densities in the mixture are well separated than if they are poorly separated. Perhaps the most troublesome implication of Table 3.3 is that, if the component densities are poorly separated, then impractically large sample sizes might be required in order to expect even moderately precise maximum-likelihood estimates. For example, Table 3.3 indicates that if one considers data from a mixture of two univariate normal densities with $\alpha_1^* = .3$, $\sigma_1^{2*} = \sigma_2^{2*} = 1$ and $\mu_1^* - \mu_2^* = 1$, then a sample size on the order of 10^6 is necessary to insure that the standard deviation of each component of the maximum-likelihood estimate is about 0.1 or less. Even if a sample of such horrendous size were available, the fact that evaluating the log-likelihood function and associated functions such as its derivatives involves summation over observations in the sample, considered together with the condition number of 5.18×10^4 for the information matrix, suggests that the computing undertaken in seeking a maximum-likelihood estimate should be carried out with great care.

Similar observations regarding the asymptotic dependence of the accuracy of maximum-likelihood estimates on sample sizes and separation of the component populations have been made by a number of authors (Mendenhall and Hader [101], Hill [73],

[2]We are grateful to the Mathematics and Statistics Division of the Lawrence Livermore National Laboratory for allowing us to use their computing facility in generating Table 3.3.

TABLE 3.3

Condition number and diagonal elements of the inverse of $I(\Phi^*)$ for a mixture of two univariate normal densities with $\alpha_1^* = .3, \sigma_1^{2*} = \sigma_2^2 = 1$.

$\mu_1^* - \mu_2^*$	κ	$I^{-1}(\alpha_1)$	$I^{-1}(\mu_1)$	$I^{-1}(\mu_2)$	$I^{-1}(\sigma_1^2)$	$I^{-1}(\sigma_2^2)$
0.2	3.06×10^{10}	4.39×10^{10}	4.86×10^{9}	8.98×10^{8}	2.15×10^{7}	4.02×10^{6}
0.5	8.05×10^{6}	5.54×10^{6}	3.81×10^{6}	7.17×10^{5}	1.04×10^{5}	2.07×10^{4}
1.0	5.18×10^{4}	8.59×10^{3}	2.32×10^{4}	4.55×10^{3}	2.58×10^{3}	578.
1.5	4.80×10^{3}	237.	1.43×10^{3}	290.	383.	95.0
2.0	1.10×10^{3}	20.4	216.	45.8	115.	31.3
3.0	187.	.874	18.9	4.81	28.2	8.83
4.0	71.7	.267	5.72	1.95	13.4	4.71
6.0	35.7	.211	3.44	1.45	7.47	3.06

Hasselblad [70], [71], Day [37], Tan and Chang [138], Dick and Bowden [43], Hosmer [74], [75], Hosmer and Dick [78]). Several of them (Mendenhall and Hader [101], Day [37], Hasselblad [71], Dick and Bowden [43], Hosmer [74]) also suggested that things are worse for small samples (less than a few hundred observations) than the asymptotic theory indicates. Hosmer [74] specifically addressed the small-sample, poor-separation case for a mixture of two univariate normals and concluded that in this case maximum-likelihood estimates "should be used with extreme caution or not at all." Dick and Bowden [43], Hosmer [75] and Hosmer and Dick [78] offered evidence which suggests that considerable improvement in the performance of maximum-likelihood estimates can result from including labeled observations in the samples by which the estimates are determined, particularly when the component densities are poorly separated. In fact, it is pointed out in [78] that most of the improvement occurs for small to moderate proportions of labeled observations in the sample.

In spite of the rather pessimistic comments above, maximum-likelihood estimates have fared well in comparisons with most other estimates for mixture density estimation problems. Day [37], Hasselblad [71], Tan and Chang [138] and Dick and Bowden [43] found maximum-likelihood estimates to be markedly superior to moment estimates in their investigations, especially in cases involving poorly separated component populations. (See also the comment by Hosmer [77] on the paper of Quandt and Ramsey [118].) Day [37] also remarked that minimum chi-square and Bayes estimates have less appeal than maximum-likelihood estimates, primarily because of the difficulty of obtaining them in most cases. James [80] and Ganesalingam and McLachlan [52] observed that their proportion estimates are less efficient than maximum-likelihood estimates; however, they also outlined circumstances in which their estimates might be preferred. On the other hand, as we remarked in §2.3, Hosmer [77] has commented that the moment generating function method of Quandt and Ramsey [118] provides estimates which may outperform maximum-likelihood estimates in the small-sample case, although Kumar et al. [91] attribute certain difficulties to this method.

4. The EM algorithm. We now derive the EM algorithm for general mixture density estimation problems and discuss its important general properties. As stated in the introduction, we feel that the EM algorithm for mixture density estimation problems is

best regarded as a specialization of the general EM algorithm formalized by Dempster, Laird and Rubin [39] for obtaining maximum-likelihood estimates from incomplete data. Accordingly, we begin by reviewing the formulation of the general EM algorithm given in [39].

Suppose that one has a measure space $\mathcal{Y}$ of "complete data" and a measurable map $\mathbf{y} \to \mathbf{x}(\mathbf{y})$ of $\mathcal{Y}$ to a measure space $\mathcal{X}$ of "incomplete data." Let $f(\mathbf{y} \mid \Phi)$ be a member of a parametric family of probability density functions defined on $\mathcal{Y}$ for $\Phi \in \Omega$, and suppose that $g(\mathbf{x} \mid \Phi)$ is a probability density function on $\mathcal{X}$ induced by $f(\mathbf{y} \mid \Phi)$. For a given $\mathbf{x} \in \mathcal{X}$, the purpose of the EM algorithm is to maximize the incomplete data log-likelihood $L(\Phi) = \log g(\mathbf{x} \mid \Phi)$ over $\Phi \in \Omega$ by exploiting the relationship between $f(\mathbf{y} \mid \Phi)$ and $g(\mathbf{x} \mid \Phi)$. It is intended especially for applications in which the maximization of the complete data log-likelihood $\log f(\mathbf{y} \mid \Phi)$ over $\Phi \in \Omega$ is particularly easy.

For $\mathbf{x} \in \mathcal{X}$, set $\mathcal{Y}(\mathbf{x}) = \{\mathbf{y} \in \mathcal{Y}: \mathbf{x}(\mathbf{y}) = \mathbf{x}\}$. The conditional density $k(\mathbf{y} \mid \mathbf{x}, \Phi)$ on $\mathcal{Y}(\mathbf{x})$ is given by $f(\mathbf{y} \mid \Phi) = k(\mathbf{y} \mid \mathbf{x}, \Phi) g(\mathbf{x} \mid \Phi)$. For Φ and Φ' in Ω, one then has

$$L(\Phi) = Q(\Phi \mid \Phi') - H(\Phi \mid \Phi'),$$

where $Q(\Phi \mid \Phi') = E(\log f(\mathbf{y} \mid \Phi) \mid \mathbf{x}, \Phi')$ and $H(\Phi \mid \Phi') = E(\log k(\mathbf{y} \mid \mathbf{x}, \Phi) \mid \mathbf{x}, \Phi')$. The general EM algorithm of Dempster, Laird and Rubin [39] is the following: Given a current approximation Φ^c of a maximizer of $L(\Phi)$, obtain a next approximation Φ^+ as follows:

1. E-step. Determine $Q(\Phi \mid \Phi^c)$.
2. M-step. Choose $\Phi^+ \in \arg\max_{\Phi \in \Omega} Q(\Phi \mid \Phi^c)$.

Here, $\arg\max_{\Phi \in \Omega} Q(\Phi \mid \Phi^c)$ denotes the set of values $\Phi \in \Omega$ which maximize $Q(\Phi \mid \Phi^c)$ over Ω. (Of course, this set must be nonempty for the M-step of the algorithm to be well defined.) If this set is a singleton, then we denote its sole member in the same way and write $\Phi^+ = \arg\max_{\Phi \in \Omega} Q(\Phi \mid \Phi^c)$. Similar notation is used without further explanation in the sequel.

From this general description, it is not clear that the EM algorithm even deserves to be called an algorithm. However, as we indicated above, the EM algorithm is used most often in applications which permit the easy maximization of $\log f(\mathbf{y} \mid \Phi)$ over $\Phi \in \Omega$. In such applications, the M-step maximization of $Q(\Phi \mid \Phi')$ over $\Phi \in \Omega$ is usually carried out with corresponding ease. In fact, as one sees in the sequel, the E-step and the M-step are usually combined into one very easily implemented step in most applications involving mixture density estimation problems. At any rate, the sense of the EM algorithm lies in the fact that $L(\Phi^+) \geqq L(\Phi^c)$. Indeed, the manner in which Φ^+ is determined guarantees that $Q(\Phi^+ \mid \Phi^c) \geqq Q(\Phi^c \mid \Phi^c)$; and it follows from Jensen's inequality that $H(\Phi^+ \mid \Phi^c) \leqq H(\Phi^c \mid \Phi^c)$. (See Dempster, Laird and Rubin [39, Thm. 1].) This fact implies that L is monotone increasing on any iteration sequence generated by the EM algorithm, which is the fundamental property of the algorithm underlying the convergence theorems given below.

To discuss the EM algorithm for mixture density estimation problems, we assume as in the preceding section that a parametric family of mixture densities of the form (1.1) is specified and that a particular $\Phi^* = (\alpha_1^*, \cdots, \alpha_m^*, \phi_1^*, \cdots, \phi_m^*)$ is the "true" parameter value to be estimated. In the usual way, we regard this family of densities as being associated with a statistical population which is a mixture of m component populations. The EM algorithm for a mixture density estimation problem associated with this family is derived by first interpreting the problem as one involving incomplete data and then obtaining the algorithm from its general formulation given above. The problem is interpreted as one involving incomplete data by regarding each unlabeled observation in the sample at hand as "missing" a label indicating its component population of origin.

It is instructive to consider the forms which the EM algorithm might take for mixture density estimation problems involving samples of the types introduced in the preceding section. We first illustrate in some detail the derivation of the function $Q(\Phi \mid \Phi')$ of the E-step of the algorithm, assuming for convenience that the sample at hand is a sample $S_1 = \{x_k\}_{k=1\ldots N}$ of Type 1 described in the preceding section. One can regard S_1 as a sample of incomplete data by considering each x_k to be the "known" part of an observation $y_k = (x_k, i_k)$, where i_k is an integer between 1 and m indicating the component population of origin. For $\Phi = (\alpha_1, \cdots, \alpha_m, \phi_1, \cdots, \phi_m) \in \Omega$, the sample variables $\mathbf{x} = (x_1, \cdots, x_N)$ and $\mathbf{y} = (y_1, \cdots, y_N)$ have associated probability density functions $g(\mathbf{x} \mid \Phi) = \Pi_{k=1}^N p(x_k \mid \Phi)$ and $f(\mathbf{y} \mid \Phi) = \Pi_{k=1}^N \alpha_{i_k} p_{i_k}(x_k \mid \phi_{i_k})$, respectively. Then for $\Phi' = (\alpha_1', \cdots, \alpha_m', \phi_1', \cdots, \phi_m') \in \Omega$, the conditional density $k(\mathbf{y} \mid \mathbf{x}, \Phi')$ is given by

$$k(\mathbf{y} \mid \mathbf{x}, \Phi') = \prod_{k=1}^{N} \frac{\alpha_{i_k}' p_{i_k}(x_k \mid \phi_{i_k}')}{p(x_k \mid \Phi')},$$

and the function $Q(\Phi \mid \Phi')$, which we denote by $Q_1(\Phi \mid \Phi')$, is determined to be

$$\begin{aligned} Q_1(\Phi \mid \Phi') &= \sum_{i_1=1}^{m} \cdots \sum_{i_N=1}^{m} \sum_{k=1}^{N} \log \alpha_{i_k} p_{i_k}(x_k \mid \phi_{i_k}) \prod_{k=1}^{N} \frac{\alpha_{i_k}' p_{i_k}(x_k \mid \phi_{i_k}')}{p(x_k \mid \Phi')} \\ &= \sum_{i=1}^{m} \sum_{k=1}^{N} \log \alpha_i p_i(x_k \mid \phi_i) \frac{\alpha_i' p_i(x_k \mid \phi_i')}{p(x_k \mid \Phi')} \qquad (4.1) \\ &= \sum_{i=1}^{m} \left[\sum_{k=1}^{N} \frac{\alpha_i' p_i(x_k \mid \phi_i')}{p(x_k \mid \Phi')} \right] \log \alpha_i + \sum_{i=1}^{m} \sum_{k=1}^{N} \log p_i(x_k \mid \phi_i) \frac{\alpha_i' p_i(x_k \mid \phi_i')}{p(x_k \mid \Phi')}. \end{aligned}$$

For samples $S_2 = \bigcup_{i=1}^m \{y_{ik}\}_{k=1,\ldots,J_i}$, $S_3 = \bigcup_{i=1}^m \{z_{ik}\}_{k=1,\ldots,K_i}$ and $S_4 = \bigcup_{i=0}^m \{w_{ik}\}_{k=1,\ldots,M_i}$ of Types 2, 3 and 4, one determines in a similar manner the respective functions $Q_2(\Phi \mid \Phi')$, $Q_3(\Phi \mid \Phi')$ and $Q_4(\Phi \mid \Phi')$ for the E-step of the EM algorithm to be

$$Q_2(\Phi \mid \Phi') = \sum_{i=1}^{m} \sum_{k=1}^{J_i} \log p_i(y_{ik} \mid \phi_i), \qquad (4.2)$$

$$Q_3(\Phi \mid \Phi') = \sum_{i=1}^{m} K_i \log \alpha_i + \sum_{i=1}^{m} \sum_{k=1}^{K_i} \log p_i(z_{ik} \mid \phi_i), \qquad (4.3)$$

$$\begin{aligned} Q_4(\Phi \mid \Phi') = &\sum_{i=1}^{m} \left[M_i + \sum_{k=1}^{M_0} \frac{\alpha_i' p_i(w_{0k} \mid \phi_i')}{p(w_{0k} \mid \Phi')} \right] \log \alpha_i \\ &+ \sum_{i=1}^{m} \left[\sum_{k=1}^{M_i} \log p_i(w_{ik} \mid \phi_i) + \sum_{k=1}^{M_0} \log p_i(w_{0k} \mid \phi_i) \frac{\alpha_i' p_i(w_{0k} \mid \phi_i')}{p(w_{0k} \mid \Phi')} \right] \end{aligned} \qquad (4.4)$$

for $\Phi = (\alpha_1, \cdots, \alpha_m, \phi_1, \cdots, \phi_m)$ and $\Phi' = (\alpha_1', \cdots, \alpha_m', \phi_1', \cdots, \phi_m')$ in Ω. We note that $Q_2(\Phi \mid \Phi')$ and $Q_3(\Phi \mid \Phi')$ are just $L_2(\Phi)$ and (except for an additive constant) $L_3(\Phi)$ given by (3.2) and (3.3), respectively; and one might well wonder why they are of interest in this context. By way of explanation, we observe that if a sample of interest is a stochastically independent union of smaller samples, then the function for the E-step of the EM algorithm which is appropriate for this sample is just the sum of the functions which are appropriate for the smaller samples. Thus, for example, if $S = S_1 \cup S_2 \cup S_3$ is a union of independent samples of Types 1, 2, and 3, then the function for the E-step appropriate for S is $Q(\Phi \mid \Phi') = Q_1(\Phi \mid \Phi') + Q_2(\Phi \mid \Phi') + Q_3(\Phi \mid \Phi')$, where $Q_1(\Phi \mid \Phi')$, $Q_2(\Phi \mid \Phi')$ and $Q_3(\Phi \mid \Phi')$ are given by (4.1), (4.2) and (4.3), respectively.

Having determined an appropriate function $Q(\Phi \mid \Phi')$ for the E-step of the EM algorithm as one or a sum of the functions $Q_i(\Phi \mid \Phi')$ defined above, one is likely to find

that the maximization problem of the M-step has a number of attractive features. It is clear from (4.1), (4.2), (4.3) and (4.4) that this maximization problem separates into two maximization problems, the first involving the proportions $\alpha_1, \cdots, \alpha_m$ alone and the second involving only the remaining parameters $\phi_1, \cdots, \phi_m$. Since $\log \alpha_1, \cdots, \log \alpha_m$ appear linearly in each function $Q_i(\Phi \mid \Phi')$ for $i \neq 2$, the first maximization problem has a unique solution if the sample is not strictly of Type 2, and this solution is easily and explicitly determined regardless of the functional forms of the component densities $p_i(x \mid \phi_i)$. If $\phi_1, \cdots, \phi_m$ are mutually independent variables, then the second maximization problem separates further into m component problems, each of which involves only one of the parameters ϕ_i. Both these component problems and the maximization problem for the proportions alone have the appealing property that they can be regarded as "weighted" maximum-likelihood estimation problems involving sums of logarithms weighted by posterior probabilities that sample observations belong to appropriate component populations, given the current approximate maximum-likelihood estimate of Φ^*.

To illustrate these remarks, we consider a sample $S_1 = \{x_k\}_{k=1,\cdots,N}$ of Type 1 and assume that $\phi_1, \cdots, \phi_m$ are mutually independent variables. If $\Phi^c = (\alpha_1^c, \cdots, \alpha_m^c, \phi_1^c, \cdots, \phi_m^c)$ is a current approximate maximizer of the log-likelihood function $L_1(\Phi)$ given by (3.1), then one easily verifies that the next approximate maximizer $\Phi^+ = (\alpha_1^+, \cdots, \alpha_m^+, \phi_1^+, \cdots, \phi_m^+)$ prescribed by the M-step of the EM algorithm satisfies

$$\alpha_i^+ = \frac{1}{N} \sum_{k=1}^{N} \frac{\alpha_i^c p_i(x_k \mid \phi_i^c)}{p(x_k \mid \Phi^c)}, \tag{4.5}$$

$$\phi_i^+ \in \arg\max_{\phi_i \in \Omega_i} \sum_{k=1}^{N} \log p_i(x_k \mid \phi_i) \frac{\alpha_i^c p_i(x_k \mid \phi_i^c)}{p(x_k \mid \Phi^c)} \tag{4.6}$$

for $i = 1, \cdots, m$. Note that each weight $\alpha_i^c p_i(x_k \mid \phi_i^c)/p(x_k \mid \Phi^c)$ is the posterior probability that x_k originated in the ith component population, given the current approximate maximum-likelihood estimate Φ^c.

In addition to prescribing each α_i^+ and ϕ_i^+ as the solution of a heuristically appealing weighted maximum-likelihood estimation problem, there are other attractions to (4.5) and (4.6). For example, (4.5) insures that the next approximate proportions α_i^+ inherit from the current approximate proportions α_i^c the property of being nonnegative and summing to 1. Furthermore, although there is no guarantee that the maximization problems (4.6) will have nice properties in general, it happens that each ϕ_i^+ is usually easily (even uniquely and explicitly) determined by (4.6) in most applications of interest, especially in those applications in which each component density $p_i(x \mid \phi_i)$ is one of the common parametric densities for which ordinary (labeled-sample) maximum-likelihood estimates of ϕ_i are uniquely and explicitly determined. As an illustration, consider the case in which some $p_i(x \mid \phi_i)$ is a multivariate normal density, i.e., $p_i(x \mid \phi_i)$ and ϕ_i are given by

$$p_i(x \mid \phi_i) = \frac{1}{(2\pi)^{n/2}(\det \Sigma_i)^{1/2}} e^{-1/2(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i)}, \qquad \phi_i = (\mu_i, \Sigma_i), \tag{4.7}$$

where $\mu_i \in R^n$ and Σ_i is a positive-definite symmetric $n \times n$ matrix. For a given $\phi_i^c = (\mu_i^c, \Sigma_i^c)$, the unique solution $\phi_i^+ = (\mu_i^+, \Sigma_i^+)$ of (4.6) is given by

$$\mu_i^+ = \left\{ \sum_{k=1}^{N} x_k \frac{\alpha_i^c p_i(x_k \mid \phi_i^c)}{p(x_k \mid \Phi^c)} \right\} \Bigg/ \left\{ \sum_{k=1}^{N} \frac{\alpha_i^c p_i(x_k \mid \phi_i^c)}{p(x_k \mid \Phi^c)} \right\}, \tag{4.8}$$

$$\Sigma_i^+ = \left\{\sum_{k=1}^{N} (x_k - \mu_i^+)(x_k - \mu_i^+)^T \frac{\alpha_i^c p_i(x_k \mid \phi_i^c)}{p(x_k \mid \Phi^c)}\right\} \Big/ \left\{\sum_{k=1}^{N} \frac{\alpha_i^c p_i(x_k \mid \phi_i^c)}{p(x_k \mid \Phi^c)}\right\}. \tag{4.9}$$

(The factors α_i^c have been left in the numerators and denominators of these expressions for aesthetic reasons only.) Note that Σ_i^+ is positive-definite symmetric with probability 1 if $N > n$.

What convergence properties hold for a sequence of iterates generated by applying the EM algorithm to a mixture density estimation problem? If nothing in particular is said about the parametric family of interest, then the properties which can be specified are essentially those obtained by specializing the convergence results associated with the EM algorithm for general incomplete data problems. The convergence results below are formulated so that they are valid for the EM algorithm in general. Points relating to these results which are of particular interest in the mixture density context are made in remarks following the theorems.

The first theorem is a global convergence result for sequences generated by the EM algorithm. It essentially summarizes the results of Wu [155] for the general EM algorithm and of Redner [121] for the more specialized case of the EM algorithm applied to a mixture of densities from exponential families. Similar but more restrictive results have been formulated for the general EM algorithm by Boyles [18] and for a special application of the EM algorithm by Vardi [149]. Statements (i), (ii) and (iii) of the theorem are valid for any sequence and are stated here as a convenience because of their usefulness in applications. Statements (iv), (v) and (vi) are based on the fact reviewed earlier and reiterated in the statement of the theorem that the log-likelihood function increases monotonically on a sequence generated by the EM algorithm. Through the use of this fact, the theorem can be related to general results in optimization theory such as the convergence theorems of Zangwill [162, pp. 91, 128, 232] concerning point-to-set maps which increase an objective function. In fact, such general results were used explicitly by Wu [155] and Vardi [149].

THEOREM 4.1. *Suppose that for some* $\Phi^{(0)} \in \Omega$, $\{\Phi^{(j)}\}_{j=0,1,2,\ldots}$ *is a sequence in* Ω *generated by the* EM *algorithm, i.e., a sequence in* Ω *satisfying*

$$\Phi^{(j+1)} \in \arg\max_{\Phi \in \Omega} Q(\Phi \mid \Phi^{(j)}), \qquad j = 0, 1, 2, \cdots,$$

where $Q(\Phi \mid \Phi')$ *is the function determined in the* E*-step of the* EM *algorithm. Then the log-likelihood function* $L(\Phi)$ *increases monotonically on* $\{\Phi^{(j)}\}_{j=0,1,2,\ldots}$ *to a (possibly infinite) limit* L^*. *Furthermore, denoting the set of limit points of* $\{\Phi^{(j)}\}_{j=0,1,2,\ldots}$ *in* Ω *by* $\mathcal{L}$, *one has the following:*

(i) $\mathcal{L}$ *is a closed set in* Ω.

(ii) *If* $\{\Phi^{(j)}\}_{j=0,1,2,\ldots}$ *is contained in a compact subset of* Ω, *then* $\mathcal{L}$ *is compact.*

(iii) *If* $\{\Phi^{(j)}\}_{j=0,1,2,\ldots}$ *is contained in a compact subset of* Ω *and* $\lim_{j\to\infty} \|\Phi^{(j+1)} - \Phi^{(j)}\| = 0$ *for a norm* $\|\cdot\|$ *on* Ω, *then* $\mathcal{L}$ *is connected as well as compact.*

(iv) *If* $L(\Phi)$ *is continuous in* Ω *and* $\mathcal{L} \neq \varnothing$, *then* L^* *is finite and* $L(\hat{\Phi}) = L^*$ *for* $\hat{\Phi} \in \mathcal{L}$.

(v) *If* $Q(\Phi \mid \Phi')$ *and* $H(\Phi \mid \Phi') = Q(\Phi \mid \Phi') - L(\Phi)$ *are continuous in* Φ *and* Φ' *in* Ω, *then each* $\hat{\Phi} \in \mathcal{L}$ *satisfies* $\hat{\Phi} \in \arg\max_{\Phi \in \Omega} Q(\Phi \mid \hat{\Phi}')$.

(vi) *If* $Q(\Phi \mid \Phi')$ *and* $H(\Phi \mid \Phi')$ *are continuous in* Φ *and* Φ' *in* Ω *and differentiable in* Φ *at* $\Phi = \Phi' = \hat{\Phi} \in \mathcal{L}$, *then* $L(\Phi)$ *is differentiable at* $\Phi = \hat{\Phi}$ *and the likelihood equations* $\nabla_\Phi L(\Phi) = 0$ *are satisfied by* $\Phi = \hat{\Phi}$.

Proof. The monotonicity of $L(\Phi)$ on $\{\Phi^{(j)}\}_{j=0,1,2,\ldots}$ has already been established; the

existence of a (possibly infinite) limit L^* follows. Statement (i) holds since closedness is a general property of sets of limit points. To obtain (ii), note that if $\{\Phi^{(j)}\}_{j=0,1,2,\ldots}$ is contained in a compact subset of Ω, then $\mathcal{L}$ is a closed subset of this compact subset and, hence, is compact. To prove (iii), suppose that $\{\Phi^{(j)}\}_{j=0,1,2,\ldots}$ is contained in a compact subset of Ω, that $\lim_{j\to\infty} \|\Phi^{(j+1)} - \Phi^{(j)}\| = 0$, and that $\mathcal{L}$ is not connected. Since $\mathcal{L}$ is compact, there is a minimal distance between distinct components of $\mathcal{L}$, and the fact that $\lim_{j\to\infty} \|\Phi^{(j+1)} - \Phi^{(j)}\| = 0$ implies that there is an infinite subsequence of $\{\Phi^{(j)}\}_{j=0,1,2,\ldots}$ whose members are bounded away from $\mathcal{L}$. This subsequence lies in a compact set, and so it has limit points. Since these limit points cannot be in $\mathcal{L}$, one has a contradiction.

Statement (iv) follows immediately from the monotonicity of $L(\Phi)$ on $\{\Phi^{(j)}\}_{j=0,1,2,\ldots}$. To prove (v), suppose that $Q(\Phi \mid \Phi')$ and $H(\Phi \mid \Phi')$ are continuous in Φ and Φ' in Ω and that one can find some $\hat{\Phi} \in \mathcal{L}$ and $\Phi \in \Omega$ for which $Q(\Phi \mid \hat{\Phi}) > Q(\hat{\Phi} \mid \hat{\Phi})$. Then for every j,

$$L(\Phi^{(j+1)}) = Q(\Phi^{(j+1)} \mid \Phi^{(j)}) - H(\Phi^{(j+1)} \mid \Phi^{(j)}) \geqq Q(\Phi \mid \Phi^{(j)}) - H(\Phi^{(j)} \mid \Phi^{(j)})$$

by the M-step determination of $\Phi^{(j+1)}$ and Jensen's inequality. Since $Q(\Phi \mid \Phi')$ and $H(\Phi \mid \Phi')$ are continuous, it follows by taking limits along a subsequence converging to $\hat{\Phi}$ that

$$L^* \geqq Q(\Phi \mid \hat{\Phi}) - H(\hat{\Phi} \mid \hat{\Phi}) > Q(\hat{\Phi} \mid \hat{\Phi}) - H(\hat{\Phi} \mid \hat{\Phi}) = L(\hat{\Phi}) = L^*,$$

which is a contradiction. To establish (vi), suppose that $Q(\Phi \mid \Phi')$ and $H(\Phi \mid \Phi')$ are continuous in Φ and Φ' in Ω and differentiable in Φ at $\Phi = \Phi' = \hat{\Phi} \in \mathcal{L}$. Then $L(\Phi) = Q(\Phi \mid \hat{\Phi}) - H(\Phi \mid \hat{\Phi})$ is differentiable at $\Phi = \hat{\Phi}$, and, since $\hat{\Phi} \in \arg\max_{\Phi\in\Omega} Q(\Phi \mid \hat{\Phi})$ by (v) and $\hat{\Phi} \in \arg\max_{\Phi\in\Omega} H(\Phi \mid \hat{\Phi})$ by Jensen's inequality, one has $\nabla_\Phi L(\hat{\Phi}) = 0$. This completes the proof.

Statement (iii) of Theorem 4.1 has precedent in such results as Ostrowski [108, Thm. 28.1]. It is usually satisifed in practice, especially in the mixture density context. Indeed, it often happens that each Φ^+ is uniquely determined by the EM algorithm as a function of Φ^c which is continuous in Ω. For example, one sees that $\alpha_1^+, \cdots, \alpha_m^+$ are determined in this way by (4.5) whenever each $p_i(x \mid \phi_i)$ depends continuously on ϕ_i. In addition, each ϕ_i is likely to be determined in this way by (4.6) whenever each $p_i(x \mid \phi_i)$ is one of the common parametric densities for which ordinary maximum-likelihood estimates are determined as a continuous function of ϕ_i; see, for example, (4.8) and (4.9). If Φ^+ is determined in this way from Φ^c and if the conditions of (v) are also satisfied, then each $\hat{\Phi} \in \mathcal{L}$ is a fixed point of a continuous function. It follows that if in addition $\{\Phi^{(j)}\}_{j=0,1,2,\ldots}$ is contained in a compact subset of Ω, then the elements of a "tail" sequence $\{\Phi^{(j)}\}_{j=J,J+1,\ldots}$ can all be made to lie arbitrarily close to the compact set $\mathcal{L}$ by taking J sufficiently large and, hence, $\lim_{j\to\infty} \|\Phi^{(j+1)} - \Phi^{(j)}\| = 0$ by the uniform continuity near $\mathcal{L}$ of the function determining $\Phi^{(j+1)}$ from $\Phi^{(j)}$.

It is useful to expand a little on the interpretation of statement (vi) in the mixture density context. Assuming that each $p_i(x \mid \phi_i)$ is differentiable with respect to ϕ_i, that the parameters ϕ_i are unconstrained in Ω_i and mutually independent, and, for convenience, that the sample of interest is of Type 1, one can reasonably interpret the likelihood equations $\nabla_\Phi L(\Phi) = 0$ in the sense of (3.9) at a point $\hat{\Phi} = (\hat{\alpha}_1, \cdots, \hat{\alpha}_m, \hat{\phi}_1, \cdots, \hat{\phi}_m) \in \mathcal{L}$ which is such that each $\hat{\alpha}_i$ is positive. Now it is certainly possible for some $\hat{\alpha}_i$ to be zero for $\hat{\Phi} \in \mathcal{L}$, in which case (3.9) might not be valid. Fortunately, (3.5) and (3.8) provide a better interpretation than (3.9) of the likelihood equations in the mixture density context, which is valid whether each $\hat{\alpha}_i$ is positive or not. Indeed, if the conditions of (v) hold, then it follows from (4.5) that the equations (3.8) are satisfied on $\mathcal{L}$. Thus in the mixture density context under the present assumptions, (vi) should be replaced with the following:

(vi)′ *If* $Q(\Phi \mid \Phi')$ *and* $H(\Phi \mid \Phi')$ *are continuous in* Φ *and* Φ' *in* Ω *and differentiable in*

$\phi_1, \cdots, \phi_m$ at $\Phi = \Phi' = \hat{\Phi} \in \mathcal{L}$, then $L(\Phi)$ is differentiable in $\phi_1, \cdots, \phi_m$ at $\Phi = \hat{\Phi}$ and the likelihood equations (3.5) *and* (3.8) *are satisfied by* $\hat{\Phi}$.

To illustrate the application of Theorem 4.1, we consider the problem of estimating the proportions in a mixture under the assumption that each component density $p_i(x \mid \phi_i)$ is known (and denoted for the present purposes by $p_i(x)$ for simplicity). The theorem below is a global convergence result for the EM algorithm applied to this problem. For convenience in presenting the theorem, it is assumed that the sample at hand is a sample $S_1 = \{x_k\}_{k=1,\cdots,N}$ of Type 1. Similar results hold for other cases in which the sample at hand is one or a union of the types considered in the preceding section. For this problem, one has simply $\Phi = (\alpha_1, \cdots, \alpha_m)$; and it is, of course, always understood that $\Sigma_{i=1}^m \alpha_i = 1$ and $\alpha_i \geqq 0, i = 1, \cdots, m$, for all such Φ. We remark that the condition of the theorem on the matrix of second derivatives of $L_1(\Phi)$ is quite reasonable. This matrix is always defined and negative semidefinite whenever $p(x_k \mid \Phi) \neq 0$ for $k = 1, \cdots, N$; and if $p_1(x), \cdots, p_m(x)$ are linearly independent nonvanishing functions on the support of the underlying measure on R^n appropriate for p, then with probability 1 it is defined and negative definite for all Φ whenever N is sufficiently large.

THEOREM 4.2. *Suppose that the matrix of second derivatives of $L_1(\Phi)$ is defined and negative definite for all Φ. Then there is a unique maximum-likelihood estimate, and for any $\Phi^{(0)} = (\alpha_1^{(0)}, \cdots, \alpha_m^{(0)})$ with $\alpha_i^{(0)} > 0$ for $i = 1, \cdots, m$, the sequence $\{\Phi^{(j)} = (\alpha_1^{(j)}, \cdots, \alpha_m^{(j)})\}_{j=0,1,2,\ldots}$ generated by the* EM *algorithm, i.e., determined inductively by*

$$\alpha_i^{(j+1)} = \frac{1}{N} \sum_{k=1}^{N} \frac{\alpha_i^{(j)} p_i(x_k)}{p(x_k \mid \Phi^{(j)})}, \qquad i = 1, \cdots, m,$$

converges to the maximum-likelihood estimate.

Proof. It follows from Theorem 4.1 and the subsequent remarks that the set of limit points of $\{\Phi^{(j)}\}_{j=0,1,2,\ldots}$ is a compact, connected subset of the simplex of proportion vectors Φ on which the likelihood equations (3.8) are satisfied. Since the matrix of second derivatives of $L_1(\Phi)$ is negative definite, $L_1(\Phi)$ is strictly concave. It follows that there is a unique maximum-likelihood estimate and, furthermore, that the likelihood equations (3.8) have at most one solution on the interior of each face of the proportion simplex. Consequently, each component of the set of solutions of the likelihood equations consists of a single point, and $\{\Phi^{(j)}\}_{j=0,1,2,\ldots}$ must converge to one such point. But if $\{\Phi^{(j)}\}_{j=0,1,2,\ldots}$ is convergent, then its limit must be the maximum-likelihood estimate by Peters and Coberly [112, Thm. 2] or Peters and Walker [114, Thm. 1].

Despite the usefulness of Theorem 4.1 in characterizing the set of limit points of an iteration sequence generated by the EM algorithm, it leaves unanswered the questions of whether such a sequence converges at all and, if it does, whether it converges to a maximum-likelihood estimate. In an attempt to provide reasonable sufficient conditions under which the answer to these questions is "yes", we offer the local convergence theorem below.

THEOREM 4.3. *Suppose that Conditions* 1 *through* 4 *of* §3 *are satisfied in Ω, and let Ω' be a compact subset of Ω which contains Φ^* in its interior and which is such that $p(x \mid \Phi) = p(x \mid \Phi^*)$ almost everywhere in x for $\Phi \in \Omega'$ only if $\Phi = \Phi^*$. Suppose further that with probability* 1, *the function $Q(\Phi \mid \Phi')$ of the* E*-step of the* EM *algorithm is continuous in Φ and Φ' in Ω' and both $Q(\Phi \mid \Phi')$ and the log-likelihood function $L(\Phi)$ are differentiable in Φ for Φ and Φ' in Ω' whenever N is sufficiently large. Finally, for $\Phi^{(0)}$ in Ω' denote by $\{\Phi^{(j)}\}_{j=0,1,2,\ldots}$ a sequence generated by the* EM *algorithm in Ω', i.e., a sequence in Ω' satisfying*

$$\Phi^{(j+1)} \in \arg\max_{\Phi \in \Omega'} Q(\Phi \mid \Phi^{(j)}), \qquad j = 0, 1, 2, \cdots.$$

Then with probability 1, *whenever N is sufficiently large, the unique strongly consistent maximum-likelihood estimate* Φ^N *is well defined in* Ω' *and* $\Phi^N = \lim_{j\to\infty} \Phi^{(j)}$ *whenever* $\Phi^{(0)}$ *is sufficiently near* Φ^N.

Proof. It follows from Theorems 3.1 and 3.2 that with probability 1, N can be taken sufficiently large that the unique strongly consistent maximum-likelihood estimate Φ^N is well defined, lies in the interior of Ω' and is the unique maximizer of $L(\Phi)$ in Ω'. Also with probability 1, we can assume that N is sufficiently large that $Q(\Phi \mid \Phi')$ is continuous in Φ and Φ' in Ω', and $Q(\Phi \mid \Phi')$ and $L(\Phi)$ are differentiable in Φ for Φ and Φ' in Ω'. Since $L(\Phi)$ is continuous, one can find a neighborhood Ω'' of Φ^N of the form

$$\Omega'' = \{\Phi \in \Omega' : L(\Phi) \geqq L(\Phi^N) - \varepsilon\}$$

for some $\varepsilon > 0$ which lies in the interior of Ω' and which is such that Φ^N is the only solution of the likelihood equations contained in it. If $\Phi^{(0)}$ lies in Ω'', then $\{\Phi^{(j)}\}_{j=0,1,2,\ldots}$ must also lie in Ω'' since $L(\Phi)$ is monotone increasing on $\{\Phi^{(j)}\}_{j=0,1,2,\ldots}$. It follows that each limit point of $\{\Phi^{(j)}\}_{j=0,1,2,\ldots}$ lies in Ω'' and, by statement (vi) of Theorem 4.1, also satisfies the likelihood equations. Since Φ^N is the only solution of the likelihood equations in Ω'', one concludes that $\Phi^N = \lim_{j\to\infty} \Phi^{(j)}$.

As in the case of Theorem 4.1, Theorem 4.3 is stated so that it is valid for the EM algorithm in general. It should be noted, however, that Theorem 4.3 makes heavy use of Theorems 3.1 and 3.2 as well as Theorem 4.1, and so for mixture density estimation problems, it pertains as it stands, strictly speaking, to the case to which Theorems 3.1 and 3.2 apply, namely that in which the sample at hand is of Type 1 and $L(\Phi) = L_1(\Phi)$ is given by (3.1) and $Q(\Phi \mid \Phi') = Q_1(\Phi \mid \Phi')$, by (4.1). Of course, Theorems 3.1 and 3.2 and, therefore, Theorem 4.3 can be modified to treat mixture density estimation problems involving samples of other types.

5. The EM algorithm for mixtures of densities from exponential families. Almost all mixture density estimation problems which have been studied in the literature involve mixture densities whose component densities are members of exponential families. As it happens, the EM algorithm is especially easy to implement on problems involving densities of this type. Indeed, in an application of the EM algorithm to such a problem, each successive approximate maximum-liklihood estimate Φ^+ is uniquely and explicitly determined from its predecessor Φ^c, almost always in a continuous manner. Furthermore, a sequence of iterates produced by the EM algorithm on such a problem is likely to have relatively nice convergence properties.

In this section, we first determine the special form which the EM algorithm takes for mixtures of densities from exponential families. We then look into the desirable properties of the algorithm and sequences generated by it which are apparent from this form. Finally, we discuss several specific examples of the EM algorithm for component densities from exponential families which are commonly of interest.

A very brief discussion of exponential families of densities is in order. For an elaboration on the topics touched on here, the reader is referred to the book of Barndorff-Nielsen [7]. A parametric family of densities $q(x \mid \theta)$, $\theta \in \tilde{\Omega} \subseteq R^k$, on R^n is said to be an *exponential family* if its members have the form

$$q(x \mid \theta) = a(\theta)^{-1} b(x) e^{\theta^T t(x)}, \qquad x \in R^n, \tag{5.1}$$

where $b : R^N \to R^1$, $t : R^n \to R^k$ and $a(\theta)$ is given by

$$a(\theta) = \int_{R^n} b(x)\, e^{\theta^T t(x)}\, d\mu$$

for an appropriate underlying measure μ on R^n. It is, of course, assumed that $b(x) \geqq 0$ for

all $x \in R^n$ and that $a(\theta) < \infty$ for $\theta \in \Omega$. Note that every member of an exponential family has the same support in R^n, namely that of the function $b(x)$.

The representation (5.1) of the members of an exponential family, in which the parameter θ appears linearly in the argument of the exponential function, is called the "natural" parametrization; and θ is called the "natural" parameter. If the set $\tilde{\Omega}$ is open and convex and if the component functions of $t(x)$ together with the function which is identically 1 on R^n are linearly independent functions on the intersection of the supports of $b(x)$ and μ, then there is another parametrization of the members of the family, called the "expectation" or "mean value" parametrization, in terms of the "expectation" parameter

$$\phi = E(t(X) \mid \theta) = \int_{R^n} t(x) q(x \mid \theta)\, d\mu.$$

Indeed, under these conditions on $\tilde{\Omega}$ and $t(x)$, one can show that

$$[E(t(X) \mid \theta') - E(t(X) \mid \theta)]^T(\theta' - \theta) > 0$$

whenever $\theta' \neq \theta$, and it follows that the assignment $\theta \rightarrow \phi = E(t(X) \mid \theta)$ is one-to-one and onto from $\tilde{\Omega}$ to an open set $\Omega \subseteq R^k$. In fact, the correspondence $\theta \leftrightarrow \phi = E(t(X) \mid \theta)$ is a both-ways continuously differentiable mapping between $\tilde{\Omega}$ and Ω. (See Barndorff-Nielsen [7, p. 121].) So under these conditions on $\tilde{\Omega}$ and $t(x)$, one can represent the members of the family as

$$p(x \mid \phi) = q(x \mid \theta(\phi)) = a(\phi)^{-1} b(x) e^{\theta(\phi)^T t(x)}, \qquad x \in R^n, \tag{5.2}$$

for $\phi \in \Omega$, where $\theta(\phi)$ satisfies $\phi = E(t(X) \mid \theta(\phi))$ and $a(\theta(\phi))$ is written as $a(\phi)$ for convenience. Note that $p(x \mid \phi)$ is continuously differentiable in ϕ, since $q(x \mid \theta)$ is continuously differentiable in θ and $\theta(\phi)$ is continuously differentiable in ϕ.

Now suppose that a parametric family of mixture densities of the form (1.1) is given, with $\Phi^* = (\alpha_1^*, \cdots, \alpha_m^*, \phi_1^*, \cdots, \phi_m^*)$ the "true" parameter value to be estimated; and suppose that each component density $p_i(x \mid \phi_i)$ is a member of an exponential family. Specifically, we assume that each $p_i(x \mid \phi_i)$ has the "expectation" parametrization for $\phi_i \in \Omega_i \subseteq R^{n_i}$ given by

$$p_i(x \mid \phi_i) = a_i(\phi_i)^{-1} b_i(x) e^{\theta_i(\phi_i)^T t_i(x)}, \qquad x \in R^n,$$

for appropriate a_i, b_i, t_i and θ_i. We further assume that it is valid to reparametrize $p_i(x \mid \phi_i)$ in terms of the "natural" parameter $\theta_i = \theta_i(\phi_i)$ in the manner of the above discussion.

To investigate the special form and properties of the EM algorithm for the given family of mixture densities, we assume that $\phi_1, \cdots, \phi_m$ are mutually independent variables and consider for convenience a sample $S_1 = \{x_k\}_{k=1,\ldots,N}$ of Type 1. (A discussion similar to the following is valid mutatis mutandis for samples of other types.) If $\Phi^c = (\alpha_1^c, \cdots, \alpha_m^c, \phi_1^c, \cdots, \phi_m^c)$ is a current approximate maximizer of the log-likelihood function $L_1(\Phi)$ given by (3.1), then the next approximate maximizer $\Phi^+ = (\alpha_1^+, \cdots, \alpha_m^+, \phi_1^+, \cdots, \phi_m^+)$ prescribed by the M-step of the EM algorithm satisfies (4.5) and (4.6). For $i = 1, \cdots, m$, what ϕ_i^+ satisfy (4.6)? If one replaces each $p_i(x_k \mid \phi_i)$ in the sum in (4.6) by its expression in the "natural" parameter θ_i, differentiates *with respect to* θ_i, equates the sum of derivatives to zero, and finally restores the "expectation" parametrization, then one sees that the unique ϕ_i^+ which satisfies (4.6) is given explicitly by

$$\phi_i^+ = \left\{ \sum_{k=1}^{N} t_i(x_k) \frac{\alpha_i^c\, p_i(x_k \mid \phi_i^c)}{p(x_k \mid \Phi^c)} \right\} \Bigg/ \left\{ \sum_{k=1}^{N} \frac{\alpha_i^c p_i(x_k \mid \phi_i^c)}{p(x_k \mid \Phi^c)} \right\}. \tag{5.3}$$

(As in the case of (4.8) and (4.9), the factors α_i^c are left in the numerator and denominator for aesthetic reasons only.)

Not only are (4.5) and (5.3) easily evaluated and heuristically appealing formulas for determining Φ^+ from Φ^c, they also provide the key to a global convergence analysis of iteration sequences generated by the EM algorithm in the case at hand which goes beyond Theorem 4.1. Theorem 5.1 below summarizes such an analysis. In order to make the theorem complete and self-contained, some of the general conclusions of Theorem 4.1 are repeated in its statement.

THEOREM 5.1. *Suppose that* $\{\Phi^{(j)} = (\alpha_1^{(j)}, \cdots, \alpha_m^{(j)}, \phi_1^{(j)}, \cdots, \phi_m^{(j)})\}_{j=0,1,2,\ldots}$ *is a sequence in* Ω *generated by the* EM *iteration* (4.5) *and* (5.3). *Then* $L_1(\Phi)$ *increases monotonically on* $\{\Phi^{(j)}\}_{j=0,1,2,\ldots}$ *to a (possibly infinite) limit* L^*. *Furthermore, for each* i, $\{\phi_i^{(j)}\}_{j=1,2,\ldots}$ *is contained in the convex hull of* $\{t_i(x_k)\}_{k=1,\ldots,N}$. *Consequently, the set* $\overline{\mathcal{L}}$ *of all limit points of* $\{\Phi^{(j)}\}_{j=0,1,2,\ldots}$ *is compact, and the likelihood equations* (3.5) *and* (3.8) *are satisfied on* $\mathcal{L} = \overline{\mathcal{L}} \cap \Omega$. *If* $\mathcal{L} \neq \varnothing$, *then* L^* *is finite and each* $\hat{\Phi} \in \mathcal{L}$ *satisfies* $L(\hat{\Phi}) = L^*$ *and is a fixed point of the* EM *iteration. Finally, if* $\mathcal{L} = \overline{\mathcal{L}} \subseteq \Omega$, *then* $\mathcal{L}$ *is connected as well as compact.*

Remark. If the convex hull of $\{t_i(x_k)\}_{k=1,\ldots,N}$ is contained in Ω_i for each i, then $\varnothing \neq \mathcal{L} = \overline{\mathcal{L}} \subseteq \Omega$ and all of the conditional conclusions of Theorem 5.1 hold. The convex hull of $\{t_i(x_k)\}_{k=1,\ldots,N}$ is indeed contained in Ω_i for each i in many (but not all) applications. (See the examples at the end of this section.)

Proof. One sees from (5.3) that for each i, ϕ_i^+ is always a convex combination of the values $\{t_i(x_k)\}_{k=1,\ldots,N}$, and it follows that $\{\phi_i^{(j)}\}_{j=0,1,2,\ldots}$ is contained in the convex hull of $\{t_i(x_k)\}_{k=1,\ldots,N}$ for each i. Since these convex hulls are compact sets, one concludes that $\overline{\mathcal{L}}$ is compact.

Now each density $p_i(x \mid \phi_i)$ is continuously differentiable in ϕ_i on Ω_i, and so it is clear from (3.1) and (4.1) that $L_1(\Phi)$ and $Q_1(\Phi \mid \Phi')$ are continuous in Φ and Φ' and differentable in $\phi_1, \cdots, \phi_m$ in Ω. Furthermore, it is apparent from (4.5) and (5.3) that Φ^+ depends continuously on Φ^c, and one sees from the discussion following Theorem 4.1 that $\lim_{j\to\infty} \|\Phi^{(j+1)} - \Phi^{(j)}\| = 0$ if $\mathcal{L} = \overline{\mathcal{L}} \subseteq \Omega$. In light of these points, one verifies the remaining conclusions of Theorem 5.1 via a straightforward application of Theorem 4.1 (including statement (vi)'), and the proof is complete.

One can also exploit (4.5) and (5.3) to obtain a local convergence result which goes beyond Theorem 4.3 for mixture density estimation problems of the type now under consideration. Theorem 5.2 and its proof below provide not only a stronger local convergence statement than Theorem 4.3 for a sequence of iterates produced by the EM algorithm, but also a means of both quantifying the speed of convergence of the sequence and gaining insight into properties of the mixture density which affect the speed of convergence. This theorem is essentially the generalization of Redner [120] of the local convergence results of Peters and Walker [113] for mixtures of multivariate normal densities, and its proof closely parallels the proofs of those results. In the general case of incomplete data from exponential families, Sundberg [134], [135] has obtained consistency results for maximum-likelihood estimates and local convergence results for the EM algorithm. His results are similar to those below in the mixture density case, although the reader of [135] is referred to [133] for proofs of the local convergence results. A proof of Theorem 5.2 is included here since, to the best of our knowledge, no proof of a local convergence result of this type has previously appeared in the generally available literature.

THEOREM 5.2. *Suppose that the Fisher information matrix* $I(\Phi)$ *given by* (2.5.1) *is positive definite at* Φ^* *and that* $\Phi^* = (\alpha_1^*, \cdots, \alpha_m^*, \phi_1^*, \cdots, \phi_m^*)$ *is such that* $\alpha_i^* > 0$ *for* $i = 1, \cdots, m$. *For* $\Phi^{(0)}$ *in* Ω, *denote by* $\{\Phi^{(j)}\}_{j=0,1,2,\ldots}$ *the sequence in* Ω *generated by*

the EM *iteration* (4.5) *and* (5.3). *Then with probability* 1, *whenever* N *is sufficiently large, the unique strongly consistent solution* $\Phi^N = (\alpha_1^N, \cdots, \alpha_m^N, \phi_1^N, \cdots, \phi_m^N)$ *of the likelihood equations is well defined and there is a certain norm* $\|\cdot\|$ *on* Ω *in which* $\{\Phi^{(j)}\}_{j=0,1,2,\ldots}$ *converges linearly to* Φ^N *whenever* $\Phi^{(0)}$ *is sufficiently near* Φ^N, *i.e., there is a constant* λ, $0 \leqq \lambda < 1$, *for which*

$$\|\Phi^{(j+1)} - \Phi^N\| \leqq \lambda \|\Phi^{(j)} - \Phi^N\|, \qquad j = 0, 1, 2, \cdots, \tag{5.4}$$

whenever $\Phi^{(0)}$ *is sufficiently near* Φ^N.

Proof. One sees not only that Condition 2 of §3 is satisfied, but also, by restricting Ω to be a small neighborhood of Φ^* if necessary, that Condition 1 holds as well for the family of mixture densities under consideration. It follows from Theorem 3.1 that, with probability 1, Φ^N is well defined whenever N is sufficiently large and converges to Φ^* as N approaches infinity. It must be shown that, with probability 1, whenever N is sufficiently large, there is a norm $\|\cdot\|$ on Ω and a constant λ, $0 \leqq \lambda < 1$, such that (5.4) holds whenever $\Phi^{(0)}$ is sufficiently near Φ^N. Toward this end, we observe that the EM iteration of interest is actually a functional iteration $\Phi^+ = G(\Phi^c)$, where $G(\Phi)$ is the function defined in the obvious way by (4.5) and (5.3). Note that $G(\Phi)$ is continuously differentiable in Ω and that any $\hat{\Phi}$ which satisfied the likelihood equations (3.5) and (3.8) (and $\hat{\Phi} = \Phi^N$ in particular) is a *fixed point* of $G(\Phi)$, i.e., $\hat{\Phi} = G(\hat{\Phi})$. Consequently one can write

$$\Phi^+ - \Phi^N = G(\Phi^c) - G(\Phi^N) = G'(\Phi^N)(\Phi^c - \Phi^N) + O(\|\Phi^c - \Phi^N\|^2) \tag{5.5}$$

for any Φ^c in Ω near Φ^N and any norm $\|\cdot\|$ on Ω, where $G'(\Phi^N)$ denotes the Fréchet derivative of $G(\Phi)$ evaluated at Φ^N. (For questions concerning Fréchet derivatives, see, for example, Luenberger [95].) We will complete the proof by showing that with probability 1, $G'(\Phi^N)$ converges as N approaches infinity to an operator which has operator norm less than 1 with respect to a certain norm on Ω.

For convenience, we introduce the following notation for $i = 1, \cdots, m$:

$$\beta_i(x) = p_i(x \mid \phi_i^N)/p(x \mid \Phi^N),$$

$$\sigma_i = \int_{R^n} [t_i(x) - \phi_i^N][t_i(x) - \phi_i^N]^T p_i(x \mid \phi_i^N)\, d\mu,$$

$$\gamma_i(x) = \sigma_i^{-1}[t_i(x) - \phi_i^N].$$

Regarding an element $\Phi \in \Omega$ as an $(m + \Sigma_{i=1}^m n_i)$-vector in the natural way, one can show via a very tedious calculation that $G'(\Phi^N)$ has the $(m + \Sigma_{i=1}^m n_i) \times (m + \Sigma_{i=1}^m n_i)$ matrix representation

$$G'(\Phi^N) = \operatorname{diag}\left(1, \cdots, 1, \frac{1}{N}\sum_{k=1}^N \beta_1(x_k)\sigma_1\,\gamma_1(x_k)\gamma_1(x_k)^T, \cdots, \right.$$

$$\left. \frac{1}{N}\sum_{k=1}^N \beta_m(x_k)\sigma_m\gamma_m(x_k)\gamma_m(x_k)^T\right)$$

$$- \operatorname{diag}(\alpha_1^N, \cdots, \alpha_m^N, \alpha_1^{N^{-1}}\sigma_1, \cdots, \alpha_m^{N^{-1}}\sigma_m)\left\{\frac{1}{N}\sum_{k=1}^N V(x_k)V(x_k)^T\right\},$$

where

$$V(x) = (\beta_1(x), \cdots, \beta_m(x), \alpha_1\beta_1(x)\gamma_1(x)^T, \cdots, \alpha_m\beta_m(x)\gamma_m(x)^T)^T.$$

(Since Φ^N converges to Φ^* with probability 1, we can assume that, with probability 1, each α_i^N is nonzero whenever N is sufficiently large.) It follows from the strong law of large

numbers (see Loéve [94]) that, with probability 1, $G'(\Phi^N)$ converges to $E[G'(\Phi^*)] = I - QR$, where

$$Q = \text{diag}\,(\alpha_1^*, \cdots, \alpha_m^*, \alpha_1^{*-1}\sigma_1, \cdots, \alpha_m^{*-1}\sigma_m)$$

and

$$R = \int_{R^n} V(x)V(x)^T p(x\,|\,\Phi^*)\,d\mu.$$

It is understood that in these expressions defining Q and R, Φ^N and its components have been replaced by Φ^* and its components.

It remains to be shown that there is a norm $\|\cdot\|$ on Ω with respect to which $E[G'(\Phi^*)]$ has operator norm less than 1. Now Q and R are positive-definite symmetric operators with respect to the Euclidean inner product, and so QR is a positive-definite symmetric operator with respect to the inner product $\langle\cdot,\cdot\rangle$ defined by $\langle U, W\rangle = U^TQ^{-1}W$ for $(m + \Sigma_{i=1}^m n_i)$-vectors U and W. Consequently, to prove the theorem, it suffices to show that the operator norm of QR with respect to the norm defined by $\langle\cdot,\cdot\rangle$ is less than or equal to 1.

Since QR is positive-definite symmetric with respect to $\langle\cdot,\cdot\rangle$, we need only show that $\langle U, QRU\rangle \leqq \langle U, U\rangle$ for an arbitrary $(m + \Sigma_{i=1}^m n_i)$-vector $U = (\delta_1, \cdots, \delta_m, \psi_1^T, \cdots, \psi_m^T)^T$. One has

$$\begin{aligned}
\langle U, QRU\rangle &= U^TRU \\
&= \int_{R^n} \left\{ \sum_{i=1}^m \delta_i\beta_i(x) + \sum_{i=1}^m \psi_i^T\,[\alpha_i^*\beta_i(x)\gamma_i(x)] \right\}^2 p(x\,|\,\Phi^*)\,d\mu \\
&= \int_{R^n} \left\{ \sum_{i=1}^m [\delta_i\alpha_i^{*-1} + \psi_i^T\gamma_i(x)]\alpha_i^*\beta_i(x) \right\}^2 p(x\,|\,\Phi^*)\,d\mu \\
&\leqq \int_{R^n} \sum_{i=1}^m [\delta_i\alpha_i^{*-1} + \psi_i^T\gamma_i(x)]^2\alpha_i^* p_i(x\,|\,\phi_i^*)\,d\mu.
\end{aligned}$$

The inequality is a consequence of the following corollary of Schwarz's inequality: If $\eta_i \geqq 0$ for $i = 1, \cdots, m$ and if $\Sigma_{i=1}^m \eta_i = 1$, then $\{\Sigma_{i=1}^m \xi_i\eta_i\}^2 \leqq \Sigma_{i=1}^m \xi_i^2\eta_i$ for all $\{\xi_i\}_{i=1,\cdots,m}$. Since

$$\int_{R^n} \gamma_i(x)p_i(x\,|\,\phi_i^*)\,d\mu = 0,$$

one continues to obtain

$$\langle U, QRU\rangle \leqq \int_{R^n} \sum_{i=1}^m [\delta_i^2\alpha_i^{*-2} + \psi_i^T\gamma_i(x)\gamma_i(x)^T\psi_i]\alpha_i^* p_i(x\,|\,\phi_i^*)\,d\mu = \langle U, U\rangle.$$

This completes the proof.

It is instructive to explore the consequences of Theorem 5.2 and the developments in its proof. One sees from the proof of Theorem 5.2 that, with probability 1, for sufficiently large N and $\Phi^{(0)}$ sufficiently near Φ^N, an inequality (5.4) holds in which $\|\cdot\|$ is the norm determined by the inner product $\langle\cdot,\cdot\rangle$ defined in the proof and λ is arbitrarily close to the operator norm of $E[G'(\Phi^*)] = I - QR$ determined by $\|\cdot\|$. Since QR is positive-definite symmetric with respect to $\langle\cdot,\cdot\rangle$, this operator norm is just $\rho(I - QR)$, the spectral radius or largest absolute value of an eigenvalue of $I - QR$. Thus, with probability 1, one can obtain a quantitative estimate of the speed of convergence to Φ^N for large N of a sequence generated by the EM iteration (4.5) and (5.3) by taking $\lambda \approx \rho(I - QR)$ in (5.4).

What properties of the mixture density influence the speed of convergence to Φ^N of an EM iteration sequence for large N? Careful inspection shows that if the component populations in the mixture are "well separated" in the sense that

$$\frac{p_i(x \mid \phi_i^*)}{p(x \mid \Phi^*)} \frac{p_j(x \mid \phi_j^*)}{p(x \mid \Phi^*)} \approx 0 \quad \text{for } x \in R^n, \quad \text{whenever } i \neq j,$$

then $QR \approx I$. It follows that $\rho(I - QR) \approx 0$, and an EM iteration sequence which converges to Φ^N exhibits rapid linear convergence. On the other hand, if the component populations in the mixture are "poorly separated" in the sense that, say, the ith and jth component populations are such that

$$\frac{p_i(x \mid \phi_i^*)}{p(x \mid \Phi^*)} \approx \frac{p_j(x \mid \phi_j^*)}{p(x \mid \Phi^*)} \quad \text{for } x \in R^n,$$

then R is nearly singular. One concludes that $\rho(I - QR) \approx 1$ in this case and that slow linear convergence of an EM iteration sequence to Φ^N can be expected.

In the interest of obtaining iteration sequences which converge more rapidly than EM iteration sequences, Peters and Walker [113], [114] and Redner [120] considered iterative methods which proceed at each iteration in the EM direction with a step whose length is controlled by a parameter ε. In the present context, these methods take the form

$$\Phi^+ = F_\varepsilon(\Phi^c) \equiv (1 - \varepsilon)\Phi^c + \varepsilon G(\Phi^c), \tag{5.6}$$

where $G(\Phi)$ is the EM iteration function defined by (4.5) and (5.3). The idea is to optimize the speed of convergence to Φ^N of an iteration sequence generated by such a method for large N by choosing ε to minimize the spectral radius of $E[F'_\varepsilon(\Phi^*)] = I - \varepsilon QR$. As in [113], [114] and [120], one can easily show that the optimal choice of ε is always greater than one, lies near one if the component populations in the mixture are "well-separated" in the above sense, and cannot be much smaller than two if the component populations are "poorly separated" in the above sense. The extent to which the speed of convergence of an iteration sequence can be enhanced by making the optimal choice of ε in (5.6) is determined by the length of the subinterval of (0, 1] in which the spectrum of QR lies. (Greater improvements in convergence speed are realized from the optimal choice of ε when this subinterval is relatively narrow.) The applications of iterative procedures of the form (5.6) are at present incompletely explored and might well bear further investigation.

We conclude this section by briefly reviewing some of the special forms which the EM iteration takes when a particular component density $p_i(x \mid \phi_i)$ is a member of one of the common exponential families. We also comment on some convergence properties of sequences $\{\phi_i^{(j)}\}_{j=0,1,2,\ldots}$ generated by the EM algorithm in the examples considered. Hopefully, our comments will prove helpful in determining convergence properties of EM iteration sequences $\{\Phi^{(j)}\}_{j=0,1,2,\ldots}$ through the use of Theorem 5.1 or other means when all component densities are from one or more of these example families.

Example 5.1. *Poisson density.* In this example, $n = 1$ and a natural choice of Ω_i is $\Omega_i = \{\phi_i \in R^1 : 0 < \phi_i < \infty\}$. For $\phi_i \in \Omega_i$, one has

$$p_i(x \mid \phi_i) = \frac{1}{x!} e^{-\phi_i} \phi_i^x, \qquad x = 0, 1, 2, \cdots,$$

and the EM iteration (5.3) for a sample of Type 1 becomes

$$\phi_i^+ = \left\{ \sum_{k=1}^N x_k \frac{\alpha_i^c p_i(x_k \mid \phi_i^c)}{p(x_k \mid \Phi^c)} \right\} \Bigg/ \left\{ \sum_{k=1}^N \frac{\alpha_i^c p_i(x_k \mid \phi_i^c)}{p(x_k \mid \Phi^c)} \right\}.$$

Note that ϕ_i^+ is always contained in the convex hull of $\{x_k\}_{k=1,\dots,N}$, which is a compact subset of Ω_i. Therefore, the set of limit points of an EM iteration sequence $\{\phi_i^{(j)}\}_{j=0,1,2,\dots}$ is a nonempty compact subset of Ω_i.

Example 5.2. *Binomial density.* Here $n = 1$, and one naturally chooses Ω_i to be the open set $\{\phi_i \in R^1 : 0 < \phi_i < 1\}$. For $\phi_i \in \Omega_i$, $p_i(x \mid \phi_i)$ is given by

$$p_i(x \mid \phi_i) = \binom{v_i}{x} \phi_i^x (1 - \phi_i)^{v_i - x}, \qquad x = 0, 1, \cdots, v_i,$$

for a prescribed integer v_i. In this case, the EM iteration (5.3) for a sample of Type 1 becomes

$$\phi_i^+ = \left\{ \frac{1}{v_i} \sum_{k=1}^{N} x_k \frac{\alpha_i^c p_i(x_k \mid \phi_i^c)}{p(x_k \mid \Phi^c)} \right\} \Bigg/ \left\{ \sum_{k=1}^{N} \frac{\alpha_i^c p_i(x_k \mid \phi_i^c)}{p(x_k \mid \Phi^c)} \right\}.$$

Since $p_i(x \mid \phi_i)$ is nonzero only if $x = 0, 1, \cdots, v_i$, one sees from this expression that the set of limit points of an EM iteration sequence $\{\phi_i^{(j)}\}_{j=0,1,2,\dots}$ is a nonempty compact subset of $\overline{\Omega}_i = \{\phi_i \in R^1 : 0 \leqq \phi_i \leqq 1\}$.

Example 5.3. *Exponential density.* Again, $n = 1$ and one takes $\Omega_i = \{\phi_i \in R^1 : 0 < \phi_i < \infty\}$. For $\phi_i \in \Omega_i$, one has

$$p_i(x \mid \phi_i) = \frac{1}{\phi_i} e^{-x/\phi_i}, \qquad 0 < x < \infty.$$

The EM iteration (5.3) for a sample of Type 1 now becomes

$$\phi_i^+ = \left\{ \sum_{k=1}^{N} x_k \frac{\alpha_i^c p_i(x_k \mid \phi_i^c)}{p(x_k \mid \Phi^c)} \right\} \Bigg/ \left\{ \sum_{k=1}^{N} \frac{\alpha_i^c p_i(x_k \mid \phi_i^c)}{p(x_k \mid \Phi^c)} \right\},$$

and one sees that the set of limit points of an EM iteration sequence $\{\phi_i^{(j)}\}_{j=0,1,2,\dots}$ is a nonempty compact subset of Ω_i.

Example 5.4. *Multivariate normal density.* In this example, n is an arbitrary positive integer, and ϕ_i is most conveniently represented as $\phi_i = (\mu_i, \Sigma_i)$, where $\mu_i \in R^n$ and Σ_i is a positive-definite symmetric $n \times n$ matrix. (Of course, this representation of ϕ_i is not the usual representation of the "expectation" parameter.) Then Ω_i is the set of all such ϕ_i, and $p_i(x \mid \phi_i)$ is given by (4.7). For a sample of Type 1, the EM iteration (5.3) becomes that of (4.8) and (4.9).

One can see from (4.9) that each Σ_i^+ is in the convex hull of $\{(x_k - \mu_i^+)(x_k - \mu_i^+)^T\}_{k=1,\dots,N}$, a set of rank-one matrices which, of course, are not positive definite. Thus there is no guarantee that a sequence of matrices $\{\Sigma_i^{(j)}\}_{j=0,1,2,\dots}$ produced by the EM iteration will remain bounded from below. Indeed, it has been observed in practice that sequences of iterates produced by the EM algorithm for a mixture of multivariate normal densities do occasionally converge to "singular solutions" (cf. Duda and Hart [46]), i.e., points on the boundary of Ω_i with associated singular matrices.

It was observed by Hosmer [75] that if enough labeled observations are included in a sample on a mixture of normal densities, then with probability 1, the log-likelihood function attains its maximum value at a point at which the covariance matrices are positive definite. Similarly, consideration of samples with a sufficiently large number of labeled observations alleviates with probability 1 the problem of an EM iteration sequence having "singular solutions" as limit points. For example, if one considers a sample $S = S_1 \cup S_3$ which is a stochastically independent union of a sample $S_1 = \{x_k\}_{k=1,\dots,N}$ of Type 1 and a sample $S_3 = \bigcup_{i=1}^{m} \{z_{ik}\}_{k=1,\dots,K_i}$ of Type 3, then the EM

iteration becomes

$$\alpha_i^+ = \frac{1}{N+K}\left\{\sum_{k=1}^{N}\frac{\alpha_i^c p_i(x_k|\phi_i^c)}{p(x_k|\Phi^c)} + K_i\right\},$$

$$\mu_i^+ = \left\{\sum_{k=1}^{N} x_k \frac{\alpha_i^c p_i(x_k|\phi_i^c)}{p(x_k|\Phi^c)} + \sum_{k=1}^{K_i} z_{ik}\right\} \Bigg/ \left\{\sum_{k=1}^{N}\frac{\alpha_i^c p_i(x_k|\phi_i^c)}{p(x_k|\phi^c)} + K_i\right\},$$

$$\Sigma_i^+ = \frac{\left\{\sum_{k=1}^{N}(x_k-\mu_i^+)(x_k-\mu_i^+)^T\frac{\alpha_i^c p_i(x_k|\phi_i^c)}{p(x_k|\Phi^c)} + \sum_{k=1}^{K_i}(z_{ik}-\mu_i^+)(z_{ik}-\mu_i^+)^T\right\}}{\left\{\sum_{k=1}^{N}\frac{\alpha_i^c p_i(x_k|\phi_i^c)}{p(x_k|\Phi^c)} + K_i\right\}},$$

where $K = \sum_{i=1}^{m} K_i$. One sees from the expression for Σ_i^+ that Σ_i^+ is bounded below by

$$\left\{\sum_{k=1}^{K_i}(z_{ik}-\mu_i^+)(z_{ik}-\mu_i^+)^T\right\} \Bigg/ \left\{\sum_{k=1}^{N}\frac{\alpha_i^c p_i(x_k|\phi_i^c)}{p(x_k|\Phi^c)} + K_i\right\},$$

which is in turn bounded below by

$$\frac{1}{N+K_i}\left\{\sum_{k=1}^{K_i}(z_{ik}-\bar{z})(z_{ik}-\bar{z})^T\right\},$$

where

$$\bar{z} = \frac{1}{K_i}\sum_{k=1}^{K_i} z_{ik}.$$

Now this last matrix is positive definite with probability 1 whenever $K_i > n$. Consequently, if $K_i > n$, then with probability 1, the elements of a sequence $\{\Sigma_i^{(j)}\}_{j=0,1,2,\ldots}$ produced by the EM algorithm are bounded below by a positive definite matrix; hence, such a sequence cannot have singular matrices as limit points.

6. Performance of the EM algorithm. In this concluding section, we review and summarize features of the EM algorithm having to do with its effectiveness in practice on mixture density estimation problems. As always, it is understood that a parametric family of mixture densities of the form (1.1) is of interest, and that a particular $\Phi^* = (\alpha_1^*, \cdots, \alpha_m^*, \phi_1^*, \cdots, \phi_m^*)$ is the "true" parameter value to be estimated.

In order to provide some perspective, we begin by offering a brief description of the most basic forms of several alternative methods for numerically approximating maximum-likelihood estimates. In describing these methods, it is assumed for convenience that the sample at hand is a sample $S_1 = \{x_k\}_{k=1,\ldots,N}$ of Type 1 described in §3 and that one can write Φ as a vector $\Phi = (\xi_1, \cdots, \xi_v)^T$ of unconstrained scalar parameters at points of interest in Ω. Each of the methods to be described seeks a maximum-likelihood estimate by attempting to determine a point $\hat{\Phi}$ such that

$$\nabla_\Phi L_1(\hat{\Phi}) = 0, \tag{6.1}$$

where $L_1(\Phi)$ is the log-likelihood function given by (3.1). The features of the methods which concern us here are their speed of convergence, the computation and storage required for their implementation, and the extent to which their basic forms need to be modified in order to make them effective and trustworthy in practice.

The first of the alternative methods to be described is Newton's method. It is the method on which all but the last of the other methods reviewed here are modeled, and it is as follows: Given a current approximation Φ^c of a solution of (6.1), determine a next approximation Φ^+ by

$$\Phi^+ = \Phi^c - H(\Phi^c)^{-1}\nabla_\Phi L_1(\Phi^c). \tag{6.2}$$

The function $H(\Phi)$ in (6.2) is the Hessian matrix of $L_1(\Phi)$ given by (3.10).

Under reasonable assumptions on $L_1(\Phi)$, one can show that a sequence of iterates $\{\Phi^{(j)}\}_{j=0,1,2,\ldots}$ produced by Newton's method enjoys quadratic local convergence to a solution $\hat{\Phi}$ of (6.1) (see, for example, Ortega and Rheinboldt [107], or Dennis and Schnabel [42]). This is to say that, given a norm $\|\cdot\|$ on Ω, there is a constant β such that if $\Phi^{(0)}$ is sufficiently near $\hat{\Phi}$, then an inequality

$$\|\Phi^{(j+1)} - \hat{\Phi}\| \leq \beta \|\Phi^{(j)} - \hat{\Phi}\|^2 \tag{6.3}$$

holds for $j = 0, 1, 2, \cdots$. Quadratic convergence is ultimately very fast, and it is regarded as the major strength of Newton's method. Unfortunately, there are aspects of Newton's method which are associated with potentially severe problems in some applications. For one thing, Newton's method requires at each iteration the computation of the $v \times v$ Hessian matrix and the solution of a system of v linear equations (at a cost of $O(v^3)$ arithmetic operations in general) with this Hessian as the coefficient matrix; thus the computation required for an iteration of Newton's method is likely to become expensive very rapidly as m, n and N grow large. (It should also be mentioned that one must allow for the storage of the Hessian or some set of factors of it.) For another thing, Newton's method in its basic form (6.2) requires for some problems an impractically accurate initial approximate solution $\Phi^{(0)}$ in order for a sequence of iterates $\{\Phi^{(j)}\}_{j=0,1,2,\ldots}$ to converge to a solution of (6.1). Consequently, in order to be regarded as an algorithm which is safe and effective on applications of interest, the basic form (6.2) is likely to require augmentation with some procedure for enhancing the global convergence behavior of sequences of iterates produced by it. Such a procedure should be designed to insure that a sequence of iterates not only converges, but also does not converge to a solution of (6.1) which is not a local maximizer of $L_1(\Phi)$.

A broad class of methods which are based on Newton's method are quasi-Newton methods of the general form

$$\Phi^+ = \Phi^c - B^{-1}\nabla_\Phi L_1(\Phi^c), \tag{6.4}$$

in which B is regarded as an approximation of $H(\Phi^c)$. Methods of the form (6.4) which are particularly successful are those in which the approximation $B \approx H(\Phi^c)$ is maintained by doing a *secant update* of B at each iteration (see Dennis and Moré [41] or Dennis and Schnabel [42]). In the applications of interest here, such updates are typically realized as rank-one or (more likely) rank-two changes in B. Methods employing such updates have the advantages over Newton's method of not requiring the evaluation of the Hessian matrix at each iteration and of being implementable in ways which require only $O(v^2)$ arithmetic operations to solve the system of v linear equations at each iteration. The price paid for these advantages is that the full quadratic convergence of Newton's method is lost; rather, under reasonable assumptions on $L_1(\Phi)$, a sequence of iterates $\{\Phi^{(j)}\}_{j=0,1,2,\ldots}$ produced by one of these methods can only be shown to exhibit local superlinear convergence to a solution $\hat{\Phi}$ of (6.1), i.e., one can only show that if a norm $\|\cdot\|$ on Ω is given and if $\Phi^{(0)}$ is sufficiently near $\hat{\Phi}$ (and an initial approximate Hessian $B^{(0)}$ is sufficiently near $H(\hat{\Phi})$), then there exists a sequence $\{\beta_j\}_{j=0,1,2,\ldots}$ which converges to zero and is such

that

$$\|\Phi^{(j+1)} - \hat{\Phi}\| \leqq \beta_j \|\Phi^{(j)} - \hat{\Phi}\|$$

for $j = 0, 1, 2, \cdots$. Like Newton's method, methods of the general form (6.4), including those employing secant updates, are likely to require augmentation with safeguards to enhance global convergence properties and to insure that iterates do not converge to solutions of (6.1) which are not local maximizers of $L_1(\Phi)$.

There is a particular method of the form (6.4) which is specifically formulated for solving likelihood equations. This is the method of scoring, mentioned earlier in connection with the work of Rao [119] and reviewed in a general setting by Kale [84], [85]. (Kale [84], [85] also discusses modifications of Newton's method and the method of scoring in which the Hessian matrix or an approximation of it is held fixed for some number of iterations in the hope of reducing overall computational effort.) In the method of scoring, one ideally chooses B in (6.4) to be

$$B = -NI(\Phi^c), \tag{6.5}$$

where $I(\Phi)$ is the Fisher information matrix given by (2.5.1). Since the computation of $I(\Phi^c)$ is likely to be prohibitively expensive for most mixture density problems, a more appealing choice of B than (6.5) might be the sample approximation

$$B = -\sum_{k=1}^{N} [\nabla_\Phi \log p(x_k|\Phi^c)][\nabla_\Phi \log p(x_k|\Phi^c)]^T. \tag{6.6}$$

The choice (6.6) can be justified in the following manner: The Hessian $H(\Phi)$ is given by

$$\begin{aligned} H(\Phi) = &-\sum_{k=1}^{N} [\nabla_\Phi \log p(x_k|\Phi)][\nabla_\Phi \log p(x_k|\Phi)]^T \\ &+ \sum_{k=1}^{N} \frac{1}{p(x_k|\Phi)} \nabla_\Phi \nabla_\Phi^T p(x_k|\Phi). \end{aligned} \tag{6.7}$$

Now the second sum in (6.7) has zero expectation at $\Phi = \Phi^*$; furthermore, since the terms $\nabla_\Phi \log p(x_k|\Phi)$ must be computed in order to obtain $\nabla_\Phi L_1(\Phi)$, the first sum in (6.7) is available at the cost of only $O(Nv^2)$ arithmetic operations, while determining the second sum is likely to involve a great deal more expense. Thus (6.6) is a choice of B which is readily available at relatively low cost and which is likely to constitute a major part of $H(\Phi^c)$ when N is large and Φ^c is near Φ^*. It is clear from this discussion that the method of scoring with B given by (6.6) is an analogue for general maximum-likelihood estimation of the Gauss–Newton method for nonlinear least-squares problems (see Ortega and Rheinboldt [107]). If the computation of $I(\Phi^c)$ is not too expensive, then the choice of B given by (6.5) can be justified in much the same way.

The method of scoring in its basic form requires $O(Nv^2)$ arithmetic operations to evaluate B given by (6.6) and $O(v^3)$ arithmetic operations to solve the system of v linear equations implicit in (6.4). Since these $O(Nv^2)$ arithmetic operations are likely to be considerably less expensive than the evaluation of the full Hessian given by (6.7), the cost of computation per iteration of the method of scoring lies between that of a quasi-Newton method employing a low-rank secant update and that of Newton's method. Under reasonable assumptions on $L_1(\Phi)$, one can show that, with probability 1, if a solution $\hat{\Phi}$ of (6.1) is sufficiently near Φ^* and if N is sufficiently large, then a sequence of iterates $\{\Phi^{(j)}\}_{j=0,1,2,\ldots}$ generated by the method of scoring with B given by either (6.5) or (6.6) exhibits local linear convergence to $\hat{\Phi}$; i.e., there is a norm $\|\cdot\|$ on Ω and a constant λ,

$0 \leqq \lambda < 1$, for which

$$\|\Phi^{(j+1)} - \hat{\Phi}\| \leqq \lambda \|\Phi^{(j)} - \hat{\Phi}\|, \qquad j = 0, 1, 2, \cdots, \tag{6.8}$$

whenever $\Phi^{(0)}$ is sufficiently near $\hat{\Phi}$. If $\hat{\Phi}$ is very near Φ^* and if N is very large, then this convergence should be fast, i.e., (6.8) should hold for a small constant λ. Like Newton's method and all methods of the general form (6.4), the method of scoring is likely to require augmentation with global convergence safeguards in order to be considered trustworthy and effective.

The last methods to be touched on in this review are conjugate gradient methods. These methods are conceptually quite different from Newton's method and have the particular virtue of not requiring the storage or maintenance of an approximate Hessian. Thus they may be the methods of choice when the number of independent variables is so large that storage is at a premium. The original conjugate gradient method was derived to solve linear equations with positive-definite symmetric coefficient matrices, and a number of extensions and variations of this basic method now exist for both linear and nonlinear problems. Precise general theoretical convergence properties of conjugate gradient methods are not easily stated, but in practice they can be regarded as nearly always linearly convergent for nonlinear problems (see Gill, Murray and Wright [53]). In the nonlinear case, the amount of work required for an iteration is difficult to predict because an approximate maximizer or minimizer must be found along a direction of search. Overall, they appear to be less efficient on small problems than methods modeled after Newton's method, but potentially of great value (and perhaps the only useful methods) for very large problems. The reader is referred to Ortega and Rheinboldt [107] and Gill, Murray and Wright [53] for an expanded discussion of the forms and properties of conjugate gradient methods. Additional references are offered by Dennis and Schnabel [42], as well as by [107] and [53].

Having reviewed the above alternative methods, we return now to the EM algorithm and summarize its attractive features. Its most appealing general property is that it produces sequences of iterates on which the log-likelihood function increases monotonically. This monotonicity is the basis of the general convergence theorems of §4, and these theorems reinforce a large body of empirical evidence to the effect that the EM algorithm does not require augmentation with elaborate safeguards, such as those necessary for Newton's method and quasi-Newton methods, in order to produce iteration sequences with good global convergence characteristics. To be clear on the meaning of "good global convergence characteristics," let us specifically say that barring very bad luck or some local pathology in the log-likelihood function, one can expect an EM iteration sequence to converge to a local maximizer of the log-likelihood function. We do not intend to suggest that the EM algorithm is especially adept at finding global maximizers, but at this stage of the art, there are no well tested, general purpose algorithms which can reliably and efficiently find global maximizers. There has been a good bit of work directed toward the development of such algorithms, however; see Dixon and Szegö [44].

More can be said about the EM algorithm for mixtures of densities from exponential families under the assumption that $\phi_1, \cdots, \phi_m$ are mutually independent variables. One sees from (4.5) and (5.3) and similar expressions for samples of types other than Type 1 that it is unlikely that any other algorithm would be nearly as easy to encode on a computer or would require as little storage. In view of (4.5) and (5.3), it also seems that any constraints on Φ are likely to be satisfied, or at least nearly satisfied for large samples. For example, it is clear from (4.9) that each Σ_i^+ generated by the EM algorithm for a mixture of multivariate normal densities is symmetric and, with probability 1, positive-

definite whenever $N > n$. Certainly the mixing proportions generated by (4.5) are always nonnegative and sum to 1. It is also apparent from (4.5) and (5.3) that the computational cost of each iteration of the EM algorithm is low compared to that of the alternative methods reviewed above. In the case of a mixture of multivariate normal densities, for example, the EM algorithm requires $O(mn^2N)$ arithmetic operations per iteration, compared to at least $[O_1(m^2n^4N) + O_2(m^3n^6)]$ for Newton's method and the method of scoring and $[O_1(mn^2N) + O_2(m^2n^4)]$ for a quasi-Newton method employing a low-rank secant update. (All of these methods require the same number of exponential function evaluations per iteration.) Arithmetic per iteration for the three latter methods can, of course, be reduced by retaining a fixed approximate Hessian for some number of iterations at the risk of increasing the total number of iterations.

In spite of these attractive features, the EM algorithm can encounter problems in practice. The source of the most serious practical problems associated with the algorithm is the speed of convergence of sequences of iterates generated by it, which can often be annoyingly or even hopelessly slow. In the case of mixtures of densities from exponential families, Theorem 5.2 suggests that one can expect the convergence of EM iteration sequences to be linear, as opposed to the (very fast) quadratic convergence associated with Newton's method, the (fast) superlinear convergence associated with a quasi-Newton method employing a low-rank secant update, the (perhaps fast) linear convergence of the method of scoring, and the linear convergence of conjugate gradient methods. The discussion following Theorem 5.2 suggests further that the speed of this linear convergence depends in a certain sense on the separation of the component populations in the mixture. To demonstrate the speed of this linear convergence and its dependence on the separation of the component populations, we again consider the example of a mixture of two univariate normal densities (see (1.3) and (1.4)).

Table 6.1 summarizes the results of a numerical experiment involving a mixture of two univariate normal densities for the choices of Φ^* appearing in Table 3.3. (These choices were obtained as before by taking $\alpha_1^* = .3$, $\sigma_1^{2*} = \sigma_2^{2*} = 1$, and varying the mean separation $\mu_1^* - \mu_2^*$. For convenience, we took $\mu_2^* = -\mu_1^*$.) In this experiment, a Type 1 sample of 1000 observations on the mixture was generated for each choice of Φ^*; and a sequence of iterates was produced by the EM algorithm (see (4.5), (4.8) and (4.9)) from starting values $\alpha_1^{(0)} = \alpha_2^{(0)} = .5$, $\mu_1^{(0)} = 1.5\mu_1^*$, $\mu_2^{(0)} = 1.5\mu_2^*$ and $\sigma_1^{2(0)} = \sigma_2^{2(0)} = .5$. An accurate determination of the limit of the sequence was made in each case, and observations were made of the iteration numbers at which various degrees of accuracy were first obtained. These iteration numbers are recorded in Table 6.1 beneath the corresponding degrees of accuracy; in the table, "E" denotes the largest absolute value of the components of the difference between the indicated iterate and the limit. In addition, the spectral radius of the derivative of the EM iteration function at the limit was calculated in each case (cf. Theorem 5.2 and the following discussion). These spectral radii, appearing in the column headed by "ρ" in Table 6.1, provide quantitative estimates of the factors by which errors are reduced from one iteration to the next in each case. Finally, to give an idea of the point in an iteration sequence at which numerical error first begins to affect the theoretical performance of the algorithm, we observed in each case the iteration numbers at which loss of monotonicity of the log-likelihood function first occurred; these iteration numbers appear in Table 6.1 in the column headed by "LM."

In preparing Table 6.1, all computing was done in double precision on an IBM 3032. Eigenvalues were calculated with EISPACK subroutines TREDI and TQLI, and normally distributed data was obtained by transforming uniformly distributed data generated by the subroutine URAND of Forsythe, Malcolm and Moler [48] based on suggestions of Knuth [90].[3]

TABLE 6.1

Results of applying the EM *algorithm to a problem involving a Type* 1 *sample on a mixture of two univariate normal densities with* $\alpha_1^* = .3, \sigma_1^{2*} = \sigma_2^{2*} = 1$.

$\mu_1^* - \mu_2^*$	$E < 10^{-1}$	$E < 10^{-2}$	$E < 10^{-3}$	$E < 10^{-4}$	$E < 10^{-5}$	$E < 10^{-6}$	$E < 10^{-7}$	$E < 10^{-8}$	LM	ρ
0.2	2078	2334	2528	2717	2906	3095	3283	3472	3056	.9879
0.5	710	852	985	1117	1249	1381	1513	1643	1361	.9827
1.0	349	442	526	610	693	777	861	949	779	.9728
1.5	280	414	537	660	783	906	1028	1151	887	.9814
2.0	126	281	432	582	732	883	1033	1183	846	.9849
3.0	2	31	62	93	124	155	185	216	173	.9280
4.0	1	6	16	25	35	44	54	63	55	.7864
6.0	1	1	2	3	4	5	7	8	8	.2143

A number of comments about the contents of Table 6.1 are in order. First, it is clear from the table that an exorbitantly large number of EM iterations may be required to obtain a very accurate numerical approximation of the maximum-likelihood estimate if the sample is from a mixture of poorly separated component populations. However, in such a case, one sees from Table 3.3 that the variance of the estimate is likely to be such that it may be pointless to seek very much accuracy in a numerical approximation. Second, we remark on the pleasing consistency between the computed values of the spectral radius of the derivative of the EM iteration function and the differences between the iteration numbers needed to obtain varying degrees of accuracy. What we have in mind is the following: If the errors among the members of a linearly convergent sequence are reduced more or less by a factor of ρ, $0 \leq \rho < 1$, from one iteration to the next, then the number of iterations Δk necessary to obtain an additional decimal digit of accuracy is given approximately by $\Delta k \approx -\log 10/\log \rho$. This relationship between Δk and ρ is borne out very well in Table 6.1. This fact strongly suggests that after a number of EM iterations have been made, the errors in the iterates lie almost entirely in the eigenspace corresponding to the dominant eigenvalue of the derivative of the EM iteration function. We take this as evidence that one might very profitably apply simple relaxation-type acceleration procedures such as those of Peters and Walker [113], [114] and Redner [120] to sequences of iterates generated by the EM algorithm.

Third, in all of the cases listed in Table 6.1 except one, we observed that over 95 percent of the change in the log-likelihood function between the starting point and the limit of the EM iteration sequence was realized after only five iterations, regardless of the number of iterations ultimately required to approximate the limit very closely. (The exceptional case is that in which $\mu_1^* - \mu_1^* = 1.0$; in that case, about 83 percent of the change in the log-likelihood function was observed after five iterations.) This suggests to us that even when the component populations in a mixture are poorly separated, the EM algorithm can be expected to produce in a very small number of iterations parameter values such that the mixture density determined by them reflects the sample data very

[3]We are grateful to the Mathematics and Statistics Department of the University of New Mexico for providing the computing support for the generation of Table 6.1.

well. Fourth, it is evident from Table 6.1 that elements of an EM iteration sequence continue to make steady progress toward the limit even after numerical error has begun to interfere with the theoretical properties of the algorithm.

Fifth, the apparently anomalous decrease in ρ occurring when $\mu_1^* - \mu_2^*$ decreases from 2.0 to 1.0 happened concurrently with the iteration sequence limit of the proportion of the first population in the mixture becoming very small. (Such very small limit proportions continued to be observed in the cases $\mu_1^* - \mu_2^* = 0.5, 0.2$.) We do not know whether this decrease in the limit proportion of the first population indicates a sudden movement of the maximum-likelihood estimate as $\mu_1^* - \mu_2^*$ drops below 2.0, or whether the iteration sequence limit is something other than the maximum-likelihood estimate in the cases in which $\mu_1^* - \mu_2^*$ is less than 2.0. Finally, we also conducted more than 60 trials similar to those reported in Table 6.1, except with samples of 200 rather than 1000 generated observations on the mixture. The results were comparable to those given in Table 6.1. It should be mentioned, however, that the EM iteration sequences obtained using samples of 200 observations did occasionally converge to "singular solutions," i.e., limits associated with zero component variances. Convergence to such "singular solutions" did not occur among the relatively small number of trials involving samples of 1000 observations.

At present, the EM algorithm is being widely applied, not only to mixture density estimation problems, but also to a wide variety of other problems as well. We would like to conclude this survey with a little speculation about the future of the algorithm. It seems likely that the EM algorithm in its basic form will find a secure niche as an algorithm useful in situations in which some resources are limited. For example, the limited time which an experimenter can afford to spend writing programs coupled with a lack of available library software for safely and efficiently implementing competing methods could make the simplicity and reliability of the EM algorithm very appealing. Also, the EM algorithm might be very well suited for use on small computers for which limitations on program and data storage are more stringent than limitations on computing time.

Although meaningful comparison tests have not yet been made, it seems doubtful to us that the unadorned EM algorithm can be competitive as a general tool with well-designed general optimization algorithms such as those implemented in good currently available software library routines. Our doubt is based on the intolerably slow convergence of sequences of iterates generated by the EM algorithm in some applications. On the other hand, it is entirely possible that the EM algorithm could be modified to incorporate procedures for accelerating convergence, and that such modification would enhance its competitiveness. It is also possible that an effective hybrid algorithm might be constructed which first takes advantage of the good global convergence properties of the EM algorithm by using it initially and then exploits the rapid local convergence of Newton's method or one of its variants by switching to such a method later. Our feeling is that time might well be spent on research addressing these possibilities.

REFERENCES

[1] M. A. ACHESON AND E. M. MCELWEE, *Concerning the reliability of electron tubes,* The Sylvania Technologist, 4 (1951), pp. 105–116.

[2] J. AITCHISON AND S. D. SILVEY, *Maximum-likelihood estimation of parameters subject to restraints,* Ann. Math. Statist., 3 (1958), pp. 813–828.

[3] J. A. ANDERSON, *Multivariate logistic compounds,* Biometrika, 66 (1979), pp. 17–26.

[4] N. ARLEY AND K. R. BUCH, *Introduction to the Theory of Probability and Statistics,* John Wiley, New York, 1950.

[5] G. A. BAKER, *Maximum-likelihood estimation of the ratio of the two components of non-homogeneous populations,* Tohoku Math. J., 47 (1940), 304–308.

[6] O. BARNDORFF-NIELSEN, *Identifiability of mixtures of exponential families,* J. Math. Anal. Appl., 12 (1965), pp. 115–121.

[7] ———, *Information and Exponential Families in Statistical Theory,* John Wiley, New York, 1978.

[8] L. E. BAUM, T. PETRIE, G. SOULES AND N. WEISS, *A maximization technique occuring in the statistical analysis of probabilistic functions of Markov chains,* Ann. Math. Statist., 41 (1970), pp. 164–171.

[9] J. BEHBOODIAN, *On a mixture of normal distributions,* Biometrika, 57 Part 1 (1970), 215–217.

[10] ———, *Information matrix for a mixture of two normal distributions,* J. Statist. Comput. Simul., 1 (1972), pp. 295–314.

[11] C. T. BHATTACHARYA, *A simple method of resolution of a distribution into Gaussian components,* Biometrics, 23 (1967), pp. 115–137.

[12] W. R. BLISCHKE, *Moment estimators for the parameters of a mixture of two binomial distributions,* Ann. Math. Statist., 33 (1962), pp. 444–454.

[13] ———, *Mixtures of discrete distributions,* in Proc. of the International Symposium on Classical and Contagious Discrete Distributions, Pergamon Press, New York, 1963, pp. 351–372.

[14] ———, *Estimating the parameters of mixtures of binomial distributions,* J. Amer. Statist. Assoc., 59 (1964), pp. 510–528.

[15] D. C. BOES, *On the estimation of mixing distributions,* Ann. Math. Statist., 37 (1966), pp. 177–188.

[16] ———, *Minimax unbiased estimator of mixing distribution for finite mixtures,* Sankhyā A, 29 (1967), pp. 417–420.

[17] K. O. BOWMAN AND L. R. SHENTON, *Space of solutions for a normal mixture,* Biometrika, 60 (1973), pp. 629–636.

[18] R. A. BOYLES, *On the convergence of the* EM *algorithm,* J. Royal Statist. Soc. Ser. B, 45 (1983), pp. 47–50.

[19] C. BURRAU, *The half-invariants of the sum of two typical laws of errors, with an application to the problem of dissecting a frequency curve into components,* Skand. Aktuarietidskrift, 17 (1934), pp. 1–6.

[20] R. M. CASSIE, *Some uses of probability paper in the analysis of size frequency distributions,* Austral. J. Marine and Freshwater Res., 5 (1954), pp. 513–523.

[21] R. CEPPELINI, S. SINISCALCO AND C. A. B. SMITH, *The estimation of gene frequencies in a random-mating population,* Ann. Human Genetics, 20 (1955), pp. 97–115.

[22] K. C. CHANDA, *A note on the consistency and maxima of the roots of the likelihood equations,* Biometrika, 41 (1954), pp. 56–61.

[23] W. C. CHANG, *The effects of adding a variable in dissecting a mixture of two normal populations with a common covariance matrix,* Biometrika 63, (1976), pp. 676–678.

[24] ———, *Confidence interval estimation and transformation of data in a mixture of two multivariate normal distributions with any given large dimension,* Technometrics, 21 (1979), pp. 351–355.

[25] C. V. L. CHARLIER, *Researches into the theory of probability,* Acta Univ. Lund. (Neue Folge. Abt. 2), 1 (1906), pp. 33–38.

[26] C. V. L. CHARLIER AND S. D. WICKSELL, *On the dissection of frequency functions,* Arkiv. for Matematik Astronomi Och Fysik, 18 (6) (1924), Stockholm.

[27] T. CHEN, *Mixed-up frequencies in contingency tables,* Ph.D. dissertation, Univ. of Chicago, Chicago, 1972.

[28] K. CHOI, *Estimators for the parameters of a finite mixture of distributions,* Ann. Inst. Statist. Math., 21 (1969), pp. 107–116.

[29] K. CHOI AND W. B. BULGREN, *An estimation procedure for mixtures of distributions,* J. Royal Statist. Soc. Ser. B, 30 (1968), pp. 444–460.

[30] A. C. COHEN, Jr., *Estimation in mixtures of discrete distributions,* in Proc. of the International Symposium on Classical and Contagious Discrete Distributions, Pergamon Press, New York, 1963, pp. 351–372.

[31] ———, *A note on certain discrete mixed distributions,* Biometrics, 22 (1966), pp. 566–571.

[32] ———, *Estimation in mixtures of two normal distributions,* Technometrics, 9 (1967), pp. 15–28.

[33] *Communications in Statistics,* Special Issue on Remote Sensing, Comm. Statist. Theor. Meth., A5 (1976).

[34] P. W. COOPER, *Some topics on nonsupervised adaptive detection for multivariate normal distributions,* in Computer and Information Sciences, II, J. T. Tou, ed., Academic Press, New York, 1967, pp. 143–146.

[35] D. R. COX, *The analysis of exponentially distributed lifetimes with two types of failure,* J. Royal Statist. Soc., 21 B (1959), pp. 411–421.

[36] H. CRAMÉR, *Mathematical Methods of Statistics,* Princeton Univ. Press, Princeton, NJ, 1946.
[37] N. E. DAY, *Estimating the components of a mixture of normal distributions,* Biometrika, 56 (1969), pp. 463–474.
[38] J. J. DEELY AND R. L. KRUSE, *Construction of sequences estimating the mixing distribution,* Ann. Math. Statist., 39 (1968), pp. 286–288.
[39] A. P. DEMPSTER, N. M. LAIRD AND D. B. RUBIN, *Maximum-likelihood from incomplete data via the* EM *algorithm,* J. Royal Statist. Soc. Ser. B (methodological), 39 (1977), pp. 1–38.
[40] J. E. DENNIS, JR., *Algorithms for nonlinear fitting,* in Proc. of the NATO Advanced Research Symposium, Cambridge Univ., Cambridge, England, July, 1981.
[41] J. E. DENNIS, JR., AND J. J. MORÉ, *Quasi-Newton methods, motivation and theory,* this Review, 19 (1977), pp. 46–89.
[42] J. E. DENNIS, JR., AND R. B. SCHNABEL, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations,* Prentice-Hall, Englewood Cliffs, NJ, 1983.
[43] N. P. DICK AND D. C. BOWDEN, *Maximum-likelihood estimation for mixtures of two normal distributions,* Biometrics, 29 (1973), pp. 781–791.
[44] L. C. W. DIXON AND G. P. SZEGÖ, *Towards Global Optimization,* Vols. 1, 2, North-Holland, Amsterdam, 1975, 1978.
[45] G. DOETSCH, *Zerlegung einer Funktion in Gausche Fehlerkurven und zeitliche Zuruckverfolgung eines Temperaturzustandes,* Math. Z., 41 (1936), pp. 283–318.
[46] R. O. DUDA AND P. E. HART, *Pattern Classification and Scene Analysis,* John Wiley, New York, 1973.
[47] B. S. EVERITT AND D. J. HAND, *Finite Mixture Distributions,* Chapman and Hall, London, 1981.
[48] G. E. FORSYTHE, M. A. MALCOLM AND C. B. MOLER, *Computer Methods for Mathematical Computations,* Prentice-Hall, Englewood Cliffs, NJ 1977.
[49] E. B. FOWLKES, *Some methods for studying the mixture of two normal (log normal) distributions,* J. Amer. Statist. Assoc., 74 (1979), pp. 561–575.
[50] J. G. FRYER AND C. A. ROBERTSON, *A comparison of some methods for estimating mixed normal distributions,* Biometrika, 59 (1972), pp. 639–648.
[51] K. FUKUNAGA, *Introduction to Statistical Pattern Recognition,* Academic Press, New York, 1972.
[52] S. GANESALINGAM AND G. J. MCLACHLAN, *Some efficiency results for the estimation of the mixing proportion in a mixture of two normal distributions,* Biometrics, 37 (1981), pp. 23–33.
[53] P. E. GILL, W. MURRAY AND M. H. WRIGHT, *Practical Optimization,* Academic Press, London, 1981.
[54] L. A. GOODMAN, *The analysis of systems of qualitative variables when some of the variables are unobservable: Part* I-A *modified latent structure approach,* Amer. J. Sociol., 79 (1974), pp. 1179–1259.
[55] V. H. GOTTSCHALK, *Symmetric bimodal frequency curves,* J. Franklin Inst., 245 (1948), pp. 245–252.
[56] J. GREGOR, *An algorithm for the decomposition of a distribution into Gaussian components,* Biometrics, 25 (1969), pp. 79–93.
[57] N. T. GRIDGEMAN, *A comparison of two methods of analysis of mixtures of normal distributions,* Technometrics, 12 (1970), pp. 823–833.
[58] E. J. GUMBEL, *La dissection d'une repartition,* Annales de l'Universite de Lyon, (3), A (1939), pp. 39–51.
[59] A. K. GUPTA AND T. MIYAWAKI, *On a uniform mixture model,* Biometrical J., 20(1978), pp. 631–638.
[60] L. F. GUSEMAN, JR., AND J. R. WALTON, *An application of linear feature selection to estimation of proportions,* Comm. Statist. Theor. Meth., A6 (1977), pp. 611–617.
[61] ———, *Methods for estimating proportions of convex combinations of normals using linear feature selection,* Comm. Statist. Theor. Meth., A7 (1978), pp. 1439–1450.
[62] S. J. HABERMAN, *Log-linear models for frequency tables derived by indirect observations: Maximum-likelihood equations,* Ann. Statist., 2 (1974), pp. 911–924.
[63] ———, *Iterative scaling procedures for log-linear models for frequency tables derived by indirect observation,* Proc. Amer. Statist. Assoc. (Statist. Comp. Sect. 1975), (1976), pp. 45–50.
[64] ———, *Product models for frequency tables involving indirect observation,* Ann. Statist., 5 (1977), pp. 1124–1147.
[65] A. HALD, *The compound hypergeometric distribution and a system of single sampling inspection plans based on prior distributions and costs,* Technometrics, 2 (1960), pp. 275–340.
[66] P. HALL, *On the non-parametric estimation of mixture proportions,* J. Royal Statist. Soc. Ser. B, 43 (1981), pp. 147–156.
[67] J. P. HARDING, *The use of probability paper for the graphical analysis of polynomial frequency distributions,* J. Marine Biological Assoc., 28 (1949), pp. 141–153.
[68] J. A. HARTIGAN, *Clustering Algorithms,* John Wiley, New York, 1975.
[69] M. J. HARTLEY, *Comment on* [118], J. Amer. Statist. Assoc., 73 (1978), pp. 738–741.

[70] V. HASSELBLAD, *Estimation of parameters for a mixture of normal distributions,* Technometrics, 8 (1966), pp. 431–444.

[71] ———, *Estimation of finite mixtures of distributions from the exponential family,* J. Amer. Statist. Assoc., 64 (1969), pp. 1459–1471.

[72] R. J. HATHAWAY, *Constrained maximum-likelihood estimation for a mixture of m univariate normal distributions.* Statistics Tech. Rep. 92, 62F10-2, Univ. of South Carolina, Columbia, SC, 1983.

[73] B. M. HILL, *Information for estimating the proportions in mixtures of exponential and normal distributions,* J. Amer. Statist. Assoc., 58 (1963), pp. 918–932.

[74] D. W. HOSMER, JR., *On* MLE *of the parameters of a mixture of two normal distributions when the sample size is small,* Comm. Statist., 1 (1973), 217–227.

[75] ———, *A comparision of iterative maximum-likelihood estimates of the parameters of a mixture of two normal distributions under three different types of sample,* Biometrics, 29 (1973), pp. 761–770.

[76] ———, *Maximum-likelihood estimates of the parameters of a mixture of two regression lines,* Comm. Statist., 3 (1974), pp. 995–1006.

[77] ———, *Comment on* [118], J. Amer. Statist. Assoc., 73 (1978), pp. 741–744.

[78] D. W. HOSMER, JR., AND N. P. DICK, *Information and mixtures of two normal distributions,* J. Statist. Comput. Simul., 6 (1977), pp. 137–148.

[79] V. S. HUZURBAZAR, *The likelihood equation, consistency and the maxima of the likelihood function,* Annals of Eugenics London, 14 (1948), pp. 185–200.

[80] I. R. JAMES, *Estimation of the mixing proportion in a mixture of two normal distributions from simple, rapid measurements,* Biometrics, 34 (1978), pp. 265–275.

[81] S. JOHN, *On identifying the population of origin of each observation in a mixture of observations from two normal populations,* Technometrics, 12 (1970), pp. 553–563.

[82] ———, *On identifying the population of origin of each observation in a mixture of observations from two gamma populations,* Technometrics, 12 (1970), pp. 565–568.

[83] A. B. M. L. KABIR, *Estimation of parameters of a finite mixture of distributions,* J. Royal Statist. Soc. Ser. B (methodological), 30 (1968), pp. 472–482.

[84] B. K. KALE, *On the solution of the likelihood equation by iteration processes,* Biometrika, 48 (1961), pp. 452–456.

[85] ———, *On the solution of likelihood equations by iteration processes. The multiparametric case,* Biometrika, 49 (1962), pp. 479–486.

[86] D. KAZAKOS, *Recursive estimation of prior probabilities using a mixture,* IEEE Trans. Inform. Theory, IT-23 (1977), pp. 203–211.

[87] J. KIEFER AND J. WOLFOWITZ, *Consistency of the maximum-likelihood estimation in the presence of infinitely many incidental parameters,* Ann. Math. Statist., 27 (1956), pp. 887–906.

[88] N. M. KIEFER, *Discrete parameter variation: Efficient estimation of a switching regression model,* Econometrica, 46 (1978), pp. 427–434.

[89] ———, Comment on [118], J. Amer. Statist. Assoc., 73 (1978), pp. 744–745.

[90] D. E. KNUTH, *Seminumerical algorithms,* in The Art of Computer Programming, Vol. 2, Addison-Wesley, Reading, MA, (1969).

[91] K. D. KUMAR, E. H. NICKLIN AND A. S. PAULSON, *Comment on* [118], J. Amer. Statist. Assoc., 74 (1979), pp. 52–55.

[92] N. LAIRD, *Nonparametric maximum-likelihood estimation of a mixing distribution,* J. Amer. Statist. Assoc., 73 (1978), pp. 805–811.

[93] P. F. LAZARSFELD AND N. W. HENRY, *Latent Structure Analysis,* Houghton Mifflin Company, Boston, 1968.

[94] M. LOÉVE, *Probability Theory,* Van Nostrand, New York, 1963.

[95] D. G. LUENBERGER, *Optimization by Vector Space Methods,* John Wiley, New York, 1969.

[96] P. D. M. MACDONALD, *Comment on "An estimation procedure for mixtures of distribution" by Choi and Bulgren,* J. Royal Statist. Soc. Ser. B, 33 (1971), pp. 326–329.

[97] ———, *Estimation of finite distribution mixtures,* in Applied Statistics, R. P. Gupta, ed., North-Holland Publishing Co., Amsterdam, 1975.

[98] C. L. MALLOWS, review of [100], Biometrics, 18 (1962), p. 617.

[99] J. S. MARITZ, *Empirical Bayes Methods,* Methuen and Co., London, 1970.

[100] P. MEDGYESSY, *Decomposition of Superpositions of Distribution Functions,* Publishing House of the Hungarian Academy of Sciences, Budapest, 1961.

[101] W. MENDENHALL AND R. J. HADER, *Estimation of parameters of mixed exponentially distributed failure time distributions from censored life test data,* Biometrika, 45 (1958), pp. 504–520.

[102] H. MUENCH, *Probability distribution of protection test results,* J. Amer. Statist. Assoc., 31 (1936), pp. 677–689.

[103] ———, *Discrete frequency distributions arising from mixtures of several single probability values,* J. Amer. Statist. Assoc., 33 (1938), pp. 390–398.

[104] P. L. ODELL AND J. P. BASU, *Concerning several methods for estimating crop acreages using remote sensing data,* Comm. Statist. Theor. Meth., A5 (1976), pp. 1091–1114.

[105] P. L. ODELL AND R. CHHIKARA, *Estimation of a large area crop acreage inventory using remote sensing technology* in Annual Report: Statistical Theory and Methodology for Remote Sensing Data Analysis, Rep. NASA/JSC-09703, Univ. of Texas at Dallas, Dallas, TX, 1975.

[106] T. ORCHARD AND M. A. WOODBURY, *A missing information principle: theory and applications,* in Proc. of the 6th Berkeley Symposium on Mathematical Statistics and Probability, 1972, vol. 1, pp. 697–715.

[107] J. M. ORTEGA AND W. C. RHEINBOLDT, *Iterative Solution of Nonlinear Equations in Several Variables,* Academic Press, New York, 1970.

[108] A. M. OSTROWSKI, *Solutions of Equations and Systems of Equations,* Academic Press, New York, 1966.

[109] K. PEARSON, *Contributions to the mathematical theory of evolution,* Phil. Trans. Royal Soc., 185A (1894), pp. 71–110.

[110] ———, *On certain types of compound frequency distributions in which the components can be individually described by binomial series,* Biometrika, 11 (1915–17), pp. 139–144.

[111] K. PEARSON AND A. LEE, *On the generalized probable error in multiple normal correlation,* Biometrika, 6 (1908–09), pp. 59-68.

[112] B. C. PETERS, JR., AND W. A. COBERLY, *The numerical evaluation of the maximum-likelihood estimate of mixture proportions,* Comm. Statist. Theor. Meth., A5 (1976), pp. 1127–1135.

[113] B. C. PETERS, JR., AND H. F. WALKER, *An iterative procedure for obtaining maximum-likelihood estimates of the parameters for a mixture of normal distributions,* SIAM J. Appl. Math., 35 (1978), pp. 362–378.

[114] ———, *The Numerical evaluation of the maximum-likelihood estimate of a subset of mixture proportions,* SIAM J. Appl. Math., 35 (1978), pp. 447–452

[115] H. S. POLLARD, *On the relative stability of the median and the arithmetic mean, with particular reference to certain frequency distributions which can be dissected into normal distributions,* Ann. Math. Statist., 5 (1934), pp. 227–262.

[116] E. J. PRESTON, *A graphical method for the analysis of statistical distributions into two normal components,* Biometrika, 40 (1953), pp. 460–464.

[117] R. E. QUANDT, *A new approach to estimating switching regressions,* J. Amer. Statist. Assoc., 67 (1972), pp. 306–310.

[118] R. E. QUANDT AND J. B. RAMSEY, *Estimating mixtures of normal distributions and switching regressions,* J. Amer. Statist. Assoc., 73 (1978), pp. 730–738.

[119] C. R. RAO, *The utilization of multiple measurements in problems of biological classification.* J. Royal Statist. Soc. Ser. B, 10 (1948), pp. 159–193.

[120] R. A. REDNER, *An iterative procedure for obtaining maximum liklihood estimates in a mixture model,* Rep. SR-T1-04081, NASA Contract NAS9-14689, Texas A&M Univ., College Station, TX, Sept., 1980.

[121] ———, *Maximum-likelihood estimation for mixture models,* NASA Tech. Memorandum, to appear.

[122] ———, *Note on the consistency of the maximum-likelihood estimate for nonidentifiable distributions,* Ann. Statist., 9 (1981), pp. 225–228.

[123] P. R. RIDER, *The method of moments applied to a mixture of two exponential distributions,* Ann. Math. Statist., 32 (1961), pp. 143–147.

[124] ———, *Estimating the parameters of mixed Poisson, binomial and Weibull distributions by the method of moments,* Bull. Internat. Statist. Inst., 39 Part 2 (1962), pp. 225–232.

[125] C. A. ROBERTSON AND J. G. FRYER, *The bias and accuracy of moment estimators,* Biometrika, 57 Part 1 (1970), pp. 57–65.

[126] J. W. SAMMON, JR., *An adaptive technique for multiple signal detection and identification,* in Pattern Recognition, L. N. Kanal, ed., Thompson Book Co., London, 1968, pp. 409–439.

[127] W. SCHILLING, *A frequency distribution represented as the sum of two Poisson distributions,* J. Amer. Statist. Assoc., 42 (1947), pp. 407–424.

[128] L. SIMAR, *Maximum-liklihood estimation of a compound Poisson process,* Ann. Statist., 4 (1976), pp. 1200–1209.

[129] J. SITTIG, *Superpositie van twee frequentieverdelingen,* Statistica, 2 (1948), pp. 206–227.

[130] D. F. STANAT, *Unsupervised learning of mixtures of probability functions,* in Pattern Recognition, L. N. Kanal, ed., Thompson Book Co., London, 1968, pp. 357–389.

[131] G. W. STEWART, *Introduction to Matrix Computations,* Academic Press, New York, 1973.
[132] B. STRÖMGREN, *Tables and diagrams for dissecting a frequency curve into components by the half-invariant method,* Skand. Aktuarietidskrift, 17 (1934), pp. 7–54.
[133] R. SUNDBERG, *Maximum-likelihood theory and applications for distributions generated when observing a function of an exponential family variable,* Doctoral thesis, Inst. Math. Stat., Stockholm Univ., Stockholm, Sweden, 1972.
[134] ———, *Maximum liklihood theory for incomplete data from an exponential family,* Scand. J. Statist., 1 (1974), pp. 49–58.
[135] ———, *An iterative method for solution of the likelihood equations for incomplete data from exponential families,* Comm. Statist. Simulation Comput., B5, (1976), pp. 55–64.
[136] G. M. TALLIS AND R. LIGHT, *The use of fractional moments for estimating the parameters of a mixed exponential distribution,* Technometrics, 10 (1968), pp. 161–175.
[137] W. Y. TAN AND W. C. CHANG, *Convolution approach to genetic analysis of quantitative characters of self-fertilized population,* Biometrics, 28 (1972), pp. 1073–1090.
[138] ———, *Some comparisons of the method of moments and the method of maximum-likelihood in estimating parameters of a mixture of two normal densities,* J. Amer. Statist. Assoc., 67 (1972), pp. 702–708.
[139] R. D. TARONE AND G. GRUENHAGE, *A note on the uniqueness of roots of the likelihood equations for vector-valued parameters,* J. Amer. Statist. Assoc., 70 (1975), pp. 903–904.
[140] M. TARTER AND A. SILVERS, *Implementation and applications of bivariate Gaussian mixture decomposition,* J. Amer. Statist. Assoc., 70 (1975), pp. 47–55.
[141] H. TEICHER, *On the mixture of distributions,* Ann. Math. Statist., 31 (1960), pp. 55–73.
[142] ———, *Identifiability of mixtures,* Ann. Math. Statist., 32 (1961), pp. 244–248.
[143] ———, *Identifiability of finite mixtures,* Ann. Math. Statist., 34 (1963), pp. 1265–1269.
[144] ———, *Identifiability of mixtures of product measures,* Ann. Math. Statist., 38 (1967), pp. 1300–1302.
[145] D. M. TITTERINGTON, *Some problems with data from finite mixture distributions,* Mathematics Research Center, University of Wisconsin-Madison, Technical Summary Report 2369, Madison, WI, 1982.
[146] ———, *Minimum distance non-parametric estimation of mixture proportions,* J. Royal Statist. Soc. Ser. B, 45 (1983), pp. 37–46.
[147] J. D. TUBBS AND W. A. COBERLY, *An empirical sensitivity study of mixture proportion estimators,* Comm. Statist. Theor. Meth., A5 (1976), pp. 1115–1125.
[148] J. VAN RYZIN, ed., *Classification and Clustering:* Proc. of an Advanced Seminar Conducted by the Mathematics Research Center, The University of Wisconsin, Madison (May, 1976), Academic Press, New York, 1977.
[149] Y. VARDI, *Nonparametric estimation in renewal processes,* Ann. Statist., 10 (1982), pp. 772–785.
[150] A. WALD, *Note on the consistency of the maximum-likelihood estimate,* Ann. Math. Statist., 20 (1949), pp. 595–600.
[151] H. F. WALKER, *Estimating the proportions of two populations in a mixture using linear maps,* Comm. Statist. Theor. Meth. A9 (1980), pp. 837–849.
[152] K. WEICHSELBERGER, *Uber ein graphisches Verfahren zur Trennung von Mischverteilungen und zur Identifikation kupierter Normalverteilungen bie grossem Stichprobenumfang,* Metrika, 4 (1951), pp. 178–229.
[153] J. H. WOLFE, *Pattern clustering by multivariate mixture analysis,* Multivariate Behavioral Res., 5 (1970), pp. 329–350.
[154] J. WOLFOWITZ, *The minimum distance method,* Ann Math. Statist., 28 (1957), pp. 75–88
[155] C.-F. WU, *On the convergence of the* EM *algorithm,* Ann. Statist., (1983), to appear.
[156] S. J. YAKOWITZ, *A consistent estimator for the identification of finite mixtures,* Ann. Math. Stat., 40 (1969), pp. 1728–1735.
[157] ———, *Unsupervised learning and the identification of finite mixtures,* IEEE Trans. Inform. Theory, IT-16 (1970), pp. 330–338.
[158] S. J. YAKOWITZ AND J. D. SPRAGINS, *On the identifiability of finite mixtures,* Ann. Math. Statist., 39 (1968), pp. 209–214.
[159] T. Y. YOUNG AND T. W. CALVERT, *Classification, Estimation, and Pattern Recognition,* American Elsevier, New York, 1973.
[160] T. Y. YOUNG AND G. CORALUPPI, *Stochastic estimation of a mixture of normal density functions using an information criterion,* IEEE Trans. Inform. Theory, IT-16 (1970), pp. 258–263.
[161] S. ZACKS, *The Theory of Statistical Inference,* John Wiley, New York, 1971.
[162] W. ZANGWILL, *Nonlinear Programming: A Unified Approach,* Prentice-Hall, Englewood Cliffs, NJ (1969).