The Annals of Statistics
2003, Vol. 31, No. 1, 201–224

NONPARAMETRIC ESTIMATION OF COMPONENT DISTRIBUTIONS IN A MULTIVARIATE MIXTURE

BY PETER HALL AND XIAO-HUA ZHOU

Australian National University, and Australian National University and VA Puget Sound Health Care System and University of Washington

Suppose k-variate data are drawn from a mixture of two distributions, each having independent components. It is desired to estimate the univariate marginal distributions in each of the products, as well as the mixing proportion. This is the setting of two-class, fully parametrized latent models that has been proposed for estimating the distributions of medical test results when disease status is unavailable. The problem is one of inference in a mixture of distributions without training data, and until now it has been tackled only in a fully parametric setting. We investigate the possibility of using nonparametric methods. Of course, when $k = 1$ the problem is not identifiable from a nonparametric viewpoint. We show that the problem is "almost" identifiable when $k = 2$; there, the set of all possible representations can be expressed, in terms of any one of those representations, as a two-parameter family. Furthermore, it is proved that when $k \geq 3$ the problem is nonparametrically identifiable under particularly mild regularity conditions. In this case we introduce root-n consistent nonparametric estimators of the $2k$ univariate marginal distributions and the mixing proportion. Finite-sample and asymptotic properties of the estimators are described.

1. Introduction and summary. In the problem of determining accuracy of diagnostic tests, such as those associated with receiver operating characteristic (ROC) curves, it is clearly advantageous to know the true disease status (present or absent) for each patient, independently of the patient's test results. See, for example, Metz (1978). Such perfectly classified data would form so-called training samples that could be used in subsequent statistical analyses of the tests, for example, in studies of the distributions of test results.

In the case of many diseases, however, it is difficult or impossible to establish a definitive diagnosis. A perfect "gold standard," as it is often called, may not exist or may be too expensive or impractical to attain. This is especially true for complex medical conditions in the usual clinical practice setting, for example, the condition of myocardial infarction. In such contexts it has been suggested [see, e.g., Rindskopf and Rindskopf (1986) and Hui

Received May 2000; revised May 2002.

AMS 2000 subject classifications. Primary 62G05; secondary 62G70.

Key words and phrases. Biased bootstrap, distribution estimation, empirical likelihood, identification, latent model, multivariate analysis, nonparametric maximum likelihood, root-n consistency.

and Zhou (1998)] that a two-term multivariate mixture model, with the terms corresponding to "disease absent" and "disease present," respectively, be fitted to unclassified data on the results of diagnostic tests. Distributional properties of individual tests may thereby be determined directly, without the need for training samples.

In this approach the results of k tests applied to a particular patient are assumed to be stochastically independent, conditional on the disease status of the patient. Thus, the k-variate dataset $\mathcal{X} = \{X_1, \ldots, X_n\}$ is considered to have been drawn from a two-term mixture distribution,

$$F(x) = \pi \prod_{j=1}^{k} F_{j1}(x_j) + (1 - \pi) \prod_{j=1}^{k} F_{j2}(x_j). \tag{1.1}$$

The first term, say, corresponds to patients without the disease, the second to patients with the disease, and the jth component of the k-vector X_i represents the result of the jth test applied to the ith patient. In the representation at (1.1), π denotes the mixture proportion, F_{jr} is the univariate distribution function of the jth marginal in the rth population Π_r, and $x = (x_1, \ldots, x_k)$. We wish to estimate π and F_{jr}, for $1 \le j \le k$ and $r = 1, 2$, using only the data $\mathcal{X}$. Validity of the assumption of independent marginals will be discussed in Section 2.3.

In the present paper we consider nonparametric methods for solving this problem. Clearly, when $k = 1$ neither π nor F_{jr} is identifiable nonparametrically. However, the issue of identifiability is much less clear for $k \ge 2$. We shall show in Section 4 that when $k = 2$, π and F_{jr} are in fact not identifiable. Indeed, if $k = 2$ and F has the form (1.1) for a particular vector

$$Q = (\pi, F_{11}, \ldots, F_{k1}, F_{12}, \ldots, F_{k2}), \tag{1.2}$$

then there is in general a continuum of distinct Q's such that (1.1) holds for the same F on the left-hand side of (1.1). In a nonparametric setting the continuum of potential component distributions when $k = 2$ is determined, as a functional of the component distributions arising from any particular version of (1.1) for a fixed F, by two independent, scalar parameters.

However, for $k \ge 3$ and under mild regularity conditions, (1.1) does determine Q uniquely, up to the dichotomy obtained by interchanging the two products on the right-hand side of (1.1). Moreover, it is possible to consistently, and often root-n consistently, estimate all the components of Q using purely nonparametric methods. These results will be detailed in Section 4. They apparently do not have straightforward generalizations to mixtures of three or more products of independent components. However, in the medical diagnosis context of our work such cases do not seem to be as important as the two-product case.

The model at (1.1) can be considered to be for mixtures with "fixed effects." A density form of the random effects version of the model would be

$$f(x) = \pi \int \left\{ \prod_{j=1}^{k} f_{j1}(x_j|\lambda) \right\} \phi(\lambda)\, d\lambda + (1-\pi) \int \left\{ \prod_{j=1}^{k} f_{j2}(x_j|\lambda) \right\} \phi(\lambda)\, d\lambda, \tag{1.3}$$

where ϕ represents the density of the random effect Λ, λ is a realized value of λ, and for each λ and for $1 \le j \le k$ and $r = 1, 2$, $f_{jr}(\cdot|\lambda)$ is a univariate probability density. We may consider f and $f_{jr}(\cdot|\cdot)$ to be analogues, in the random effect case, of the densities of F and F_{jr}, the latter appearing at (1.1). We shall show that regardless of the value of k, and even if each ϕ and $f_{jr}(\cdot|\cdot)$ has a known parametric form, the parameters may not be consistently estimated from an infinite amount of data on f. Likewise, no matter how large the value of k, $f_{jr}(\cdot|\cdot)$ generally cannot be estimated using any nonparametric procedure. Our counterexamples illustrating these points are mixtures of distributions with mixture-distribution marginals, and are not pathological. Inference in the random effects case is becoming popular [see, e.g., Qu, Tan and Kutner (1996), Hadgu and Qu (1998) and Qu and Hadgu (1998)].

These and other theoretical results will be outlined in Section 4, with technical arguments deferred to Section 5. Our methods will be described in Section 2, and their numerical properties summarised in Section 3.

There is of course an extensive literature on parametric inference in mixture models. We mention here only the monographs of Titterington, Smith and Makov (1985), McLachlan and Basford (1988) and Everitt and Hand (1981), and some of the main types of estimators that have been proposed: maximum likelihood [e.g., Cohen (1967), Lindsay (1983a, b) and Redner and Walker (1984)], minimum chi-square [e.g., Day (1969)], method of moments [e.g., Lindsay and Basak (1993)], Bayesian approaches [e.g., Day (1969)], and techniques based on the moment generating function [e.g., Quandt and Ramsey (1978)].

The majority of existing nonparametric techniques rely at least in part on training data; see, for example, Murray and Titterington (1978), Hall (1981), Titterington (1983), Hall and Titterington (1984, 1985), Cerrito (1992), Shahshahani and Landgrebe (1994), Lancaster and Imbens (1996) and Qin (1998, 1999). Very little is known of the potential for consistent nonparametric inference in mixtures without training data, and that problem motivates the present paper.

On the subject of identifiability, Teicher (1967) has shown that in a parametric setting the k-variate model at (1.1) is identifiable if and only if each univariate submodel,

$$F_j(x_j) = \pi\, F_{j1}(x_j) + (1-\pi)\, F_{j2}(x_j), \qquad 1 \le j \le k,$$

is identifiable. However, this result does not apply in a nonparametric setting.

2. Methodology and assumptions.

2.1. *Estimation of marginal distributions.* We suggest a purely nonparametric method for estimating the vector Q at (1.2), describing the mixture proportion π and marginal distributions F_{jr} in the model at (1.1). To ensure the model is identifiable in a nonparametric sense we assume $k \geq 3$. See Section 4 for discussion of this issue.

Let $\mathcal{X} = \{X_1, \ldots, X_n\}$ denote data generated by the model at (1.1), and write $X_i = (X_{i1}, \ldots, X_{ik})$. Let p_{ijr}, for $1 \leq i \leq n$, $1 \leq j \leq k$ and $r = 1, 2$, be nonnegative numbers constrained to satisfy

$$p_{ijr} \geq 0 \qquad \text{for each } i, j, r \quad \text{and} \quad \sum_{i=1}^{n} p_{ijr} = 1 \qquad \text{for each } j, r. \tag{2.1}$$

The jth entry in the vector X_i will be ascribed probability mass p_{ijr} when X_i is used to estimate the distribution of the rth component in the mixture, for $1 \leq i \leq n$, $1 \leq j \leq k$ and $r = 1, 2$. The estimates that result [see, e.g., (2.2)] are closely related to weighted- or biased-bootstrap estimates of distributions [see, e.g., Efron (1981), Barbe and Bertail (1995) and Hall and Presnell (1999)], where a standard bootstrap estimate in which each datum is weighted equally is replaced by one where the data are given different weights. It also has connections to the methods of nonparametric and empirical likelihood; see Laird (1978) and Qin (1998, 1999) for applications of those ideas in the setting of inference for mixtures.

Put $p_{jr} = (p_{1jr}, \ldots, p_{njr})$, let p denote the vector of length $2kn$ obtained by concatenating the components of p_{jr} (for $1 \leq j \leq k$ and $r = 1, 2$), let X denote a generic X_i, and let $x_j < y_j$ for $1 \leq j \leq k$. The probability $F_{jr}(x_j, y_j)$ that the jth component of X lies in the interval $(x_j, y_j]$, conditional on X coming from population Π_r, will be estimated by adding the weights associated with Π_r and with data X_i that lie in this interval:

$$\widetilde{F}_{jr}(x_j, y_j) = \widetilde{F}_{jr}(x_j, y_j)(p_{jr}) = \sum_{i\,:\,x_j \leq X_{ij} \leq y_j} p_{ijr}. \tag{2.2}$$

This is a standard weighted-bootstrap estimator tailored to the present setting.

Put $y = (y_1, \ldots, y_k)$. The probability $F(x, y)$ that X lies in the rectangle $\mathcal{R}(x, y) = \prod_j (x_j, y_j]$ is estimated by $\widetilde{F}(x, y)$, which is defined to equal the proportion of the data in $\mathcal{X}$ that lies in $\mathcal{R}(x, y)$. This is a conventional unweighted bootstrap estimator of a distribution function. In particular, unlike $\widetilde{F}_{jr}$, the value of $\widetilde{F}$ does not depend on the p_{ijr}'s; neither does it take any account of the model structure at (1.1). We estimate the weights p_{ijr}, and the mixture proportion π, by minimizing the distance between $\widetilde{F}$ and the version of the distribution at (1.1) in which each F_{jr} is replaced by $\widetilde{F}_{jr}$.

That is, we fit the $2k(n-1)+1$ independent parameters among p and π by minimizing

$$S(p,\pi)=\iint_{\mathcal{S}}\Bigg[\widetilde{F}(x,y)-\Bigg\{\pi\prod_{j=1}^{k}\widetilde{F}_{j1}(x_j,y_j)+(1-\pi)\prod_{j=1}^{k}\widetilde{F}_{j2}(x_j,y_j)\Bigg\}\Bigg]^2 w(x,y)\,dx\,dy, \tag{2.3}$$

where w denotes a nonnegative weight function and $\mathcal{S}$ is the subset of the space of $2k$-vectors $(x_1,\dots,x_k,y_1,\dots,y_k)$ for which $x_j<y_j$ for $1\le j\le k$.

Let

$$(\hat{p},\hat{\pi})=\arg\min S(p,\pi), \tag{2.4}$$

where the minimization is conducted under the constraints at (2.1). Write $\hat{p}_{jr}$ for the corresponding estimator of p_{jr}, and put $\widehat{F}_{jr}(x)=\widetilde{F}_{jr}(-\infty,x_j)(\hat{p}_{jr})$ and

$$\widehat{F}(x)=\hat{\pi}\prod_{j=1}^{k}\widehat{F}_{j1}(x_j)+(1-\hat{\pi})\prod_{j=1}^{k}\widehat{F}_{j2}(x_j).$$

We shall show in Section 4 that under mild regularity conditions, $\hat{\pi}$, $\widehat{F}_{jr}$ and $\widehat{F}$ are root-n consistent for π, F_{jr} and F, respectively, where F is as defined at (1.1).

2.2. *Iterative algorithm.* We suggest two methods for computing the starting point in an iterative scheme for minimizing $S(p,\pi)$. The first is founded on maximum likelihood and the EM algorithm, and the second uses a "majority vote" idea based on inexplicit information contained in the data. As is commonly the case even for maximum likelihood inference for mixture distributions, the criterion function $S(p,\pi)$ generally has a number of local extrema. Randomly perturbing the starting point and recomputing the local minimum is recommended as a means of overcoming this problem.

In our first method we assume an approximate parametric model, such as Normal with means μ_{jr} and covariances σ_{jr}, in the obvious notation. Define D_i to equal 1 if the ith observation belongs to the first population, and to equal zero otherwise. Since D_i is unobserved we treat it as missing, and use the EM algorithm to compute maximum likelihood estimates of the vector θ of the unknown parameters π, μ_{jr} and σ_{jr}. Use the bracketed superscript notation $^{(m)}$ to denote values of parameter estimates after the mth iteration of the EM algorithm. Then, estimates on the next iteration are

$$\hat{\pi}^{(m+1)}=n^{-1}\sum_{i=1}^{n}q_{i1},\qquad \widehat{\mu}_{jr}^{(m+1)}=\frac{\sum_i q_{ir}^{(m)}x_{ij}}{\sum_i q_{ir}^{(m)}},$$

$$(\hat{\sigma}_{jr}^{(m+1)})^2=\frac{\sum_i q_{ir}^{(m)}(x_{ij}-\widehat{\mu}_{jr}^{(m+1)})^2}{\sum_i q_{ir}^{(m)}},$$

where

$$q_{ir}^{(m)} = \frac{\prod_j \phi(x_{ij} \mid \widehat{\mu}_{jr}^{(m)}, \hat{\sigma}_{jr}^{(m)}) \hat{\pi}^{(m)}}{\sum_r \prod_j \phi(x_{ij} \mid \widehat{\mu}_{jr}^{(m)}, \hat{\sigma}_{jr}^{(m)}) \hat{\pi}^{(m)}}$$

and ϕ denotes a Normal probability density with the indicated mean and covariance parameters. We iterate until convergence is achieved.

The EM algorithm requires initial values for parameter estimates. We choose the initial values of $\widehat{\mu}_{j1}$ and $\widehat{\mu}_{j2}$ to both equal the mean of $x_{1j}, \ldots, x_{nj}$, and the initial values of σ_{j1} and σ_{j2} to both equal the corresponding sample variance. After computing the maximum likelihood estimates we use Fisher's linear discriminant function, LDF, to classify each observation into one of the two populations [see O'Neill (1978)]. The LDF is defined by

$$L(x_i) = \sum_{k=1}^{k} (\widehat{\mu}_{2j} - \widehat{\mu}_{1j}) \hat{\tau}_j^{-2} x_{ij} + \log\{(1 - \hat{\pi})/\hat{\pi}\} - \tfrac{1}{2} \sum_{k=1}^{k} (\widehat{\mu}_{2j}^2 - \widehat{\mu}_{1j}^2)\, \hat{\tau}_j^{-2},$$

where $\hat{\tau}_j^2 = (\hat{\sigma}_{j1}^2 + \hat{\sigma}_{j2}^2)/2$. Let N_r (unknown) denote the number of the observations in the sample belonging to the rth population. We approximate N_1 by $n\hat{\pi}$ and N_2 by $n - N_1$. Our initial values for the components of p are $\hat{p}_{ij2}^{(0)} = 1/N_2$ if $L(x_i) > 0$ and zero otherwise, and $\hat{p}_{ij1}^{(0)} = 1/N_1$ if $L(x_i) < 0$ and zero otherwise.

To describe our second approach to choosing the initial values we note that in some applications, such as studies of the accuracy of diagnostic tests, x_{ij} may be thought of as a physician's degree of suspicion about the presence of disease. Hence, a higher value of x_{ij} indicates a greater "likelihood" of disease than a lower value. Therefore, using this information we can choose the initial estimates based on a majority rule. Given a cutoff point c, if the majority of $x_{i1}, \ldots, x_{ik}$ are greater than c then we classify the ith subject as diseased ($D_i = 1$, say); otherwise, we classify it as nondiseased ($D_i = 0$). After classifying all observations into each population, we approximate N_1 by the number of D_i's with $D_i = 1$, and N_2 by $n - N_1$. Our initial values for the components of p are $\hat{p}_{ij2}^{(0)} = 1/N_2$ if $D_i = 0$, and zero otherwise; and $\hat{p}_{ij1}^{(0)} = 1/N_1$ if $D_i = 1$ and zero otherwise. By altering c we obtain different starting values.

2.3. *Independence assumption.* Suppose an "item" (for example, a piece of engineering equipment or a hospital patient) is subjected to a battery of k tests, and that the distribution of the vector response, X, is a functional of the value Λ of a random variable. The response is broadly classified as one of two types, A or B, say, representing "pass" or "fail" in the case of a piece of equipment, or "healthy" or "ill" in the case of a patient subjected to medical tests.

The variable Λ would be the random effect in a random effects model. It would generally be unobservable, and would reflect the response of the item in

more detailed terms than is encompassed by the simple classification A or B. In particular, in a medical setting Λ might represent physiological information that could be obtained only by invasive tests. If Λ were known it would be used to assign the item to category A or B according as $\Lambda \in \mathcal{I}_A$ or $\Lambda \in \mathcal{I}_B$, respectively, where $\mathcal{I}_A$ and $\mathcal{I}_B$ denoted complementary sets. For example, if Λ were a scalar then it might be appropriate to classify the item as type A if $\Lambda \leq \lambda_0$, say, and as type B otherwise.

We may write $X = JU + (1-J)V$ where J is the indicator of the event $\Lambda \in \mathcal{I}_A$, and the distributions of U and V are those of X conditional on $\Lambda \in \mathcal{I}_A$ and $\Lambda \in \mathcal{I}_B$, respectively. While the distribution of X conditional on $\Lambda = \lambda$, for any particular value λ, might reasonably be modelled by a vector of independently distributed components, conditional on $\Lambda \in \mathcal{I}_A$ or $\Lambda \in \mathcal{I}_B$ the marginal distributions may be far from independent.

Nevertheless, when both $\mathcal{I}_A$ and $\mathcal{I}_B$ are singletons, the distributions of U and V factorize into the products of their respective marginals. In a medical context this property has been discussed by, for example, Rindskopt and Rindskopt (1986) and Hui and Zhou (1998), who treated methods that use the independence assumption as well as methods that do not. A significant number of statistical techniques for dealing with the problem of an imperfect gold standard assume independence; see, for example, Thompson and Walter and Irwig (1988), Walter (1988), Valenstein (1990) and Torrance-Rynard and Walter (1988).

The case where $\mathcal{I}_A$ is a singleton but $\mathcal{I}_B$ contains more than one element is also reasonable. In a medical setting it models test results for a healthy patient by white noise rather than by a response having a systematic trend depending on Λ. In this case the distribution of U will have independent marginals, although that of V may not. Suppose this setting prevails, and we have approximations to the mixture proportion π and the means and variances of the populations of type A and type B patients. This information may come from either previous experience or training data. It is readily used to approximate a k-variate change of scale, and rotation, that produce a transformation of the mixture distribution to one with uncorrelated marginals in both mixture components. Provided uncorrelated marginals are not far from being independent, this approach (when combined with methods discussed in Sections 2.1 and 2.2) offers a means of accessing information about component distributions in mixture data without making parametric assumptions.

The remaining case, where both $\mathcal{I}_A$ and $\mathcal{I}_B$ contain more than one element, seems difficult to accommodate using our approach. There the marginal distributions of neither type A nor type B items will necessarily factorize, and it is unclear how to ensure that the mixture distribution is identifiable in a nonparametric sense, let alone how to estimate it nonparametrically.

Of course, the difficulty here is created by the fact that binning a class of fixed effects models which, individually, satisfy the independence condition, produces a model which fails to satisfy the condition. The same "averaging" property can prevent identifiability of the conventional random effects model, even in

parametric settings. This is because the mixture distribution produced by a random effects model is in reality based on certain moments with respect to the distribution of the random effect, and there may exist more than one random effect distribution, or more than one set of marginal distributions for the mixture components, that produces the same moments. We shall give an example in Section 4.2.

3. Simulation study. We assessed numerical properties of the proposed method in a simulation study. There we generated 250 datasets, each of size $n = 500$ (excepting at the very end of this section, where we took $n = 1000$), from a trivariate normal mixture,

$$F(x) = \pi N_3(\mu, I) + (1 - \pi)N_3(0, I). \tag{3.1}$$

Thus, $k = 3$ in the context of (1.1). The value of π was varied. For the results reported here we took μ equal to either $(0.5, 1.0, 2.5)$ or $(2.0, 2.5, 3.0)$, representing relatively "close" or "distant" component distributions, respectively. Analogous results, providing a spectrum of performance between these two extremes, are obtained in intermediate cases.

To apply the method suggested in Section 2 we first used Gauss–Hermite quadrature to approximate the three-dimensional integral in the definition of $S(p, \pi)$, and then employed an IMSL Fortran subroutine, BCONF, to find the value of (p, π) that minimized $S(p, \pi)$. Initial values were chosen using the majority rule described in Section 2, taking $c = 0.5$. For the parametric maximum likelihood method we used the EM algorithm to compute estimates of the component distribution functions and the mixing proportion.

We calculated empirical approximations to bias and mean squared error (MSE) of both parametric and nonparametric estimators of π in order to assess their performance. For the parametric and nonparametric estimators of the component distribution functions we assessed performance in terms of mean integrated squared error (MISE), integrated squared bias (ISB) and integrated variance (IVAR).

The first and second rows of panels in Figure 1 display results for estimation of π in "close" and "distant" settings, respectively. In each row of panels the first panel gives biases of the parametric and nonparametric estimators of π, and the second gives MSE. It can be seen from the figure that there is not a great deal to choose between the estimators of π on the basis of bias. Moreover, while the parametric approach consistently has less MSE in the "close" setting, the situation is reversed in the "distant" case.

In the "close" and "distant" settings, respectively, Figures 2 and 3 graph MISE for both parametric and nonparametric estimators of the marginal distributions. In each figure the three plots in the first row of panels give results for estimators of the three respective marginal distributions of the first component in the mixture at (3.1). The three plots in the second row correspond to the three marginals of the

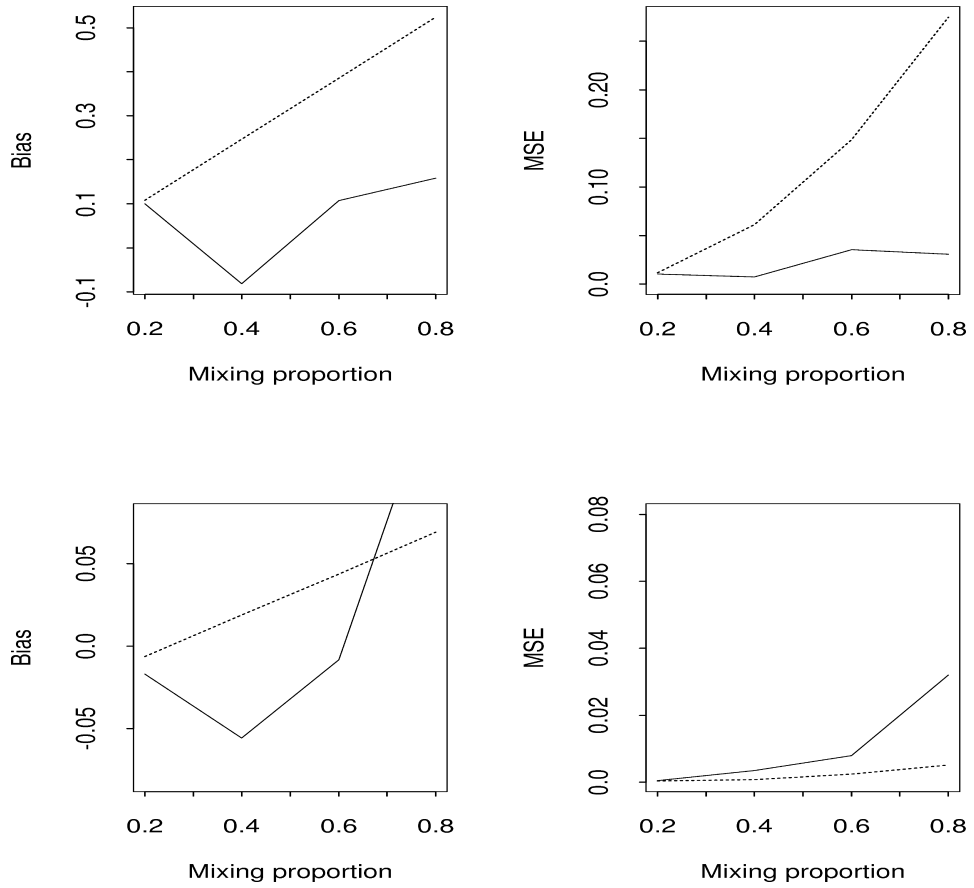

FIG. 1. *Bias and mean squared error* (*MSE*) *of parametric and nonparametric estimators of the mixing proportion*, *for sample size* 500. (*Results for parametric and nonparametric estimators are depicted by solid and dotted lines*, *respectively. The first row of panels shows bias and MSE curves for the* "*close*" *distribution setting, while the second row is for the* "*distant*" *setting.*)

second component. The second component has identically distributed marginals, whereas the marginals of the first component have different means.

Figures 2 and 3 argue that when estimating the first two distributions in the first component, the parametric method bests its nonparametric competitor in the "close" case. However, this order is reversed for the third distribution in the "close" setting. There is relatively little to choose between parametric and nonparametric methods when estimating marginal distributions in the first component in the "distant" setting, although the nonparametric approach has an edge when the first component is relatively common (i.e., π is close to 1).

However, the nonparametric method performs particularly well when used to estimate marginal distributions of the second component in either the "close" or

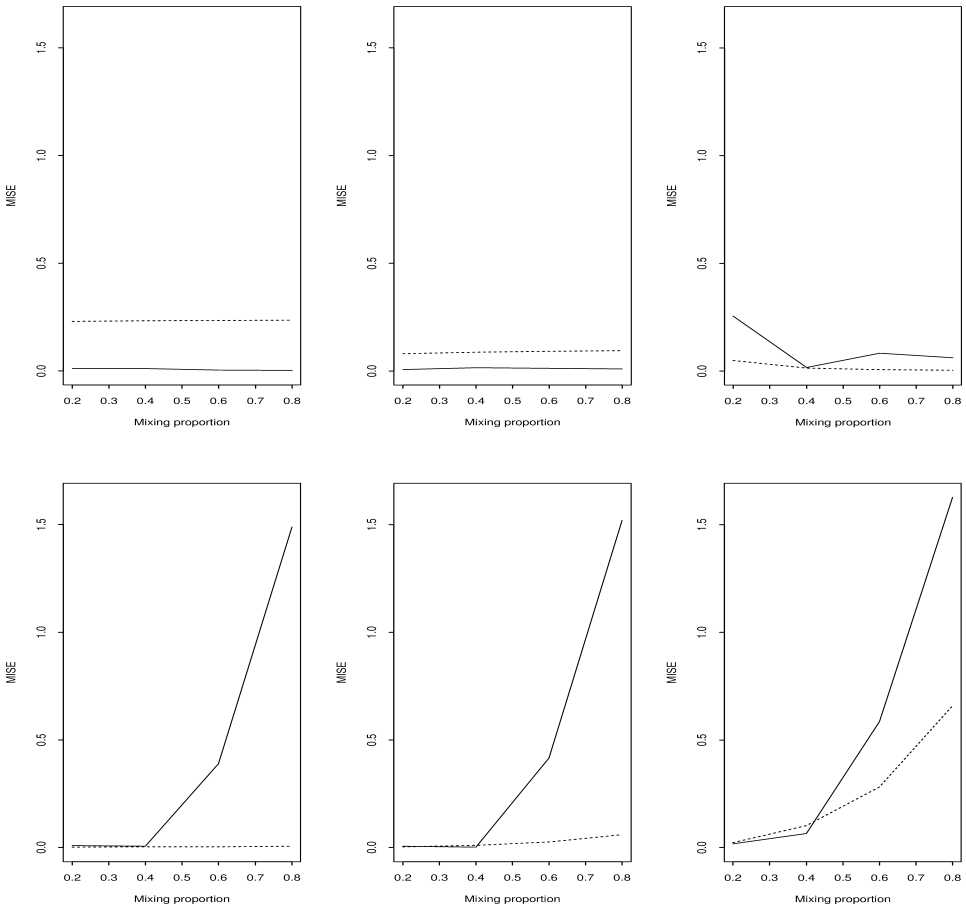

FIG. 2. *Mean integrated squared errors of marginal distribution estimators in the "close" setting, for sample size* 500. (*The three plots in the first* [*respectively, second*] *row of panels depict MISE, graphed against* π, *for estimates of the three marginal distributions of the first* [*second*] *component in the mixture at* (3.1), *when the two components are "close." Results for parametric and nonparametric estimators are depicted by solid and dotted lines, respectively.*)

"distant" settings. When the second component is commonly observed (i.e., π is close to 0) the methods perform similarly, but the nonparametric technique rapidly comes into its own when the second component is rare. It can have substantially less MISE than its parametric counterpart.

Any difficulties experienced by either the parametric or the nonparametric method are caused primarily by bias. Indeed, analogues of the plots in Figures 2 and 3 for ISB, rather than MISE, show curves whose shapes are virtually identical to those for their respective counterparts in Figures 2 and 3. Reflecting this property, and with one exception mentioned below, plots of IVAR on the scale of Figures 2 and 3 are barely distinguishable from 0.

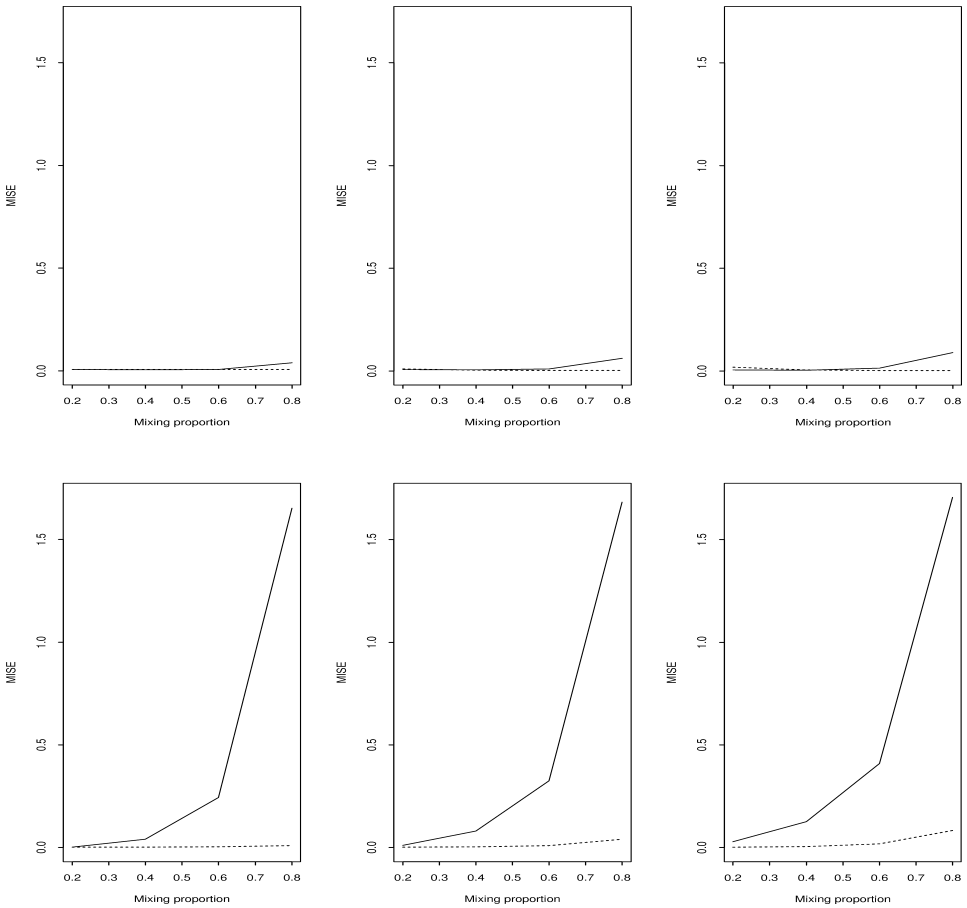

FIG. 3. *Mean integrated squared errors of marginal distribution estimators in the* "*distant*" *setting, for sample size* 500. [*Graphs are as for Figure* 2 *except that now the two components at* (3.1) *are* "*distant.*"]

The exceptional case is that of estimating marginal distributions in the second component of the mixture, in the "close" setting. The corresponding IVAR curves, analogues of the curves in the bottom row of panels in Figure 2, are shown in Figure 4. It can be seen that the variance contribution to MISE for the parametric estimator increases from 0 as π increases through values in the mid and upper range. This increase is also observed in the MISE and ISB plots, but unlike those cases the IVAR curve stays relatively low as π increases to 1. A very slight tendency to do the same thing can be noticed for IVAR curves in the "distant" setting, for the second component of the mixture. By way of contrast, the IVAR curves for the nonparametric estimator remain very flat, virtually at 0, across the range of values of π.

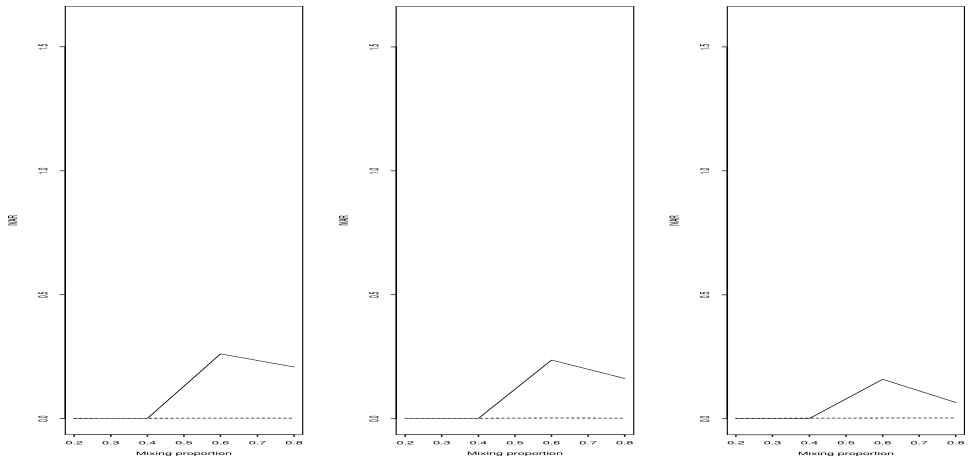

FIG. 4. *Integrated variances of marginal distribution estimators in the "close" setting, for sample size* 500. [*The plotted curves represent IVARs for the three marginal distributions in the second component at* (3.1), *when the two components are "close."*]

In summary, and in the context of marginal distribution estimation, the parametric method is sometimes (but not always) superior when both techniques perform relatively well. However, when both have difficulty the nonparametric method comes into its own, and can perform substantially better than its parametric counterpart. It rarely performs worse in that case.

Relatively speaking, difficulties in estimating the mixing proportion are less pronounced than those when estimating marginal distributions, and the two methods perform similarly there. In that case the parametric method tends to be superior, but not universally so.

All the foregoing results are for $n = 500$. To assess performance in the "close" setting for a larger sample size, we took $n = 1000$ and repeated the earlier simulations. We found that performance of the nonparametric approach improved very little with a mere doubling of sample size, and in particular it still struggled when estimating the mixing proportion. For $n = 1000$ the bias of the parametric estimator of π was negative over almost all of the interval $[0, 1]$, and its mean squared error slightly improved. Its performance relative to that of the nonparametric estimator of π improved.

Although, in the "close" setting, the nonparametric estimators of marginal distributions hardly improved when $n = 1000$, their parametric competitors made up significant ground. In particular, the latter's performance for medium to large values of π improved considerably, especially in the case of the second component. As a result, only for π towards the upper end of the interval $[0, 1]$ did the nonparametric estimator have advantages when $n = 1000$, and then only occasionally. Figure 5 illustrates this point.

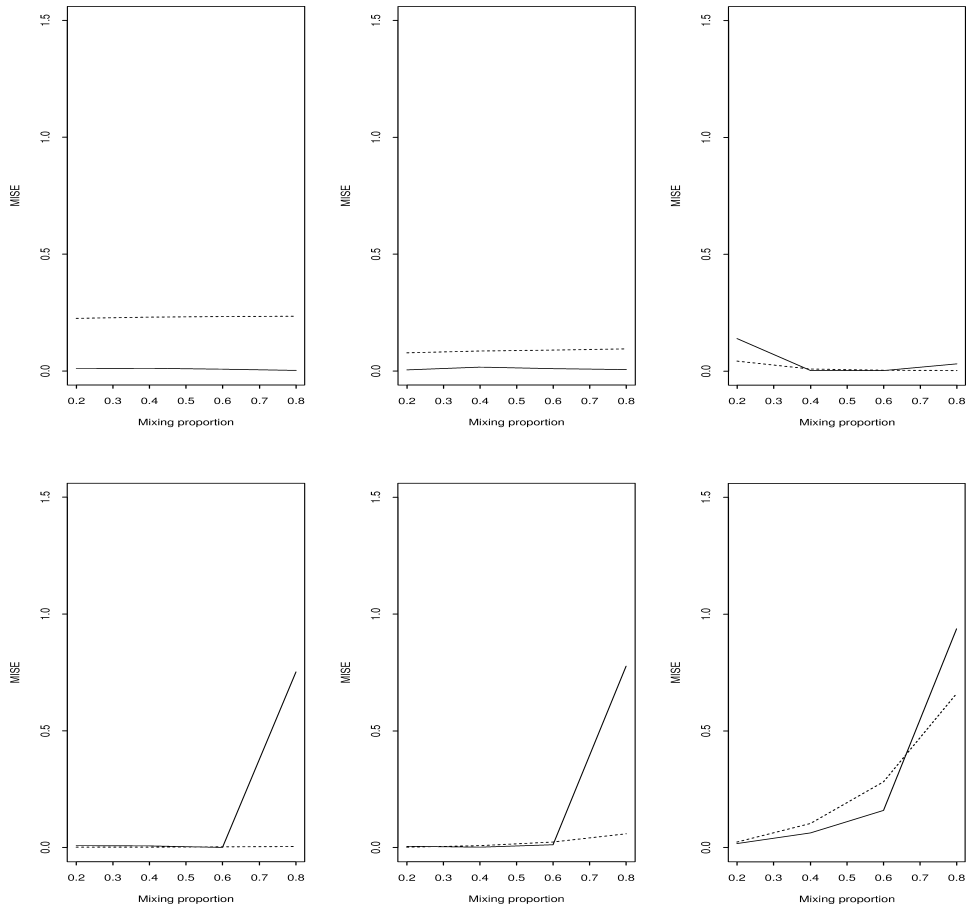

FIG. 5. *Mean integrated squared errors of marginal distribution estimators in the "close" setting, for sample size* 1000. (*Graphs are as for Figure* 2 *except that now sample size has doubled.*)

4. Theoretical properties.

4.1. *Identifiability of the standard mixture model.* In this section we address identifiability issues associated with the model at (1.1). For simplicity we work with density versions of the model, and for notational convenience we write a_j and b_j, rather than f_{j1} and f_{j2}, for the densities corresponding to distributions F_{j1} and F_{j2}, respectively. The notation $f_{j_1 j_2}$ can then be used for another purpose.

Thus, $a_1, \ldots, a_k$ and $b_1, \ldots, b_k$ denote probability densities, $0 < \pi < 1$, and we define the mixture density f to be the k-fold partial derivative of F at (1.1), differentiated once with respect to each x_i. Thus,

$$f(x_1, \ldots, x_k) = \pi \prod_{j=1}^{k} a_j(x_j) + (1-\pi) \prod_{j=1}^{k} b_j(x_j), \tag{4.1}$$

where $k \geq 2$. Given $1 \leq j_1, j_2, j \leq k$, with $j_1 \neq j_2$, write $f_{j_1 j_2}$ and f_j for the marginal densities corresponding to components j_1 and j_2 together, and j, respectively, of the distribution with density f.

Our first result describes the extent to which pairs of components may be identified from their respective marginals derived from f. To simplify notation we represent a general pair (j_1, j_2), for which $j_1 \neq j_2$, by simply $(1, 2)$.

THEOREM 4.1. *Suppose f is given by* (4.1). *Then the function $f_{12} - f_1 f_2$ factorizes. That is, there exist functions g_1, g_2 such that*

$$f_{12}(x_1, x_2) - f_1(x_1)\, f_2(x_2) = g_1(x_1)\, g_2(x_2). \tag{4.2}$$

If in addition f_{12} is not identically equal to $f_1 f_2$ then we may write $a_j = f_j + \alpha_j g_j$ and $b_j = f_j + \beta_j g_j$, where the constants α_j, β_j have the properties: (a) $\pi\alpha_j + (1-\pi)\beta_j = 0$ *for* $j = 1, 2$, *and* (b) $\alpha_1\alpha_2 = (\beta_1\beta_2)^{-1} = (1-\pi)/\pi$.

Thus, the right-hand side of (4.2) may be written as $\pi^2(1-\pi)^{-2}(a_1 - b_1) \times (a_2 - b_2)$. Of course, g_1 and g_2 are determined by (4.2) only up to constant multiples, and $g_j = c_j(a_j - b_j)$ where c_j is a constant. If a_j, b_j and a_j^0, b_j^0 are two pairs of candidates for the marginal densities then necessarily, $a_j^0 - b_j^0 = d_j(a_j - b_j)$ for a constant d_j. Hence, $d_1 d_2 = 1$. If $k \geq 3$ then this inequality holds for at least three analogues of the pair $(1, 2)$. This property may be used as the basis for a proof that, under a mild additional assumption, the model is identifiable provided $k \geq 3$; see Theorem 4.3.

Note that (4.2) implies $\int g_j = 0$, and that result (a) in Theorem 4.1 implies $\operatorname{sgn}\alpha_j = -\operatorname{sgn}\beta_j$ and

$$\pi = \frac{|\beta_j|}{|\alpha_j| + |\beta_j|}, \tag{4.3}$$

from which it follows that the ratio on the right-hand side does not depend on j. Since $a_j - b_j = (\alpha_j - \beta_j)g_j$ then we may think of the denominator at (4.3), $|\alpha_j| + |\beta_j| = |\alpha_j - \beta_j|$, as a measure of how far the jth component densities are apart.

In the case $k = 2$ the densities a_1, a_2, b_1, b_2 and the mixture proportion π are determined only up to the properties described in Theorem 4.1, as our next result, a converse to Theorem 4.1, shows.

THEOREM 4.2. *Assume $k = 2$ and f is a bivariate density, let f_1 and f_2 denote the respective marginal densities, and suppose the factorization at* (4.2) *holds for functions g_1, g_2. If $\alpha_1, \alpha_2, \beta_1, \beta_2$ is any sequence of constants such that* (i) $\operatorname{sgn}\alpha_j = -\operatorname{sgn}\beta_j$, (ii) $|\beta_j|/(|\alpha_j| + |\beta_j|)$ *does not depend on j, and* (iii) *the functions $a_j = f_j + \alpha_j g_j$ and $b_j = f_j + \beta_j g_j$ are nonnegative, then a_1, a_2, b_1, b_2 are probability densities, and* (4.1) *holds for these functions and for π defined by* (4.3).

Theorem 4.1 implies that the family of densities $a_j = f_j + \alpha_j g_j$ and $b_j = f_j + \beta_j g_j$, such that (4.1) holds, is completely determined by just two scalar parameters. Indeed, if nonzero numbers α_1, β_1 are given and satisfy $\operatorname{sgn}\alpha_1 = -\operatorname{sgn}\beta_1$, then α_2, β_2 and π are determined:

$$\alpha_2 = -\beta_1^{-1}, \qquad \beta_2 = -\alpha_1^{-1} \quad \text{and} \quad \pi = \beta_1/(\beta_1 - \alpha_1). \tag{4.4}$$

Additionally, the values of α_1 and β_1 are governed by requirement (iii) of Theorem 4.2. This generally constrains α_1 and β_1 to lie in intervals, the size and location of which may be deduced from properties, particularly those of the tails, of any particular sequence of densities a_1, a_2, b_1, b_2 for which (4.1) holds in the case $k = 2$. Therefore, there is generally a two-parameter continuum of solutions $(\pi, a_1, a_2, b_1, b_2)$ to (4.1), if we regard the density f there as given.

To appreciate the origins of (4.4), observe that by (b) of Theorem 4.1, $\pi\alpha_1\alpha_2 = 1 - \pi$ and $(1 - \pi)\beta_1\beta_2 = \pi$, whence

$$\pi\alpha_1\alpha_2 + (1 - \pi)\beta_1\beta_2 = 1. \tag{4.5}$$

By (a) of Theorem 4.1, $\pi\alpha_1 = -(1 - \pi)\beta_1$, and so by (4.5), $(1 - \pi)\beta_1(-\alpha_2 + \beta_2) = 1$. But, again by (a), $1 - \pi = -\alpha_2/(\beta_2 - \alpha_2)$, and so $-\alpha_2\beta_1 = 1$, which is the first result at (4.4). The second result there follows similarly, and the third result is a consequence of (a) of Theorem 4.1.

Next we give conditions under which the representation (4.1) is unique for $k \geq 3$. Suppose f admits the representation (4.1). We shall say that f is *irreducible* if none of its bivariate marginals factorises into the product of univariate marginals.

THEOREM 4.3. *If f admits the representation* (4.1), *with $k \geq 3$, and if f is irreducible, then, modulo replacing $\pi, a_1, \ldots, a_k$ by $1-\pi, b_1, \ldots, b_k$, respectively, the mixture proportion π and the densities a_j and b_j are uniquely determined by* (4.1). *In particular, if it is stipulated that $0 < \pi < \frac{1}{2}$, then $a_1, \ldots, a_k, b_1, \ldots, b_k$ are uniquely determined as functionals of f.*

The irreducibility condition is sufficient not only for the representation to be unique for a particular value of k, but also for it to be unique for any distribution F that is obtained by integrating out at most $k - 3$ of the k marginals. It is also necessary, in the sense that if it fails then f factorizes into the product of its marginal distributions. See the Appendix for details. Moreover, Theorem 4.3 applies directly to the case where some or all of the marginal distributions are discrete. To appreciate why, it suffices to note that, without loss of generality, a discrete marginal distribution is supported on the integers, and that there is a one-to-one correspondence between such distributions and the set of histogram densities in which each histogram block is of unit width and centred at an integer.

Theorem 4.3 may be derived as suggested in the paragraph following the statement of Theorem 4.1. In more detail, assume for simplicity that $k = 3$. By

considering the bivariate subproblems defined by the pairs $(j_1, j_2) = (1, 2)$, $(2, 3)$ and $(3, 1)$, and noting result (4.4) and the fact that f is irreducible, it may be shown that $f = h_1 + h_2$ where $h_1 = (c_1\alpha_1 - \alpha_1^{-1}) \prod g_j$, $h_2 = \prod f_j + c_1 f_1 g_2 g_3 + f_2 g_1 g_3 + f_3 g_1 g_2$, and c_1 is a constant determined by f alone and not depending on the particular decomposition of f at (4.1). We rewrite the equation as $h_1 = f - h_2$; the right-hand side here depends only on f, whereas the left-hand side depends, through α_1, on the particular representation chosen. Since f is irreducible then the latter dependence is nondegenerate. Therefore, $c_1\alpha_1 - \alpha_1^{-1}$ must equal a particular constant, c_2 say, completely determined by f. Therefore, α_1 is given by the quadratic equation $c_1\alpha_1^2 - c_2\alpha_2 - 1 = 0$. The two solutions in α_1 of this equation correspond to the two representations obtained by interchanging π, a_1, a_2, a_3 and $1 - \pi, b_1, b_2, b_3$ at (4.1).

4.2. *Nonidentifiability of the random effects model.* Here we give counterexamples that verify claims made in Section 1 about nonidentifiability in the random effects model at (1.3). It suffices to show that the first product-term in that model, which we write here as the density

$$\psi(x) = \int \left\{ \prod_{j=1}^{k} a_j(x_j|\lambda) \right\} \phi(\lambda)\, d\lambda,$$

is not identifiable.

For each $1 \le j \le k$, let s_j and t_j denote univariate probability density functions. In a parametric setting they could be assumed completely known, although they would be unknown in a nonparametric view of the problem. Let $q_j(\lambda)$ denote a function that takes values only in the interval $(0, 1)$, and put

$$a_j(x_j|\lambda) = q_j(\lambda) s_j(x_j) + \{1 - q_j(\lambda)\} t_j(x_j). \tag{4.6}$$

Then for each λ, $a_j(\cdot|\lambda)$ is a proper probability density. Moreover,

$$\psi(x) = {\sum}' \mu_{j_1 \cdots j_k} s_{j_1}(x_{j_1}) \cdots s_{j_\ell}(x_{j_\ell}) t_{j_{\ell+1}}(x_{j_{\ell+1}}) \cdots t_{j_k}(x_{j_k}), \tag{4.7}$$

where (a) $0 \le \ell \le k$, (b) $(j_1, \ldots, j_k)$ denotes a permutation of $1, \ldots, k$ with the properties $j_1 < \cdots < j_\ell$ and $j_{\ell+1} < \cdots < j_k$,

$$\text{(c)} \quad \mu_{j_1 \cdots j_k} = \int q_{j_1}(\lambda) \cdots q_{j_\ell}(\lambda) \{1 - q_{j_{\ell+1}}(\lambda)\} \cdots \{1 - q_{j_k}(\lambda)\} \phi(\lambda)\, d\lambda,$$

and (d) ${\sum}'$ denotes summation over ℓ and $(j_1, \ldots, j_k)$ satisfying (a) and (b).

Clearly, ψ depends on the functions q_j and the density ϕ only through the moments $\mu_{j_1 \cdots j_k}$. Even if ϕ is known completely, for example if it is the standard Normal density, there exists an infinity of functions $q_1, \ldots, q_k$ satisfying the constraint $0 < q_j < 1$ and producing the same moment sequence $\mu_{j_1 \cdots j_k}$. This makes the component densities nonidentifiable, even for known ϕ, if one takes a nonparametric view of the mixture estimation problem.

In a parametric setting, for a fixed, known ϕ such as standard Normal, and with known functions s_j and t_j at (4.6), one can readily construct functions $q_j = q_j(\cdot|\theta)$ which satisfy $0 < q_j < 1$ and depend in a smooth, nondegenerate way on a parameter vector θ, but for which each moment $\mu_{j_1 \cdots j_k}$ in the series at (4.7) does not depend on θ. In this case the model for a_j is fully parametrized, but θ is not identifiable from data on ψ.

4.3. *Properties of estimators in the standard mixture model.* We show that the estimators proposed in Section 2 are consistent for an irreducible distribution and its component parts. Moreover, we prove that in many cases the rate of convergence is $O_p(n^{-1/2})$, where n denotes sample size. To avoid ambiguity in the order of the two terms at (1.1) we assume $0 < \pi < \frac{1}{2}$, and impose this condition too on the estimator $\hat{\pi}$; this may be interpreted as replacing $\hat{\pi}$ by $1 - \hat{\pi}$, and interchanging $\widehat{F}_{j1}$ and $\widehat{F}_{j2}$, if we estimate π to be a number exceeding $\frac{1}{2}$.

THEOREM 4.4. *Let the weight function w at* (2.3) *be a continuous density supported on all of $2k$-dimensional space. Assume too that f admits the representation* (4.1), *with $k \geq 3$, that f is irreducible, and that $0 < \pi < \frac{1}{2}$. Then the k-variate distribution function F corresponding to f, as well as the mixture proportion π and the component distribution functions, are strongly, uniformly consistently estimated by the procedure introduced in Section* 2. *Furthermore, if the distributions at* (1.1) *are compactly supported then $\hat{\pi} - \pi = O_p(n^{-1/2})$ and all the distribution estimators are root-n consistent in L^2, in the sense that*

$$\int \{\widehat{G}(u) - G(u)\}^2 \, du = O_p(n^{-1}), \tag{4.8}$$

where G may represent either F, that is, the full mixture distribution, or any one of its components, that is, F_{j1} or F_{j2}.

Result (4.8) admits generalizations to the case of noncompactly supported distributions, but they require conditions on the tails of the component distributions.

5. Technical arguments.

5.1. *Proof of Theorem* 4.1. Since Theorem 4.1 deals only with bivariate submodels of the full distribution then we may, without loss of generality, assume $k = 2$. Then (4.1) becomes

$$\pi a_1 a_2 + (1 - \pi) b_1 b_2 = f, \tag{5.1}$$

and, integrating with respect to the component complementary to j, we deduce that

$$\pi a_j + (1 - \pi) b_j = f_j. \tag{5.2}$$

Now, (5.2) implies that $b_j = (1-\pi)^{-1}(f_j - \pi a_j)$. Substituting for b_1 and b_2 in (5.1), and simplifying, we deduce that

$$\frac{\pi}{1-\pi}(a_1 - f_1)(a_2 - f_2) = f - f_1 f_2, \tag{5.3}$$

which proves (4.2), of which the version here is $f - f_1 f_2 = g_1 g_2$ where $g_1 g_2$ denotes the left-hand side of (5.3).

Henceforth we assume that f is not identically equal to $f_1 f_2$. Result (5.3) also implies that for constants α_1, α_2,

$$a_j = f_j + \alpha_j g_j \qquad \text{where } \alpha_1 \alpha_2 = \frac{1-\pi}{\pi}. \tag{5.4}$$

A symmetric argument gives, for constants β_1, β_2,

$$b_j = f_j + \beta_j g_j \qquad \text{where } \beta_1 \beta_2 = \frac{\pi}{1-\pi}. \tag{5.5}$$

Property (b) in Theorem 4.1 follows from (5.4) and (5.5).

Substituting for a_1, a_2, b_1, b_2 in (5.1) using (5.4) and (5.5), and simplifying, we obtain

$$\{\pi\alpha_1 + (1-\pi)\beta_1\} f_2 g_1 + \{\pi\alpha_2 + (1-\pi)\beta_2\} f_1 g_2 = 0. \tag{5.6}$$

Integrating over the first component, and noting that by (4.2), $\int g_1 = 0$, we deduce that $\{\pi\alpha_2 + (1-\pi)\beta_2\} g_2 = 0$, and so either $\pi\alpha_2 + (1-\pi)\beta_2 = 0$ or $g_2 \equiv 0$. The latter property implies, by (5.4) and (5.5), that $a_2 = b_2 = f_2$, and hence by (5.1) and (5.2) that $f = f_1 f_2$, which is contrary to assumption. Therefore, $\pi\alpha_j + (1-\pi)\beta_j = 0$ for $j = 2$, and likewise the version for $j = 1$ holds. This establishes part (a) of the theorem.

Note that by (a),

$$\pi = \frac{\beta_j}{\beta_j - \alpha_j} \quad \text{and} \quad 1 - \pi = \frac{-\alpha_j}{\beta_j - \alpha_j}. \tag{5.7}$$

The denominators here cannot vanish, since to do so would imply $\alpha_j = \beta_j$ which, because α_j and β_j cannot both vanish, because f is not identically equal to $f_1 f_2$, leads to a contradiction via (a) of the theorem. It follows from (a), (5.7) and the fact that π and $1 - \pi$ are both strictly positive that β_j and α_j must have opposite signs, which completes the proof of Theorem 4.1.

5.2. *Proof of Theorem* 4.2. If (4.2) holds then $\int g_j = 0$ for each j, and so by (iii) of Theorem 4.2, a_1, a_2, b_1, b_2 are probability densities. Substituting into (4.1), the right-hand side becomes

$$f_1 f_2 + g_1 g_2 + \{\pi\alpha_1 + (1-\pi)\beta_1\} f_2 g_1 + \{\pi\alpha_2 + (1-\pi)\beta_2\} f_1 g_2. \tag{5.8}$$

Defining π by (4.3), and using assumptions (i) and (ii), we may prove that $\pi = \beta_j/(\beta_j - \alpha_j)$, whence it follows that $\pi\alpha_j + (1-\pi)\beta_j = 0$. Therefore the quantity at (5.8) equals $f_1 f_2 + g_1 g_2$, which by (4.2) is just f. This proves the theorem.

5.3. *Proof of Theorem* 4.4. Let N_r equal the number of X_i's that come from Π_r, and put $\tilde{\pi} = N_r/n$ and

$$\tilde{p}_{ijr} = \begin{cases} 1/N_r, & \text{if } X_i \in \Pi_r, \\ 0, & \text{otherwise,} \end{cases}$$

for $1 \le j \le k$. Let $\tilde{p}$ denote the concatenation of values of $\tilde{p}_{ijr}$. Given two functions $G_1(x, y)$ and $G_2(x, y)$ of the type $\widetilde{F}(x, y)$, define

$$\|G_1 - G_2\|^2 = \int \{G_1(x, y) - G_2(x, y)\}^2 \, w(x, y) \, dx \, dy.$$

Put $\widehat{F}(x, y) = \int_{\mathcal{R}(x,y)} d\widehat{F}(x)$, $D_0 = \|\widetilde{F} - F\|$ and

$$D(p, \pi; F) = \left\| \pi \prod_{j=1}^{k} \widetilde{F}_{j1}(x_j, y_j) + (1 - \pi) \prod_{j=1}^{k} \widetilde{F}_{j2}(x_j, y_j) - F \right\|.$$

In this formula the dependence of $D(p, \pi; F)$ on p is through the fact that each $\widetilde{F}_{jr}$, on the right-hand side, depends on p; see (2.2).

Recall from (2.4) that $(\hat{p}, \hat{\pi})$ minimizes $D(p, \pi, \widetilde{F})$, and so $D(\hat{p}, \hat{\pi}; \widetilde{F}) \le D(\tilde{p}, \tilde{\pi}; \widetilde{F})$. Therefore,

$$\begin{aligned} \|\widehat{F} - F\| &\le D(\hat{p}, \hat{\pi}; \widetilde{F}) + D_0 \le D(\tilde{p}, \tilde{\pi}; \widetilde{F}) + D_0 \\ &\le D(\tilde{p}, \tilde{\pi}; F) + 2 D_0 \equiv T, \end{aligned} \tag{5.9}$$

say, where the first and third inequalities are consequences of triangle relations. It is readily proved that T converges to 0 with probability 1, which, since F is continuous and monotone, implies strong uniform consistency of $\widehat{F}$ for F.

To prove strong uniform consistency of $\hat{\pi}$, $\widehat{F}_{jr}$ for π, F_{jr}, respectively, suppose that to the contrary this result is false. Then there exists an event $\mathcal{E}$, with $P(\mathcal{E}) > 0$, such that on $\mathcal{E}$, at least one of $\hat{\pi}$, $\widehat{F}_{jr}$ does not converge to its counterpart among π, F_{jr}. Using Helly's extraction principle, for each sample-space point $\omega \in \mathcal{E}$ we may choose a subsequence along which $\hat{\pi}$ and $\widehat{F}_{jr}$, the latter for each j and r, converge to proper limits $\hat{\pi}^\omega$ and $\widehat{F}^\omega_{jr}$, say, at least one of the latter being distinct from its counterpart among π and F_{jr}. Note that $\widehat{F}^\omega_{jr}$ may be a subdistribution function. Either for this reason or because, while each $\widehat{F}^\omega_{jr}$ is a distribution function, either not all these functions are equal to the respective F_{jr}'s or $\hat{\pi}^\omega \ne \pi$, and noting that the representation (1.1) is uniquely determined by F, the corresponding mixture distribution or subdistribution function $\widehat{F}^\omega$, defined by

$$\widehat{F}^\omega(x) = \hat{\pi}^\omega \prod_{j=1}^{k} \widehat{F}^\omega_{j1}(x_j) + (1 - \hat{\pi}^\omega) \prod_{j=1}^{k} \widehat{F}^\omega_{j2}(x_j),$$

is not identical to F. However, $\widehat{F}^\omega$ is a limit point of the sequence of functions $\widehat{F}$ when the sample-space point is ω. And it follows, using the argument in the

previous paragraph, that $\widehat{F}$ converges to F with probability 1 conditional on $\omega \in \mathcal{E}$. Therefore, $\widehat{F}^{\omega} = F$ with probability 1, conditional on $\omega \in \mathcal{E}$. This contradiction proves that each of $\hat{\pi}$, $\widehat{F}_{jr}$ converges almost surely to the corresponding π, F_{jr}. Finally, note that since F_{jr} is continuous then convergence of $\widehat{F}_{jr}$ to F_{jr}, in the sense of weak convergence of distributions, for any particular sample-space point ω, implies uniform convergence.

If the component distributions are all compactly supported, then T, defined at (5.9), equals $O_p(n^{-1})$. This implies (4.8) in the case $G = F$. That result will imply (4.8) when G is a component distribution and will also imply the property $\hat{\pi} - \pi = O_p(n^{-1/2})$, if we prove that when

$$H(x) = q \prod_{j=1}^{k} H_{j1}(x_j) + (1-q) \prod_{j=1}^{k} H_{j2}(x_j)$$

is a mixture representation alternative to that at (1.1), we have

$$(\pi - q)^2 + \sum_{k=1}^{k} \int \{(F_{j1} - H_{j1})^2 + (F_{j2} - H_{j2})^2\} = O\left\{ \int (F - H)^2 \right\} \tag{5.10}$$

as $H \to F$ in L^2.

Redefine $\| \cdot \|$ to be the L^2 norm, that is, $\|F - H\| = \{\int (F - H)^2\}^{1/2}$. In this notation, (5.10) may be interpreted as implying that, for example, $|\pi - q| = O(\|F - H\|)$ as $H \to F$. This is in turn a smoothness condition on the functional q of H, a little weaker than the existence of a Fréchet derivative.

To establish (5.10), write $K = K[F]$ to denote that a given j-variate function K, where $0 \le j \le k$, is a functional of F. (The case $j = 0$ corresponds to $K[F]$ being a constant when F is fixed.) We shall say that the functional "is Lipschitz" if it satisfies $\|K[F] - K[H]\| = O(\|F - H\|)$ as $H \to F$ in L^2. If (5.10) were untrue then it would fail if the left-hand side were replaced by $(\pi - q)^2$, or if it were replaced by $\int (F_{jr} - H_{jr})^2$ for some pair (j, r). That is, we see that if (5.10) were to fail then either (i) $|\pi - q| / \|F - H\|$ would be unbounded, or (ii) $\|F_{jr} - H_{jr}\| / \|F - H\|$ would be unbounded for some pair (j, r), as $H \to F$ in L_2. If (i) were true it would imply that (iii) $\pi = \pi[F]$ was not Lipschitz, while (ii) would imply that (iv) $F_{jr} = F_{jr}[F]$ was not Lipschitz. We shall contradict (iv) by proving its complement: (v) F_{jr} is Lipschitz. Similarly, (iii) may be shown to be contradicted.

Assume for the sake of simplicity that $k = 3$, and let g_1, g_2 and g_3 be functions such that

$$f_{j_1 j_2}(x_{j_1}, x_{j_2}) - f_{j_1}(x_{j_1})\, f_{j_2}(x_{j_2}) = g_{j_1}(x_{j_1})\, g_{j_2}(x_{j_2}) \tag{5.11}$$

for $(j_1, j_2) = (1, 2)$, $(2, 3)$ and $(3, 1)$ [cf. (4.2)]. Put $G_j(x_j) = \int_{u \le x_j} g_j(u)\, du$, with the constant of proportionality in the definition of g_j, for $j = 1, 2, 3$, determined by the condition $\int G_1^2 = 1$. Write A_j and $B f_{j_1 j_2}$ for the univariate

marginal distribution function of F corresponding to the jth coordinate, and the bivariate marginal distribution function of F corresponding to coordinates (j_1, j_2), respectively. Then by (5.11),

$$B_{j_1 j_2}(x_{j_1}, x_{j_2}) - A_{j_1}(x_{j_1}) A_{j_2}(x_{j_2}) = G_{j_1}(x_{j_1}) G_{j_2}(x_{j_2}). \tag{5.12}$$

It follows directly from the definitions of A_j and $B_{j_1 j_2}$ that both are Lipschitz. Hence, by (5.12), G_j is also Lipschitz. [Of course, in order to obtain G_{j_1} from (5.12) we simply take x_{j_2} to be a value for which $G_{j_2}(x_{j_2}) \neq 0$. It follows from the irreducibility condition that this is possible.]

By Theorem 4.1, $F_{jr} = A_j + c_{jr} G_j$ where c_{jr} is a scalar. From this result, and the Lipschitz property of G_j deduced in the previous paragraph, we see that F_{jr} is Lipschitz too, provided c_{jr} is as well. However, we know from the proof of Theorem 4.3 (outlined two paragraphs below the statement of that theorem) that c_{jr} is determined as the solution of a quadratic equation. It is straightforward to prove, first, that each coefficient in the equation is Lipschitz, and thence that c_{jr} also has that property. This completes the proof of result (v), stated two paragraphs above.

APPENDIX

Nature of the irreducibility condition. The irreducibility condition implies that for any particular version of the representation on the right-hand side of (4.1), no function $a_j - b_j$ vanishes identically. To appreciate why, let us suppose without loss of generality that this fails for $j = 3$; that is, $a_3 \equiv b_3 \equiv A$, say. Then we may write

$$\begin{aligned} f(x_1, x_2, x_3) &= \pi \prod_{j=1}^{3} a_j(x_j) + (1 - \pi) \prod_{j=1}^{3} b_j(x_j) \\ &= \left\{ \pi \prod_{j=1}^{2} a_j(x_j) + (1 - \pi) \prod_{j=1}^{2} b_j(x_j) \right\} A(x_3). \end{aligned}$$

Now integrate over either x_1 or x_2; choosing x_2 we obtain

$$f(x_2, x_3) = \{\pi\, a_2(x_2) + (1 - \pi)\, b_2(x_2)\} A(x_3),$$

which implies that the bivariate density of (X_2, X_3) factorizes into the product of its marginals. This violates the irreducibility condition.

Next we prove that if the irreducibility condition fails then $f(x_1, \ldots, x_k)$ factorizes into the product of its k univariate marginals, and so (4.1) is not uniquely determined with respect to π. To understand why, note that if irreducibility fails, then without loss of generality the bivariate density of (X_1, X_2) factorizes into the product of its marginals. We shall indicate this by writing $f(x_1, x_2) = A_1(x_1) A_2(x_2)$, where $A_1(x_1)$ and $A_2(x_2)$ are the marginal densities. Consider

any trivariate density which includes the above two marginals; without loss of generality it is the density of (X_1, X_2, X_3), and equals

$$f(x_1, x_2, x_3) = \pi \prod_{j=1}^{3} a_j(x_j) + (1-\pi) \prod_{j=1}^{3} b_j(x_j).$$

Integrating over x_3 we deduce that

$$f(x_1, x_2) = \pi \prod_{j=1}^{2} a_j(x_j) + (1-\pi) \prod_{j=1}^{2} b_j(x_j) = A_1(x_1) A_2(x_2).$$

From this it follows that $a_j = b_j = A_j$ for $j = 1, 2$. Therefore,

$$f(x_1, x_2, x_3) = A_1(x_1) A_2(x_2) \{\pi a_3(x_3) + (1-\pi) b_3(x_3)\}.$$

This of course implies that $f(x_1, x_2, x_3)$ factorizes as the product of its three univariate marginals, and arguing in the same manner we may prove by induction that $f(x_1, \ldots, x_k)$ factorizes into the product of its k univariate marginals.

Acknowledgments. We are grateful for the helpful and constructive comments of three reviewers.

REFERENCES

BARBE, P. and BERTAIL, P. (1995). *The Weighted Bootstrap.* Springer, Berlin.

CERRITO, P. B. (1992). Using stratification to estimate multimodal density functions with applications to regression. *Comm. Statist. Simulation Comput.* **21** 1149–1164.

COHEN, A. C. (1967). Estimation in mixtures of two normal distributions. *Technometrics* **9** 15–28.

DAY, N. E. (1969). Estimating the components of a mixture of normal distributions. *Biometrika* **56** 463–474.

EFRON, B. (1981). Nonparametric standard errors and confidence intervals (with discussion). *Canad. J. Statist.* **9** 139–172.

EVERITT, B. S. and HAND, D. J. (1981). *Finite Mixture Distributions.* Chapman and Hall, London.

HADGU, A. and QU, Y. (1998). A biomedical application of latent models with random effects. *Appl. Statist.* **47** 603–616.

HALL, P. (1981). On the nonparametric estimation of mixture proportions. *J. Roy. Statist. Soc. Ser. B* **43** 147–156.

HALL, P. and PRESNELL, B. (1999). Intentionally biased bootstrap methods. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **61** 143–158.

HALL, P. and TITTERINGTON, D. M. (1984). Efficient nonparametric estimation of mixture proportions. *J. Roy. Statist. Soc. Ser. B* **46** 465–473.

HALL, P. and TITTERINGTON, D. M. (1985). The use of uncategorized data to improve the performance of a nonparametric estimator of a mixture density. *J. Roy. Statist. Soc. Ser. B* **47** 155–163.

HUI, S. L. and ZHOU, X. H. (1998). Evaluation of diagnostic tests without gold standards. *Statist. Methods Medical Res.* **7** 354–370.

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

LANCASTER, T. and IMBENS, G. (1996). Case-control studies with contaminated controls. *J. Econometrics* **71** 145–160.

LINDSAY, B. G. (1983a). The geometry of mixture likelihoods: A general theory. *Ann. Statist.* **11** 86–94.

LINDSAY, B. G. (1983b). The geometry of mixture likelihoods. II. The exponential family. *Ann. Statist.* **11** 783–792.

LINDSAY, B. G. and BASAK, P. (1993). Multivariate normal mixtures: A fast consistent method of moments. *J. Amer. Statist. Assoc.* **88** 468–476.

MCLACHLAN, G. J. and BASFORD, K. E. (1988). *Mixture Models. Inference and Applications to Clustering.* Dekker, New York.

METZ, C. E. (1978). Basic principles of ROC analysis. *Seminars in Nuclear Medicine* **8** 283–298.

MURRAY, G. D. and TITTERINGTON, D. M. (1978). Estimation problems with data from a mixture. *Appl. Statist.* **27** 325–334.

O'NEILL, T. J. (1978). Normal discrimination with unclassified observations. *J. Amer. Statist. Assoc.* **73** 821–826.

QIN, J. (1998). Semiparametric likelihood based method for goodness of fit tests and estimation in upgraded mixture models. *Scand. J. Statist.* **25** 681–691.

QIN, J. (1999). Empirical likelihood ratio based confidence intervals for mixture proportions. *Ann. Statist.* **27** 1368–1384.

QU, Y. and HADGU, A. (1998). A model for evaluating sensitivity and specificity for correlated diagnostic tests in efficacy studies with an imperfect reference test. *J. Amer. Statist. Assoc.* **93** 920–928.

QU, Y., TAN, M. and KUTNER, M. H. (1996). Random effects models in latent class analysis for evaluating accuracy of diagnostic tests. *Biometrics* **52** 797–810.

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

REDNER, R. A. and WALKER, H. F. (1984). Mixture densities, maximum likelihood and the EM algorithm. *SIAM Rev.* **26** 195–239.

RINDSKOPF, D. and RINDSKOPF, W. (1986). The value of latent class analysis in medical diagnosis. *Statist. Medicine* **5** 21–27.

SHAHSHAHANI, B. M. and LANDGREBE, D. A. (1994). The effect of unlabeled samples in reducing the small sample-size problem and mitigating the Hughes phenomenon. *IEEE Trans. Geosci. Remote Sensing* **32** 1087–1095.

TEICHER, H. (1967). Identifiability of mixtures of product measures. *Ann. Math. Statist.* **38** 1300–1302.

THOMPSON, W. D. and WALTER, S. D. (1988). A reappraisal of the kappa coefficient. *J. Clinical Epidemiol.* **41** 949–958.

TITTERINGTON, D. M. (1983). Minimum-distance non-parametric estimation of mixture proportions. *J. Roy. Statist. Soc. Ser. B* **45** 37–46.

TITTERINGTON, D. M., SMITH, A. F. M. and MAKOV, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions.* Wiley, Chichester.

TORRANCE-RYNARD, V. L. and WALTER, S. D. (1988). Effects of dependent errors in the assessment of diagnostic test performance. *Statist. Medicine* **16** 2157–2175.

VALENSTEIN, P. N. (1990). Evaluating diagnostic tests with imperfect standards. *Amer. J. Clinical Pathology* **93** 252–258.

WALTER, S. D. and IRWIG, L. M. (1988). Estimation of test error rates, disease prevalence and relative risk from misclassified data: A review. *Journal of Clinical Epidemiology* **41** 923–937.

CENTRE FOR MATHEMATICS
AND ITS APPLICATIONS
AUSTRALIAN NATIONAL UNIVERSITY
CANBERRA, ACT 0200
AUSTRALIA
E-MAIL: halpstat@pretty.anu.edu.au

NORTHWEST HRS&D CENTER OF EXCELLENCE
VA PUGET SOUND HEALTH CARE SYSTEM
UNIVERSITY OF WASHINGTON
1160. S. COLUMBIAN WAY
SEATTLE, WASHINGTON 98108
E-MAIL: Andrew.Zhou@med.va.gov