The Annals of Statistics
1988, Vol. 16, No. 4, 1623–1634

MIXTURES OF DISTRIBUTIONS: A TOPOLOGICAL APPROACH

BY L. A. LI AND N. SEDRANSK

State University of New York at Albany and Yale University

Identifiability problems have previously precluded a general approach to testing the hypothesis of a "pure" distribution against the alternative of a mixture of distributions. Three types of identifiability are defined, and it is shown that B-identifiability allows a Bayesian solution to the testing problem. First, an equivalence relation is defined over parametrizations of probability functions. Then the projection onto the quotient space is shown to give a B-identifiable parametrization. Bayesian inference proceeds using the Bayes factor as a "test" criterion.

1. Introduction. The history of estimation of mixtures is long, dating from Karl Pearson's work in 1894 studying 1000 forehead breadths of Naples crabs from a mixture of two species. For convenience, it can be divided into several groups of methods for estimation of mixtures: moments methods, maximum likelihood methods, Bayes estimates, informal graphical techniques and some others including θ-efficient estimates and spectral decomposition. [Appropriate bibliographies can be found in Li and Sedransk (1982) and Everitt and Hand (1981).] Despite the laborious computations involved, work on estimation of mixtures has led to results which are satisfactory to varying degrees.

By comparison, few results are available for hypothesis testing. The problem is easily stated for the simplest case: Test the hypothesis that the sample comes from a single ("pure") population against the alternative that the sample is a mixture of two (or more) populations when the parameters as well as the mixing rates are unknown. More generally stated, the null hypothesis is that the sample comes from a mixture of k or fewer populations and the alternative is that the mixture is of $k + 1$ or more populations, where k is specified but all parameter values and mixing rates are unknown.

Informal diagnostic tools for the detection of mixtures have been developed using the sample histogram or probability plotting [see Hazen (1914), Everitt (1978) and Fowlkes (1979)]. These methods have also been criticized for being ad hoc and potentially misleading [Murphy (1964) and Cox (1966)]. A number of authors have derived tests for the presence of a mixture of specified components [see Baker (1958), Tiago de Oliveira (1965) and Binder (1978)] or special cases [see Johnson (1973) and Thomas (1969)]; and the difficulties in testing for the presence of a mixture have often been pointed out [for example, Everitt and Hand (1981)]. The inherent problem of identifiability is crucial. As will be seen later, this identifiability problem is also implicit in the estimation of mixtures and is the source of difficulty in getting solutions to converge in some estimation problems.

Received March 1985; revised February 1988

AMS 1980 *subject classifications*. Primary 62E10; secondary 62F03, 62F15.

Key words and phrases. Identifiability, hypothesis testing, detecting mixtures, quotient space.

In this article a resolution of the identifiability problem is given, based on an identification map (or equivalently, the use of a particular quotient space). This explains the convergence problem in the estimation of mixtures and leads to a method for detecting the presence of a mixture. First two types of mixtures are defined, and conditions for identifiability in several senses are reviewed. Then a more stringent definition of identifiability is shown to be necessary for solution of the hypothesis testing problem. The identification map (projection onto a particular quotient space) renders unidentified parameter spaces identifiable; and inferential procedures are given for the new identifiable (quotient) spaces.

2. Mixtures and identifiability. A (type I) mixture is defined to be a mixture of probability density functions from the same family. (A mixture of density functions from several families is a type II mixture, which is discussed briefly in Section 5.) That is, let $f(\mathbf{x}|\theta)$ be a probability density function with respect to m-dimensional Lebesgue measure for $\mathbf{x} = (x_1, x_2, \dots, x_m)$ on $\mathbb{R}^m$ given the parameter $\theta \in \Theta$, where Θ is a measurable subset of $\mathbb{R}^l$. If Q is a probability measure such that the parameter space Θ includes the support of Q, then the probability density function with respect to m-dimensional Lebesgue measure

$$f(\mathbf{x}) = \int f(\mathbf{x}|\theta)Q(d\theta)$$

is called a mixture density function (obtained by mixing probability density functions from the same family $\{f(\mathbf{x}|\theta)\colon \theta \in \Theta\}$) for the random vector $\mathbf{X} = (X_1, X_2, \dots, X_m)$ on $\mathbb{R}^m$, and Q is called a mixing probability measure.

Hence, if the mixing probability measure Q assigns positive probability p_i to the point $\theta_i \in \Theta$, (l-dimensional), $i = 1, 2, \dots, k$, where $\theta_i \neq \theta_j$ for $i \neq j$, the mixture density function for the random vector $\mathbf{X}$ on $\mathbb{R}^m$ in (2.1) is of the form

$$f(\mathbf{x}) = \sum_{i=1}^{k} p_i f(\mathbf{x}|\theta_i), \tag{2.1}$$

where $\theta_i \in \Theta$, $0 < p_i < 1$ for $i = 1, 2, \dots, k$, $\theta_i \neq \theta_j$ for $i \neq j$, and $\sum_{i=1}^{k} p_i = 1$. More precisely, $f(\mathbf{x})$ is called a type I (k) mixture if it is obtained by mixing exactly k probability density functions from the same family $\{f(\mathbf{x}|\theta)\colon \theta \in \Theta\}$ and is of the form (2.1). A type I ($\leq k$) mixture is obtained by mixing at most k components.

The problem of identifiability of mixture density functions has been extensively studied by Teicher (1960, 1961, 1963, 1967), Yakowitz and Spragins (1968) and Chandra (1977). The notion of identifiability for mixtures used by Teicher et al. is denoted here by T-identifiability.

Definition 2.1. A class of mixture density functions $f(\mathbf{x})$ is said to be T-identifiable if there is one-to-one correspondence between the probability measure, which is determined by one and only one mixture density function in this class, and the corresponding mixing probability measure Q for that mixture density function.

A sufficient condition, denoted here by (C.0), for T-identifiability, is given by Yakowitz and Spragins (1968), for type I (k) mixtures.

(C.0) The mixture comes from a family $\{f(\mathbf{x}|\boldsymbol{\theta})\colon \boldsymbol{\theta} \in \Theta\}$ which is linearly independent (a.s. on $\mathbb{R}^m$) over $\mathbb{R}$.

Some weaker conditions than (C.0) can be shown to be necessary and/or sufficient for T-identifiability. A sufficient condition for T-identifiability of type I ($\leq k$) mixture is

(C.1) The mixture comes from a family $\{f(\mathbf{x}|\boldsymbol{\theta})\colon \boldsymbol{\theta} \in \Theta\}$, where every subset with at most $2k$ elements is linearly independent (a.s. on $\mathbb{R}^m$) over $\mathbb{R}$.

Necessary and sufficient conditions for T-identifiability for type I ($\leq k$) and type I (k) mixtures are given in (C.2) and (C.3), respectively.

(C.2) The mixture comes from a family $\{f(\mathbf{x}|\boldsymbol{\theta})\colon \boldsymbol{\theta} \in \Theta\}$, where every positive linear combination of at most k elements is unique.

(C.3) The mixture comes from a family $\{f(\mathbf{x})|\boldsymbol{\theta})\colon \boldsymbol{\theta} \in \Theta\}$, where every positive linear combination of exactly k elements is unique.

Obviously, (C.0) ⇒ (C.1) ⇒ (C.2) ⇒ (C.3).

A natural definition of identifiability differing from T-identifiability is P-identifiability.

Definition 2.2. A parametrization for a class of mixture density functions is said to be P-identifiable if there is one-to-one correspondence between the mixture density function in this class and its representing parameter.

In general, a T-identifiable class of all type I (k) mixtures parametrized by $(p_1, p_2, \dots, p_{k-1}, \boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_k) \in (0,1)^{k-1} \times \Theta^k$ is not P-identifiable, since a type I (k) mixture can also be written as

$$f(\mathbf{x}) = \sum_{i=1}^{k} p_{\pi(i)} f(\mathbf{x}|\boldsymbol{\theta}_{\pi(i)}),$$

for any permutation π of $\{1, 2, \dots, n\}$.

However, restricted to a suitable subset Δ of $(0,1)^{k-1} \times \Theta^k$, the T-identifiable class of type I (k) mixtures is P-identifiable. For example, take

$$\Delta = \Bigg\{ (p_1, p_2, \dots, p_{k-1}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \dots, \boldsymbol{\theta}_k) | 0 < p_i < 1 \text{ for } i = 1, 2, \dots, k-1, \\ \sum_{i=1}^{k-1} p_i < 1,\ \boldsymbol{\theta}_i \in \Theta \text{ for } i = 1, 2, \dots, k, \text{ with } \boldsymbol{\theta}_1 \ll \boldsymbol{\theta}_2 \ll \cdots \ll \boldsymbol{\theta}_k \Bigg\}, \tag{2.2}$$

where " $\ll$ " is some ordering and $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_k$ are distinct.

If $f(\mathbf{x}|\boldsymbol{\theta})$ is a continuous function of its (l-dimensional) parameter $\boldsymbol{\theta}$, then the mixture density functions of the form $f(\mathbf{x}) = \sum_{i=1}^{k} p_i f(\mathbf{x}|\boldsymbol{\theta}_i)$ are also continuous functions of the parameter $(p_1, \dots, p_{k-1}, \theta_1, \dots, \theta_k)$ on $\mathbb{R}^{kl+k-1}$. Hence, if the probability density functions being mixed are continuous on their parameters, the type I mixture is continuous on $\bar{\Delta} = \Delta \cup \partial\Delta$ (where $\partial\Delta$ is the boundary of Δ).

Now, if (C.3) holds but (C.2) fails, then there may exist a type I mixture of k components identical to a type I mixture obtained by mixing fewer than k components. That is, a P-identifiability problem occurs on $\bar{\Delta}$. This creates a crucial difficulty in testing for the presence of a mixture.

Moreover, if the class of type I $(\leq k)$ mixtures is not T-identifiable despite T-identifiability of the class of type I (k) mixtures, the mapping from $\bar{\Delta}$ to the space of probability density functions need not be bicontinuous. If (C.3) holds but (C.2) fails, let $\tilde{\lambda}$ be a point of $\partial\Delta$ and let λ be a point of $\bar{\Delta}$ such that both represent the same mixture density function $\tilde{f}$. Then there exist disjoint neighborhoods $\tilde{N}$ of $\tilde{\lambda}$ and N of λ. Consider the mixture density functions corresponding to parameter values along a path through $\tilde{N}$ to the point $\tilde{\lambda}$ and then along a path from λ through N. As the mixture density functions vary smoothly, the parameters must "jump" from $\tilde{N}$ to N at the parameter values for $\tilde{f}$.

Alternatively, suppose that condition (C.2) holds. Then the class of all type I (k) mixtures parametrized by $(p_1, \dots, p_{k-1}, \theta_1, \dots, \theta_k) \in \bar{\Delta}$ is P-identifiable, although the class of all type I $(\leq k)$ mixtures need not be P-identifiable.

To obtain a better parametrization, consider the following definition of identifiability.

DEFINITION 2.3. A parametrization for a class of mixture density functions is said to be B-identifiable if the correspondence between a mixture density function in this class and its representing parameter is bicontinuous.

Note that B-identifiability is a more strict definition of identifiability than either P-identifiability or T-identifiability. Even if condition (C.2) holds and if the probability density functions being mixed are continuous on their parameters, the parametrization for the class of all type I (k) mixtures parametrized by $(p_1, \dots, p_{k-1}, \theta_1, \dots, \theta_k) \in \bar{\Delta}$ need not be B-identifiable. The B-identifiability problems occur on $\partial\Delta$.

3. Parameter spaces. Suppose that the null hypothesis is that the random vector $\mathbf{X}$ does *not* have a mixture density function; that is, the probability density function of an observation $\mathbf{x}$ is simply $f(\mathbf{x}|\theta)$ for some parameter θ in $\Theta \subset \mathbb{R}^l$. Then the class of all probability density functions under the null hypothesis is the family $\mathscr{F}_0 = \{f(\mathbf{x}|\theta)\colon \theta \in \Theta\}$.

The alternative hypothesis is that the random vector $\mathbf{X}$ has a type I (k) mixture density function for some parameter $(p_1, p_2, \dots, p_{k-1}, \theta_1, \theta_2, \dots, \theta_k)$ in $\Delta \subset \mathbb{R}^{kl+k-1}$. Then the class of all probability density functions under the alternative hypothesis is $\mathscr{F}_1$, the class of all type I (k) mixtures. Since $\mathscr{F}_0$ is naturally parametrized by Θ and $\mathscr{F}_1$ is parametrized by Δ, it is intuitively appealing to think of Θ as the parameter space for the null hypothesis and Δ as the parameter space for the alternative hypothesis. However, these are two distinct spaces; so some other parameter space must be used to test the null hypothesis of no mixture.

To construct an appropriate parameter space, first the relationship of $\mathscr{F}_1$ to $\mathscr{F}_0$ and the induced relationship of Δ to $\partial\Delta$ are considered. Then, using a correspondence between $\partial\Delta$ and Θ and identifying points in $\Delta \cup \partial\Delta$ which represent a single (unmixed) density function, an appropriate space is found.

Note first that for a type I (k) mixture,

$$f(\mathbf{x}) = \sum_{i=1}^{k} p_i f(\mathbf{x}|\boldsymbol{\theta}_i); \quad \text{and as } p_i \text{ approaches 1 for some } i,$$

$f(\mathbf{x})$ approaches a probability density function $f(\mathbf{x}|\boldsymbol{\theta}_i)$ in $\mathscr{F}_0$. Note also that $\mathscr{F}_1 \cup \mathscr{F}_0$ is (arcwise) connected. Thus for any probability density function in $\mathscr{F}_0$, there is a boundary point of Δ, $\tilde{\lambda} \in \partial\Delta$, to represent it. One useful characterization of such a $\tilde{\lambda}$ is

$$\tilde{\lambda} = \left(\tilde{p}_1, \tilde{p}_2, \dots, \tilde{p}_{k-1}, \tilde{\boldsymbol{\theta}}_1, \tilde{\boldsymbol{\theta}}_2, \dots, \tilde{\boldsymbol{\theta}}_k\right), \tag{3.1}$$

where for some $\boldsymbol{\theta} \in \Theta$, $\tilde{p}_i > 0$ only if $\tilde{\boldsymbol{\theta}}_i = \boldsymbol{\theta}$. Let $S = \{\tilde{\lambda}\}$; then $S \subset \partial\Delta$.

The following lemma summarizes the T-, P- and B-identifiability of these parametrizations.

Lemma 3.1. *Assume T-identifiability holds for the class of all type* I ($\leq k$) *mixtures. Then*

(i) *the parametrization for $\mathscr{F}_1$ by Δ is P-identifiable;*

(ii) *the parametrization for $\mathscr{F}_0$ by S is not P-identifiable;*

(iii) *the parametrization for $\mathscr{F}_0 \cup \mathscr{F}_1$ is neither P-identifiable nor B-identifiable.*

The statistical identifiability problem for testing for the presence of a mixture is the existence of multiple representations (points in S) for a single probability density function in $\mathscr{F}_0$. Consider a mapping which identifies these to a single point.

Theorem 3.1. *Suppose that T-identifiability holds for the class of all type I* ($\leq k$) *mixtures, and that Θ is arcwise connected. If a mapping $\sim$ from $T = \Delta \cup S$ onto $\tilde{T}$ satisfies*

(i) $\sim$ *is continuous on T,*

(ii) $\sim$ *is one-to-one on Δ,*

(iii) $\tilde{\Delta} \cap \tilde{S} = \varnothing$,

(iv) $\sim$ *maps two points in S satisfying* (3.1) *to the same point in $\tilde{S}$ if and only if they represent the same probability density function, and maps any two points not satisfying* (3.1) *for the same $\boldsymbol{\theta} \in \Theta$ into distinct points in $\tilde{S}$,*

then the new parametrization by $\sim (p_1, p_2, \dots, p_{k-1}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \dots, \boldsymbol{\theta}_k)$ is P-identifiable.

PROOF. By Lemma 3.1, the original parametrization for $\mathscr{F}_1$ by $(p_1, p_2, \dots, p_{k-1}, \theta_1, \theta_2, \dots, \theta_k) \in T$ is P-identifiable. Because $\sim$ is one-to-one on Δ, the new parametrization for $\mathscr{F}_1$ by $\sim (p_1, p_2, \dots, p_{k-1}, \theta_1, \theta_2, \dots, \theta_k) \in \tilde{T}$ is also P-identifiable. Since there is no density function in $\mathscr{F}_0$ giving an alternative representation for a density in $\mathscr{F}_1$, there is no P-identifiability problem between Δ and S; and as $\tilde{\Delta} \cap \tilde{S} = \varnothing$, there is no P-identifiability problem between $\tilde{\Delta}$ and $\tilde{S}$ either. Hence, the only possible P-identifiability problem is on S. But $\sim$ maps all points in S satisfying (3.1) for any particular $\theta \in \Theta$ to one point in $\tilde{S}$, so the P-identifiability problem no longer exists on $\tilde{S}$. Therefore, the new parametrization by $\sim (p_1, p_2, \dots, p_{k-1}, \theta_1, \theta_2, \dots, \theta_k)$ is P-identifiable. □

The existence of such an identification mapping $\sim$ satisfying requirements (i)–(iv), is demonstrated with the following construction of one such mapping in two steps.

Define the mapping φ from T into $\mathbb{R}^{kl+k-1}$ by

$$\begin{aligned}\varphi(&p_1, p_2, \dots, p_{k-1}, \theta_1, \theta_2, \dots, \theta_k)\\ &= \Big(p_1, p_2, \dots, p_{k-1}, (p_1 p_2 \cdots p_k) \cdot (\theta_1 - \theta_2),\\ &\qquad (p_1 p_2 \cdots p_k) \cdot (\theta_3 - \theta_2), \dots, (p_1 p_2 \cdots p_k) \cdot (\theta_k - \theta_{k-1}), \sum_{i=1}^{k} p_i \theta_i\Big),\end{aligned} \tag{3.2}$$

where $p_k = 1 - \sum_{i=1}^{k-1} p_i$. Then φ is C^∞-differentiable on T and homeomorphic on Δ. Further, $\varphi\Delta \cap \varphi S = \varnothing$ and $\varphi S \subset \partial\varphi\Delta$.

Provided that the class of all type I ($\le k$) mixtures is T-identifiable, the reparametrized space $\varphi\Delta$ for the alternative hypothesis is P-identifiable. Although the reparametrized spaces φS and φT for the null hypothesis and for the alternative hypothesis are not P-identifiable, there is no P-identifiability problem between $\varphi\Delta$ and φS.

Next, define a second mapping ψ from φT into $\mathbb{R}^{kl+k-1}$ by

$$\begin{aligned}\psi(&p_1, p_2, \dots, p_{k-1}, \theta_1, \theta_2, \dots, \theta_k)\\ &= \Big(\frac{1}{k} + p_1\|\theta_1\|^2, \frac{1}{k} + p_2\|\theta_2\|^2, \dots, \frac{1}{k} + p_{k-1}\|\theta_{k-1}\|^2,\\ &\qquad p_1\theta_1, p_2\theta_2, \dots, p_{k-1}\theta_{k-1}, \theta_k\Big),\end{aligned} \tag{3.3}$$

where $p_k = 1 - \sum_{i=1}^{k-1} p_i$. Then ψ is C^∞-differentiable on φT and homeomorphic on $\varphi\Delta$. Further, $\psi \circ \varphi\Delta \cap \psi \circ \varphi S = \varnothing$ and $\psi \circ \varphi S \subset \partial\psi \circ \varphi\Delta$.

Finally, take the mapping $\sim$ to be $\psi \circ \varphi$. Then the parametrization for testing the hypothesis by $\sim (p_1, p_2, \dots, p_{k-1}, \theta_1, \theta_2, \dots, \theta_k) \in \tilde{T}$ is P-identifiable; and $\tilde{S}$ is the parameter space for the null hypothesis while $\tilde{A}$ is the parameter space for the alternative hypothesis.

The mapping constructed in Theorem 3.1 is a projection onto a quotient space defined in the following way. Let $\mathscr{P}(\mathbf{x})$ be the family of all distinct probability density functions under consideration (mixed and unmixed). Then the function g maps T into $\mathscr{P}$; and $\sim$ is an equivalence relation defined for $t_1, t_2 \in T$ by $t_1 \sim t_2$ if and only if $g(t_1) = g(t_2)$. Thus $\tilde{T}$ is the quotient space (often written $T/\sim$) with the natural topology being the quotient topology. This new (identifiable) parametrization defines $\tilde{g}$ which maps $\tilde{T}$ into $\mathscr{P}$. The continuity of g induces continuity of $\tilde{g}$. Thus $\tilde{g}$ is P-identifiable.

The bicontinuity of $\tilde{g}$, and hence the B-identifiability of $\tilde{T}$, however, depend upon the mixture density family of functions g, that is, whether the parametrization is either open or closed.

THEOREM 3.2. *If g is a continuous mapping onto $\mathscr{P}$ and $\sim$ is the equivalence relation defined for $t_1, t_2 \in \tilde{T}$ by $t_1 \sim t_2$ if and only if $g(t_1) = g(t_2)$, then $\tilde{T}$ is B-identifiable if and only if g is either open or closed.*

PROOF. A particular case of Theorem 11.2 in Munkres (1975). □

COROLLARY. *If the class of all type* I $(\leq k)$ *mixtures is T-identifiable, if g is either open or closed and if $\sim$ satisfies conditions* (i)–(iv) *in Theorem* 3.1, *then if $\mathscr{P}$ is Hausdorff* [*as with the usual* (*weak or Prohorov*) *topologies*], *so is $\tilde{T}$ and the parametrization by $\sim (p_1, p_2, \dots, p_{k-1}, \theta_1, \theta_2, \dots, \theta_k)$ is B-identifiable.*

(In point of fact, the condition of T-identifiability in this corollary is not essential. However, the equivalence relation would no longer be one-to-one on Δ in the absence of T-identifiability.)

Hence, the particular mapping $\sim = \psi \circ \varphi$ defined in (3.2) and (3.3) gives an appropriate parametrization for testing the null hypothesis of no mixture:

$$\sim (p_1, p_2, \dots, p_{k-1}, \theta_1, \theta_2, \dots, \theta_k) \in \tilde{T}.$$

A stronger condition which often is easily checked is sufficient for B-identifiability.

COROLLARY. *Suppose that the class of all type* I $(\leq k)$ *mixtures is T-identifiable, that Θ is compact and that $\mathscr{P} = \mathscr{F}_0 \cup \mathscr{F}_1$ is Hausdorff. If a mapping $\sim$ from $T = \Delta \cup S$ onto $\tilde{T}$ satisfies* (i)–(iv) *in Theorem* 3.1 *and also satisfies*

(v) *the parametrization of $\mathscr{P}$ by T is continuous,*

then the parametrization by $\sim (p_1, p_2, \dots, p_{k-1}, \theta_1, \theta_2, \dots, \theta_k) \in \tilde{T}$ is B-identifiable.

4. Bayesian inference about a mixture. Return now to the problem of determining whether or not there is a mixture. Borrowing both notation and vocabulary from frequentist theory, let H_0 denote the null hypothesis that the sample comes from an unmixed distribution; and let H_1 denote the alternative hypothesis that the sample comes from a type I (k) mixture distribution.

It is clear that a traditional (absolutely) continuous prior with respect to $(kl + k - 1)$-dimensional Lebesgue measure on $\tilde{T}$ is improper for this problem since the $(kl + k - 1)$-dimensional Lebesgue measure of $\tilde{S}$, the parameter space for the null hypothesis, is 0. Thus neither positive prior nor posterior knowledge about $\tilde{S}$ can be expressed using $(kl + k - 1)$-dimensional Lebesgue measure.

Hence, a mixture prior distribution of different dimensionalities is used,

$$P = \alpha P_0 + (1 - \alpha)P_1, \tag{4.1}$$

where $0 \le \alpha \le 1$, and P_0, P_1 are probability measures on the parameter spaces for the null and alternative hypotheses, respectively.

For the null hypothesis of no mixture, $\tilde{S}$ has dimension l; and P_0 is an l-dimensional continuous probability measure. For the alternative hypothesis of type I (k) mixture, $\tilde{\Delta}$ has dimension $kl + k - 1$; and, for simplicity, let P_1 be a $(kl + k - 1)$-dimensional continuous probability measure. Define π_1 to be the $(kl + k - 1)$-dimensional density function of P_1 on $\tilde{\Delta}$ and to be 0 otherwise. Thus for every Borel subset B of $\mathbb{R}^{kl+k-1}$,

$$P_1(B) = \int \pi_1 B \, dH^{kl+k-1}, \tag{4.2}$$

where $\int dH^{kl+k-1}$ is the integral taken with respect to $(kl + k - 1)$-dimensional measure, or equivalently (on $\mathbb{R}^{kl+k-1}$),

$$P_1(B) = \int \pi_1 B \, dL^{kl+k-1}. \tag{4.3}$$

Similarly, define π_0 to be an l-dimensional density function for P_0 such that $\pi_0 = 0$ off $\tilde{S}$. Then for $\Gamma(\boldsymbol{\theta})$, a parametrization of $\tilde{S}$,

$$P_0(B) = \int \pi_0 \Gamma(\boldsymbol{\theta}) |HJ_\Gamma(\boldsymbol{\theta})| \Gamma^{-1}(B) \, d\boldsymbol{\theta}, \tag{4.4}$$

where $HJ_\Gamma(\boldsymbol{\theta})$ is the Jacobian of Γ. A simple choice of $\Gamma(\boldsymbol{\theta}) = (1/k, 1/k, \dots, 1/k, \mathbf{0}, \mathbf{0}, \dots, \mathbf{0}, \boldsymbol{\theta})$ reduces the expression in (4.4) to

$$P_0(B) = \int \pi_0 \left(\frac{1}{k}, \frac{1}{k}, \dots, \frac{1}{k}, \mathbf{0}, \mathbf{0}, \dots, \mathbf{0}, \boldsymbol{\theta} \right) \Gamma^{-1}(B) \, d\boldsymbol{\theta}. \tag{4.5}$$

Given data, $D = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$, where $\mathbf{x}_i$ is a realization of the random vector $\mathbf{X}$ on $\mathbb{R}^m$, $i = 1, \dots, n$, the likelihood under the null hypothesis is

$$\prod_{i=1}^{n} f(\mathbf{x}_i | \boldsymbol{\theta}), \quad \text{where } \boldsymbol{\theta} \in \Theta,$$

and the likelihood under the alternative hypothesis is

$$\prod_{i=1}^{n} \left[\sum_{j=1}^{k} p_j f(\mathbf{x}_i | \boldsymbol{\theta}_j) \right],$$

where $\boldsymbol{\theta}_j \in \Theta$ for $j = 1, 2, \dots, k$, $\boldsymbol{\theta}_1 < \boldsymbol{\theta}_2 < \cdots < \boldsymbol{\theta}_k$, and $0 < p_j < 1$, $\sum_{j=1}^{k} p_j = 1$.

Both the Bayes factor and the posterior odds ratio have commonly been used in Bayesian decision making. The Bayes factor comparing these null and alternative hypotheses is

$$B_{01} = \frac{P(H_0|D)}{P(H_1|D)} \bigg/ \frac{P(H_0)}{P(H_1)} = \frac{P(D|H_0)}{P(D|H_1)}, \tag{4.6}$$

where H_0, H_1 denote the null and alternative hypotheses, respectively.

Substitution of (4.3) and (4.5) into (4.6) gives

$$\begin{aligned} B_{01} = {} & \int \prod_{i=1}^{n} f(\mathbf{x}_i|\boldsymbol{\theta}) \pi_0\left(\frac{1}{k}, \frac{1}{k}, \dots, \frac{1}{k}, 0, 0, \dots, 0, \boldsymbol{\theta}\right) d\boldsymbol{\theta} \\ & \div \int \prod_{i=1}^{n} \left[\sum_{j=1}^{k} p_j f(\mathbf{x}_i|\boldsymbol{\theta}_j)\right] \pi_1(s_1, s_2, \dots, s_{k-1}, \mathbf{t}_1, \mathbf{t}_2, \dots, \mathbf{t}_k) \\ & \times d(s_1, \dots, s_{k-1}, \mathbf{t}_1, \dots, \mathbf{t}_k), \end{aligned} \tag{4.7}$$

where

$$(p_1, p_2, \dots, p_{k-1}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \dots, \boldsymbol{\theta}_k) = \sim^{-1}(s_1, s_2, \dots, s_{k-1}, \mathbf{t}_1, \mathbf{t}_2, \dots, \mathbf{t}_k)$$

in the denominator, $p_k = 1 - \sum_{i=1}^{k-1} p_i$ and $\sim^{-1} = \varphi^{-1} \circ \psi^{-1}$ [φ and ψ are defined in (3.2) and 3.3)].

The posterior odds ratio is defined for this problem as

$$K_{01} = \frac{P(H_0|D)}{P(H_1|D)} \tag{4.8}$$

which, with substitution of (4.5) and (4.3), gives

$$\begin{aligned} K_{01} = {} & \alpha \int \prod_{i=1}^{n} f(\mathbf{x}_i|\boldsymbol{\theta}) \pi_0\left(\frac{1}{k}, \frac{1}{k}, \dots, \frac{1}{k}, 0, 0, \dots, 0, \boldsymbol{\theta}\right) d\boldsymbol{\theta} \\ & \div (1-\alpha) \int \prod_{i=1}^{n} \left[\sum_{j=1}^{k} p_j f(\mathbf{x}_i|\boldsymbol{\theta}_j)\right] \pi_1(s_1, s_2, \dots, s_{k-1}, \mathbf{t}_1, \mathbf{t}_2, \dots, \mathbf{t}_k) \\ & \times d(s_1, \dots, s_{k-1}, \mathbf{t}_1, \dots, \mathbf{t}_k). \end{aligned} \tag{4.9}$$

The Bayes factor or the posterior odds ratio is then interpreted in the usual way. That is, a large value of B_{01} or K_{01} supports the null hypothesis while a value less than 1 supports the alternative hypothesis. [See Spiegelhalter and Smith (1982) for discussion of the interpretation of Bayes factor values.]

From the initial formulation of the problem using Lebesgue $(kl + k - 1)$-dimensional measure, Hausdorff measure is a natural but not essential choice for defining integrals. The role of Hausdorff integrals in defining conditional probabilities is discussed in Li and Sedransk (1984).

5. Discussion. The method of construction illustrated in the preceding section is a general one in that it relies only upon the projection $\sim$ onto the quotient space $\tilde{T}$ and on properties of the family of probability density functions being mixed. Thus, this approach can also be applied to mixtures of distributions from different families (type II mixtures). That is, let $f_j(\mathbf{x}|\boldsymbol{\theta}^j)$ be a probability density function with respect to m-dimensional Lebesgue measure for $\mathbf{x} = (x_1, x_2, \dots, x_m)$ on $\mathbb{R}^m$ given the parameter $\theta^j \in \Theta_j$, where Θ_j is a measurable subset of $\mathbb{R}^{l_j}$, $j = 1, 2, \dots, k$. If the mixing probability measure Q assigns positive probability p_j to the point $\theta^j \in \Theta_j$ for $j = 1, 2, \dots, k$, then the mixture density function for the random vector $\mathbf{X} = (X_1, X_2, \dots, X_m)$ on $\mathbb{R}^m$ is of the form

$$f(\mathbf{x}) = \sum_{j=1}^{k} p_j f_j(\mathbf{x}|\theta^j). \tag{5.1}$$

where $\theta^j \in \Theta_j$, $0 < p_j < 1$ for $j = 1, 2, \dots, k$, and $\sum_{i=1}^{k} p_j = 1$.

In general, exact analogues of the results for type I mixtures hold for type II mixtures [with the exception that the parametrization of a T-identifiable class of all type II (k) mixtures parametrized by $(p_1, \dots, p_{k-1}, \theta^1, \dots, \theta^k) \in (0,1)^{k-1} \times \Theta_1 \times \cdots \times \Theta_k$ *is* P-identifiable; see Li (1983)]. Thus, the procedures to detect a type II mixture or to determine whether a mixture has k components [against the alternative of $k'(< k)$ components] are variants of the procedure given in the preceding section where an appropriate choice of the mapping $\sim$ is made in each case. Furthermore, if the class of mixtures is enlarged to include mixtures obtained by simultaneously mixing probability density functions from the same and different families,

$$f(\mathbf{x}) = \sum_{j=1}^{k} \sum_{i=1}^{k_j} p_{ij} f_j(\mathbf{x}|\theta_{ij}),$$

where $0 < p_{ij} < 1$, $\theta_{ij} \in \Theta_j$ for $i = 1, 2, \dots, k_j$, $j = 1, 2, \dots, k$, and $\sum_{i,j} p_{ij} = 1$, the same procedure can be applied, although $\sim$ will in general be a more complicated function.

Recognizing the necessity of B-identifiability, or equivalently, using the quotient space $\tilde{T}$, also resolves some confusion arising from simulation studies of the estimation problem for mixtures. For example, in estimating (p, θ_1, θ_2) for a type I (2) mixture, Chiang (1951), Robertson and Fryer (1972) and Tan and Chang (1972), have variously observed that: (i) the estimation algorithm begins looping as p approaches 1, (ii) the estimated values for (p, θ_1, θ_2) are most accurate when p is not close to either 1 or 0 and (iii) the variance of the estimator increases as $\theta_1 - \theta_2$ approaches 0. All these phenomena result from use of a non-B-identifiable parametrization.

To see this, consider the estimation problem when $\theta_1 - \theta_2 < \delta$, for some small $\delta > 0$; $0 < p < 1$; $\theta_1, \theta_2 \in \mathbb{R}$. Then there exists $\theta' \in \mathbb{R}$ such that (p, θ_1, θ_2) is in an ε-neighborhood, N_1, of the boundary point (p, θ', θ'). A single density $f(x|\theta')$ is multiply represented by (p, θ', θ'), $(0, \theta_1^*, \theta')$ and $(1, \theta', \theta_2^{**})$ for all choices of $p, \theta_1^*, \theta_2^{**}$. Consider two points, one in N_2 and one in N_3, where N_2 and N_3 are ε-neighborhoods of $(0, \theta_1^*, \theta')$ and $(1, \theta', \theta_2^{**})$, respectively. As probability functions vary smoothly through the probability density function $f(x|\theta')$, the corresponding parameters can jump discontinuously among N_1, N_2, N_3. Since most parametric estimates (either method of moments or maximum likelihood) are obtained as continuous functions of the probability density function, these will also show the discontinuous behavior (as with the parameters) near boundary values. The looping for values of p close to 1 is just the jumping among neighborhoods of various representations of $f(x|\theta')$. The inaccuracy of estimated values for (p, θ_1, θ_2) occurs when the algorithm finally does converge in some neighborhood other than N_2 (or N_3) of a representation of $f(x|\theta')$. The increased variance of the estimated (p, θ_1, θ_2) occurs when $\theta_1 - \theta_2$ is close to 0 because although the estimated density function is close to $f(x|\theta')$, the estimated (p, θ_1, θ_2) is in some other neighborhood than N_1. This behavior would argue against an iteration termination criterion based on difference of consecutive iterates of parameter estimates (a strategy widely used with the EM algorithm). A more reasonable termination criterion might be based on the similarity of the densities corresponding to consecutive parameter iterates. However, difficulties of this kind are naturally obviated when looking directly at the quotient space $\tilde{T}$.

Acknowledgments. The authors wish to thank the several referees and the Associate Editor for their careful reading of this paper and particularly for their constructive suggestions.

REFERENCES

BAKER, G. A. (1958). Empiric investigation of a test of homogeneity for populations composed of normal distributions. *J. Amer. Statist. Assoc.* **53** 551–557.

BINDER, D. A. (1978). Bayesian cluster analysis. *Biometrika* **65** 31–38.

CHANDRA, S. (1977). On the mixtures of probability distributions. *Scand. J. Statist.* **4** 105–112.

CHIANG, C. L. (1951). On the design of mass medical surveys. *Human Biol.* **23** 242–271.

COX, D. R. (1966). Notes on the analysis of mixed frequency distributions. *British J. Math. Statist. Psych.* **19** 39–47.

EVERITT, B. S. (1978). *Graphical Techniques for Multivariate Data.* Heinemann, London.

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

FOWLKES, E. B. (1979). Some methods for studying the mixture of two normal distributions. *J. Amer. Statist. Assoc.* **74** 561–575.

HAZEN, A. (1914). Storage to be provided in impounding reservoirs for municipal water supply. *Trans. Amer. Soc. Civil Engr.* **77** 1539–1659.

JOHNSON, N. L. (1973). Some simple tests of mixtures with symmetrical components. *Comm. Statist. A—Theory Methods* **1** 17–25.

LI, L. A. (1983). Decomposition theorems, conditional probability, and finite mixtures distributions. Thesis, State Univ. New York, Albany.

LI, L. A. and SEDRANSK, N. (1982). Inference about the presence of a mixture. Technical Report, State Univ. New York, Albany.

LI, L. A. and SEDRANSK, N. (1984). A generalized definition of conditional probability. Technical Report, State Univ. New York, Albany.

MUNKRES, J. R. (1975). *Topology.* Prentice-Hall, Englewood Cliffs, N.J.

MURPHY, E. A. (1964). One cause? Many causes? The argument from the bimodal distribution. *J. Chron. Dis.* **17** 301–324.

PEARSON, K. (1894). Contribution to the mathematical theory of evolution. *Philos. Trans. Roy. Soc. London Ser. A* **185** 71–110.

ROBERTSON, C. A. and FRYER, J. G. (1972). A comparison of some methods for estimating mixed normal distributions. *Biometrika* **59** 639–648.

SPIEGELHALTER, D. J. and SMITH, A. F. M. (1982). Bayes factors for linear and log-linear models with vague prior information. *J. Roy. Statist. Soc. Ser. B* **44** 377–387.

TAN, W. Y. and CHANG, W. C. (1972). 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** 702–708.

TEICHER, H. (1960). On the mixture of distributions. *Ann. Math. Statist.* **31** 55–73.

TEICHER, H. (1961). Identifiability of mixtures. *Ann. Math. Statist.* **32** 244–248.

TEICHER, H. (1963). Identifiability of finite mixtures. *Ann. Math. Statist.* **34** 1265–1269.

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

THOMAS, E. A. C. (1969). Distribution free tests for mixed probability distributions. *Biometrika* **56** 475–484.

TIAGO DE OLIVEIRA, J. (1965). Some elementary tests for mixtures of discrete distributions. In *Classical and Contagious Discrete Distributions* (*Proc. Internat. Symposium, McGill Univ., 1963*) 379–384. Statistical Publishing Society, Calcutta.

YAKOWITZ, S. J. and SPRAGINS, J. D. (1968). On the identifiability of finite mixtures. *Ann. Math. Statist.* **39** 209–214.

INSTITUTE OF STATISTICS
ACADEMIA SINICA
TAIPEI, TAIWAN
REPUBLIC OF CHINA

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF IOWA
IOWA CITY, IOWA 52242