The Annals of Mathematical Statistics
1969, Vol. 40, No. 5, 1728–1735

A CONSISTENT ESTIMATOR FOR THE IDENTIFICATION OF FINITE MIXTURES

By S. Yakowitz

University of Arizona

0. Summary. Henry Teicher [10] has initiated a systematic study called "identifiability of finite mixtures" (these terms to be defined in Section 1) which has significance in several areas of statistics. [10] gives a sufficiency condition that a family $\mathfrak{F}$ of cdf's (cumulative distribution functions) generate identifiable finite mixtures, and consequently establishes that finite mixtures of the one-dimensional Gaussian or gamma families are identifiable. From [9] it is known that the Poisson family generates identifiable finite mixtures, and the binomial and uniform families do not. In [11], Teicher proves that the class of mixtures of n products of any identifiable one-dimensional family is likewise identifiable (and that the analogous statement for finite mixtures is valid). Spragins and I have shown [13] that the finite mixtures on a family of cdf's is identifiable if and only if $\mathfrak{F}$ is linearly independent in its span over the real numbers, and that $\mathfrak{F}$ generates identifiable finite mixtures if $\mathfrak{F}$ is any of the following: the n-dimensional normal family, the union of the n-dimensional normal family and the family of n products of one dimensional exponential distributions, the Cauchy family, the negative binomial family, and the translation parameter family generated by any one dimensional cdf. (In this last case, our proof directly generalizes to any n-dimensional translation parameter family.)

In view of the fact that many of the important distribution families have been seen to give identifiable finite mixtures, it would seem appropriate to seek methods for performing this identification, and therefore the intention of this paper is to reveal (Section 2) a general algorithm for construction of a consistent estimator. In Section 3 we demonstrate that the algorithm is effective for all the identifiable families mentioned above. Our results, in addition to having application to an interesting problem in communication theory [6], can be used to extend the empiric Bayes approach to a certain type of decision problem. Section 4 will discuss the details of this application.

1. Introduction. Let

$$\mathfrak{F} = \{F(x; \alpha) : \alpha \, \varepsilon \, R_1^m\} \tag{1}$$

be a family of n-dimensional cdf's indexed by a parameter α whose domain is parameter space R_1^m, a Borel subset of Euclidean m-space. Define $\mathcal{G}$ to be all discrete m-dimensional cdf's G which have only finitely many mass points and whose measures μ_G assign probability 1 to R_1^m. Q is defined to be the mapping

$$Q(G) = \int_{R_1^m} F(x; \alpha)\, dG(\alpha) = \sum F(x; \alpha_i) p(\alpha_i) \quad \alpha_i \text{: a mass point of } G. \tag{2}$$

$p(\alpha_i)$ is the mass that G assigns to $\alpha_i \, \varepsilon \, R_1^m$. The subscripting in (2) is not in-

Received 13 December 1967; revised 25 February 1969.

tended to imply anything about the cardinality of $\mathfrak{F}$. A cdf such as $Q(G)$ is termed a *finite mixture* on $\mathfrak{F}$. The cdf G is called a *mixing distribution*. From the description of the set $\mathcal{G}$ it is evident that $\mathfrak{H} = Q(\mathcal{G})$ is the convex hull of $\mathfrak{F}$. It is consistent with earlier papers to say that the set $\mathfrak{H}$ of finite mixtures of $\mathfrak{F}$ is *identifiable* if Q has an inverse, denoted Q^{-1}. That is, $\mathfrak{H}$ is identifiable if and only if the equation $\sum_{i=1}^{N} p_i F(x; \alpha_i) = \sum_{i=1}^{M} p_i' F(x; \alpha_i')$ implies that $N = M$, and for each i, $1 \leqq i \leqq N$, there is some j, $1 \leqq j \leqq N$, such that $p_i = p_j'$ and $\alpha_i = \alpha_j'$.

The question of identifiability of finite mixtures arises, for example, in the situation in which a finite set of experiments $\{\mathbf{E}_1, \cdots \mathbf{E}_N\}$ gives rise to a sequence of rv's $\{X_i\}$ as follows: At each observation time i with probability p_j, at the exclusion of the other experiments, experiment $\mathbf{E}_j$, (whose cdf is $F_j(x)$) is selected and an observation x is made. x is taken to be the ith element of an observation of the sequence $\{X_i\}$. The statistician does not know the parameters $p_1, \cdots p_N$ nor the cdf's $F_1, \cdots, F_N$ or even the value of N. He is told that the distributions $F_1, \cdots, F_N$ are distinct and are all members of a specified family $\mathfrak{F}$ as defined in (1). The *identification problem* is the problem of determining the number N, the parameters $p_1, \cdots, p_N$, and the distributions $F_1, \cdots, F_N$ solely on the basis of knowledge of $\mathfrak{F}$ and an observation of the sequence $\{X_i\}$ determined as we have described. It is to be emphasized that the statistician is never told which experiment was performed at any time.

It is evident that a solution exists to the identification problem only if $\mathfrak{F}$ generates identifiable mixtures. For if Q^{-1} is not defined (Q not one-to-one), then observations of $\{X_i\}$, being distributed as $Q(G) = Q(G')$ for some G, G' are of no avail in deciding whether G or G' is the actual mixing distribution.

The purpose of this paper is to show that whenever a parametric family of distribution functions yields identifiable finite mixtures and is weakly continuous with respect to the parameter (that is $(\theta_i \to \theta)$ implies $(F_{\theta_i} \Rightarrow F_\theta)$), then by the algorithm we propose, a consistent estimator for the mixing distribution G can be found.

Our algorithm bears some resemblance to the Wolfowitz distance method [12] (which, however, is not directly applicable to the identification problem). We intend to explore this relation in a future note to this journal.

Choi [2] has given an estimator for the mixing distribution which is effective if the statistician knows in advance the number of experiments giving rise to the mixture. In some cases, Choi has been able to establish a rate of convergence of his estimator. Boes [1] has published a method for estimating the mixing distribution when $\mathfrak{F}$ is a finite set. The author is indebted to the referee for calling his attention to these last two papers.

2. Construction of an estimator for an unknown mixing distribution in $\mathcal{G}$. For the various sets of cdf's which enter our discussions, it will serve our purposes to use an n-dimensional version of the Lévy metric:

(3) $L(F, G) =$ infimum over all positive numbers ϵ which for all $x \,\varepsilon\, R^n$ satisfy the inequality

$$F(x - (\epsilon, \epsilon, \cdots, \epsilon)) - \epsilon \leqq G(x) \leqq F(x + (\epsilon, \epsilon, \cdots, \epsilon)) + \epsilon.$$

In this regard, a fact which will be useful in the sequel is

LEMMA. *L convergence is equivalent to weak convergence, and L distance never exceeds Prohorov distance.*

PROOF. It is immediate from the definition of L that if F_i converges in L to G, then F_i converges pointwise to G at every point of continuity of the latter cdf, i.e. F_i converges weakly to G.

Prohorov [7] proved that the metric P defined below is equivalent to weak convergence.

$$(4)\quad P(G, F) = \max \{\epsilon_{FG}, \epsilon_{GF}\}, \qquad \text{where}$$

$$(5)\qquad \epsilon_{FG} = \inf \epsilon \text{ such that for all closed sets } A,\ \mu_F(A) \leqq \mu_G(A^\epsilon) + \epsilon,$$

where $A^\epsilon \equiv \{x : \|x - y\| < \epsilon \text{ for some } y \varepsilon A\}$. ϵ_{GF} is defined similarly. μ_F is the measure induced by F.

For any $q > P(F, G)$ and all $x \varepsilon R^n$,

$$(6)\quad F(x) = \mu_F(-\infty, x] \leqq \mu_G(-\infty, x + \bar{q}) + q \leqq \mu_G(-\infty, x + \bar{q}] + q = G(x + \bar{q}) + q,$$

where $\bar{q} = (q, q, \cdots, q)$, and $(-\infty, x)$ denotes the points in R^n whose coordinates are less than the respective coordinates of x. The closed n-dimensional interval is designated in the same fashion.

For any $q > P(F, G)$ and all $x \varepsilon R^n$,

$$(7)\quad G(x - \bar{q}) = \mu_G(-\infty, x - \bar{q}] \leqq \mu_F(-\infty, x) + q \leqq \mu_F(-\infty, x] + q = F(x) + q.$$

This gives $G(x - \bar{q}) - q \leqq F(x)$, and from (6) and (7) we conclude that $P(F, G) \geqq L(F, G)$, and further, that weak convergence implies convergence L.

THEOREM 1. *If F_θ is continuous (L) with respect to its parameter θ, then Q is continuous.*

PROOF. In the sequel, if F is any cdf, ψ_F denotes its characteristic function. Also $\psi_\theta(t)$ denotes the characteristics function of F_θ. Suppose $G_j \rightarrow_L G$ (and thus $G_j \Rightarrow G$), and that F_θ is continuous (L) with respect to its parameter θ. By the lemma and the equivalence of weak convergence and pointwise convergence of characteristic functions, $\psi_\theta(t)$, for t fixed, is a continuous bounded function of θ. Then for each $t \varepsilon R^n$,

$$(8)\qquad \int \psi_\theta(t)\, dG_j(\theta) \rightarrow \int \psi_\theta(t)\, dG(\theta).$$

The above equation shows that the characteristic functions of the sequence $\{Q(G_j)\}$ converge pointwise to $\psi_{Q(G)}$, implying that $Q(G_i) \Rightarrow Q(G)$. In view of the lemma, we thus have

$$(9)\qquad L(Q(G_i), Q(G)) \rightarrow 0.$$

Let n be any positive integer and B be any subset of parameter space, R_1^m, and define $\mathcal{G}(B, n)$ to be the members of $\mathcal{G}$ which have at most n mass points, all of which are contained in B and all of which are at least a distance $1/n$ apart. Then we have the following.

THEOREM 2. *If B is a compact subset of R_1^m, then $\mathcal{G}(B, n)$ is compact.*

PROOF. Since $(\mathcal{G}(B, n), L)$ is a metric space, it suffices to show that any sequence $\{G_j\}$ $(G_j \varepsilon \mathcal{G}(G, n), n \geqq 1)$ has a limit point $G \varepsilon \mathcal{G}(B, n)$. As in the proof to Theorem 1, it is convenient to regard G_j as its sequence of mass point pairs $G_j = \{(p_i(j), \theta_i(j))\}_{i=1}^{v(j)}$, $(v(j) \leqq n)$. Here the subscripting is arbitrary. Let p_1 be any limit point of $\{p_1(j)\}_{j=1}^{\infty}$ and ζ_1 a sequence of increasing positive integers such that $p_1(\zeta_1(j)) \rightarrow p_1$. θ_1 is any limit point of $\{\theta_1(\zeta_1(j))\}$ and ζ_2 an increasing subsequence of the integers in the range of ζ_1 such that $\theta_1(\zeta_2(j)) \rightarrow \theta_1$.

We continue this construction until we have found a sequence ζ_{2n} such that

$$(10) \qquad p_i(\zeta_{2n}(j)) \rightarrow p_i \quad \text{and} \quad \theta_i(\zeta_{2n}(j)) \rightarrow \theta_i, \qquad (1 \leqq i \leqq n).$$

The construction terminates at $v < n$ if only finitely many elements of $\{G_{\zeta_{2v}(j)}, j \geqq 1\}$ have $v + 1$ mass points. Define $G = \{(p_i, \theta_i)\}_{i=1}^{v}$ where either $v = n$ or (as just described) $v < n$ and observe that $G \varepsilon \mathcal{G}(B, n)$, i.e., $\sum_{i=1}^{n} p_i = 1$, $\|\theta_i - \theta_k\| \geqq 1/n$ if $i \neq k$, and $\theta_i \varepsilon B$. $\theta_i \varepsilon B$ because, by hypothesis, B is compact, and the other two assertions are a consequence of the fact that since for each j, $G_{\zeta_{2n}(j)} \varepsilon \mathcal{G}(B, n)$, for all j,

$$(11) \qquad \textstyle\sum_{i=1}^{n} p_i(\zeta_{2n}(j)) = 1, \quad \text{thus} \quad \sum_{i=1}^{n} p_i = 1, \quad \text{and}$$

$$\|\theta_i(\zeta_{2n}(j)) - \theta_k(\zeta_{2n}(j))\| \geqq 1/n, \quad \text{thus} \quad \|\theta_i - \theta_k\| \geqq 1/n.$$

It is evident that (10) implies that $G_{\zeta_{2n}(j)} \rightarrow_L G$, where at last we have returned to viewing G, G_j as cdf's rather than sequences of mass point pairs.

A final task in preparation for explaining the construction of the estimator for mixing distributions is to recall some facts about convergence of empiric distribution functions. If $\{x_i\}$, $(1 \leqq i \leqq v)$, is a sequence of v observations on an n-dimensional random variable, the associated empiric distribution function (df) is defined, for each n-tuple x, by: $vF_v(x)$ = number of observations in $\{x_i\}$ $(1 \leqq i \leqq v)$ such that no coordinate of x_i is greater than the corresponding coordinate of x. In Theorem 3, below, $\|H\| \equiv \sup_x |H(x)|$.

THEOREM 3. *Let n be any positive integer. F_v is the empiric* df *constructed from v independent n-dimensional* rv's *identically distributed according to the* cdf F. *If c and d are any positive numbers, one may compute a number $N(c, d)$ such that if $v > N(c, d)$, for all F,*

$$(12) \qquad P[\|F_v - F\| > c] < d.$$

The number $N(c, d)$ does not depend on F, but does depend on the dimension of the random variables X_i. Kiefer and Wolfowitz [4] have proved something much stronger than Theorem 3. Let n be any positive integer. There exist positive numbers c_0, c' such that for every n-dimensional cdf F, every positive integer v, and every positive number r,

$$(13) \qquad P[v^{\frac{1}{2}} \|F_v - F\| > r] < c_0 \exp(-c'r^2).$$

They give an effective procedure for bounding c' from above, and mention that c_0 can also effectively be bounded, thus justifying our assertion that $N(c, d)$ can be found. In the special case that $n = 1$, Massey [5] gives a recurrence relation from which the smallest possible $N(c, d)$ may be computed, in the case that F is known to be continuous. (This number is an upper bound in case F is not continuous.)

Algorithm for construction of a consistent estimator for a finite mixing distribution.

Conditions. $\mathfrak{F} = \{F_\theta : \theta \varepsilon R_1^m\}$. F_θ is continuous (L) with respect to θ, and $\mathfrak{F}$ generates identifiable finite mixtures. R_1^m is the limit of a monotonically increasing (by containment) sequence $\{B_j\}$ of compact subsets of R^m.

Step 1. For each positive integer j, find a number δ_j such that if H, $H' \varepsilon Q(\mathcal{G}(B_j, j))$ and $L(H, H') \leqq \delta_j$, then $L(Q^{-1}(H), Q^{-1}(H')) < 1/j$. (For each j, a number δ_j as described exists because Q is continuous (Theorem 1) and $\mathcal{G}(B_j, j)$ is compact (Theorem 2). Q^{-1} exists as $\mathfrak{F}$ is hypothesized to generate identifiable mixtures. Thus Q^{-1} is uniformly continuous on the compact set $Q(\mathcal{G}(B_j, j))$.

Step 2. For each j, find a number N_j $(N_j > N_{j-1})$ such that, in the notation of Theorem 3, for dimension n the same as the cdf's in $\mathfrak{F}$,

$$P[\|F_v - F\| > \delta_j/2] < 1/j \tag{14}$$

if $v > N_j$. (Theorem 3 allows us to conclude that N_j, as described, can be computed. In fact, $N(\frac{1}{2}\delta_j, 1/j)$ suffices.)

Step 3. For each positive number $v > N_1$, our estimate G_v of the mixing df G is constructed as follows. Find j such that $N_j < v \leqq N_{j+1}$. Compute the empiric df H_v associated with v observations distributed according to the finite mixture $Q(G)$. Choose, if possible (otherwise choose G_v arbitrarily), any member G_v of $\mathcal{G}(B_j, j)$ such that

$$L(Q(G_v), H_v) \leqq \tfrac{1}{2}\delta_j. \tag{15}$$

Then G_v is the estimate for G.

The assertion that G_v, so determined, is a consistent estimate for G (i.e., for any positive number ϵ, $\lim_{v\to\infty} P[L(G_v, G) > \epsilon] = 0$) admits the following justification. Eventually (say M), $G \varepsilon \mathcal{G}(B_j, j)$ if $j \geqq M$. Always, $L(F, G) \leqq \|F - G\|$ and thus if

$$\|H_v - Q(G)\| \leqq \tfrac{1}{2}\delta_j \tag{16}$$

some element (G itself, for instance) of $\mathcal{G}(B_j, j)$ will satisfy condition (15). By construction of N_j, (16) will happen with probability $>1 - 1/j$ (if $j > M$). Thus, if $v > N_j$, with probability at least $1 - 1/j$, (15) and (16) will simultaneously be satisfied, and

$$\begin{aligned} L(Q(G), Q(G_v)) &\leqq L(Q(G), H_v) + L(H_v, Q(G_v)) \\ &\leqq \|Q(G) - H_v\| + L(H_v, Q(G_v)) \leqq \tfrac{1}{2}\delta_j + \tfrac{1}{2}\delta_j = \delta_j \end{aligned} \tag{17}$$

As G, G_v are both in $\mathcal{G}(B_j, j)$, by construction of $\{\delta_j\}$, (17) implies that $L(G, G_v) < 1/j$. In summary, for $v > \max\{N_j, M\}$,

$$P[L(G, G_v) > 1/j] < 1/j, \tag{18}$$

which completes the demonstration.

Without altering the conditions of the algorithm, it can be adjusted to provide a sequence of estimators $\{G_j\}$ such that with probability 1, $G_j \to_L G$. Observe that the results of Kiefer and Wolfowitz [4] which we already cited also imply that one can compute $M(c, d)$ such that $P[\sup_{n>M(c,d)} \|F_n - F\| > c] < d$.

If $M_j = M(\frac{1}{2}\delta_j, 2^{-j})$ replaces N_j in step 2 and $M' = \max\{M_j, M\}$ (where M is as defined above), then

$$\begin{aligned} &P[\sup_{n>M'} L(G_n, G) > 1/j] \\ &\quad \leqq P[\textstyle\bigcup_{k\geqq j} [L(Q(G_n), Q(G)) \geqq \delta_k, \quad \text{for any } n \quad (M_k \leqq n \leqq M_{k+1})]] \\ &\quad \leqq \textstyle\sum_{k\geqq j} P[\sup_{n\leqq M_k} L(Q(G_n), Q(G)) > \delta_k] \leqq \sum_{k\geqq j} 2^{-k} = 2^{-j+1}. \end{aligned}$$

3. Distribution families which satisfy the conditions of the algorithm. All parameter spaces this author has encountered are limits of increasing sequences of compact sets, and thus we focus our attention on the condition that F_θ be continuous (L) with respect to θ. Our plan is to show that the multi-dimensional normal family has this property. The demonstration proceeds in such a manner that the reader will see how it can be directly and successfully applied to showing that all the other families mentioned in Section 1 (as yielding identifiable finite mixtures) satisfy the condition.

The n-dimensional normal distribution with covariance matrix Λ, mean vector θ has the characteristic function

$$\psi_{\Lambda,\theta}(t) = \exp(-\tfrac{1}{2}t\Lambda t' + it\theta), \qquad t \in R^n \tag{19}$$

which, for t a fixed real n-tuple, is obviously continuous with respect to (Λ, θ). That is, if $(\Lambda_i, \theta_i) \to (\Lambda, \theta)$, (convergence, of course, being in the Euclidean norm of $R^{n(n+3)/2}$) then for t fixed, $\psi_{\Lambda_i,\theta_i}(t) \to \psi_{\Lambda,\theta}(t)$, convergence being in the complex norm. We conclude the demonstration by observing that pointwise convergence of characteristic functions to a continuous function is equivalent to weak convergence of the corresponding cdf's and that weak convergence (by virtue of our lemma) is equivalent to convergence (L). It is trivial to similarly verify that the other distribution families mentioned in Section 1 have characteristic functions which are pointwise continuous with respect to their parameters.

4. Application to empirical Bayes decision problems. The basic component of an empirical Bayes decision problem is a sequence $\{(X_i, \boldsymbol{\theta}_i)\}$ of independent, identically distributed rv's $(X, \boldsymbol{\theta})$, (X and $\boldsymbol{\theta}$ are not independent, however). If θ_i is the outcome of $\boldsymbol{\theta}_i$, then X_i has the cdf F_{θ_i}. $\boldsymbol{\theta}$ has the cdf G. The statistician does not know G, and he is permitted to observe only the X_i's. He does know the parametric family $\mathfrak{F} = \{F_\theta : \theta \in R_1^m\}$, and that G is a member of a specified set $\mathcal{G}$ of cdf's. At each time n $(n = 1, 2, \cdots)$, the statistician is to choose an action a_n from some (time-independent) set $\mathcal{A}$. His choice may depend on $\{x_j\}_{j=1}^n$ (thus, we write $a_n = t_n(\bar{x}_n)$, $\bar{x}_n = \{x_i\}_{i=1}^n$). The loss which results from that decision is $L(a_n, \theta_n)$.

The statistician is never permitted to know the outcomes θ_n, but under certain circumstances, it is possible to find a sequence $\{t_i\}$ of decision functions such that, regardless of the value $G \in \mathcal{G}$, as n increases, the expected risk converges to the Bayes risk relative to G. Robbins [5] has defined the sequence $\{t_i\}$ of decision functions for an empirical Bayes decision problem to be *asymptotically optimal* if, for all $G \in \mathcal{G}$, with probability 1,

$$\lim_{n\to\infty} E_G[L(t_n(\bar{x}_n), \boldsymbol{\theta}_n)] = B(G), \tag{20}$$

where $B(G)$ denotes the Bayes risk relative to G. Further, Robbins has provided several techniques for constructing asymptotically optimal sequences. From our viewpoint, a particularly interesting procedure is one in which an estimate G_n for G is constructed (from $\bar{x}_n$) such that $G_n \Rightarrow G$. Then one constructs $\{t_n\}$ so that t_n is Bayes relative to G_n. Under very weak conditions on the loss function, it is easily concluded that $\{t_n\}$ is asymptotically optimal. One method proposed is effective if $\mathcal{G}$ is the set of all cdf's on parameter space and arbitrary mixtures on $\mathfrak{F}$ are identifiable, i.e. $(\int F_\theta\, dG = \int F_\theta\, dG') \Rightarrow (G = G')$. Deely and Kruse, following a suggestion of H. Robbins, have devised [3] a method for constructing $\{G_n\}$ by solving, after each observation x_n, a two-person zero-sum game. The methods given for construction of $\{G_n\}$ for this situation cannot be carried over directly to the case that $\mathfrak{F}$ generates identifiable finite mixtures, (but unidentifiable arbitrary mixtures) and $\mathcal{G}$ is as in Section 1 of this paper, because they depend on limits of sequences, under weak convergence, being in $\mathcal{G}$, a conclusion which does not hold for $\mathcal{G}$ the set of discrete df's with only finitely many mass points.

In view of the fact that some important families, while not generating identifiable arbitrary mixtures, generate identifiable finite mixtures (e.g. the normal family with mean and variance both considered as parameters), it is interesting that the analysis of this paper provides a scheme for finding a sequence $\{G_n\}$ which converges weakly to G under the circumstances that $\mathcal{G}$ is the set of all discrete df's with only finitely many mass points. In summary, for empirical Bayes decision problems in which $\mathcal{G}$ is known to generate only finite mixtures, a rather comprehensive theory is available. In [10], [9], and [13], one finds useful criteria for establishing whether $\mathfrak{F}$ generates identifiable finite mixtures. In case the answer is affirmative, one may expect that the algorithm of this paper will be useful in finding the required sequence $\{G_i\}$.

Acknowledgments. The incentive for thinking through the matters of this paper was a seminar invitation extended by the Information Science Group of the Electrical Engineering Department at Purdue University. In this regard, I am grateful to Professor Patrick for his discussions and interest. Professor Studden suggested the connection between my studies and [8].

REFERENCES

[1] Boes, D. (1966). On the estimation of mixing distributions. *Ann. Math. Statist.* **37** 177–188.

[2] Choi, K. (1969). Estimators for parameters of a finite mixture of distributions. To appear in *Ann. Inst. Statist. Math.* (Japan).

[3] Deeley, J. and Kruse, R. (1968). Construction of sequences estimating the mixing distribution. *Ann. Math. Statist.* **39** 286–288.

[4] Kiefer, J. and Wolfowitz, J. (1958). On the deviations of the empiric distribution function of vector chance variables. *Trans. Amer. Math. Soc.* **87** 173–186.

[5] Massey, R. (1950). A note on the estimation of a distribution function by confidence limits. *Ann. Math. Statist.* **21** 116–119.

[6] Patrick, E. and Hancock, J. (1966). Nonsupervised sequential classification and recognition of patterns. *IEEE Transactions on Information Theory* **IT12** 362–372.

[7] Prohorov, Yu. (1956). Convergence of random processes and limit theorems, *Theor. Probability Appl.* **1** 157–214.

[8] Robbins, H. (1964). The empirical Bayes approach to statistical decision problems. *Ann. Math. Statist.* **35** 1–20.

[9] Teicher, H. (1961). Identifiability of mixtures. *Ann. Math. Statist.* **32** 244–248.

[10] Teicher, H. (1963). Identifiability of finite mixtures. *Ann. Math. Statist.* **34** 1265–1269.

[11] Teicher, H. (1967). Identifiability of mixtures of product measures. *Ann. Math. Statist.* **38** 1300–1302.

[12] Wolfowitz, J. (1957). The minimum distance method. *Ann. Math. Statist.* **28** 75–88.

[13] Yakowitz, S. and Spragins, J. (1968). On the identifiability of finite mixtures. *Ann. Math. Statist.* **39** 209–214.