The Annals of Probability
1977, Vol. 5, No. 2, 200-209

ESTIMATION OF A MIXING DISTRIBUTION FUNCTION[1]

BY J. R. BLUM AND V. SUSARLA[2]

University of Wisconsin, Milwaukee

Let $\mathscr{A} = \{f(\cdot, \theta): \theta \in J\}$, J an interval, be a family of univariate probability densities (with respect to Lebesgue measure) on an interval I. First, a necessary and sufficient condition is proved for $\mathscr{A}$ to be identifiable whenever $\mathscr{A} \subset C_0(J)$, the class of continuous functions on J vanishing at ∞. If f_G is a G-mixture of the densities in $\mathscr{A}$ with G unknown, an estimator G_n based on f_G and $\mathscr{B} = \{f(x, \cdot): x \in I\}$ is provided such that $G_n \to_w G$ under certain conditions on $\mathscr{A}$. If $X_1, \cdots, X_n$ are i.i.d. random variables from f_G, an estimator $\hat{G}_n$ is provided such that $G_n(X_1, \cdots, X_n, \cdot) \to_w G(\cdot)$ almost surely under certain conditions on $\mathscr{A}$ and G. Furthermore, it is shown that $|f_{G_n}(x) - f_G(x)| \to 0$ a.s. and in L_2 with rates like $O(n^{-C})$ $(C > 0)$ under certain conditions on the density estimator $\hat{f}_G(x)$ involved in the definition of $\hat{G}_n$. The conditions of various theorems are verified in the case of location parameter and scale parameter families of densities.

1. Introduction and summary. Let f be a Borel measurable function from $I \times J$ to $(0, \infty)$ such that $\int_I f(x, \theta)\, dx = 1$ for each θ in J where I and J are intervals contained in $R = (-\infty, \infty)$ and $\mathscr{B}$ and $\mathscr{A}$ be the collections of sections of f with the first coordinate (in I) and the second coordinate (in J) fixed respectively. For a probability distribution function G on J, let

$$f_G(x) = \int_J f(x, \theta)\, dG(\theta), \qquad x \text{ in } I. \tag{1.1}$$

We provide an equivalent condition for the identifiability of $\mathscr{A}$ (for the definition of identifiability, see (A1)) in Section 2. In Section 3, we consider the problem of estimating G in terms of f_G and $\mathscr{B}$. To obtain an estimate G_n of G, we solve a system of equalities and inequalities and then show that G_n converges weakly to $(\to_w)$ G under some conditions on $\mathscr{B}$. If G and f_G are unknown, but i.i.d. random variables $X_1, \cdots, X_n, \cdots$ are observable (this is the standard empirical Bayes situation of Robbins [4], described in Section 4), then we construct (in Section 4) estimates $\hat{G}_n(X_1, \cdots, X_n, \cdot)$ which $\to_w G(\cdot)$ almost surely (a.s.) under some conditions on $\mathscr{B}$. It is then immediate that $\int_J \theta f(x, \theta)\, dG_n(\theta) \to \int_J \theta f(x, \theta)\, dG(\theta)$ a.s. whenever $\theta f(x, \theta) \in C(J)$. Furthermore, it is shown that our method of construction of G_n provides rates for a.s. and L_2 convergences of $f_{\hat{G}_n}(x) - f_G(x)$ to zero for each x under some additional conditions on $\mathscr{B}$. In Section 5, all the above results are shown to hold for location and scale parameter families of Lebesgue densities under rather weak conditions.

Received April 29, 1975; revised July 26, 1976.

[1] The revision of this paper was supported by the NSF Grant MCS76-05952.

[2] Research supported by a grant from the Graduate School of the University of Wisconsin, Milwaukee, and by H.E.W., N.I.H. Grant #1 ROl GM23129-01.

AMS 1970 *subject classifications.* Primary 60F05; Secondary 62099.

Key words and phrases. Identifiability, weak convergence, mixing distribution, empirical Bayes.

The results of Section 3 are not only of mathematical interest, but also provide an intuitive basis for the results of Section 4. In Sections 4 and 5, we take $I = J = R$ as other cases can be treated with obvious modifications of the method presented here. Throughout, G is assumed to be a distribution function with support in J. The estimator and its properties are compared with three other estimators for G in Section 6. In Section 4, we discuss the application of the main result of this paper to empirical Bayes estimation problems.

2. Identifiability. For the distribution function G in (1.1) to be estimable in terms of f_G and $\mathscr{B}$, it is obvious that the following condition should be satisfied.

(A1) $$f_G(x) = f_H(x) \qquad \text{for all} \quad x \quad \text{in} \quad I \Leftrightarrow H - G = 0 .$$

This condition is called the identifiability (of $\mathscr{A}$) condition. (For example, see Teicher [7].)

With $C_0(J)$ denoting the Banach space of continuous functions on the interval J which vanish at ∞ and normed by

(2.1) $$||g|| = \sup\{|g(y)| \,|\, y \ \text{in} \ J\},$$

we obtain

THEOREM 2.1. *Let* $\mathscr{B} \subset C_0(J)$. *Then* (A1) *holds if and only if* $\mathscr{B}$ *generates* $C_0(J)$ *in the supremum norm* (2.1).

PROOF. Let (A1) hold. Let B be the closed subspace generated by $\mathscr{B}$. If $B \neq C_0(J)$, then there exists a g in $C_0(J) - B$ and a bounded linear functional Φ on $C_0(J)$ such that $\Phi(g) = 1$ and $\Phi(f^*) = 0$ for f^* in B. Also, by the Riesz representation theorem, there exist nondecreasing, nonnegative functions K_1 and K_2 of bounded variations on J such that

$$\Phi(f) = \int_J f(y)\, d(K_1 - K_2)(y) \qquad \text{for} \quad f \quad \text{in} \quad C_0(J) .$$

Since $\Phi(f^*) = 0$ for f^* in B, it follows that $\int_J f(x, \theta)\, dK_1(\theta) = \int_J f(x, \theta)\, dK_2(\theta)$ for all x in I which, by (A1), implies that $K_1 - K_2 =$ constant. But then, this implies that $\Phi(g) = \int_J g(y)\, d(K_1 - K_2)(y) = 0$ which is a contradiction since $\Phi(g) = 1$. Hence $\mathscr{B}$ generates $C_0(J)$.

Conversely, let $\mathscr{B}$ generate $C_0(J)$ and (1.1) hold at G and H. We show that $G - H = 0$. By assumption

(2.2) $$\int_J f(x, \theta)\, dG(\theta) = \int_J f(x, \theta)\, dH(\theta) \qquad \text{for all} \quad x \quad \text{in} \quad I .$$

Since $\mathscr{B}$ generates $C_0(J)$ in the supremum norm, (2.2) can be extended to

(2.3) $$\int_J g(\theta)\, dG(\theta) = \int_J g(\theta)\, dH(\theta) \qquad \text{for all} \quad g \quad \text{in} \quad C_0(J) .$$

Since $\Phi(g) = \int_J g(\theta)\, dG(\theta)$ is a bounded linear functional on $C_0(J)$ whenever G is of bounded variation on J, the uniqueness part of the Riesz representation theorem and (2.3) show that $G - H =$ constant. This completes the proof of the theorem since G and H are distribution functions on J.

3. Construction of an estimator of G in (1.1). In this section, we define an estimator G_n ((3.6)) of G in terms of f_G and $\mathscr{B}$. We consider in detail the case $I = J = R$ only and point out the required changes if I or (or and) J is an (are) interval(s). Throughout this section, the integration is over $(-\infty, \infty)$, and the limits are as $n \to \infty$ unless otherwise stated.

For a fixed partition

$$\theta_{n,-1}(=-\infty) < \theta_{n,0}(=-n) < \theta_{n,1} < \cdots < \theta_{n,m(n)}(=n) < \theta_{n,m(n)+1} = \infty \tag{3.1}$$

with

$$\delta_n = \max\{\theta_{n,j} - \theta_{n,j-1} \mid j = 1, \cdots, m(n)\} \to 0 \tag{3.2}$$

and for x in R, and for $l = -1, \cdots, m(n)$, let

$$M_{n,l}(x) = \sup\{f(x, \theta) \mid \theta_{n,l} \leqq \theta \leqq \theta_{n,l+1}\}, \tag{3.3}$$

and

$$m_{n,l}(x) = \inf\{f(x, \theta) \mid \theta_{n,l} \leqq \theta \leqq \theta_{n,l+1}\}. \tag{3.4}$$

Let $p_n = \{p_{n,-1}, \cdots, p_{n,m(n)}\}$ be such that

$$\begin{aligned} &\text{(i)} \quad p_{n,l} \geqq 0 \quad \text{and} \quad \textstyle\sum_{l=-1}^{m(n)} p_{n,l} = 1, \\ &\text{(ii)} \quad \textstyle\sum_{l=-1}^{m(n)} p_{n,l} M_{n,l}(x) \geqq f_G(x) \quad \text{and} \\ &\text{(iii)} \quad \textstyle\sum_{l=-1}^{m(n)} p_{n,l} m_{n,l}(x) \leqq f_G(x) \end{aligned} \tag{3.5}$$

where (ii) and (iii) hold for x in $\{\theta_{n,0}, \cdots, \theta_{n,m(n)}\}$.

Let $P_n = \{p_n \mid p_n \text{ is a solution of (3.5)}\}$. That P_n is not empty follows since one such solution is given by $p_{n,l} = \int_{\theta_{n,l}}^{\theta_{n,l+1}} dG$ for $l = -1, \cdots, m(n)$. For any p_n in P_n, define

$$\begin{aligned} G_n(y) &= 0 && y < \theta_{n,0} \\ &= p_{n,-1} + p_{n,0} && \theta_{n,0} \leqq y < \theta_{n,1} \\ &= \textstyle\sum_{j=1}^{l} p_{n,l} && \theta_{n,l} \leqq y < \theta_{n,l+1}, \quad l = 1, \cdots, m(n). \end{aligned} \tag{3.6}$$

Clearly G_n is a discrete distribution function on R.

NOTE. The solution of (3.5) is a simple linear programming problem and there are efficient computational algorithms available for the solution of such inequalities. (3.5) can be solved theoretically for p_n without the assumption that x is in $\{\theta_{n,0}, \cdots, \theta_{n,m(n)}\}$, but such a solution might be difficult to obtain.

The result leading to $G_n \to_w G$ is

THEOREM 3.1. *Let* $f(x, \cdot) \in C_0(R)$,

$$\lim_{x' \to x} \sup_\theta |f(x, \theta) - f(x', \theta)| = 0, \quad \textit{and} \tag{A2}$$

$$\textit{for each } \varepsilon > 0, \ \exists\, \delta, \delta' > 0 \ni |x' - x| < \delta \textit{ and } |\theta' - \theta| < \delta \Rightarrow |f(x', \theta') - f(x', \theta)| < \varepsilon. \tag{A3}$$

Then

$$\int f(x, \theta)\, dG_n(\theta) \to \int f(x, \theta)\, dG(\theta) = f_G(x).$$

PROOF. Without loss of generality, let $|x| < n$. By the choice of the partition (3.1) and (3.2), there exists a sequence $\{\theta_{n,j(n)}\}$ such that $\theta_{n,j(n)} \to x$. We also observe that $f(x, \cdot) \in C_0(R)$ and (A2) imply that

(3.7) for each $\varepsilon > 0$, $\exists\, \delta, M > 0 \ni |x' - x| \leqq \delta$, $|\theta| > M \Rightarrow f(x', \theta) < \varepsilon$.

By the definitions of p_n and G_n given in (3.4) and (3.6) respectively,

$$\sum_{l=-1}^{m(n)} p_{n,l}\, m_{n,l}(\theta_{n,j(n)}) \leqq \int f(\theta_{n,j(n)}, \theta)\, dG_n(\theta) \leqq \sum_{l=-1}^{m(n)} p_{n,l}\, M_{n,l}(\theta_{n,j(n)}) . \tag{3.8}$$

Now observe that $0 \leqq D_n$ (= the difference between the extreme sides of (3.8))

$$\begin{aligned} &\leqq \sum_{l=-1}^{m(n)} p_{n,l}\{M_{n,l}(\theta_{n,j(n)}) - m_{n,l}(\theta_{n,j(n)})\} \\ &\leqq \sup\{|f(x', \theta) - f(x', \theta')| \,|\, |x - x'| \leqq \delta_n, |\theta - \theta'| \leqq \delta\} \\ &\quad + \sup\{f(x', \theta) \,|\, |x' - x| \leqq \delta_n, |\theta| \geqq n\} \end{aligned}$$

by the choice of the partition $\{\theta_{n,-1}, \cdots, \theta_{n,m(n)+1}\}$ and the sequence $\{\theta_{n,j(n)}\}$. This last expression (and hence D_n) $\to 0$ due to (A3) and (3.7). Hence, since the lhs of (3.8) $\leqq f_G(\theta_{n,j(n)}) \leqq$ rhs of (3.8) due to (ii) and (iii) of (3.5),

$$\int f(\theta_{n,j(n)}, \theta)\, dG_n(\theta) - f_G(\theta_{n,j(n)}) \to 0 . \tag{3.9}$$

But $f_G(\theta_{n,j(n)}) \to f_G(x)$ by (A2) since $\theta_{n,j(n)} \to x$. For the same reason, $\int f(\theta_{n,j(n)}, \theta)\, dG_n - \int f(x, \theta)\, dG_n(\theta) \to 0$. This completes the proof in view of (3.9).

COROLLARY 3.1. *Let* $\mathscr{B} \subset C_0(R)$, (A1); (A2) *and* (A3) *hold for each* x *in* R. *Then* $G_n \to_w G$. *If, in addition,* $\theta f(x, \theta) \in C(R)$, *then* $\int \theta f(x, \theta)\, dG_n(\theta) \to \int \theta f(x, \theta)\, dG(\theta)$.

PROOF. By Theorem 3.1,

$$\int f(x, \theta)\, dG_n(\theta) \to \int f(x, \theta)\, dG(\theta) \quad \text{for each } x \text{ in } R . \tag{3.10}$$

Since $\mathscr{B} \subset C_0(R)$ and (A1) holds, $\mathscr{B}$ generates $C_0(R)$ in the supremum norm (2.1) by Theorem 2.1. Therefore, (3.10) can be extended to $\int g(\theta)\, dG_n(\theta) \to \int g(\theta)\, dG(\theta)$ for each g in $C_0(R)$ which is equivalent to the first result. The second result is a consequence of the first result since $\theta f(x, \theta) \in C(R)$.

REMARK 3.1. If $\bar{J} = [a, b]$ and $\bar{I} = [c, d]$ with $-\infty < a, b, c,$ and $d < \infty$, then take $\theta_{n,-1} = a < \theta_{n,0} < \cdots < \theta_{n,m(n)} < \theta_{n,m(n)+1} = b$ with $\delta_n = \max\{|\theta_{n,j} - \theta_{n,j-1}| \,|\, j = 0, 1, \cdots, m(n) + 1\} \to 0$ and solve (3.5) at the nth stage when x is in $\{x_1, x_2, \cdots, x_{m(n)+1}\}$ where $\{x_1, x_2, \cdots\}$ is dense in I.

4. Estimation of G when f_G is unknown. In this section, assume that the distribution function G and f_G are unknown, $I = J = R$ and that $X_1, \cdots, X_n$ are i.i.d. random variables with common density f_G. We exhibit $\hat{G}_n(\cdot)$ (= $\hat{G}_n(X_1, \cdots, X_n, \cdot)$) such that $\hat{G}_n \to_w G$ almost surely (a.s.). An application of and motivation for the results of this section is given in the lengthy Remark 4.2.

Let $\hat{f}_G(x)$ $(= f_G(X_1, \cdots, X_n, x))$ be an estimator of $f_G(x)$ such that

(A4) $$||\hat{f}_G(\cdot) - f_G(\cdot)|| \to 0 \quad \text{a.s.}$$

where $||\ ||$ denotes the sup norm. For each fixed n, let $\hat{P}_n$ be the class of solutions obtained for (3.5) when f_G in (ii) and (iii) is replaced by $\hat{f}_G - \varepsilon$ and $\hat{f}_G + \varepsilon$ respectively where $\varepsilon\ (= \varepsilon_n)$ is the smallest positive number for which the class $\hat{P}_n$ is not empty. This method of choosing $\hat{P}_n$ does not require ε to be known in advance, $\hat{P}_n$ is well defined for each n, and the method involves a linear programming problem. Whenever the sample sequence is in the a.s. event A guaranteed to exist by (A4), the $\varepsilon\ (= \varepsilon_n)$ corresponding to that sample sequence at stage n converges to zero for the following reason. Let $\varepsilon_n{}^* = 2||\hat{f}_G(x_1, \cdots, x_n, \cdot) - f_G(\cdot)||$. Then $\varepsilon_n{}^* \to 0$ by assumption. Moreover, $\hat{P}_n$ is not empty for large n since $||\hat{f}_G(x_1, \cdots, x_n, \cdot) - f_G(\cdot)|| < \varepsilon_n{}^*$ implies that $\hat{p}_n = \{\hat{p}_{n,-1}, \cdots, \hat{p}_{n,m(n)}\}$, with $\hat{p}_{n,l} = \int_{\theta_{n,l}^-}^{\theta_{n,l+1}} dG$ for $l = -1, \cdots, m(n)$, is a solution belonging to $\hat{P}_n$. Since $\varepsilon\ (= \varepsilon_n) < \varepsilon_n{}^*$, $\varepsilon \to 0$. Define $\hat{G}_n(\cdot)$ $(= G_n(X_1, \cdots, X_n, \cdot))$ by

(4.1) $\hat{G}_n(y) = G_n(y)$ of (3.6) with $p_{n,l}$ replaced by $\hat{p}_{n,l}$ where $\hat{p}_n = \{\hat{p}_{n,-1}, \cdots, \hat{p}_{n,m(n)}\}$ is in $\hat{P}_n$.

With the above notation, we obtain the following two theorems. The first theorem is an analogue of Theorem 3.1 for $\hat{G}_n$. The second theorem provides rates of convergence for $f_{\hat{G}_n}(x) - f_G(x) \to 0$ a.s. and $\to 0$ in L_2.

THEOREM 4.1. *Let* (A1) *and* (A4) *and for each* x *in* R, *the conditions of Theorem* 3.1 *hold. Then* $\hat{G}_n \to_w G$ a.s. *If, in addition,* $\theta f(x, \theta) \in C(R)$, *then* $\int \theta f(x, \theta)\, d\hat{G}_n(\theta) \to \int \theta f(x, \theta)\, dG(\theta)$ a.s.

PROOF. Let the sample sequence $\{x_1, \cdots, x_n, \cdots\}$ be a fixed point in the a.s. event A guaranteed to exist by (A4). We show that $\hat{G}_n(x_1, \cdots, x_n, \cdot) \to_w G(\cdot)$.

Unless otherwise stated, let x be fixed. Without loss of generality, let $n > |x|$ and let ε_n be as in the discussion preceding (4.1) and let $\varepsilon_n{}^* = ||\hat{f}_G(x_1, \cdots, x_n, \cdot) - f_G(\cdot)||$. As in the proof of Theorem 3.1, let $\theta_{n,j(n)} \to x$. By the construction of $\hat{p}_n$ and G_n,

(4.2) $$\sum_{l=-1}^{m(n)} \hat{p}_{n,l}\, m_{n,l}(\theta_{n,j(n)}) \leqq \int f(\theta_{n,j(n)}, \theta)\, d\hat{G}_n(\theta) = f_{\hat{G}_n}(\theta_{n,j(n)}) \leqq \sum_{l=-1}^{m(n)} \hat{p}_{n,l}\, M_{n,l}(\theta_{n,j(n)}) .$$

The difference between the extreme sides of (4.2) goes to zero due to (A2) and (A3) as in the proof of Theorem 3.1. Also, by the construction preceding (4.1), and the assumption on $\hat{f}_G$, the lhs of (4.2) $\geqq \hat{f}_G(\theta_{n,j(n)}) - \varepsilon_n \geqq f_G(\theta_{n,j(n)}) - \varepsilon_n - \varepsilon_n{}^*$ and the rhs of (4.2) $\leqq f_G(\theta_{n,j(n)}) + \varepsilon_n + \varepsilon_n{}^*$ for large n. Now recall that $0 \leqq \varepsilon_n \leqq \varepsilon_n{}^* \to 0$. Hence $\int f(\theta_{n,j(n)}, \theta)\, d\hat{G}_n(\theta) \to \lim f_G(\theta_{n,j(n)}) = f_G(x)$ since (A2) implies that f_G is continuous at x and $\theta_{n,j(n)} \to x$. For the same reasons, $\int f(\theta_{n,j(n)}, \theta)\, d\hat{G}_n(\theta) - \int f(x, \theta)\, d\hat{G}_n(\theta) \to 0$. Therefore,

(4.3) $$\int f(x, \theta)\, d\hat{G}(\theta) \to f_G(x) = \int f(x, \theta)\, dG(\theta) \qquad \text{for all } x \text{ in } R .$$

Now the conditions $B \subset C_0(R)$ and (A1) imply (as in the proofs of Theorem

3.1 and Corollary 3.1) that (4.3) can be extended to $\int g(\theta)\, d\hat{G}_n(\theta) \to \int g(\theta)\, dG(\theta)$ for all g in $C_0(R)$ which is equivalent to $\hat{G}_n(x_1, \cdots, x_n, \cdot) \to_w G(\cdot)$. Since $\{x_1, \cdots, x_n, \cdots\}$ is an arbitrary point in A with $P(A) = 1$, the proof of the first part of the theorem is complete. The second part follows from the first part since $\theta f(x, \theta) \in C(R)$.

One advantage of our method of construction of $\hat{G}_n$ is that the rate results of $\hat{f}_G$ can be used to obtain the corresponding rate results for $f_{\hat{G}_n}$ as the following theorem shows. Recall that δ_n in defined by (3.2).

THEOREM 4.2. *Let the conditions of Theorem* 4.1 *hold and let* (A4) *hold with rate* $O(\alpha_n)$ *with* $\alpha_n \downarrow 0$. *If*

$$\text{(A5)} \qquad \sup_{|x'-x|<\delta_n} \sup_\theta \{|f(x, \theta) - f(x', \theta)|\} < \gamma_n$$

with $\gamma_n \downarrow 0$ *as* $\delta_n \downarrow 0$, *then* $[\max\{\alpha_n, \gamma_n\}]^{-1}$. $|f_{\hat{G}_n}(x) - f_G(x)| = O(1)$ a.s. (a.s. *set is independent of* x, *but* $O(1)$ *could depend on* x). *Additionally, if*

$$\text{(A6)} \qquad \sup\{E(\hat{f}_G(x') - f_G(x'))^2 | x^2 - x| < \delta_n\} = O(\beta_n^2)\,,$$

then $[\max\{\alpha_n^2, \beta_n^2, \gamma_n^2\}]^{-1} E[(f_{\hat{G}_n}(x) - f_G(x))^2] = O(1)$. (*Again,* $O(1)$ *could depend on* x.)

NOTE. (A5) is actually (A2) with a rate of convergence property.

PROOF. Let $\{\theta_{n,j(n)}\}$ be such that $|x - \theta_{n,j(n)}| < \delta_n$. By Theorem 4.1, $\hat{G}_n \to_w G$ a.s. The results now follow from the following set of inequalities:

$$\begin{aligned} |f_{\hat{G}_n}(x) - f_G(x)| &\leqq |f_{\hat{G}_n}(\theta_{n,j(n)}) - f_G(\theta_{n,j(n)})| \\ &\quad + |f_{\hat{G}_n}(x) - f_{\hat{G}_n}(\theta_{n,j(n)})| + |f_G(x) - f_G(\theta_{n,j(n)})| \\ &\leqq |f_{\hat{G}_n}(\theta_{n,j(n)}) - f_G(\theta_{n,j(n)})| + 2\gamma_n \end{aligned} \tag{4.4}$$

where the second inequality follows from (A5). Now observe that

$$\begin{aligned} &|f_{\hat{G}_n}(\theta_{n,j(n)}) - f_G(\theta_{n,j(n)})| \\ &\quad \leqq |f_{\hat{G}_n}(\theta_{n,j(n)}) - \hat{f}_G(\theta_{n,j(n)})| + |\hat{f}_G(\theta_{n,j(n)}) - f_G(\theta_{n,j(n)})| \\ &\quad \leqq 2\varepsilon_n + |\hat{f}_G(\theta_{n,j(n)}) - f_G(\theta_{n,j(n)})| \end{aligned} \tag{4.5}$$

where the last inequality follows from the construction of $\hat{G}_n$ (for example, see the argument following (4.2)). Now the first result follows from (4.4), (4.5) and (A4) while the second result follows from (4.4), (4.5) and (A6).

REMARK 4.1. If I and/or J are finite intervals, then apply the modifications suggested in Remark 3.1.

REMARK 4.2. Here, we discuss an application of Theorem 4.1 to the standard empirical Bayes decision problem of Robbins [4]. In an empirical Bayes decision problem, there is a sequence of i.i.d. vectors $\{(\theta_n, X_n)\}$ where $\theta_n \sim_{\text{i.i.d.}} G$, an unknown distribution and given $\theta_n = \theta$, $X_n \sim f(\cdot, \theta)(\in \mathscr{A})$. X_n is observable while θ_n is not. The empirical Bayes problem involves exhibiting $\{t_n(X_1, \cdots, X_n)\}$ such that the Bayes risk of using t_n in deciding about θ_n less the minimum Bayes risk of deciding (using X_n) about θ_n converges to zero, hopefully with a rate.

Robbins [4] named such rules as asymptotically optimal empirical Bayes rules (a.o.e.B.). In this situation, one can use Theorem 4.1 as follows: Use $X_1, \cdots, X_n$ to estimate G by $\hat{G}_n$ as in Theorem 4.2. Then take $t_{n+1} = t_{n+1}(X_1, \cdots, X_{n+1})$ as the Bayes rule of deciding (using X_{n+1}) about θ_{n+1} when the prior distribution is $\lambda_n \Phi + (1 - \lambda_n)\hat{G}_n$ where Φ is the standard normal distribution function and $0 < \lambda_n \downarrow 0$ as $n \uparrow \infty$. Such a rule $\{t_n\}$ cannot only be shown to be a.o.e.B., but also componentwise admissible under fairly general conditions on $\mathscr{A}$, G and the loss function involved in the definition of the Bayes risk. For example, in the problem of empirical Bayes squared error loss estimation of θ, the above method and the dominated convergence theorem provide a.o.e.B. estimators which are component admissible (with θ restricted to $[a, b]$) provided G is in the class of all distributions with support in $[a, b]$, $-\infty < a < b < \infty$. The compactness of the support of G is not an unrealistic assumption. If the prior distribution does not have a compact support, the asymptotic optimality of the above procedure can be obtained by appealing to an unpublished lemma of Le Cam and Scheffé's theorem. All these details, which are too long, will appear elsewhere. Before closing, we note that one has to use both parts of Theorem 4.1 namely, the convergence of $f_{\hat{G}_n}$ to f_G and that of $\int \theta f(\cdot, \theta)\, d\hat{G}_n$ to $\int \theta f(\cdot, \theta)\, dG$ to obtain the above empirical Bayes results.

To obtain rate of convergence results in the above empirical Bayes estimation problem, we make the following change. Instead of solving the equations as described in the paragraph preceeding (4.1), solve the equations (3.5) with f_G in (ii) and (iii) replaced by $\hat{f}_G - \eta$ and $\hat{f}_G + \eta$ respectively along with the following equations:

$$\text{(iv)} \quad \sum_{l=-1}^{m(n)} p_{n,l} \sup \{\theta f(x, \theta) \mid \theta_{n,l} \leqq \theta \leqq \theta_{n,l+1}\} \geqq \hat{h}_G(x) - \eta$$

$$\text{(v)} \quad \sum_{l=-1}^{m(n)} p_{n,l} \inf \{\theta f(x, \theta) \mid \theta_{n,l} \leqq \theta \leqq \theta_{n,l+1}\} \leqq \hat{h}_G(x) + \eta$$

where $\hat{h}_G(\cdot)$ is an estimator of $h_G(\cdot) = \int \theta f(\cdot, \theta)\, dG$ and η is the smallest positive number for which the above five equations (i) through (v) can be solved simultaneously. Such a solution, as in Theorem 4.2, will lead to simultaneous rates for the mean square convergences of $\hat{f}_G$ and $\hat{h}_G$ to f_G and h_G respectively. In turn, these mean square convergence results can be applied to obtain rates in the above empirical Bayes estimation problem along with componentwise admissibility since the function to be estimated is simply $h_G(\cdot)/f_G(\cdot)$ based on $X_1, \cdots, X_n$. This method of obtaining componentwise admissible procedures has been used in a nonparametric context in Susarla and Phadia [6].

5. Examples. We consider two examples, one involving a location parameter family of densities on $I = R(-\infty, \infty)$ and the other involving a scale parameter family of densities on $I = [0, \infty)$. All densities are wrt Lebesgue measure on the real line or on $[0, \infty)$.

To consider the location parameter case, assume that

$$(5.1) \qquad h \text{ is a continuous density with } h(x) \to 0 \text{ as } |x| \to \infty .$$

If

$$f(x, \theta) = h(x - \theta), \qquad -\infty < \theta, x < \infty, \tag{5.2}$$

then we have

THEOREM 5.1. *If f is defined via* (5.2) *and satisfies* (A1), *then G_n of* (3.6) $\to_w G$. *If, in addition,*

$$\sup\{|h'(t)| \mid t \in R\} < \infty \tag{5.3}$$

$$\hat{f}_G(x) = (na_n)^{-1} \textstyle\sum_{j=1}^{n} k((x - X_j)/a_n) \tag{5.4}$$

where $X_1, \cdots, X_n$ are i.i.d. $f_G(f_G(x) = \int h(x - \theta)\, dG(\theta))$, *$k$ is the standard normal density and $a_n^4 = n^{-1}$, then $\hat{G}_n$ of* (4.1) $\to_w G$ a.s. *provided ε_n of* (A4) $= n^{-c}$ *with* $0 < 4c < 1$. *If $\delta_n = O(n^{-\gamma})$ with $\gamma > 1$, then $|\hat{f}_{\hat{G}_n}(x) - f_G(x)| = O(n^{-c})$* a.s. *Moreover, $E[(f_{\hat{G}_n}(x) - f_G(x))^2] = O(n^{-\min\{2c, 1-2c\}})$.*

PROOF. The first part of the theorem follows from Corollary 3.1 upon observing that (5.1) implies the conditions (A2) and (A3) and that $\mathscr{B} \subset C_0(R)$.

For the second result, observe that (5.1) and (5.3), respectively, imply that f_G and f_G' are bounded. Therefore Corollary 2.6 with $r = 0$ of Schuster [5] obtains that

$$n^c \|\hat{f}_G - f_G\| \to 0 \quad \text{a.s.}$$

where $\|\ \|$ stands for the supremum norm and $0 < 4c < 1$. Thus (A4) also holds with $\varepsilon_n = n^{-c}$. Now Theorem 4.1 obtains the result $\hat{G}_n \to_w G$ a.s.

The third part of the theorem follows since

$$\sup_{|x'-x|<\delta_n} \sup_\theta \{|f(x, \theta) - f(x', \theta)|\} \leqq \delta_n \|h'\|$$

implying (A5) with $\gamma_n = \delta_n = n^{-\gamma}$. To obtain the L_2 convergence result, we verify (A6) with $\beta_n^2 = n^{-\min\{2c, 1-2c\}}$ as follows:

$$E[(\hat{f}_G(x') - f_G(x'))^2] = \operatorname{Var}(\hat{f}_G(x')) + (E[\hat{f}_G(x')] - f_G(x'))^2. \tag{5.5}$$

By the definition of $\hat{f}_G$ and Lemma 2.3 of Schuster [5], $|E[\hat{f}_G(x')] - f_G(x')| \leqq c_1 a_n$ for some constant c_1, and since k is bounded by unity and since $X_1, \cdots, X_n$ are i.i.d., $\operatorname{Var}(\hat{f}_G(x')) \leqq (na_n^2)^{-1}$. Hence, since $a_n = n^{-c}$, (5.5) $= O(n^{-\min\{2c, 1-2c\}})$. This verifies (A6) since $0 < 4c < 1$ and so the last result follows.

For considering the scale parameter case, assume that h is a continuous density on $[0, \infty)$ with

$$\begin{aligned} &\text{(i)} \quad \sup\{yh(y) \mid y \geqq n\} \to 0, \qquad \sup\{|h'(y)| \mid y \geqq 0\} < \infty, \\ &\text{(ii)} \quad \sup\{y|h'(y)| \mid y \geqq 0\} < \infty \qquad \text{and} \\ &\text{(iii)} \quad \sup\{y^2|h'(y)| \mid y \geqq 0\} < \infty. \end{aligned} \tag{5.6}$$

If

$$f(x, \theta) = \theta h(x\theta) \qquad \text{for} \quad x, \theta > 0, \tag{5.7}$$

then we have the following theorem whose proof is omitted since it is similar to that of Theorem 5.1.

THEOREM 5.2. *If f is defined via* (5.7) *and satisfies* (A1), *then G_n of* (3.6) $\to_w G$. *If, in addition,*

$$\int \theta^2 \, dG(\theta) < \infty \tag{5.8}$$

and $\hat{f}_G$ is defined by (5.4) *where X_j are* i.i.d. $f_G (= \int_0^\infty \theta h(x\theta) \, dG(\theta))$, *then $\hat{G}_n$ of* (4.1) $\to_w G$ a.s. *provided ε_n of* (A5) $= n^{-c}$ *with* $0 < 4c < 1$. *If $\delta_n = O(n^{-\gamma})$ with $\gamma > 1$ then $|f_{\hat{G}_n}(x) - f_G(x) = O(n^{-c})$* a.s. *Moreover $E[(f_{\hat{G}_n}(x) - f_G(x))^2] = O(n^{-\min\{2c, 1-2c\}})$.*

REMARK 5.1. Theorem 5.1 includes the family of normal densities indexed by the mean and with known variance while Theorem 5.2 includes the family of scale parameter exponential distributions with the second moment of the mixing distribution finite.

REMARK 5.2. The results of this paper can be extended when both the arguments x and θ are vectors and can be applied to mixtures of discrete probability distributions with appropriate changes. It is well known that the family of binomial distributions $\{B(n, p) \mid 0 \leqq p \leqq 1\}$ is not identifiable. That this is the case can be readily seen from Theorem 2.1 since the class of polynomials of degree at most n does not generate $C_0[0, 1]$.

6. Some other estimators and comparison with our estimator. We briefly describe three methods of estimation of G and compare their results with those presented here. In the method by Deely and Kruse [2], the finite interval Λ (on which G is assumed to have support) is partitioned by the points $\lambda_{1n}, \cdots, \lambda_{nn}$ so that there is a sequence $\{\mathscr{G}_n\}$ of classes of distributions such that the support of each distribution in $\mathscr{G}_n$ is in $\{\lambda_{1n}, \cdots, \lambda_{nn}\}$ and for every G with support in Λ, there exists a sequence $\{G_n\}$ with G_n in $\mathscr{G}_n$ and $G_n \to_w G$. Then their method chooses a G_n^* in $\mathscr{G}_n$ which minimizes the sup distance $\|F_n - F_H\|$ where H is in $\mathscr{G}_n$, $nF_n(\cdot) = \sum_{j=1}^n I_{[X_j \leqq \cdot]}$ and $F_H(\cdot) = \int F(\cdot, \theta) \, dH$ where $F(\cdot, \theta)$ is a distribution function for each θ. They point out that their method involves finding an optimal strategy in a game with a payoff matrix which depends on F_n, and $\lambda_{1n}, \cdots, \lambda_{nn}$. They point out that Λ can be taken to be R. Choi [1] uses the Wolfowitz distance function $d(\hat{G}, G) = \int (\hat{G}(x) - G(x))^2 \, d\hat{G}(x)$ and in the words of Deely and Kruse [2], the computational feasibility of Choi's method is not clearly established. Moreover, Choi's [1] method needs the solution of a dynamic programming problem, and considers only finite mixtures. Meeden [3] constructs a probability distribution on $\mathscr{G}$, the class of all probability distributions on $[0, \infty)$ and then show that the Bayes estimate based on the first n observations corresponding to the constructed prior converges $\to_w$ to the true element G_0 in $\mathscr{G}$. Again the solution of finding estimates by Meeden's [3] method appears as hard as we have in the paper. Our estimators have the simplicity that they need only a linear programming computation (see the note following (3.6)), have some distance properties (Theorem 4.2) and will give componentwise admissible empirical Bayes estimators with and without rates with a small amount

of extra work if the support of the prior is in a compact set. It is not clear how one can recover rate results for the density $\hat{f}_G$ and $\hat{h}_G$ from the weak convergence results of the above three authors.

Acknowledgments. The authors are grateful to the referee and to the Associate Editor for their suggestions. The Associate Editor pointed out the improved definition of $\hat{G}_n$ (see (4.1)) over our previous version and several changes which improved the presentation of the paper.

REFERENCES

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

[2] DEELY, J. J. and KRUSE, R. L. (1968). Construction of sequences estimating the mixing distribution. *Ann. Math. Statist.* **39** 286–288.

[3] MEEDEN, G. (1972). Bayes estimation of the mixing distribution, the discrete case. *Ann. Math. Statist.* **43** 1993–1999.

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

[5] SCHUSTER, E. F. (1969). Estimation of a probability density and its derivatives. *Ann. Math. Statist.* **40** 1187–1195.

[6] SUSARLA, V. and PHADIA, E. G. (1976). Empirical Bayes testing of a distribution function with Dirichlet process priors. *Comm. Statist.-Theoretical Methods* **5** 455–469.

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

NATIONAL SCIENCE FOUNDATION
STATISTICS SECTION
WASHINGTON, D.C. 20550

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF WISCONSIN
MILWAUKEE, WISCONSIN 53201