The Annals of Statistics
1988, Vol. 16, No. 4, 1609–1622

ESTIMATION OF THE MIXING DISTRIBUTION FOR A NORMAL MEAN WITH APPLICATIONS TO THE COMPOUND DECISION PROBLEM[1]

BY DAVID EDELMAN

Columbia University

A procedure for estimating the mixing distribution for a normal mean is presented. The estimator is shown to be consistent regardless of the mixing distribution, and it is suggested that the estimation techniques may be applied in a similar manner to estimate mixing distributions for a wide class of location parameter families. Implied marginal density and derivative estimates for the normal case are shown to be consistent, converging at near-optimal rates. In addition, empirical Bayes rules for the quadratic loss mean estimation problem which are derived from the estimated mixing distributions are shown to be asymptotically optimal in average compound risk and Bayes risk (if it is defined), provided the means are assumed to lie within a fixed range. A brief discussion of computational issues related to this estimation procedure is also included, along with some small-sample simulation results for the compound decision problem.

1. Introduction. In his original paper on empirical Bayes, Robbins (1956) presented an open problem [see also Robbins (1950)]:

For $(X_1, \theta_1), \ldots, (X_n, \theta_n)$ independent, indentically distributed 2-vectors, with the conditional distribution of X_i given θ_i normal with mean θ_i and variance 1, and $(\theta_1, \ldots, \theta_n)$ being regarded as a random sample from some mixing distribution with distribution function $G(\theta)$, is it possible to estimate $G(\theta)$ in general?

Following the results of Kiefer and Wolfowitz (1956), Laird (1978), Lindsay (1983a, b) and others have explored the properties of maximum likelihood estimators for the general problem of estimating a mixing distribution, but the complicated nature of the resulting likelihood function renders this approach somewhat involved numerically for problems of large sample size. Robbins (1964) proposed estimation of a mixing distribution by finding the mixture which is closest to the empirical distribution function in the sense of supreme distance and established strong convergence of the resulting estimator, but did not establish any convergence rate for it. Convergence rates for various estimation methods have been established recently by Ritov (1987) and others.

For the related mean estimation problem, Singh (1979) has proposed estimators based on kernel techniques which have asymptotically optimal mean-squared error for the Bayes problem (which therefore should be asymptotically optimal for the compound decision problem as well), though (aside from being complicated) no one of these estimators can have convergence rate which is arbitrarily

Received June 1986; revised March 1988.

[1]Partially completed while visiting University of British Columbia and Rutgers University.

AMS 1980 *subject classifications*. Primary 62C12; secondary 62C25.

Key words and phrases. Compound decision theory, density estimation, empirical Bayes estimation.

close to the optimal. George (1986) has presented multiple shrinkage estimators having minimax compound squared-error loss, but these cannot be asymptotically optimal for the Bayes problem unless the prior distribution is a mixture of a (prespecified) finite number of normal distributions with a particular configuration.

In what follows, a simple procedure will be presented for estimating the empirical distribution function of a set of normal means which will be shown to be consistent for any sequence of parameters, with a uniform bound given (in terms of the sample size) for the convergence of the integrated mean-squared error of estimation. While the proof of consistency will not rely on the assumption of an underlying distribution generating the means, the estimator will be shown to converge to a fixed mixing distribution if one exists. It will be argued that it is possible to use the same methods to estimate mixing distributions for a wide class of location families, and that the convergence proofs are virtually identical to those for the normal case. Next, it will be shown that, for normal mixtures, consistent estimation of a mixing distribution leads to marginal density and derivative estimates which are consistent (converging at near-optimal rates), and to empirical Bayes rules which are asymptotically optimal in average compound risk in some cases.

2. Some convergence theorems. The results of this section pertain to an estimator of an empirical distribution function of normal means, and the comparison of functionals of the estimator to the corresponding functionals of the true empirical distribution function.

The estimation procedure to be explored is based on the minimization of the integral-squared distance between the empirical distribution function of observations of n normal variables and a generic n-mixture of normal cumulative distribution functions. While all of the results of this section pertain to this particular estimation procedure, several possible variations will be discussed subsequently.

Let $X_1, X_2, \ldots, X_n$ be independent random variables with cumulative distribution functions $\Phi(x_1 - \theta_1), \Phi(x_2 - \theta_2), \ldots, \Phi(x_n - \theta_n)$, with $\Phi(z) = \int_{-\infty}^{z} e^{-t^2/2}\, dt/\sqrt{2\pi}$, and let $E_n(x)$ and $G_n(\theta)$ denote the empirical distribution functions of $X_1, X_2, \ldots, X_n$ and $\theta_1, \theta_2, \ldots, \theta_n$, respectively.

THEOREM 1. *There exists* $\tilde{\theta} = (\tilde{\theta}_1, \tilde{\theta}_2, \ldots, \tilde{\theta}_n)$ *satisfying*

$$\int_{-\infty}^{\infty} \left\{ \frac{1}{n} \sum_{j=1}^{n} \Phi\left(x - \tilde{\theta}_j\right) - E_n(x) \right\}^2 dx \leq \int_{-\infty}^{\infty} \left\{ \frac{1}{n} \sum_{j=1}^{n} \Phi\left(x - \theta_j'\right) - E_n(x) \right\}^2 dx$$

for all $\theta' \in R^n$ *and*

$$\lim_{n \to \infty} E \int_{-\infty}^{\infty} \left\{ \tilde{G}_n(\theta) - G_n(\theta) \right\}^2 d\theta = 0,$$

where $\tilde{G}_n(\cdot)$ *is the empirical distribution function of* $\tilde{\theta}_1, \ldots, \tilde{\theta}_n$.

PROOF. Let $(x_1, x_2, \ldots, x_n)$ be a realization of $(X_1, X_2, \ldots, X_n)$, with order statistics $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$, and let $\theta_{(1)}, \theta_{(2)}, \ldots, \theta_{(n)}$ denote the order statistics of $\theta_1, \theta_2, \ldots, \theta_n$. Then since

$$\int_{-\infty}^{\infty} \left\{ \frac{1}{n} \sum_{j=1}^{n} \Phi(x - \theta_j) - E_n(x) \right\}^2 dx$$

increases without bound as $\theta_{(1)} \to -\infty$ or as $\theta_{(n)} \to \infty$, the problem may be reduced to finding the minimum of a continuous function over a compact set, and therefore the integral must assume a minimum.

Henceforth it will be assumed that $\tilde{\theta}$ minimizes the integral. By Plancherel's identity [Seeley (1966)]

$$\int_{-\infty}^{\infty} \{\tilde{G}_n(\theta) - G_n(\theta)\}^2 d\theta = \frac{1}{2\pi} \int_{-\infty}^{\infty} |\psi_{\tilde{G}_n}(t) - \psi_{G_n}(t)|^2 \frac{1}{t^2} dt,$$

where $\psi_{\tilde{G}_n}(\cdot)$ and $\psi_{G_n}(\cdot)$ denote the characteristic functions of $\tilde{G}_n(\cdot)$ and $G_n(\cdot)$, respectively.

Next,

$$\int_{-\infty}^{\infty} |\psi_{\tilde{G}_n}(t) - \psi_{G_n}(t)|^2 \frac{1}{t^2} dt < e^{T^2} \int_{-T}^{T} |\psi_{\tilde{G}_n}(t) - \psi_{G_n}(t)|^2 \frac{1}{t^2} e^{-t^2} dt + 2 \int_{T}^{\infty} \frac{4}{t^2} dt,$$

which is less than or equal to

$$e^{T^2} \int_{-\infty}^{\infty} |\psi_{\tilde{G}_n}(t) - \psi_{G_n}(t)|^2 \frac{1}{t^2} e^{-t^2} dt + \frac{8}{T},$$

which [setting $F_n(x; \mathbf{a}) = (1/n)\Sigma_1^n \Phi(x - a_j)$] is equal to

$$2\pi e^{T^2} \int_{-\infty}^{\infty} \{F_n(x; \tilde{\theta}) - F_n(x; \theta)\}^2 dx + \frac{8}{T},$$

the latter following from the fact that $\psi_{\tilde{G}_n}(t) e^{-t^2/2}$ and $\psi_{G_n}(t) e^{-t^2/2}$ are the characteristic functions of $F_n(x; \tilde{\theta})$ and $F_n(x; \theta)$, respectively. Next, since $(a + b)^2 \leq 2a^2 + 2b^2$,

$$\int_{-\infty}^{\infty} \{F_n(x; \tilde{\theta}) - F_n(x; \theta)\}^2 dx$$

$$\leq 2 \int_{-\infty}^{\infty} \{F_n(x; \tilde{\theta}) - E_n(x)\}^2 dx + 2 \int_{-\infty}^{\infty} \{F_n(x; \theta) - E_n(x)\}^2 dx,$$

which is less than or equal to

$$4 \int_{-\infty}^{\infty} \{F_n(x; \theta) - E_n(x)\}^2 dx$$

(from the definition of $\tilde{\theta}$). Hence

$$E \int_{-\infty}^{\infty} \{F_n(x; \tilde{\theta}) - F_n(x; \theta)\}^2 dx \leq 4E \int_{-\infty}^{\infty} \{F_n(x; \theta) - E_n(x)\}^2 dx,$$

the latter equaling

$$\frac{4}{n} \int_{-\infty}^{\infty} \Phi(x)\{1 - \Phi(x)\} dx < \frac{4\sqrt{2}}{n\sqrt{\pi}},$$

and if $T = \sqrt{\alpha \log n}$ for some $\alpha < 1$,

$$E\int_{-\infty}^{\infty} \{\tilde{G}_n(\theta) - G_n(\theta)\}^2 \, d\theta < \frac{4\sqrt{2}}{n^{1-\alpha}\sqrt{\pi}} + \frac{4}{\pi\sqrt{\alpha \log n}},$$

which tends to 0 as $n \to \infty$. □

It should be noted that this proof establishes convergence of the integral at a rate of $1/\sqrt{\log n}$ as $n \to \infty$.

COROLLARY 1 (Convergence in measure).

$$E\{\tilde{G}_n(\theta) - G_n(\theta)\}^2 \to_{\mathcal{M}} 0$$

[*for any* $\varepsilon > 0$ *the Lebesgue measure of* $\{\theta\colon E\{\tilde{G}_n(\theta) - G_n(\theta)\}^2 > \varepsilon\}$ *tends to* 0 *as* $n \to \infty$.]

COROLLARY 2. *If* $\theta_1, \ldots, \theta_n$ *may be regarded as a random sample from a distribution* $G(\theta)$ *with density* $g(\theta)$, *bounded and continuous on* $(-\infty, \infty)$, *and* $\int |\theta| g(\theta)\, d\theta \le M$, *some* $M < \infty$, *then*

$$\lim_{n\to\infty} E\left[\sup_{\theta} \{\tilde{G}_n(\theta) - G(\theta)\}^2\right] = 0.$$

PROOF. First note that

$$\begin{aligned}
&\int_{-\infty}^{\infty} E\{\tilde{G}_n(\theta) - G(\theta)\}^2 \, d\theta \\
&\quad \le 2\int_{-\infty}^{\infty} E\{\tilde{G}_n(\theta) - G_n(\theta)\}^2 \, d\theta + 2\int_{-\infty}^{\infty} E\{G_n(\theta) - G(\theta)\}^2 \, d\theta \\
&\quad = 2\int_{-\infty}^{\infty} E\{\tilde{G}_n(\theta) - G_n(\theta)\}^2 \, d\theta + \frac{2}{n}\int_{-\infty}^{\infty} G(\theta)\{1 - G(\theta)\} \, d\theta.
\end{aligned}$$

Integration by parts yields

$$\int_0^{\infty} G(\theta)\{1 - G(\theta)\} \, d\theta \le \int_0^{\infty} \{1 - G(\theta)\} \, d\theta = \int_0^{\infty} \theta g(\theta) \, d\theta$$

and

$$\int_{-\infty}^{0} G(\theta)\{1 - G(\theta)\} \, d\theta \le \int_{-\infty}^{0} G(\theta) \, d\theta = \int_{-\infty}^{0} (-\theta) g(\theta) \, d\theta,$$

so

$$\int_{-\infty}^{\infty} G(\theta)\{1 - G(\theta)\} \, d\theta \le \int_{-\infty}^{\infty} |\theta| g(\theta) \, d\theta \le M,$$

implying that

$$\int_{-\infty}^{\infty} E\{\tilde{G}_n(\theta) - G(\theta)\}^2 \, d\theta = O\left(\frac{1}{\sqrt{\log n}}\right) + O\left(\frac{1}{n}\right).$$

Next, let $\gamma = \sup g(\theta)$. If $\tilde{G}_n(\theta_0) > G(\theta_0)$, then for $\theta > \theta_0$,

$$\{\tilde{G}_n(\theta) - G(\theta)\}^+ \geq \{\tilde{G}_n(\theta_0) - G(\theta_0) - \gamma(\theta - \theta_0)\}^+$$

[where $(a)^+ = a$ if $a > 0$ and $(a)^+ = 0$ otherwise] and

$$\int_{-\infty}^{\infty} \left[\{\tilde{G}_n(\theta) - G(\theta)\}^+\right]^2 d\theta \geq \int_{\theta_0}^{\infty} \left[\{\tilde{G}_n(\theta_0) - G(\theta_0) - \gamma(\theta - \theta_0)\}^+\right]^2 d\theta$$

$$= \frac{1}{3\gamma}\{\tilde{G}_n(\theta_0) - G(\theta_0)\}^3.$$

Similarly, if $\tilde{G}_n(\theta_0) < G(\theta_0)$, then for $\theta < \theta_0$,

$$\{G(\theta) - \tilde{G}_n(\theta)\}^+ \geq \{G(\theta_0) - \tilde{G}_n(\theta_0) - \gamma|\theta - \theta_0|\}^+$$

and

$$\int_{-\infty}^{\infty} \left[\{G(\theta) - \tilde{G}_n(\theta)\}^+\right]^2 d\theta \geq \int_{-\infty}^{\theta_0} \left[\{G(\theta_0) - \tilde{G}_n(\theta_0) - \gamma|\theta - \theta_0|\}^+\right]^2 d\theta$$

$$= \frac{1}{3\gamma}\{G(\theta_0) - \tilde{G}_n(\theta_0)\}^3.$$

Therefore,

$$\sup_\theta |\tilde{G}_n(\theta) - G(\theta)|^3 \leq 3\gamma \int_{-\infty}^{\infty} \{\tilde{G}_n(\theta) - G(\theta)\}^2 d\theta$$

so that

$$E\left[\sup_\theta |\tilde{G}_n(\theta) - G(\theta)|^3\right] = O\left(\frac{1}{\sqrt{\log n}}\right).$$

This completes the proof of the corollary. □

The previous theorem may be generalized to include a broad class of location mixtures by replacing the function $\Phi(\cdot)$ by any distribution function $H(\cdot)$ which is strictly increasing on $(-\infty, \infty)$, has finite first moment, and which has characteristic function with strictly positive modulus, via a virtually identical proof. Since the results to follow are more difficult to extend, this generalization will not be pursued further here.

In what follows, it will be shown that the estimation procedure proposed in Theorem 1 may be used to produce estimators of the distribution function $F_n(x; \theta)$ and its derivatives with respect to x, $F_n^{(m)}(x; \theta)$, with near-optimal convergence rates.

THEOREM 2. *For any nonnegative integer m and constant $\alpha \in (0, 1)$ ($\alpha \in (0, 1]$ for $m = 0$) there exists constants $C_{m,\alpha}$ and $D_{m,\alpha}$ such that for any sequence $\theta_1, \theta_2, \ldots,$*

$$E\int_{-\infty}^{\infty} \{F_n^{(m)}(x; \tilde{\theta}) - F_n^{(m)}(x; \theta)\}^2 dx < \frac{C_{m,\alpha}}{n^\alpha}$$

and

$$E|F_n^{(m)}(x; \tilde{\theta}) - F_n^{(m)}(x; \theta)| < \frac{D_{m,\alpha}}{n^{\alpha/2}}.$$

PROOF. By Plancherel's identity,

$$E\int_{-\infty}^{\infty}\left\{F_n^{(m)}(x;\tilde{\theta}) - F_n^{(m)}(x;\theta)\right\}^2 dx = \frac{1}{2\pi}E\int_{-\infty}^{\infty}|\psi_{\tilde{F}_n}(t) - \psi_{F_n}(t)|^2 t^{2m-2}\,dt,$$

which is less than

$$T^{2m}E\int_{-\infty}^{\infty}\left\{F_n(x;\tilde{\theta}) - F_n(x;\theta)\right\}^2 dx + \frac{4}{\pi}\int_T^{\infty} t^{2m-2}e^{-t^2}\,dt$$

[noting that $|\psi_{\tilde{F}_n}(t) - \psi_{F_n}(t)|^2 < 4e^{-t^2}$]. Integration by parts shows this to be equal to

$$T^{2m}E\int_{-\infty}^{\infty}\left\{F_n(x;\tilde{\theta}) - F_n(x;\theta)\right\}^2 dx + \frac{2}{\pi}T^{2m-3}e^{-T^2} + O(T^{2m-5}e^{-T^2})$$

(as $T \to \infty$) and, setting $T = \sqrt{\log n}$ and remembering that

$$E\int_{-\infty}^{\infty}\left\{F_n(x;\tilde{\theta}) - F_n(x;\theta)\right\}^2 dx < \frac{4\sqrt{2}}{n\sqrt{\pi}}$$

(which was established in the proof of Theorem 1), the relation

$$E\int_{-\infty}^{\infty}\left\{F_n^{(m)}(x;\tilde{\theta}) - F_n^{(m)}(x;\theta)\right\}^2 dx = O\left\{\frac{(\log n)^m}{n}\right\} + O\left\{\frac{(\log n)^{m-3/2}}{n}\right\}$$

follows. This completes the proof of the first result.

The second result may be arrived at in a similar fashion. First, the Fourier inversion formula implies that

$$E|F_n^{(m)}(x;\tilde{\theta}) - F_n^{(m)}(x;\theta)| \le \frac{1}{2\pi}E\int_{-\infty}^{\infty}|\psi_{\tilde{F}_n}(t) - \psi_{F_n}(t)|\,|t|^{m-1}\,dt,$$

which is less than or equal to

$$\frac{1}{2\pi}T^m E\int_{-T}^{T}|\psi_{\tilde{F}_n}(t) - \psi_{F_n}(t)|\,\left|\frac{1}{t}\right|dt + \frac{2}{\pi}\int_T^{\infty} t^{m-1}e^{-t^2/2}\,dt,$$

which (by Schwarz's inequality and integration by parts) is less than or equal to

$$\frac{1}{2\pi}T^{m+1}\left[\frac{1}{T}\int_{-T}^{T}\left\{E|\psi_{\tilde{F}_n}(t) - \psi_{F_n}(t)|^2\right\}^{1/2}dt\right] + O\left(T^{m-2}e^{-T^2/2}\right),$$

which (again by Schwarz's inequality) in turn is less than or equal to

$$\frac{1}{2\pi}T^{m+1}\left\{\frac{1}{T}\int_{-\infty}^{\infty}E|\psi_{\tilde{F}n}(t) - \psi_{F_n}(t)|^2\,dt\right\}^{1/2} + O\left(T^{m-2}e^{-T^2/2}\right)$$

$$= \frac{1}{2\pi}T^{m+1}\left[\frac{1}{T}\int_{-\infty}^{\infty}E\left\{F_n(x;\tilde{\theta}) - F_n(x;\theta)\right\}^2 dx\right]^{1/2} + O\left(T^{m-2}e^{-T^2/2}\right)$$

$$= O\left(\frac{T^{m+1/2}}{\sqrt{n}}\right) + O\left(T^{m-2}e^{-T^2/2}\right).$$

Again taking $T = \sqrt{\log n}$, convergence at rate $n^{-\alpha/2}$ is established for any $\alpha \in (0, 1)$. □

COROLLARY. *Under the same hypotheses as in Theorem 2 if, in addition, $\theta_1, \ldots, \theta_n$ represents a random sample from some distribution G with $\int |\theta|\, dG(\theta) \leq M$, some $M < \infty$, then for any nonnegative integer m and constant $\alpha < 1$ ($\alpha \leq 1$ for $m = 0$) there exist constants $C'_{m,\alpha}$ and $D'_{m,\alpha}$ such that*

$$E\int_{-\infty}^{\infty} \left\{F_n^{(m)}(x; \tilde{\theta}) - F_G^{(m)}(x)\right\}^2 dx < \frac{C'_{m,\alpha}}{n^{\alpha}}$$

and

$$E|F_n^{(m)}(x; \tilde{\theta}) - F_G^{(m)}(x)| < \frac{D'_{m,\alpha}}{n^{\alpha/2}}$$

[*where* $F_G(x) = \int \Phi(x - \theta)\, dG(\theta)$].

The proofs for this special case are nearly identical to those of the theorem [with $F_n(x; \theta)$ replaced by $F_G(x)$], and will be omitted.

The results to follow establish the asymptotic optimality (in the sense of minimum average compound risk) of a sequence of decision functions $\tilde{t}_n(\cdot)$ (for estimation of a collection of normal means) which are Bayes rules with respect to $\tilde{G}_n$, the estimator of G_n (the empirical distribution function of the population of true means), and the quadratic loss function, under the assumption that the means lie within a fixed interval. Before proceeding, it should be emphasized that the function G_n is of paramount importance in the mean estimation problem, even if the sequence of means is a random sample from a distribution G, since knowledge of G_n reduces the problem to one of matching a population of n observations $x_1, \ldots, x_n$ with the n mass points of G_n. [The author (1983) refers to this as the permutation Bayes problem, since, in the absence of any other information, all $n!$ matchings may be taken as equally likely.] In this case, G furnishes no additional information, and hence may be disregarded. It is for this reason that the results presented in this paper relate primarily to G_n instead of G, although statements about G do appear in the corollaries.

Let $\mathbf{t}(\mathbf{X}) = (t_1(\mathbf{X}), \ldots, t_n(\mathbf{X}))$ be a decision rule for estimation of $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ from $\mathbf{X} = (X_1, \ldots, X_n)$, define the average quadratic compound risk for estimating $\boldsymbol{\theta}$ by $\mathbf{t}$ to be

$$R(\mathbf{t}, G_n) = \frac{1}{n}\sum_{j=1}^{n} E\left\{t_j(\mathbf{X}) - \theta_j\right\}^2,$$

and let $\mathscr{S}$ be the class of decision rules for estimation of $\boldsymbol{\theta}$ from $\mathbf{X}$ which are of the form

$$(t_1(\mathbf{X}), t_2(\mathbf{X}), \ldots, t_n(\mathbf{X})) = (t(X_1), t(X_2), \ldots, t(X_n)),$$

some $t(\cdot)$. Then for any $\mathbf{t}(\cdot) \in \mathscr{S}$, the average (quadratic) compound risk for estimating $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ is given by

$$R(\mathbf{t}, G_n) = \frac{1}{n}\sum_{j=1}^{n} E\left\{t_j(\mathbf{X}) - \theta_j\right\}^2 = \frac{1}{n}\sum_{j=1}^{n} \int_{-\infty}^{\infty} \left\{t(x) - \theta_j\right\}^2 \phi(x - \theta_j)\, dx,$$

which may be written as

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\{t(x)-\theta\}^2\phi(x-\theta)\,dx\,dG_n(\theta),$$

where $G_n(\cdot)$ is the empirical distribution function of θ. The optimal $t_n^*(\cdot)$ for this risk function is given by

$$t_n^*(x)=\frac{\int_{-\infty}^{\infty}\theta\phi(x-\theta)\,dG_n(\theta)}{\int_{-\infty}^{\infty}\phi(x-\theta)\,dG_n(\theta)}=x+\frac{F_n^{(2)}(x;\theta)}{F_n^{(1)}(x;\theta)}.$$

The following theorem pertains to the decision function

$$\tilde{\mathbf{t}}_n(\mathbf{X})=(\tilde{t}_n(X_1),\ldots,\tilde{t}_n(X_n)),$$

where

$$\tilde{t}_n(x)=x+\frac{F_n^{(2)}(x;\tilde{\theta})}{F_n^{(1)}(x;\tilde{\theta})}$$

($\tilde{\theta}$ as defined in Theorem 1), and establishes that it may be used as an approximation to $\mathbf{t}_n^*(\mathbf{X})$, with the result that the net increase in risk $R(\mathbf{t};G_n)$ for using $\tilde{\mathbf{t}}_n(\cdot)$ instead of $\mathbf{t}_n^*(\cdot)$ tends to 0 as $n\to\infty$, assuming the sequence of means to be bounded.

THEOREM 3. *For all sequences $\theta_1,\theta_2,\ldots$ with $|\theta_j|\le M$, all j, some $M<\infty$, and $\tilde{\theta}$ constrained to lie within $[-M,M]^n$, every j,*

$$\lim_{n\to\infty}\{R(\tilde{\mathbf{t}}_n;G_n)-R(\mathbf{t}_n^*;G_n)\}=0.$$

PROOF. First,

$$\begin{aligned}R(\tilde{\mathbf{t}}_n;G_n)&=\frac{1}{n}\sum_{j=1}^{n}E\{\tilde{t}_n(X_j)-\theta_j\}^2\\&=\frac{1}{n}\sum_{j=1}^{n}\Big(E\{\tilde{t}_n(X_j)-t_n^*(X_j)\}^2\\&\qquad+2E\big[\{\tilde{t}_n(X_j)-t_n^*(X_j)\}\{t_n^*(X_j)-\theta_j\}\big]+E\{t_n^*(X_j)-\theta_j\}^2\Big)\\&\le\frac{1}{n}\sum_{j=1}^{n}\Big[E|\tilde{t}_n(X_j)-t_n^*(X_j)|(2M+2\cdot 2M)+E\{t_n^*(X_j)-\theta_j\}^2\Big]\\&=\left[\frac{1}{n}\sum_{j=1}^{n}6ME|\tilde{t}_n(X_j)-t_n^*(X_j)|\right]+R(\mathbf{t}_n^*;G_n).\end{aligned}$$

Next, for any $\alpha\in(0,\frac{1}{2})$ let $c_n(\alpha)=-M+\sqrt{2\alpha\log n}$ and let $A_n^{(i)}(\alpha)=$

$\{|X_i| \le c_n(\alpha)\}$. Then

$$\frac{1}{n}\sum_{j=1}^{n} E|\tilde{t}_n(X_j) - t_n^*(X_j)|$$

$$= \frac{1}{n}\sum_{j=1}^{n}\left[\left\{E|\tilde{t}_n(X_j) - t_n^*(X_j)|I_{A_n^{(j)}(\alpha)}\right\} + \left\{E|\tilde{t}_n(X_j) - t_n^*(X_j)|I_{\bar{A}_n^{(j)}(\alpha)}\right\}\right]$$

$$\le \frac{1}{n}\sum_{j=1}^{n}\left[E\left\{\sup_{|x|\le c_n(\alpha)}|\tilde{t}_n(x) - t_n(x)|\right\} + 2M\Pr\{\bar{A}_n^{(j)}(\alpha)\}\right]$$

$$\le E\left(\sup_{|x|\le c_n(\alpha)}|\tilde{t}_n(x) - t_n(x)| + 4M[1 - \Phi\{c_n(\alpha) - M\}]\right)$$

[since $\Pr(|X_j| > x) \le 2\{1 - \Phi(x - M)\}$, each j], so

$$\frac{1}{n}\sum_{j=1}^{n} E|\tilde{t}_n(X_j) - t_n^*(X_j)| \le E\left\{\sup_{|x|\le c_n(\alpha)}|\tilde{t}_n(x) - t_n(x)|\right\} + o(1).$$

Henceforth, it will be assumed that n is large enough that $-M + \sqrt{2\alpha\log n} > 0$. Then for $|x| \le -M + \sqrt{2\alpha\log n}$

$$\frac{1}{F_n^{(1)}(x;\tilde{\theta})} \le \sqrt{2\pi}\exp\left\{\tfrac{1}{2}\left(\sqrt{2\alpha\log n}\right)^2\right\} = O(n^{\alpha})$$

as $n \to \infty$, and by Theorem 2, for $|x| \le -M + \sqrt{2\alpha\log n}$,

$$\tilde{t}_n(x) = \frac{F_n^{(2)}(x;\theta) + \left\{F_n^{(2)}(x;\tilde{\theta}) - F_n^{(2)}(x;\theta)\right\}}{F_n^{(1)}(x;\theta)\left\{1 + \dfrac{F_n^{(1)}(x;\tilde{\theta}) - F_n^{(1)}(x;\theta)}{F_n^{(1)}(x;\theta)}\right\}}$$

$$= \frac{F_n^{(2)}(x;\theta) + o_p(n^{-\alpha_1})}{F_n^{(1)}(x;\theta)\{1 + o_p(n^{-\alpha_2})\}},$$

any $\alpha_1 < \frac{1}{2}$ and $\alpha_2 < \frac{1}{2} - \alpha$, implying

$$\left[\text{since}\left|\frac{F_n^{(2)}(x;\theta)}{F_n^{(1)}(x;\theta)}\right| < c_n(\alpha)\right]$$

that for $0 < \alpha_1 < \frac{1}{2}$ and $0 < \alpha_2 < \frac{1}{2} - \alpha$,

$$\sup_{|x|<c_n(\alpha)}|\tilde{t}_n(x) - t_n^*(x)| = o_p\left\{n^{-\min(\alpha_1,\alpha_2)}(\log n)^{1/2}\right\} = o_p(1)$$

as $n \to \infty$. Since convergence in probability implies L_1-convergence for bounded random variables [remembering that $\sup_x|\tilde{t}_n(x) - t_n^*(x)| \le 2M$] we have that

$$E\left\{\sup_{|x|<c_n(\alpha)}|\tilde{t}_n(x) - t_n^*(x)|\right\} = o(1)$$

as $n \to \infty$. The net result is that

$$\frac{1}{n} \sum_{j=1}^{n} E|\tilde{t}_n(X_j) - t_n^*(X_j)| = o(1)$$

as $n \to \infty$, which completes the proof. □

COROLLARY. *If, in addition to the assumptions to the previous theorem, the sequence may be regarded as a random sample from a distribution function G with finite absolute first moment, and $t_G^*(\cdot)$ denotes the Bayes rule for estimation of θ with respect to G, then*

$$\lim_{n \to \infty} \{R(\tilde{t}_n; G) - R(t_G^*; G)\} = 0.$$

This result follows from the corollary of Theorem 2 in the same manner that Theorem 3 follows from Theorem 2.

Thus, it has been established that the expected difference between the risk for $\tilde{t}_n$ and the optimal tends to 0, provided that the sequence $\theta_1, \theta_2, \ldots$ remains bounded by some number M. Edelman (1983) has established a lower bound of essentially $1/\sqrt{n}$ for the rate of convergence of a similar risk function under these same conditions (where n is the number of problems compounded), although Theorem 2 implies that for any fixed x, $E\{\tilde{t}_n(x) - t_n^*(x)\}^2 = o(n^{-\alpha})$, every $\alpha < 1$, suggesting that the actual rate of convergence of the risk function might be closer to $1/n$. The simulation results seem to be consistent with this conjecture.

3. Computation and performance of $\tilde{G}$. As in nearly all minimization problems, the computation procedure to be presented here for evaluation of $\tilde{\theta}$ involves choice of an initial estimate, followed by iteration to a local minimum. In this case, the Newton–Raphson algorithm is particularly simple to employ, the derivatives being simple to calculate and the data set being a sensible and convenient starting point. Uniqueness may be checked by verifying that various starting values lead to the same local minimum. (It should be mentioned, however, that it is not possible to determine with certainty whether or not a particular local minimum is truly a global minimum in this manner.) Simulation examples seem to suggest that a local minimum may generally be attained in this fashion within three steps, but for problems of sample size $n > 50$, other search methods not requiring inversion of an $n \times n$ matrix (such as gradient search) might be expected to be more efficient computationally. Also, the author has yet to encounter an example in which iteration from different starting points has led to distinct local minima.

The following tables summarize the results of several simulation studies performed in order to investigate the small-sample properties of the estimator $\tilde{t}_n(\cdot)$, which the author (1983) refers to as the empirical permutation Bayes estimator, as compared to the properties of the permutation Bayes estimator $t_n^*(\cdot)$ (so named because it is Bayes with respect to the prior distribution

TABLE 1
Simulated mean-squared error for estimation of normal means

Normal prior, $n = 10$			
σ_θ	0	1	2
t_n^*	0.00(0.00)	0.42(0.02)	0.64(0.03)
t_n'	0.20(0.02)	0.62(0.02)	0.84(0.03)
$\tilde{t}_n$	0.24(0.02)	0.75(0.03)	1.0(0.03)

resulting from knowledge of the population of means without the knowledge of the correspondence between means and observations), and the linear empirical Bayes estimator

$$t_n'(X_i) = \bar{X} + \left(1 - \frac{1}{s_X^2}\right)^+ (X_i - \bar{X}), \qquad \bar{X} = \frac{1}{n}\sum_1^n X_j,$$

$$s_X^2 = \frac{1}{n-1}\sum_1^n (X_j - \bar{X})^2$$

[$(\cdot)^+$ again denoting the positive part of its argument]. [For a discussion of this estimator see Robbins (1983) and James and Stein (1958).]

For each column in Table 1 the means $\theta_1, \ldots, \theta_{10}$ have been simulated as normal variables with standard deviation σ_θ, and then normal errors (independent with mean 0 and variance 1) were added to the means to produce the observations $x_1, \ldots, x_{10}$. [For a discussion of generation methods and seed numbers, see Edelman (1983).] In each case, the average squared error for estimation of the mean is given, along with one standard error in parentheses. Table 2 is of the same format, with sample size $n = 20$.

Tables 3 and 4 are of the same format as Tables 1 and 2, but refer to the example in which the means are generated according to a symmetric two-point mixing distribution with standard deviation σ_θ.

The most important features to note are the fact that for a normal mixing distribution $\tilde{t}_n$ does not appear to perform too much worse than the linear empirical Bayes estimator (which is obviously the most desirable if the mixing

TABLE 2
Simulated mean-squared error for estimation of normal means

Normal prior, $n = 20$			
σ_θ	0	1	2
t_n^*	0.00(0.00)	0.48(0.02)	0.71(0.03)
t_n'	0.09(0.01)	0.60(0.02)	0.82(0.03)
$\tilde{t}_n$	0.15(0.02)	0.65(0.02)	0.93(0.03)

TABLE 3
Simulated mean-squared error for estimation of normal means

Two-point prior, $n = 10$			
σ_θ	0	1	2
t_n^*	0.00(0.00)	0.42(0.02)	0.23(0.04)
t_n'	0.20(0.02)	0.68(0.02)	0.88(0.03)
$\tilde{t}_n$	0.24(0.02)	0.80(0.03)	0.65(0.04)

TABLE 4
Simulated mean-squared error for estimation of normal means

Two-point prior, $n = 20$			
σ_θ	0	1	2
t_n^*	0.00(0.00)	0.43(0.02)	0.25(0.04)
t_n'	0.09(0.01)	0.57(0.02)	0.83(0.03)
$\tilde{t}_n$	0.15(0.02)	0.65(0.02)	0.55(0.04)

distribution is known to be normal), but outperforms this estimator as n increases if the mixing distribution is two-point (particularly if the two modes are far apart). This suggests that unless there is particular reason to suspect a normal mixing distribution (or the sample size is very small), the empirical permutation Bayes estimator $\tilde{t}_n$ might be preferable in practice.

4. Discussion. Edelman (1983) has explored the possibility of estimating $G_n(\theta)$ by minimizing the *weighted* integral-squared distance between the empirical distribution of observations and a normal mixture, for a variety of weight functions, but it appears that changing the weight function does not noticeably improve the convergence properties. For this reason, and in light of the particularly simple form of the minimization problem for the unweighted case, the results for the weighted case are not presented here.

It is also worth noting that all of the previous results for this problem rely on the assumption that the sequence of means be a random sample (so that G_n converges), whereas the main results presented here do not require this assumption.

For the quadratic loss estimation problem, the solution presented here will be asymptotically subminimax [in the sense of Robbins (1951)] for bounded sequences $\theta_1, \theta_2, \ldots$. Specifically, if $\mathbf{t}_n$ is any sequence of decision rules for estimating the mean of a multivariate normal distribution, then

$$\lim_{n\to\infty} \sup_{\{\{\theta_j\}:\,|\theta_j|<M\}} \frac{1}{n}\sum_1^n E\big(\tilde{t}_n(X_i) - \theta_i\big)^2 \le \lim_{n\to\infty} \sup_{\{\{\theta_j\}:\,|\theta_j|<M\}} \frac{1}{n}\sum_1^n E\big(t_{n,i}(\mathbf{X}) - \theta_i\big)^2$$

and for some sequence $\theta_1, \theta_2, \ldots$,

$$\lim_{n\to\infty} \frac{1}{n} \sum_1^n E\big(\tilde{t}_n(X_i) - \theta_i\big)^2 < \lim_{n\to\infty} \frac{1}{n} \sum_1^n E\big(t_{n,i}(\mathbf{X}) - \theta_i\big)^2.$$

While simulation results seem to suggest that the average mean-squared error of $\tilde{t}_n$ is never much greater than 1, this point has not been investigated thoroughly, and deserves further study.

Also, simulation results seem to suggest that the conclusion of Theorem 3 may hold under a much broader class of mean sequences $\theta_1, \theta_2, \ldots$, such as those in which

$$\frac{1}{n} \sum_1^n \big(\theta_i - \bar{\theta}\big)^2 \le M, \quad \text{some } M < \infty, \text{ all } n,$$

and that the rate of convergence of the average compound risk to the optimal may be inversely proportional to the sample size.

At any rate, the estimator presented here [aside from being much simpler than those of Singh (1979), George (1986) and others] has been shown to be asymptotically optimal for *both* the compound decision problem *and* the Bayes problem [see Edelman (1983)], with risk which appears to converge at a near-optimal rate, so that there might be some reason to prefer it in practice.

It is hoped that the estimation methods discussed here may be applicable to a wide variety of problems involving mixing distributions and that, in particular, these methods may lead to an advance in the general theory of empirical Bayes estimation.

Acknowledgments. The author would like to thank Herbert Robbins and Steven Lalley for their helpful suggestions and comments. The author would also like to thank the Associate Editor and referees for their helpful comments and suggestions.

REFERENCES

EDELMAN, D. (1983). Empirical permutation Bayes estimation: Gaussian case. Ph.D. dissertation, Columbia Univ.

GEORGE, E. (1986). Minimax multiple-shrinkage estimators. *Ann. Statist.* **14** 188–205.

JAMES, W. and STEIN, C. (1960). Estimation with quadratic loss. *Proc. Fourth Berkeley Symp. Math. Statist. Probab.* **1** 361–379. Univ. California Press.

KIEFER, J. and WOLFOWITZ, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Ann. Math. Statist.* **27** 887–906.

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

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.

RITOV, Y. (1987). On the deconvolution of a mixture of normal distributions. Unpublished.

ROBBINS, H. (1950). A generalization of the method of maximum likelihood: Estimating a mixing distribution (Abstract). *Ann. Math. Statist.* **21** 314–315.

ROBBINS, H. (1951). Asymptotically sub-minimax solutions of compound statistical decision problems. *Proc. Second Berkeley Symp. Math. Statist. Probab.* 131–148. Univ. California Press.

ROBBINS, H. (1956). An empirical Bayes approach to statistics. *Proc. Third Berkeley Symp. Math. Statist. Probab.* **1** 157–163. Univ. California Press.

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

ROBBINS, H. (1983). Some thoughts on empirical Bayes estimation. *Ann. Statist.* **11** 713–723.

SEELEY, R.T. (1966). *An Introduction to Fourier Series and Integrals.* Benjamin, New York.

SINGH, R. (1979). Empirical Bayes estimation in Lebesgue-exponential families with rates near the best possible rate. *Ann. Statist.* **7** 890–901.

DEPARTMENT OF STATISTICS
BOX 10 MATHEMATICS
COLUMBIA UNIVERSITY
NEW YORK, NEW YORK 10027