Statistica Sinica **7**(1997), 181-193

EMPIRICAL BAYES AND COMPOUND ESTIMATION OF NORMAL MEANS

Cun-Hui Zhang

Rutgers University

Dedicated to Herbert Robbins on his 80th birthday

Abstract: This article concerns the canonical empirical Bayes problem of estimating normal means under squared-error loss. General empirical estimators are derived which are asymptotically minimax and optimal. Uniform convergence and the speed of convergence are considered. The general empirical Bayes estimators are compared with the shrinkage estimators of Stein (1956) and James and Stein (1961). Estimation of the mixture density and its derivatives are also discussed.

Key words and phrases: Asymptotic optimality, empirical Bayes, minimaxity, normal distribution, shrinkage estimate.

1. Introduction

Let (X_j, θ_j), $1 \le j \le n$, be random vectors such that conditionally on $\theta_1, \ldots, \theta_n$, $X_1, \ldots, X_n$ are independent random variables with probability density functions $f(x|\theta_j)$, where $f(\cdot|\cdot)$ is a known family of densities with respect to some σ-finite measure. We are interested in estimating θ_j based on observations X_j under the average mean squared error (MSE)

$$\frac{1}{n}\sum_{j=1}^{n} E(\widehat{\theta}_j - \theta_j)^2. \tag{1}$$

If we want to estimate θ_j by $\widehat{\theta}_j = t(X_j)$ for some Borel function $t(\cdot)$, then (1) is minimized for

$$t_n^*(x) = t^*(x; G_n) = \frac{\int \theta f(x|\theta) dG_n(\theta)}{\int f(x|\theta) dG_n(\theta)}, \tag{2}$$

where $G_n(x) = n^{-1}\sum_{j=1}^{n} P\{\theta_j \le x\}$. In an empirical Bayes (EB) setting, θ_j are assumed to be independent and identically distributed (IID) random variables with a common but unknown distribution and $t_n^*(X_i) = E(\theta_j|X_j)$ is the Bayes estimator with the additional knowledge of the prior, whereas θ_j are assumed to be unknown constants in compound estimation problems. Here and in the sequel, the distributions G_n (and therefore implicitly the probability measure P in the

EB setting) are allowed to be dependent on n. In both the EB and compound cases, the distribution G_n is unknown, and a general empirical Bayes (GEB) estimator is of the form $\widehat{\theta}_j = \widehat{t}_n(X_j)$, where $\widehat{t}_n(\cdot)$ is some estimate of $t_n^*(\cdot)$ based on observations $X_1, \ldots, X_n$. This EB approach was proposed by Robbins (1951, 1956). See also Robbins (1983). GEB estimators $\widehat{t}_n(X_j)$ are asymptotically optimal at a distribution G, if

$$\frac{1}{n}\sum_{j=1}^{n} E(\widehat{t}_n(X_j) - \theta_j)^2 - \frac{1}{n}\sum_{j=1}^{n} E(t_n^*(X_j) - \theta_j)^2 \le o(1) \tag{3}$$

as $(n, G_n) \to (\infty, G)$ in certain topology. This asymptotic optimality criterion requires local uniformity and is slightly stronger than the usual one for fixed $G = G_n$ in the EB setting.

Alternatively, we may want to consider a linear empirical Bayes (LEB) estimator which approximate the best linear estimator among $\widehat{\theta}_j = A + BX_j$, given by

$$A_n^* + B_n^* x = \frac{1}{n}\sum_{j=1}^{n} E\theta_j + \frac{\sum_{j=1}^{n} \mathrm{Cov}(\theta_j, X_j)}{\sum_{j=1}^{n} \mathrm{Var}(X_j)} \Big(x - \frac{1}{n}\sum_{j=1}^{n} EX_j\Big). \tag{4}$$

For many families $f(x|\theta)$, the constants A_n^* and B_n^* are much easier to estimate than the function $t_n^*(\cdot)$ in (2). In general, the difference in the average MSE between LEB estimators and the optimal linear estimator $A_n^* + B_n^* X_j$ converges to zero faster and more uniformly than (3). LEB estimators may have other advantages over the GEB ones. For example when $f(x|\theta) \sim N(\theta, 1)$ is the normal family with the unit variance, the estimators of Stein (1956) and James and Stein (1961) have uniformly smaller average MSE than the usual maximum likelihood estimators (MLE) $\widehat{\theta}_j = X_j$ and are therefore minimax. However, LEB estimators are in general not asymptotically optimal in the sense of (3). In view of all these, for a given sample size n and under the risk (1), shall we use GEB estimators? In this paper we provide a partial affirmative answer to this question in the canonical case, the normal family with unit variance. Our GEB estimators are asymptotically optimal at every G in the sense of (3) and asymptotically minimax. Since many asymptotically optimal GEB estimators have been constructed in the past, we shall focus on global properties in terms of (1) and uniform speed of convergence, instead of the behavior of the risk when G_n is in an infinitesimal neighborhood of a fixed G as $n \to \infty$. Uniform risk convergence of empirical Bayes estimators is useful in certain semiparametric estimation problems (cf. e.g. Lindsay (1985)).

2. Estimation of Normal Means

Suppose throughout the sequel that the conditional density of X_i given θ_i is

$$f(x|\theta_i) = \varphi(x - \theta_i), \quad \varphi(x) = (2\pi)^{-1/2} \exp(-x^2/2). \tag{5}$$

By (2) the Bayes estimator with prior G_n is

$$\widehat{\theta}_j = t_n^*(X_j), \quad t_n^*(x) = t^*(x; G_n) = x + \frac{f'(x; G_n)}{f(x; G_n)}, \quad f(x; G) = \int \varphi(x - \theta) dG(\theta). \tag{6}$$

For $\rho \geq 0$ define

$$J(\rho, G) = \int_{-\infty}^{\infty} \left\{ \frac{f'(x; G)}{f(x; G)} \right\}^2 \left\{ 2 - \frac{f(x; G)}{\max(f(x; G), \rho)} \right\} \left\{ \frac{f(x; G)}{\max(f(x; G), \rho)} \right\} f(x; G) dx. \tag{7}$$

The Bayes risk of $t_n^*(X_j)$ is (cf. Proposition 1)

$$\frac{1}{n} \sum_{j=1}^{n} E(t_n^*(X_j) - \theta_j)^2 = 1 - J(0, G_n).$$

A GEB estimator is satisfactory if its risk (1) is uniformly bounded by $1 - J(0, G_n) + \epsilon_n$ for some $\epsilon_n \to 0+$. But this is unfortunately unachievable without the knowledge of G_n.

Example 1. Let $\mathcal{G}(M)$ be the collection of all discrete distributions with sparse support set $\{a_1, \ldots, a_m\}$ for some $m \geq 1$ such that $a_\ell - a_{\ell-1} \geq M$ for all $1 \leq \ell \leq m$, $a_0 = -\infty$. Then for large enough M_n, the knowledge of G_n for some $G_n \in \mathcal{G}(M_n)$ and the observation X_j provide the exact value of θ_j, the closest a_ℓ to X_j, with large probability, so that the Bayes risk $1 - J(0, G_n)$ of (6) converges to 0 uniformly over $\mathcal{G}(M_n)$. But the minimax risk over $\mathcal{G}(M_n)$ is 1 without the knowledge of G_n.

As a remedy for this deficiency of criterion (3) under uniform convergence, we shall consider GEB estimators which approximate truncated Bayes estimators of the form

$$\widehat{\theta}_j = t_{n,\rho_n}^*(X_j), \quad t_{n,\rho_n}^*(x) = t^*(x; \rho_n, G_n) = x + \frac{f'(x; G_n)}{\max(f(x; G_n), \rho_n)}, \quad 0 \leq \rho_n \to 0, \tag{8}$$

which have the risk (cf. Proposition 1)

$$\frac{1}{n} \sum_{j=1}^{n} E(t_{n,\rho_n}^*(X_j) - \theta_j)^2 = 1 - J(\rho_n, G_n). \tag{9}$$

For suitable ρ_n, (8) is closer to the best we can do based on observations $X_1, \ldots, X_n$ than (6) in the following sense: (a) if $f(X_j; G_n)$ is not too small ($>> \rho_n$),

$t^*_{n,\rho_n}(X_j) = t^*_n(X_j)$ as we are able to pool information from nearby observations to improve the MLE X_j; and (b) if $f(X_j; G_n)$ is too small ($<< \rho_n$), $t^*_{n,\rho_n}(X_j) \approx X_j$ as there are too few observations near X_j for us to approximate the Bayes estimator (6). The truncated Bayes estimator is always between the MLE and the Bayes estimator both almost surely and in risk,

$$\{t^*_n(x) - t^*_{n,\rho_n}(x)\}\{t^*_{n,\rho_n}(x) - x\} \geq 0, \quad 1 > 1 - J(\rho_n, G_n) > 1 - J(0, G_n).$$

We shall provide GEB estimators $\widehat{\theta}_j = \widehat{t}_n(X_j)$ which uniformly approximate (8) in risk for suitable $0 < \rho_n \to 0$ in the sense that

$$\epsilon_n \stackrel{\text{def}}{=} \sup\left\{\frac{1}{n}\sum_{j=1}^n E(\widehat{t}_n(X_j) - \theta_j)^2 - \frac{1}{n}\sum_{j=1}^n E(t^*_{n,\rho_n}(X_j) - \theta_j)^2\right\} \to 0, \tag{10}$$

where the supremum is taken over all distributions G_n. Certain other desirable properties of the GEB estimators will also be presented as consequences of the main result.

A natural approximation of (8) is

$$\widehat{\theta}_j = \widehat{t}_n(X_j), \quad \widehat{t}_n(x) = x + \frac{\widehat{f}'_n(x)}{\max(\widehat{f}_n(x), \rho_n)}, \tag{11}$$

where $\widehat{f}_n(\cdot)$ can be any "good" estimate of $f(\cdot; G_n)$ based on $X_1, \ldots, X_n$. Consider kernel estimators

$$\widehat{f}_n(x) = \frac{1}{n}\sum_{j=1}^n K(X_j - x, a_n) = (2\pi)^{-1}\int_{-a_n}^{a_n} e^{-ixt}\sum_{j=1}^n \frac{e^{itX_j}}{n} dt \tag{12}$$

for some suitable $0 < a_n \to \infty$ to be given later, where

$$K(x, a) = (2\pi)^{-1}\int_{-a}^{a} e^{ixt} dt = \begin{cases} \sin(ax)/(\pi x), & x \neq 0, \\ a/\pi, & x = 0. \end{cases} \tag{13}$$

Although $\widehat{f}_n$ may take negative values, $\int \widehat{f}_n(x)dx = 1$ always holds in the Riemann sense. A reason for using this kernel is the extreme thin tail of $f^*_n(t) = \int e^{ixt} f(x; G_n)dx$, bounded by $e^{-t^2/2}$ in absolute value, as

$$E\widehat{f}^{(k)}_n(x) - f^{(k)}(x; G_n) = -(2\pi)^{-1}\int_{|t|>a_n} (-it)^k e^{-ixt} f^*_n(t)dt. \tag{14}$$

Here and in the sequel $h^{(0)} = h$ and $h^{(k)} = (\partial/\partial x)^k h$ for any function h if the derivative exists.

Our main theorem asserts that the above GEB estimator approximates the truncated Bayes estimator (8) in risk at the rate of $O(1)(\log n)^{3/2}/(\rho_n n)$ uniformly in G_n.

Theorem 1. *Let $\widehat{\theta}_j = \widehat{t}_n(X_j)$ be the GEB estimators given by* (11)-(13). *Choose $a = a_n > 0$ and $\rho = \rho_n > 0$ such that $\sqrt{2\log n} \le a = O(\sqrt{\log n})$ and $a/(\rho\sqrt{n}) = o(1)$ as $n \to \infty$. Then*

$$\frac{1}{n}\sum_{j=1}^{n} E(\widehat{t}_n(X_j) - \theta_j)^2 \le 1 - J(\rho, G_n) + (1 + o(1))\left\{\frac{a}{\sqrt{3}} + \sqrt{-\log(\rho^2)}\right\}^2 \frac{a}{\pi\rho n},$$

where the $o(1)$ depends only on (n, a, ρ). Consequently the uniform convergence (10) *holds with $\epsilon_n = O(1)(\log n)^{3/2}/(\rho n)$.*

Theorem 1 is proved in Sections 4 and 5.

Corollary 1. *The GEB estimators in Theorem* 1 *are asymptotically minimax in the sense that*

$$\sup\left\{\frac{1}{n}\sum_{j=1}^{n} E(\widehat{t}_n(X_j) - \theta_j)^2\right\} \le 1 + o(1),$$

where the supremum is taken over all distributions G_n.

For $\rho \ge 0$ and $f(x;G) = \int \varphi(x-\theta)dG(\theta)$ define

$$\Delta(\rho, G) = J(0,G) - J(\rho,G) = \int_{-\infty}^{\infty}\left\{\frac{f'(x;G)}{f(x;G)}\right\}^2\left\{1 - \frac{f(x;G)}{\max(f(x;G),\rho)}\right\}^2 f(x;G)dx. \tag{15}$$

Proposition 1. *Let $t(\cdot)$ be a Borel function and $t_n^*(\cdot)$ and $f(\cdot;G_n)$ be as in* (6). *Then*

$$\frac{1}{n}\sum_{j=1}^{n} E(t(X_j) - \theta_j)^2 = 1 - J(0, G_n) + \int \{t(x) - t_n^*(x)\}^2 f(x;G_n)dx.$$

In particular, (9) *holds.*

The first statement of Proposition 1 is a well known fact in the EB literature (cf. e.g. Robbins (1983)), and the second follows from the first and (6), (8) and (15).

Proposition 2. *Let $1 < p < \infty$, $q = p/(p-1)$ and Z be a $N(0,1)$ variable. Then*

$$\Delta(\rho, G) \le (E|Z|^{2q})^{1/q}\left\{\int f(x;G)I\{f(x;G) \le \rho\}dx\right\}^{1/p},$$

where $\Delta(\rho, G)$ and $f(\cdot;G)$ are as in (15).

Proposition 2 can be proved by the Hölder inequality and the fact that $f'(x)/f(x) = E[Z|Y = x]$ for some variable Y with density $f(\cdot) = f(\cdot;G)$. By Proposition 2, $\Delta(\rho_n, G_n) \to 0$ when $G_n \to G$ in distribution and $\rho_n \to 0$, so that Theorem 1 implies the asymptotic optimality of our GEB estimators in the sense of (3) at every G. By Corollary 1, our GEB estimators are also asymptotically minimax. But does there exist a sequence of minimax estimators which is also asymptotically optimal? We don't know the answer to this question. George (1986) considered Stein-type minimax multiple shrinkage estimators, but his estimators depend on prespecified target shrinkage regions and weights. Proposition 2 also allows us to consider certain cases where the mass of G_n escapes towards $\pm\infty$.

For the normal case (5), the best linear estimator (4) can be written as

$$\widehat{\theta}_j = A_n^* + B_n^* X_j, \quad A_n^* + B_n^* x = \mu_n + \frac{\sigma_n^2 - 1}{\sigma_n^2}(x - \mu_n),$$

and its risk is

$$\frac{1}{n}\sum_{j=1}^{n} E(A_n^* + B_n^* X_j - \theta_j)^2 = \frac{\sigma_n^2 - 1}{\sigma_n^2}, \tag{16}$$

where $\mu_n = \int x f(x; G_n)dx$ and $\sigma_n^2 = \int x^2 f(x; G_n)dx - \mu_n^2$ are respectively the mean and variance of $f(\cdot; G_n)$ in (6).

Corollary 2. *Let $\widehat{\theta}_j = \widehat{t}_n(X_j)$ be the GEB estimators in Theorem 1. Then*

$$\sup\left\{\frac{1}{n}\sum_{j=1}^{n} E(\widehat{t}_n(X_j) - \theta_j)^2 - \frac{\sigma_n^2 - 1}{\sigma_n^2}\right\} = o(1),$$

where the supremum is taken over all distributions G_n.

Corollary 2 follows from Theorem 1, Proposition 2 and the fact that

$$\int f(x; G_n) I\{f(x; G_n) \leq \rho_n\} dx \leq \sigma_n^2/M^2 + 2M\rho_n,$$

for all positive M and ρ_n. Since the risk of the centered James-Stein estimator is $(\sigma_n^2 - 1)/\sigma_n^2 + o(1)$, Corollary 2 implies that the risk of the GEB estimator is at most slightly larger than the James-Stein estimator for large n. Examples can be easily given in which the difference between the James-Stein and the GEB estimator is nearly 1 (cf. George (1986)).

Let $\mathcal{G}_m$ be the collection of all discrete distributions G supported by at most m points, $G(\theta) = \sum_{\ell=1}^{m} \pi_\ell I\{a_\ell \leq \theta\}$ for some $\pi_\ell \geq 0$ and real a_ℓ.

Corollary 3. *Let $m_n = o(1/\rho_n)$ and $\widehat{\theta}_j = \widehat{t}_n(X_j)$ be the GEB estimators in Theorem 1. Then*

$$\sup_{G_n \in \mathcal{G}_{m_n}} \left\{\frac{1}{n}\sum_{j=1}^{n} E(\widehat{t}_n(X_j) - \theta_j)^2 - 1 + J(0, G_n)\right\} = o(1).$$

Corollary 3 follows from Theorem 1, Proposition 2 and the fact that

$$\int f(x;G)I\{f(x;G)\le\rho\}dx \le m\rho M + \sum_{\pi_\ell\ge M\rho}\pi_\ell\int\varphi(x-a_\ell)I\{\pi_\ell\varphi(x-a_\ell)\le\rho\}dx$$
$$\le m\rho M+\int\varphi(x)I\{\varphi(x)\le 1/M\}dx$$

for all positive ρ and M and $G(\theta) = \sum_{\ell=1}^m \pi_\ell I\{a_\ell \le \theta\} \in \mathcal{G}_m$. It states that the GEB estimator is uniformly close to the Bayes estimator in risk if the means $\theta_1,\ldots,\theta_n$ are sampled from a set of at most m_n real numbers. Due to the condition $a_n/(\rho_n\sqrt{n}) = o(1)$ of Theorem 1, here m_n is allowed to be $o(1)\sqrt{n/\log n}$. But we are not sure whether this is the best rate for m_n.

3. Estimation of the Mixture Density

In this section we consider properties of the estimator (12) for the mixture density $f_n(\cdot) = f(\cdot;G_n)$ in (6). Let $\|h\|_p$ be the L^p norm with respect to the Lebesgue measure.

Theorem 2. *Let $\widehat{f}_n(\cdot)$ be defined by* (12) *with $a = a_n \ge \sqrt{\log n}$ and $f_n(\cdot) = f(\cdot;G_n)$ be as in* (6). *Then for $p \ge 1$ and integers $k \ge 0$*

$$\{E\|\widehat{f}_n^{(k)} - f_n^{(k)}\|_2^{2p}\}^{1/p} \le \frac{\{B_{2p}^2 + o(1)\}a^{2k+1}}{\pi(2k+1)n}$$

and

$$\{E\|\widehat{f}_n^{(k)} - f_n^{(k)}\|_\infty^{p}\}^{1/p} \le \frac{\{B_p + o(1)\}a^{k+1}}{\pi(k+1)\sqrt{n}},$$

where B_p are constants, depending on p only, such that $B_p = 1$ for $1 \le p \le 4$.

In the EB setting with IID θ_j, the estimator (12) is related to the kernel deconvolution estimator of the mixing densities considered by Carroll and Hall (1988), Carroll and Stefanski (1990), Fan (1991) and Zhang (1990). The kernel deconvolution estimator for $g_n = G_n'$ can be written as

$$\widehat{g}_n(\theta) = (2\pi)^{-1}\int_{-\infty}^{\infty} e^{-i\theta t}e^{t^2/2}\Big\{\int_{-\infty}^{\infty} e^{ixt}\widehat{f}_n(x)dx\Big\}dt,$$

motivated by the fact that the ratio of the characteristic functions of f_n and g_n is $e^{-t^2/2}$, the characteristic function of $N(0,1)$. But the optimal choice of the bandwidth $a_n = c\sqrt{\log n}$ is different: $c < 1$ for the estimation of mixing density g_n, while $c \ge 1$ for the estimation of mixture density f_n. This indicate that good estimates of the mixing density g_n may not produce good estimates of the mixture density f_n via the Fourier inversion. The rate of convergence for $\widehat{f}_n$

in Theorem 3 is slightly better than the rate obtained by Edelman (1987) who considered minimum distance estimates of f_n.

Proof of Theorem 2. Let $f_n^*(t) = \int e^{itx} f_n(x)dx$ and $Z_n(t) = n^{-1}\sum_{j=1}^n \exp(itX_j)$. Since $EZ_n(t) = f_n^*(t)$, there exist constants B_p and B_p' depending on p only such that

$$\{E|Z_n(t) - f_n^*(t)|^p\}^{1/p} \le n^{-1/2}\left(B_p + B_p' e^{-t^2/2}\right). \tag{17}$$

Here, $B_4 = B_4' = 1$ by direct computation, $B_p = B_4$ and $B_p' = B_4'$ by the Hölder inequality for $1 \le p \le 4$, and $B_p' = 0$ for some $B_p < \infty$ by the Marcinkiewicz-Zygmund inequality (Chow and Teicher (1988), page 368) for $p > 4$. Furthermore, since $\widehat{f}_n(x) = (2\pi)^{-1}\int_{-a}^a e^{-ixt} Z_n(t)dt$ by (12) and $|f_n^*(t)| \le \exp(-t^2/2)$, we have

$$\|\widehat{f}_n^{(k)} - f_n^{(k)}\|_2^2 \le \frac{1}{2\pi}\int_{-a}^a t^{2k}|Z_n(t) - f_n^*(t)|^2 dt + \frac{1}{\pi}\int_a^\infty t^{2k} e^{-t^2} dt \tag{18}$$

and

$$\|\widehat{f}_n^{(k)} - f_n^{(k)}\|_\infty \le \frac{1}{2\pi}\int_{-a}^a |t|^k |Z_n(t) - f_n^*(t)| dt + \frac{1}{\pi}\int_a^\infty |t|^k e^{-t^2/2} dt. \tag{19}$$

Putting (17)-(19) together, we obtain by the Hölder inequality

$$\begin{aligned}
&2\pi\{E\|\widehat{f}_n^{(k)} - f_n^{(k)}\|_2^{2p}\}^{1/p} \\
&\le \left\{\left(\frac{2a^{2k+1}}{2k+1}\right)^{p-1}\int_{-a}^a t^{2k} E|Z_n(t) - f_n^*(t)|^{2p} dt\right\}^{1/p} + O(a^{2k-1}/n) \\
&\le \frac{2a^{2k+1}}{2k+1}\{B_{2p}^2 + o(1)\}/n + O(a^{2k-1}/n) = \frac{2B_{2p}^2 + o(1)}{2k+1} a^{2k+1}/n
\end{aligned}$$

as $a > \sqrt{\log n}$ implies $\int_a^\infty t^{2k} e^{-t^2} dt = O(a^{2k-1}/n)$. Similarly

$$\begin{aligned}
&2\pi\{E\|\widehat{f}_n^{(k)} - f_n^{(k)}\|_\infty^p\}^{1/p} \\
&\le \left\{\left(\frac{2a^{k+1}}{k+1}\right)^{p-1}\int_{-a}^a |t|^k E|Z_n(t) - f_n^*(t)|^p dt\right\}^{1/p} + O(a^{k-1}/\sqrt{n}) \\
&\le \frac{2B_p + o(1)}{k+1} a^{k+1}/\sqrt{n}.
\end{aligned}$$

4. Proof of Theorem 1 (Part I)

Let (Y_n, λ_n) be a random vector independent of $(X_j, \theta_j), 1 \le j \le n$, such that

$$Y_n|\lambda_n \sim N(\lambda_n, 1), \quad P\{\lambda_n \le t\} = G_n(t) = \frac{1}{n}\sum_{j=1}^n P\{\theta_j \le t\}. \tag{20}$$

The proof of Theorem 1 has two steps. The first step is equivalent to the proof of the result in a sequential EB setting: estimating λ_n based on $Y_n, X_1, \ldots, X_n$. This is done here and the second step in Section 5.

The Bayes estimator of λ_n is $E\{\lambda_n|Y_n\} = t_n^*(Y_n)$ by (6) with the squared error loss, and the Bayes risk is $1 - J(0, G_n)$ by (7).

Theorem 3. *Let $\widehat{t}_n(Y_n)$ be the GEB estimator of λ_n with the $\widehat{t}_n(\cdot)$ given by* (11)-(13). *Choose $a = a_n > 0$ and $\rho = \rho_n > 0$ such that $\sqrt{\log n} \le a = O(\sqrt{\log n})$ and $a/(\rho\sqrt{n}) = o(1)$ as $n \to \infty$. Then*

$$E(\widehat{t}_n(Y_n) - \lambda_n)^2 \le 1 - J(\rho, G_n) + (2 + o(1))\{\Delta(\rho, G_n)\}^{1/2}\varphi(a)\sqrt{\frac{a}{\rho}} + (1 + o(1))\left\{\frac{a}{\sqrt{3}} + \sqrt{-\log(\rho^2)}\right\}^2 \frac{a}{\pi\rho n}, \tag{21}$$

where $J(\rho, G)$ and $\Delta(\rho, G)$ are given by (7) *and* (15) *respectively.*

Remark. If $\sqrt{2\log n} \le a = O(\sqrt{\log n})$, then the third term of the right-hand side of (21) is of smaller order than the fourth, and the statements of Theorems 1 and 3 are comparable.

Lemma 1. *Suppose $a/(\rho\sqrt{n}) = o(1)$. Let $f_n(\cdot) = f(\cdot; G_n)$ be as in* (6). *Then*

$$E\int\left\{\widehat{f}_n^{(k)}(y) - f_n^{(k)}(y)\right\}^2 \frac{\max(f_n(y), \rho)}{\max(\widehat{f}_n(y), \rho)} dy \le \frac{(1 + o(1))a^{2k+1}}{(2k+1)\pi n}.$$

Proof. Since $|\max(f_n, \rho) - \max(\widehat{f}_n, \rho)| \le |f_n - \widehat{f}_n|$,

$$\left\{\widehat{f}_n^{(k)} - f_n^{(k)}\right\}^2 \frac{\max(f_n, \rho)}{\max(\widehat{f}_n, \rho)} \le (\widehat{f}_n^{(k)} - f_n^{(k)})^2\{1 + |\widehat{f}_n - f_n|/\rho\}.$$

It follows from Theorem 2 and the condition $a/(\rho\sqrt{n}) = o(1)$ that

$$E\left\|(\widehat{f}_n^{(k)} - f_n^{(k)})\sqrt{|\widehat{f}_n - f_n|}\right\|_2^2 \le \sqrt{E\|\widehat{f}_n^{(k)} - f_n^{(k)}\|_2^4 E\|\widehat{f}_n - f_n\|_\infty^2} = O\left(\frac{a^{2k+2}}{n^{3/2}}\right) = o(a^{2k+1}\rho/n).$$

This proves the lemma as $E\|\widehat{f}_n^{(k)} - f_n^{(k)}\|_2^2 \le (1 + o(1))a^{2k+1}/\{(2k+1)\pi n\}$ by Theorem 2.

Lemma 2. *Let $f(x) = \int \varphi(x - \theta)dG(\theta)$ for some distribution function G. Then, $\{f'(x)/f(x)\}^2 \le -\log\{2\pi f^2(x)\}$ for all x, and for $\rho \le \{e\sqrt{2\pi}\}^{-1}$*

$$\{f'(x)/f(x)\}^2 f(x)/\max(f(x), \rho) \le -\log\{2\pi\rho^2\}, \quad \forall x. \tag{22}$$

Proof. Let $z = -f'(x)/f(x)$ and $dH(t) = \varphi(t)dG(t+x)/f(x)$. Then $z = \int t dH(t)$. Since $1/\varphi(t)$ is convex in t, by the Jensen inequality $1/\varphi(z) \le \int\{1/\varphi(t)\}dH(t) = 1/f(x)$, which gives

$$\{f'(x)/f(x)\}^2 = z^2 \le -\log\{2\pi f^2(x)\}.$$

For (22), we notice that $-t\log(\sqrt{2\pi}t)$ is increasing in t for $0 \le \sqrt{2\pi}t \le e^{-1}$.

Proof of Theorem 3. By definition $\widehat{t}_n - t_n^* = \widehat{f}_n'/\max(\widehat{f}_n,\rho) - f_n'/f_n = \xi_{1n} + \xi_{2n}$, where

$$\xi_{1n} = \frac{\widehat{f}_n' - f_n'}{\max(\widehat{f}_n,\rho)}, \qquad \xi_{2n} = \left(\frac{f_n'}{f_n}\right)\frac{f_n - \max(\widehat{f}_n,\rho)}{\max(\widehat{f}_n,\rho)}, \qquad f_n(x) = f(x;G_n).$$

Since Y_n is independent of $\widehat{t}_n$ and $t_n^*(Y_n)$ is the Bayes rule

$$E(\widehat{t}_n(Y_n) - \lambda_n)^2 = 1 - J(0,G_n) + E\{\widehat{t}_n(Y_n) - t_n^*(Y_n)\}^2,$$

so that by (15) it suffices to show

$$\begin{aligned} E\{\widehat{t}_n(Y_n) - t_n^*(Y_n)\}^2 &= E\int\{\xi_{1n}(y) + \xi_{2n}(y)\}^2 f_n(y)dy \\ &\le \Delta(\rho,G_n) + (2+o(1))\{\Delta(\rho,G_n)\}^{1/2}\varphi(a)\sqrt{\frac{a}{\rho}} \\ &\quad + (1+o(1))\left\{\frac{a}{\sqrt{3}} + \sqrt{-\log(\rho^2)}\right\}^2\frac{a}{\pi\rho n}. \end{aligned} \tag{23}$$

By Lemma 1

$$E\int \xi_{1n}^2(y)f_n(y)dy \le \frac{(1+o(1))a^3}{3\pi\rho n}. \tag{24}$$

Since $|f_n - \max(\widehat{f}_n,\rho)| \le |f_n - \widehat{f}_n|$ for $\max(f_n,\widehat{f}_n) \ge \rho$, it follows from Lemma 2 that

$$\xi_{2n}^2 \le -\log(2\pi\rho^2)\frac{\max(f_n,\rho)}{f_n}\left(\frac{f_n - \widehat{f}_n}{\max(\widehat{f}_n,\rho)}\right)^2 + \left(\frac{f_n'}{f_n}\right)^2\left(\frac{f_n-\rho}{\rho}\right)^2 I\{f_n < \rho\},$$

so that by Lemma 1 and (15)

$$E\int \xi_{2n}^2(y)f_n(y)dy \le (1+o(1))\frac{-\log(\rho^2)a}{\pi\rho n} + \Delta(\rho,G_n). \tag{25}$$

Since $|(f_n/c^2 - 1/c) - (f_n/\rho^2 - 1/\rho)| \le |1/c - 1/\rho|$ for $f_n \le \rho \le c$ (e.g. $c = \max(\widehat{f}_n,\rho)$),

$$\xi_{1n}\xi_{2n} \le \frac{|f_n'|}{f_n}\frac{|(\widehat{f}_n' - f_n')(f_n - \widehat{f}_n)|}{\rho\max(\widehat{f}_n,\rho)} + (\widehat{f}_n' - f_n')\frac{f_n'(f_n-\rho)}{f_n\rho^2}I\{f_n < \rho\}.$$

By the Schwarz inequality and Lemmas 1 and 2

$$E\int\Big\{\frac{|f_n'|}{f_n}\frac{|(\widehat{f}_n'-f_n')(f_n-\widehat{f}_n)|}{\rho\max(\widehat{f}_n,\rho)}\Big\}(y)f_n(y)dy\le(1+o(1))\frac{\sqrt{-\log(\rho^2)/3}a^2}{\pi\rho n}.$$

Since $|f_n^*(t)|\le\exp(-t^2/2)$, by (14)

$$\int\{E\widehat{f}_n'-f_n'\}^2(y)dy=\frac{1}{2\pi}\int_{|t|>a}|tf_n^*(t)|^2dt\le\frac{1}{\pi}\int_{t>a}t^2e^{-t^2}dt=(1+o(1))a\varphi^2(a).$$

These and (15) and the Schwarz inequality imply

$$E\int 2\xi_{1n}(y)\xi_{2n}(y)f_n(y)dy\le(2+o(1))\Big\{\frac{\sqrt{-\log(\rho^2)/3}a^2}{\pi\rho n}+\{\Delta(\rho,G_n)\}^{1/2}\varphi(a)\sqrt{\frac{a}{\rho}}\Big\}. \quad (26)$$

Hence, we have (23) and the conclusion by summing up (24)-(26).

5. Proof of Theorem 1 (Part II)

In the EB setting with IID θ_j, we may use $\widehat{\theta}_j=\tilde{t}_{n,[j]}(X_j)$ and obtain the upper bound of (1) by Theorem 3, where $\tilde{t}_{n,[j]}(\cdot)$ is the estimation of $t_n^*(\cdot)$ based on $n-1$ observations $X_1,\ldots,X_{j-1},X_{j+1},\ldots,X_n$. But this does not directly imply Theorem 1, as the θ_j are possibly dependent and not necessarily identically distributed. This technical point is handled here.

Let $X_1',\ldots,X_n'$ be random variables such that conditionally on $\theta_1,\ldots,\theta_n$, λ_n, they are independent of $X_1,\ldots,X_n,Y_n$ and distributed according to $X_j'\sim N(\theta_j,1)$. Define for $1\le j\le n$

$$\widehat{t}_{n,[j]}(x)=x+\widehat{f}_{n,[j]}'(x)/\max(\widehat{f}_{n,[j]}(x),\rho_n), \quad (27)$$

where $\widehat{f}_{n,[j]}(\cdot)$ is the estimate of f_n based on $X_1,\ldots,X_{j-1},X_j',X_{j+1},\ldots,X_n$,

$$\widehat{f}_{n,[j]}(x)=\frac{1}{n}\left\{K(X_j'-x,a_n)+\sum_{1\le\ell\le n,\ell\ne j}K(X_\ell-x,a_n)\right\}. \quad (28)$$

Lemma 3. *Let $\widehat{t}_n(\cdot)$ and $\widehat{t}_{n,[j]}(\cdot)$ be given by* (6) *and* (27) *respectively. Suppose the conditions of Theorem* 3 *hold. Then*

$$\left\{n^{-1}\sum_{j=1}^n E(\widehat{t}_{n,[j]}(X_j')-\widehat{t}_n(X_j'))^2\right\}^{1/2}\le O(1)a_n^{3/2}/(\rho_n n).$$

Proof. By (27) and (11)-(13)

$$\widehat{t}_{n,[j]}(X_j')-\widehat{t}_n(X_j')=\widehat{f}_{n,[j]}'(X_j')/\max(\widehat{f}_{n,[j]}(X_j'),\rho)-\widehat{f}_n'(X_j')/\max(\widehat{f}_n(X_j'),\rho)$$

$$= [\widehat{f}'_{n,[j]}(X'_j) - \widehat{f}'_n(X'_j)] / \max(\widehat{f}_{n,[j]}(X'_j), \rho) + \frac{[\widehat{f}'_n(X'_j)/\max(\widehat{f}_n(X'_j),\rho)][\max(\widehat{f}_n(X'_j),\rho) - \max(\widehat{f}_{n,[j]}(X'_j),\rho)]}{\max(\widehat{f}_{n,[j]}(X'_j),\rho)}.$$

Since $K'(0,a) = 0$ by (13) and $X_j - X'_j \sim N(0,2)$, it follows from definition (12) and (28) that

$$E\{\widehat{f}'_{n,[j]}(X'_j) - \widehat{f}'_n(X'_j)\}^2 \le E\{K'(X_j - X'_j, a)\}^2/n^2 \le \{1/\sqrt{4\pi}\} \int \{K'(x,a)\}^2 dx/n^2$$
$$= \{1/\sqrt{4\pi}\}(2\pi)^{-1} \int_{-a}^{a} t^2 dt/n^2 = \{2\sqrt{\pi}/3\} a^3/(2\pi n)^2.$$

By (12), (13) and (28),

$$|\max(\widehat{f}_n(X'_j),\rho) - \max(\widehat{f}_{n,[j]}(X'_j),\rho)| \le |K(X_j - X'_j, a) - a/\pi|/n \le 2a/(\pi n).$$

Putting these together, we have

$$\Big\{ n^{-1} \sum_{j=1}^n E(\widehat{t}_{n,[j]}(X'_j) - \widehat{t}_n(X'_j))^2 \Big\}^{1/2}$$
$$\le \Big\{ n^{-1} \sum_{j=1}^n E(\widehat{f}'_{n,[j]}(X'_j) - \widehat{f}'_n(X'_j))^2 \Big\}^{1/2} / \rho$$
$$+ \Big\{ n^{-1} \sum_{j=1}^n E(\widehat{f}'_n(X'_j)/\max(\widehat{f}_n(X'_j),\rho))^2 \Big\}^{1/2} 2a/(\pi\rho n)$$
$$\le \{2\sqrt{\pi}/3\}^{1/2} a^{3/2}/(2\pi n\rho) + \Big\{ n^{-1} \sum_{j=1}^n E(\widehat{f}'_n(X'_j)/\max(\widehat{f}_n(X'_j),\rho))^2 \Big\} 2a/(\pi\rho n).$$

Hence, the conclusion holds, as Theorem 3 implies

$$n^{-1} \sum_{j=1}^n E[\widehat{f}'_n(X'_j)/\max(\widehat{f}_n(X'_j),\rho)]^2$$
$$= E[\widehat{t}_n(Y_n) - Y_n]^2 \le 2E[\widehat{t}_n(Y_n) - \lambda_n]^2 + 2E[Y_n - \lambda_n]^2 \le 4 + o(1).$$

Proof of Theorem 1. It follows from Lemma 3 that

$$\Big\{ n^{-1} \sum_{j=1}^n E(\widehat{t}_n(X_j) - \theta_j)^2 \Big\}^{1/2} = \Big\{ n^{-1} \sum_{j=1}^n E[\widehat{t}_{n,[j]}(X'_j) - \theta_j]^2 \Big\}^{1/2}$$
$$\le \Big\{ n^{-1} \sum_{j=1}^n E[\widehat{t}_n(X'_j) - \theta_j]^2 \Big\}^{1/2} + \Big\{ n^{-1} \sum_{j=1}^n E[\widehat{t}_{n,[j]}(X'_j) - \widehat{t}_n(X'_j)]^2 \Big\}^{1/2}$$
$$= \{E[\widehat{t}_n(Y_n) - \lambda_n]^2\}^{1/2} + O(1) a^{3/2}/(\rho n).$$

Since $E\{\widehat{t}_n(Y_n) - \lambda_n\}^2 \le 1 + o(1)$ by Theorem 3 and $1/(\rho n) = o(1)$, this implies

$$n^{-1}\sum_{j=1}^{n} E(\widehat{t}_n(X_j) - \theta_j)^2 = E[\widehat{t}_n(Y_n) - \lambda_n]^2 + o(1)a^3/(\rho n).$$

Hence the conclusion follows from Theorem 3.

Acknowledgements

This research is partially supported by the National Security Agency and Army Research Office. The author would like to thank Herbert Robbins and William Strawderman for insightful conversations.

References

Carroll, R. J. and Hall, P. (1988). Optimal rates of convergence for deconvolving a density. *J. Amer. Statist. Assoc.* **83**, 1184-1186.

Chow, Y. S. and Teicher, H. (1988). *Probability Theory.* Springer-Verlag, New York.

Edelman, D. (1987). Estimation of the mixing distribution for a normal mean with applications to the compound decision problem. *Ann. Statist.* **16**, 1609-1622.

Fan, J. (1991). On the optimal rates of convergence for nonparametric deconvolution problems. *Ann. Statist.* **19**, 1257-1272.

George, E. I. (1986). Minimax multiple shrinkage estimation. *Ann. Statist.* **14**, 188-205.

James, W. and Stein, C. (1961). Estimation with quadratic loss. *Proc. Fourth Berkeley Symp. Math. Statist. Probab.* **1**, 361-379. Univ. of California Press, Berkeley.

Lindsay, B. G. (1985). Using empirical partially Bayes inference for increased efficiency. *Ann. Statist.* **13**, 914-931.

Robbins, H. (1951). Asymptotically subminimax solutions of compound statistical decision problems. *Proc. Second Berkeley Symp. Math. Statist. Probab.* **1**, 131-148. Univ. of California Press, Berkeley.

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

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

Stefanski, L. A. and Carroll, R. J. (1990). Deconvolvting kernel density estimators. *Statist.* **21**, 169-184.

Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. *Proc. Third Berkeley Symp. Math. Statist. Probab.* **1**, 157-163. Univ. of California Press, Berkeley.

Zhang, C.-H. (1990). Fourier methods for estimating mixing densities and distributions. *Ann. Statist.* **18**, 806-831.

Department of Statistics, Hill Center for the Mathematical Sciences, Busch Campus, Rutgers University, New Brunswick, NJ 08903, U.S.A.

(Received August 1995; accepted March 1996)