Parametric Empirical Bayes Inference: Theory and Applications

Author(s): Carl N. Morris

Source: *Journal of the American Statistical Association*, Mar., 1983, Vol. 78, No. 381 (Mar., 1983), pp. 47-55

Published by: Taylor & Francis, Ltd. on behalf of the American Statistical Association

Stable URL: https://www.jstor.org/stable/2287098

Parametric Empirical Bayes Inference: Theory and Applications

CARL N. MORRIS*

This article reviews the state of multiparameter shrinkage estimators with emphasis on the empirical Bayes viewpoint, particularly in the case of parametric prior distributions. Some successful applications of major importance are considered. Recent results concerning estimates of error and confidence intervals are described and illustrated with data.

KEY WORDS: Empirical Bayes; Parametric empirical Bayes; Empirical Bayes confidence intervals; Stein's estimator.

1. MULTIPARAMETER ESTIMATION: BORROWING STRENGTH FROM THE ENSEMBLE

Statisticians often must make many estimates of similar quantities, for example, estimates of annual income must be made for many areas, the quality of a process must be re-evaluated every month, an actuary must estimate the risks of many different groups of drivers, and so on. Thus, the situation of Table 1 often arises.

We assume that k parameters $\theta_1, \ldots, \theta_k$ must be estimated starting with independent unbiased estimates $Y_1, \ldots, Y_k$, $EY_i = \theta_i$. In cases where Y_i is normal we can write

$$Y_i \mid \theta_i \sim N(\theta_i, V_i), \quad i = 1, \ldots, k \text{ independently} \tag{1.1}$$

with $V_i = \text{var}(Y_i)$. The estimate Y_i often would be a reduction of the original data, perhaps a sample mean or some other statistic, each computed from an independent set of data. We will take the variances V_i to be known here, in order to concentrate on the main issues. Usually, however, they would have to be estimated, perhaps by the within-groups sums of squares.

If the V_i are suitably small, the statistician may be content to use the estimates Y_i, but otherwise he may seek alternative estimates $\tilde{\theta}_i$ of θ_i. Examples include $\tilde{\theta}_i = \bar{Y}$, the grand mean, or more generally,

$$\tilde{\theta}_i = z_i' b, \tag{1.2}$$

with z_i an r-dimensional column vector of regression variables, and b a vector of estimated regression coefficients. We assume $0 \le r \le k - 3$ throughout (this is required, for example, so that the expectations in (1.4), (1.17) exist and so that (1.19) be positive). The estimates (1.2) would be ideal if one knew that for every i

$$\theta_i = z_i' \beta. \tag{1.3}$$

In that case b is computed as the least squares estimator

$$b = (Z'DZ)^{-1} Z'DY \tag{1.4}$$

with Z the $k \times r$ matrix having rows z_i', D^{-1} = Diagonal $(V_1, \ldots, V_k)$, $Y = (Y_1, \ldots, Y_k)'$. Then $\tilde{\theta}_i$ is unbiased, and dominates Y_i as an estimator of θ_i for sum of squared error loss function.

Which estimate of θ_i should be preferred if (1.3) is not known to hold? In the most interesting cases a compromise estimator

$$\tilde{\theta}_i = (1 - B_i) Y_i + B_i \tilde{\theta}_i \tag{1.5}$$

will outperform either extreme choice. Here B_i is a shrinking coefficient, $0 \le B_i \le 1$, $B_i = 0$ resulting in no shrinkage of $\tilde{\theta}_i$ from Y_i towards $\tilde{\theta}_i$.

How do we choose the B_i? That depends on the variance A of the errors $\theta_i - z_i' \beta$. If $A = 0$ then (1.3) holds and $B_i = 1$ is best, but B_i should decrease to 0 as $A \to \infty$. Although A is unknown, the Y_i values contain information about A that can be used profitably.

This issue can be incorporated formally by making the assumption that

$$\theta_i \mid \beta, A_i \overset{\text{indep}}{\sim} N(z_i' \beta, A_i) \; i = 1, \ldots k \tag{1.6}$$

to complement the model (1.1). Here, we allow θ_i to have a different variance A_i for each i. Many other assumptions might be appropriate for (1.6), that the θ_i be dependent, that the θ_i be nonnormal, and so on. The complexity of such assumptions must be restricted, however, so the data Y can be used to investigate the model (1.6) and any unknown parameters.

The choice between Y_i and $\tilde{\theta}_i$ as an estimator of θ_i is, in terms of (1.6), a choice between $A_i = \infty$ and $A_i = 0$, respectively. The model (1.1), (1.6) will lead to better results because it is richer, also allowing values $0 < A_i < \infty$. Thus, the superiority of shrinkage estimators like

* Carl N. Morris is Professor, Department of Mathematics, University of Texas, Austin, TX 78712. This paper was presented at the August 16–19, 1982 meeting of the American Statistical Association, Cincinnati, Ohio, as a General Methodology Lecture. The author wishes to acknowledge interdependence between the ideas expressed here and those of his longstanding empirical Bayes collaborator, Bradley Efron, but without implying Dr. Efron's acceptance of all the paper's declarations. Research supported by the National Science Foundation under Grant MCS-8104-250.

March 1983, Volume 78, Number 381
Applications Section

Table 1. One-way Layout Showing the Statistician's Simultaneous Estimation Problem. The Parameters $\theta_1, \ldots, \theta_k$ are to be Estimated From Independent Unbiased Estimates $Y_1, \ldots, Y_k$. The Alternative Estimates $\tilde{\theta}_1, \ldots, \tilde{\theta}_k$ Probably are Biased Because They are Computed From the Y_i Values by Making Restrictive Assumptions About the Parameters, Perhaps Using Concomitant Information $z_1, \ldots, z_k$.

Parameters to be Estimated	Unbiased Estimate	Variance	Restricted Model Estimate	Concomitant Information Vector
θ_1	Y_1	V_1	$\tilde{\theta}_1$	z_1
θ_2	Y_2	V_2	$\tilde{\theta}_2$	z_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
θ_k	Y_k	V_k	$\tilde{\theta}_k$	z_k

(1.5) will be attributable mainly to the richer model. That is, the benefits are mainly due to assuming (1.6), and less critically to how one estimates B_i.

We derive from (1.1), (1.6) the posterior distribution of the unknown parameters, for $i = 1, \ldots k$,

$$\theta_i \mid Y_i, \beta, A_i \overset{\text{ind}}{\sim} N(\theta_i^*, V_i(1 - B_i)) \tag{1.7}$$

with

$$\theta_i^* = (1 - B_i)Y_i + B_i z_i'\beta \tag{1.8}$$

and

$$B_i = V_i/(V_i + A_i). \tag{1.9}$$

Were the A_i and β known, the Bayes estimator θ_i^* would be the ideal estimator for any symmetric loss function. When the A_i and β are unknown, they may be estimable from the marginal distribution of Y_i, which is readily calculated for $i = 1, \ldots k$ as

$$Y_i \mid \beta, A_i \overset{\text{ind}}{\sim} N(z_i'\beta, V_i + A_i). \tag{1.10}$$

Estimates of the shrinking coefficient B_i can be derived most simply from (1.10) when all the A_i are known to be proportional to the V_i, say $A_i = \alpha V_i$, α unknown. This usually would happen only in a carefully designed experiment, for example, if $V_i = \sigma^2/n_i$ with n_i proportional to $1/A_i$. Changing notation, the V_i may be taken to be equal if we divide Y_i and z_i by $V_i^{1/2}$. Formally, with $A_i = \alpha V_i$, which will be called the "equal variance case," (1.10) may be changed to

$$Y_i \mid \beta, A \overset{\text{ind}}{\sim} N(z_i'\beta, V + A), \tag{1.11}$$

and (1.8) becomes

$$\theta_i^* = (1 - B)Y_i + Bz_i'\beta, \tag{1.12}$$

the shrinkage $B = V/(V + A)$ not depending on i.

In the equal variances case, b in (1.4) and

$$S \equiv \sum (Y_i - z_i'b)^2 \tag{1.13}$$

are independent, complete sufficient statistics for (β, A). These yield minimum variance unbiased estimates for β and A, with $Eb = \beta$ and, because $S \sim (V + A)\chi^2_{k-r}$

$$E\hat{B} = B \quad \text{if} \quad \hat{B} \equiv (k - r - 2)V/S. \tag{1.14}$$

Insertion of $(b, \hat{B})$ for (β, B) into (1.12) gives an estimate of the Bayes rule θ_i^*, or an empirical Bayes estimator

$$\hat{\theta}_i = (1 - \hat{B})Y_i + \hat{B}\, z_i'b. \tag{1.15}$$

The compromise estimator (1.15) is Stein's celebrated estimator (James and Stein 1961). Its risk, averaging over both distributions (1.1), (1.6), that is, its empirical Bayes mean squared error, is the same for every coordinate,

$$E(\hat{\theta}_i - \theta_i)^2 = E\, E_\theta(\hat{\theta}_i - \theta_i)^2$$

$$= V(1 - B) + ((r + 2)/k)VB \tag{1.16}$$

$$= V(1 - ((k - r - 2)/k\, E\hat{B}), \tag{1.17}$$

which is less than V provided $r \leq k - 3$, as assumed. (The subscript θ on E indicates that the random variable $\theta = (\theta_1, \ldots, \theta_k)$ is fixed when one computes this expectation.) Stein's theorem simply notes that the expectation given θ

$$(1/k)\, E_\theta \sum (\hat{\theta}_i - \theta_i)^2$$

$$= V(1 - ((k - r - 2)/k)\, E_\theta\hat{B}) \tag{1.18}$$

also is less than V, meaning that (1.15) also dominates the vector Y as an estimator of $\theta = (\theta_1, \theta_2, \ldots, \theta_k)'$ in the frequentist sense for sum of squared errors. Formula (1.18) follows from (1.17) and the fact that $(b, \hat{B})$ is a complete sufficient statistic for (β, B) (Efron and Morris 1973). Unlike the empirical Bayes case (1.16), the dominance property (1.18) does not hold for individual components $E(\hat{\theta}_i - \theta_i)^2$.

Stein showed that $\hat{B}$ is further improved if $\hat{B} \leq 1$ is forced whenever S is too small. This is clear in the empirical Bayes sense because we know $B \leq 1$. In fact, the empirical Bayes viewpoint argues for changing Stein's estimator further to

$$\hat{B} = ((k - r - 2)/(k - r))\, V/(V + \hat{A}^+) \tag{1.19}$$

with

$$\hat{A} = S/(k - r) - V \tag{1.20}$$

and $\hat{A}^+ = \max(0, \hat{A})$. Note that $\hat{A}$ is unbiased for A and is the maximum likelihood estimator for A based on S alone. Formula (1.19) reduces to $\hat{B}$ in (1.14) if $\hat{A} \geq 0$. The constant $(k - r - 2)/(k - r)$ helps correct for curvature dependence of B on A. That is, $B = V/(V + A)$ is a convex nonlinear function of A and so substitution of a nearly unbiased estimate $\hat{A}$ of A into B produces an estimate of B that is biased to be too large (by Jensen's inequality). The multiplying constant $(k - r - 2)/(k - r) < 1$ in (1.19) offsets the bias.

Stein's estimator should not be used in the unequal variances case. Then, other estimators are required, and

they turn out to be more difficult to evaluate mathematically. But almost all applications involve unequal variances, so that case cannot be ignored. We discuss it in Section 5.

2. SOME SIGNIFICANT APPLICATIONS OF EMPIRICAL BAYES AND RELATED METHODOLOGIES

Statisticians and other scientists have applied shrinkage techniques similar to those of the previous section to many important problems. Some well-documented successful applications are described in Table 2, and there have been many others. In each case the investigator was aware of some deficiency in classical methods, usually that the estimates were unstable, calling for shrinkage toward some ensemble estimate ($\tilde{\theta}_i$).

For example, the per capita incomes Y_i of small (less than 500 persons) geographical areas (there are 15,000 such areas in the United States) could not be estimated well from the 20 percent income sample collected in the 1970 census. Even so, accurate income estimates of θ_i, the true per capita income, are required each year for General Revenue Sharing Program purposes. Also available were two concomitant variables: housing values for owner-occupied homes and IRS income per deduction. These produced regression estimates $\tilde{\theta}_i$ of income θ_i via (1.4). The $\tilde{\theta}_i$ are much less variable than Y_i, but are biased. Fay and Herriot (1979) developed an empirical Bayes procedure for compromising between each small area sample estimate Y_i annd the regression estimate $\tilde{\theta}_i$. As Table 2 indicates, doing this required estimating B_i when all $A_i = A$, but with differing V_i. A similar procedure is described in Section 5.

The other investigators in Table 2 also dealt with important real problems representing a wide range of applications.

Hoadley (1981) showed that Bell Telephone quality assurance decisions can be made more accurately if they use empirical Bayes procedures.

Actuaries have used credibility methods (credibility $\equiv 1 - B_i$) to do insurance ratemaking since Whitney (1918).

Rubin (1981) showed that data on all law schools can be used to improve admission policies of each law school by providing remarkably better weights for estimating first-year grade point average from the LSAT (Law School Aptitude Test) scores and undergraduate averages.

Carter and Rolph (1974) showed that the New York City Fire Department could respond to alarms more appropriately if the false alarm rates at fire alarm boxes were estimated by empirical Bayes techniques using neighborhood alarm rates too.

Efron and Morris (1975) used empirical Bayes methods to estimate toxoplasmosis rates in El Salvador cities. The modified estimators there ranked the bad cases differently than did the maximum likelihood estimator.

These investigators had to deal with the unequal variances problem not covered by Stein's method. One, Hoadley, had to develop interval estimates to determine the probability of being out of control. In most cases, normal distribution requirements were not considered too restrictive, either because a nonlinear transformation to normality could be made, or because the empirical Bayes model in Section 1 concentrates on estimating the first two moments of the prior distributions, and does not use its form. But several investigators did address the Poisson problem directly.

The most persuasive arguments for the effectiveness of the procedures actually were made via cross-validatory methods. The shrinkage techniques predicted new data in each application more accurately than did the traditional methods, and the investigators showed that better

Table 2. Some Documented Applications of Empirical Bayes and Related Methodology

Application	*Author*	*i*	Y_i, θ_i	*Technical Features*
Revenue Sharing—Census Bureau	Fay and Herriot (1979)	Small areas	Per capita census income	a, b(REG: Housing value, IRS Income per exemption).
Quality Assurance—Bell Labs.	Hoadley (1981)	Time periods	# failures,	a, b(GM), c(Poisson), d (to determine probability system is out-of-control.)
Insurance Rate-making	Many, eg. Buhlmann (1970), Morris and VanSlyke (1978)	Group of insureds or territory	Insurer's risk per unit exposure	a, b(GM) Credibility $\equiv 1 - B_i$
Law School Admissions—Educ. Test. Serv.	Rubin (1981)	Law school	Weight for LSAT score relative to GPA.	a, b(GM)
Fire Alarms—New York City	Carter and Rolph (1974)	Alarm box locations	False alarm rate	a, b(GM), c(Poisson).
Epidemiology—El Salvador	Efron and Morris (1975)	City	Toxoplasomosis prevalence rate	a

Note: Key to "Technical Features": Letters indicate presence of the following concerns. a: Unequal variances ($V_i \neq V_j$). b: Shrinking to estimated points (GM = grand mean, REG = regression surface). c: Nonnormal sampling distribution (mentioned only if alternative distribution explicitly acknowledged). d: Interval estimates.

decisions would have been made had empirical Bayes procedures been used in the past. This is not surprising because Stone (1973,1974) and Geisser (1975), have shown that Stein's estimator is nearly the best sample predictive re-use estimator for the one-way layout situation with equal numbers of observations per group.

3. EMPIRICAL BAYES STATISTICAL INFERENCE

Empirical Bayes inference concerns two stochastic processes, one for the data and one for the parameters. The data Y are assumed distributed according to a parametric family of distributions $f(y \mid \theta)$. The parameters $\theta \in \Theta$ are assumed to have some distribution π belonging to a known class Π of distributions on Θ. Inferences are to be made about the realized value of θ.

The inference procedure will depend on knowledge of Π, but may not depend upon any (unknown) $\pi \in \Pi$, and the procedure's performance will be judged by its behavior for all $\pi \in \Pi$. The inference procedure $t(y)$ is evaluated with respect to both sources of variability, f and π. The procedure $t(y)$ may be a Bayes rule with respect to a second stage prior distribution on Π, but that is not required for it to be an empirical Bayes rule. The empirical Bayes emphasis is on consideration of a wider range of models, and not on the mode of inference within richer models. The Bayes-frequency controversy thereby remains unresolved—presumably some may argue for "Bayesian empirical Bayes," and others may not.

Thus we observe

$$Y \mid \theta \sim f(y \mid \theta) \quad \text{and} \quad \theta \sim \pi, \pi \in \Pi. \tag{3.1}$$

If there is a loss function $L(\theta, a)$ for taking action a when θ is true, then the empirical Bayes risk function is

$$r(\pi, t) = E_\pi E_\theta L(\theta, t(y)), \pi \in \Pi \tag{3.2}$$

for evaluating inference procedure t.

All Section 1 examples fall into this framework with Y the vector of k sample means, $f(y \mid \theta)$ the k-dimensional density of independent normal distributions, and Π the class of *independent* normal priors with mean $z_i\beta$ and variances A_i, or $A(=A_1 = A_2 = \cdots)$. Actually Π is a parametric family of priors with parameters β, A. Formula (1.16) provides an example of $r(\pi, t)$ when $t(y)$ is Stein's estimator, as a function of β, A.

The empirical Bayes model described in (3.1) contains both the frequentist and the Bayesian models as extreme cases. Bayesian inference, at least its textbook version, corresponds to cases in which Π has but one member. Then one should use the Bayes rule for that unique, known, $\pi \in \Pi$. Frequentist theory corresponds to letting Π contain all point priors π_θ, $\theta \in \Theta$. Then $r(\pi_\theta, t)$ simplifies to the ordinary risk as a function of θ. Empirical Bayes modeling is most interesting when Π is properly included between these two extremes.

Robbins (1951,1955,1964) recognized the promise of modeling these situations, first used the term "empirical Bayes," and is responsible for the early development of the subject, principally of the asymptotic theory as $k \to \infty$. Gini (1911), whose work also is discussed by Forcina (1982), earlier provided empirical Bayes solutions for estimating a binomial parameter. Most of Robbins's work considers all (Y_i, θ_i) to be independent identically distributed $i = 1, \ldots, k$ with an unrestricted prior for θ_1 (though he occasionally has considered parametric priors (e.g., Robbins 1980)). Then in a variety of situations, Robbins showed that decision procedures t_k can be found so that

$$\lim_{k\to\infty} r(\pi, t_k) = r(\pi, t_\pi), \pi \in \Pi \tag{3.3}$$

when t_π is the Bayes rule. Thus one can do as well asymptotically as the Bayes rule, knowing nothing of the marginal distribution of θ_i.

It sometimes helps to distinguish Robbins's cases from the parametric choices for Π and then we will call them "nonparametric empirical Bayes" (NPEB). Parametric empirical Bayes (PEB) is needed to deal effectively with those cases in which k is too small for Bayes's theory to approximate well. One then cannot do as well as the Bayes rule (e.g., (3.3)), but one hopes to make substantial improvements on standard methods, as Stein's estimator does, for example.

Nonparametric empirical Bayes is asymptotically similar to Bayesian theory, but apparently not too much, for D. V. Lindley notes in Copas (1969) that "there is no one less Bayesian than an empirical Bayesian." But that earlier criticism of NPEB may apply less to parametric empirical Bayes methods, which tend to be more similar to fully Bayesian methods. And although Bayesians still abhor integrating over the distribution of Y in applications, many recognize that the marginal distribution, or predictive distribution

$$g(y \mid \pi) \equiv \int_\Theta f(y \mid \theta)\, \pi(\theta) d\theta \tag{3.4}$$

of y can be used to assess the prior π (see, e.g., Box 1980). Many put "second stage" priors on Π (Good 1965, Lindley and Smith 1972) to acknowledge uncertainty in π. Berger (1982) uses the term "robust Bayes" to refer to this problem and investigation of $r(\pi, t)$ as π varies over a restricted family Π or all priors. Others have defined "gamma minimax" for $r(\pi, t)$ (see, e.g., Jackson et al. 1970).

The pair of models for Y and θ is also represented in frequentist methodology, via random effects and variance components models and in mixing distributions. The emphasis there, however, has been on estimating the distribution π, not the θ values (there are exceptions, e.g., Harville 1980).

The empirical Bayes approach to applications seems congenial to statisticians, because formulating a second model for the parameters requires the same activity as modeling the data, to which statisticians are accustomed, and the marginal density $g(y \mid \pi)$ can help guide this process. Of course the distributions permitted for π can

be much broader than indicated above. For example, the restrictive independence and exchangeability assumptions can be replaced by more complicated correlated or autoregressive structures for the parameters, provided k is large enough.

The applications of Section 2 involve moderate values of k, so the statisticians have used parametric empirical Bayes methods. Large values of k would allow use of nonparametric empirical Bayes theories, but it may be difficult to find so many problems, each with the same distribution for the parameters. While not asymptotically optimal against all priors, parametric empirical Bayes estimators still can be asymptotically optimal against the class of "conjugate" priors. Conjugate priors enjoy an important robustness property, as follows.

Theorem 1. Suppose Y has a univariate natural exponential family with quadratic variance function (Morris 1982), and the mean θ of Y is to be estimated with quadratic loss $(\hat{\theta} - \theta)^2$. In the class of all priors Π having mean μ and variance A, the conjugate prior π_0 is "empirical Bayes minimax." That is, if $t_\pi(Y) = E_\pi\theta \mid Y$ is the Bayes rule for $\pi \in \Pi$ and $r_0 \equiv r(\pi_0, t_{\pi_0})$ then for all $\pi \in \Pi$

$$r(\pi, t_\pi) < r(\pi, t_{\pi_0}) = r_0 \equiv r(\pi_0, t_{\pi_0}) < r(\pi_0, t_\pi). \quad (3.5)$$

This result is proved in Morris (1983).

The normal, Poisson, gamma, binomial, and negative binomial distributions are covered by Theorem 1, together with their normal, gamma, reciprocal gamma, beta, and F conjugate priors. Thus Theorem 1 suggests that statisticians who choose the conjugate prior will not incur larger risks than they expect. This too should hold approximately in empirical Bayes situations when (μ, A) are unknown, but are estimated with reasonable accuracy.

We now turn to the question of empirical Bayes confidence intervals, whose definition is indicated by the general theory.

Definition. An *empirical Bayes confidence region* with respect to Π, having confidence coefficient $1 - \alpha$, is a set $C(Y) \subset \Theta$ satisfying

$$P_\pi(\theta \in C(Y)) \geq 1 - \alpha \quad \text{all} \quad \pi \in \Pi. \quad (3.6)$$

Formula (3.6) requires computing the probability with respect to the joint distribution on (Y, θ). If Π is the class of point priors on θ, π_θ, then (3.6) is the frequentist confidence region.

Section 4 will present empirical Bayes confidence intervals for the equal variances normal case that are shorter than the standard intervals for every θ_i, $i = 1, 2, \ldots, k$. No such statement is possible for frequentist intervals; that is, if one does not average over $\theta_1, \ldots, \theta_k$, it is impossible to dominate the mean squared error for every component, as (1.16) shows Stein's rule does in the empirical Bayes sense. Thus, new results are possible by taking Π to be a restrictive class of priors. Of course, if Π does not include distributions that adequately represent the actual parameters, the operating characteristics claimed will be invalid.

4. EMPIRICAL BAYES CONFIDENCE INTERVALS—THE EQUAL VARIANCE CASE

That Stein's estimator has no accepted measure of its error or associated confidence interval surely must detract from its practicality. Of course as $k \to \infty$ with r fixed it is nearly a Bayes estimator and so the variance of each $\hat{\theta}_i$ is approximately $V(1 - B)$, the Bayes posterior variance, and the Bayes credibility interval becomes an approximate empirical Bayes confidence interval. These asymptotic results have been relied upon in most attempts to assign interval estimates, but their validity must be questioned for small k.

Two other sources of variation must be accounted for when k is small or moderate: there is an increase in posterior variance due to estimating β; and a term is needed to acknowledge the fact that A is estimated with error. In the equal variances situation with $V_1 = \cdots = V_k = V$ and $A_1 = \cdots A_k = A$ (1.1), (1.6) we propose the following procedure, a slight generalization of Morris (1977,1981). Use

$$s_i^2 = V(1 - ((k - r)/k)\,\hat{B}) + v(Y_i - z_i'b)^2 \quad (4.1)$$

for the variance of Stein's estimator (1.15). Here we take $\hat{B}$ according to (1.19) and let v represent the variance of B, defined by

$$v = (2/(k - r - 2))\,\hat{B}^2. \quad (4.2)$$

Then the empirical Bayes confidence intervals proposed for Stein's estimator, modified for PEB as in (1.15), (1.19), are

$$\hat{\theta}_i \pm zs_i \quad (4.3)$$

with z determined by the $N(0,1)$ distribution; for example, $z = 1.96$ if $1 - \alpha = .95$. That is, we claim that for all $A \geq 0$, β,

$$P(\hat{\theta}_i - zs_i \leq \theta_i \leq \hat{\theta}_i + zs_i) \geq 2\Phi(z) - 1, \quad (4.4)$$

with $\Phi(z)$ the $N(0,1)$ distribution function.

Actually, (4.1) and (4.2) are approximations to the posterior variances of θ_i and B given the data Y. Although Stein's rule is not a Bayes rule, a Bayesian alternative, with slightly different formulas, may be derived (Morris 1977,1981). However, such a rule is much harder to extend to the unequal variances case, as we shall need to do in Section 5 and for most applications.

The variances (4.1) satisfy

$$\hat{R} \equiv (1/k)\sum s_i^2/V \leq 1 - ((k - r - 2)/k)\,\hat{B} \quad (4.5)$$

and so the root mean squared interval width from (4.3) is less than the corresponding frequentist interval with $V^{1/2}$ for every data set Y. Thus the intervals (4.3) generally will be shorter. In (4.5), $\hat{R}$ also arises as an estimate of the risk $E_\theta\sum(\hat{\theta}_i - \theta_i)^2/kV$.

Does (4.3) have confidence coefficient at least $1 - \alpha$

for every $k \geq r + 3$ and all $\beta \in R^r$, $A \geq 0$? There is mathematical and numerical evidence (Morris 1981, and unpublished) that this is so for the usual levels of $1 - \alpha$. No formal proof has yet been given, but if the coverage probability of (4.3) ever is less than $1 - \alpha$, it could be only by a negligible amount. The only flaw with (4.3) is that these confidence intervals are too wide, not too narrow, when S (1.13) is small, and so the coverage probability substantially exceeds $1 - \alpha$ when A is small.

The intervals (4.3) are not frequentist confidence intervals because for fixed θ they have very low probability of covering θ_1 when k is large if, for example $\theta_2 = \theta_3 = \cdots = \theta_k$ and $\theta_1 = \theta_2 + \sqrt{kV}$. Of course such an arrangement is unlikely to occur if the θ_i are independent, identically distributed as $N(\mu, A)$.

The EB confidence intervals (4.1) through (4.4) have dependent coverage probabilities because all k intervals are based on the same estimates $(b, \hat{B})$. The Bayesian theory used to suggest these intervals also easily yields confidence ellipsoids for the entire θ-vector (see, e.g., Berger 1980). Although ellipsoids could be useful, we have ignored this here because the data analyst, in choosing between empirical Bayes and frequentist methods, mainly will want EB interval estimates for each individual θ_i to replace the familiar t-intervals.

One also might suspect some skewness in the EB intervals for extreme Y_i and wish to make refinements, as in van der Merwe et al. (1981) by further relying on the underlying Bayesian theory. This is not done here for simplicity in applications and also in theoretical work when showing (4.4) holds.

We illustrate in Figure 1 the use of empirical Bayes confidence intervals to predict the season batting averages θ_i of 18 baseball players on the basis of their batting averages Y_i after the first 45 at bats, that is, in the equal variance case (Efron and Morris 1975). The 45 degree line in Figure 1 is the classical estimate Y_i marked "CLASSICAL." The shrinking rule, a modification described in Morris (1977) of Stein's estimator, has $\hat{B} = .675$. That estimation line has a much gentler slope of 18 degrees, crossing the CLASSICAL line at $\bar{Y} = .266$, the average of all 18 players. In this case the true averages θ_i were obtained and are plotted in Figure 1. The empirical Bayes line EBE lies much nearer to the least squares regression line of θ_i and Y_i, because the θ_i are much less variable than the Y_i, and so it predicts much more accurately.

The CLASSICAL confidence bands, $Y_i + z\,V^{1/2}$, also are plotted for $z = 1.00$ (68 percent confidence) and for $z = 1.96$ (95 percent confidence). They cover the right fraction of points (17 of 18 in the 95 percent case) but are quite wide because of the steep slope. The gentler slope for EBE permits shorter confidence bands, which are curved to allow for greater uncertainty at the extremes. The 95 percent empirical Bayes intervals cover all 18 points while being 37 percent shorter.

The ensemble information in this problem is strong. Thus, $Y_1 = .400$ suggests $\theta_1 = .400$ in frequentist terms, because frequency methods do not account for the fact

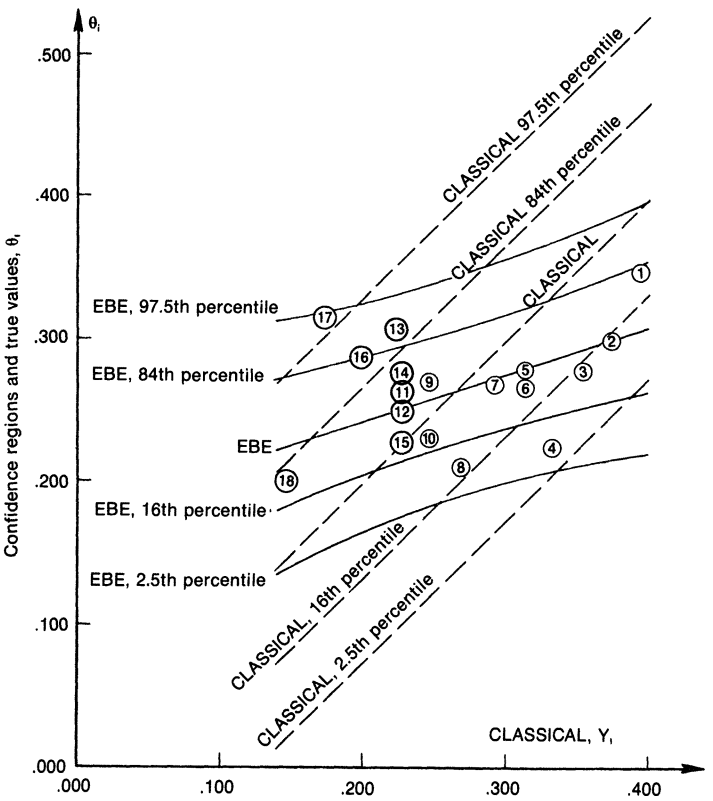

Figure 1. EIGHTEEN POINTS PLOTTED AT (Y_i, θ_i). CLASSICAL = Y_i, $Y_i \pm V^{1/2}$, and $Y_i \pm 1.96\,V^{1/2}$ (dashed lines). EBE = $\hat{\theta}_i$, EBE $\pm$ s_i, and EBE + $1.96s_i$ (solid curves)

that all 17 other players had $Y_i < .400$. The PEB interval for θ_1, which uses this information, puts .400 at the 98th percentile, an unlikely value.

On the other hand the true value $\theta_{17} = .318$ is contained in the 95 percent EB interval, which is [.152, .320], but not in the frequentist confidence interval [.046, .304] for player 17. Thus the EB intervals are smaller, but are not necessarily contained in the usual intervals. In this case .318 *was* the true average of player 17 so only the shorter EB interval correctly covers the true value. (Player 17 was Thurman Munson, who was to become a perennial all-star.)

We could say, when reporting these intervals, "Assuming the 18 players are a random sample from a population with normally distributed batting abilities, we are 95 percent EB confident that $.152 \leq \theta_{17} \leq .320$." Random sampling assures the exchangeability needed for the θ_i. The "EB" is included to warn the reader that this is not an ordinary confidence statement, although it might be left unmentioned because the "normally distributed batting abilities" phrase already indicates that the θ_i are random. Similar statements could be made for the other 17 players.

5. EMPIRICAL BAYES CONFIDENCE INTERVALS—THE UNEQUAL VARIANCE CASE

Now consider the more general case (1.1), (1.6) when the V_i may be unequal, and all $A_i = A$ are the same. The

cases $A_i = \alpha_i A$ with α_i known, seemingly more general, can be made equivalent to the case $A_i = A$ via transformation and relabeling.

Based on the model (1.7), (1.10), a reasonable estimate of θ_i or of the posterior mean θ_i^* is

$$\hat{\theta}_i = (1 - \hat{B}_i)Y_i + \hat{B}_i (z_i'\hat{\beta}). \quad (5.1)$$

Here, with $\hat{A}$ an approximately unbiased estimate of A, $\hat{\beta}$ is the weighted least squares estimate

$$\hat{\beta} \equiv (Z'DZ)^{-1}Z'DY \quad (5.2)$$

with $D = \text{Diag}(W_1, \ldots, W_k)$, the diagonal matrix of weights

$$W_i \equiv 1/(V_i + \hat{A}). \quad (5.3)$$

The variance A is estimated by

$$\hat{A} = \frac{\sum W_i\{(k/(k-r))(Y_i - z_i'\hat{\beta})^2 - V_i\}}{\sum W_i}, \quad (5.4)$$

but we force $\hat{A} \geq 0$ if (5.4) is negative. Actually $\hat{A}$, $\hat{\beta}$ must be determined by guessing at $\hat{A}$, and then computing (5.3), (5.2), and (5.4) to improve the guess. These steps may be iterated until $\hat{A}$ converges, usually taking about four to eight cycles. Any set of weights W_i in (5.4) will make $\hat{A}$ nearly biased. The estimate (5.4) of A is not the maximum likelihood estimate of A, but is related to ones used by Fay and Herriot (1979), Carter and Rolph (1974), and "RIDGM" in Dempster, Schatzoff, and Wermuth (1977). When $r = 0$, the maximum likelihood estimate of A (Efron and Morris 1973) is obtained by replacing W_i with W_i^2 in (5.4). When $r > 0$ it is better to use a restricted maximum likelihood estimator (REML) of A than the MLE. The REML is given approximately by (5.4) provided W_i is replaced with W_i^2.

Now let

$$\hat{B}_i = ((k - r - 2)/(k - r))\, V_i/(V_i + \hat{A}) \quad (5.5)$$

and

$$s_i^2 = V_i[1 - ((k - \hat{r}_i)/k)\,\hat{B}_i] + v_i(Y_i - z_i'\hat{\beta})^2. \quad (5.6)$$

In (5.6) we have

$$\hat{r}_i = kW_i[Z(Z'DZ)^{-1}Z']_{ii} \quad (5.7)$$

($\hat{r}_i$ estimates the effective value of r for component i),

$$v_i = (2/(k - r - 2))\,\hat{B}_i^2\,(\bar{V} + \hat{A})/(V_i + \hat{A}) \quad (5.8)$$

(v_i approximates the variance of $\hat{B}_i$), and

$$\bar{V} = \sum W_i V_i / \sum W_i \quad (5.9)$$

is the average variance.

The interval

$$\hat{\theta}_i \pm z\, s_i \quad (5.10)$$

is proposed as an approximate empirical Bayes confidence interval containing θ_i with probability $1 - \alpha = 2\Phi(z) - 1$. Evidence for this claim is incomplete, but is based on three observations.

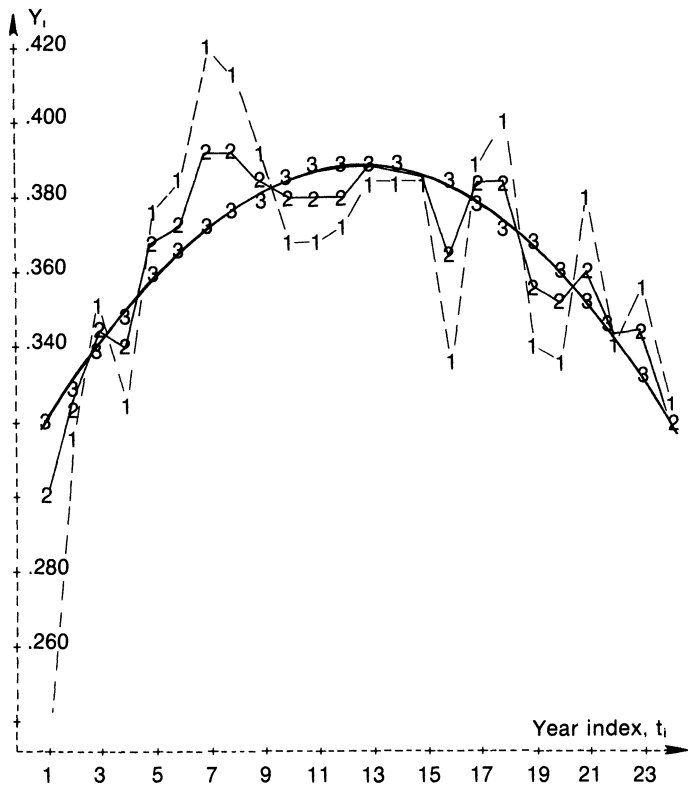

Figure 2. TY COBB'S BATTING AVERAGES, 1905–1928. 1 = observed average; 3 (Dark line) = regression estimator (r = 3 quadratic); 2 = empiricial Bayes estimate, shrinkages B_i range from 56 to 79 percent

1. The intervals (5.10) reduce in the equal variances case to those of Section 4, which have the claimed coverage probability.

2. The proposed rule has been derived by following Bayes's theory. As k increases, $\hat{A}$ coverages to A in probability, and so Bayes's theory guarantees the correct coverage probability.

3. The coverage probabilities have held up in various computer simulations.

The unequal variance estimates and intervals just proposed will be illustrated next by a second baseball example. We ask whether Ty Cobb was ever a "true" .400 hitter for one season. Cobb had the highest lifetime batting average of all time (.367) and in 1911, 1912, and 1922 he batted over .400. Only a handful of other players ever have exceeded .400 for an entire season, no one since 1941. If there ever was a true .400 hitter, it probably was Ty Cobb. (Hornsby is the only other serious candidate.)

Figure 2 shows Cobb's batting averages Y_i for the $k = 24$ years of his career. He batted .420 in his best year, 1911, but that is only 1.04 standard deviations above .399 for his 591 at bats. Thus we ask whether Cobb was lucky that year with $Y_7 > .400$ but $\theta_7 < .400$.

A quadratic curve fits Cobb's career performance pretty well and we take that as the prior mean

$$E\theta_i = z_i'\beta = \beta_0 + \beta_1 t_i + \beta_2 t_i^2 \quad (5.11)$$

with $r = 3$ parameters, t_i the year index, $z_i' = (1, t_i, t_i^2)$, and $\beta' = (\beta_0, \beta_1, \beta_2)$ estimated according to (5.2).

From binomial considerations we take $V_i = .367\,(.633)/n_i$, n_i = at bats. Then $\bar{V} = (.022)^2$ and $\hat{A} = (.015)^2$ from (5.9) and (5.4). Shrinking is typically 62 percent of the way towards the quadratic curve,, with $.56 \leq \hat{B}_i \leq .79$ always. The risk estimate is $\hat{R} = .51$, about half that of Y_i. The typical s_i is .016 with $.014 \leq s_i \leq .026$ ($n_1 = 150$ is quite small, giving $s_1 = .026$).

The jagged line in Figure 2 shows the empirical Bayes estimates. In every case $\hat{\theta}_i < .400$, with a high in 1911 of .392, so in no year was Cobb more likely than not to be a true .400 hitter.

We estimate the probability that $\theta_i \geq .400$ by

$$P(\theta_i \geq .400 \mid Y) = \Phi((\hat{\theta}_i - .400)/s_i) \qquad (5.12)$$

The largest of these numbers is 34 percent, in 1911. The probability that Cobb was a true .400 hitter in at least one unidentified year is about 88 percent, assuming the 24 probabilities in (5.12) can be combined by assuming independence. When this analysis was redone with $r = 5$ (a quartic polynomial for the prior mean fits better) the 88 percent dropped to an 84 percent probability.

6. CONCLUDING REMARKS

In some real problems the statistician will have a model $f(y \mid \theta)$ for the data but he may be undecided between two hypotheses concerning the possible values of the parameter θ, say between H_1, a high-dimensional space, and H_0, a low-dimensional subspace. His dilemma will be most acute if the parameter estimates under H_1 have large variances. Resolution may lie in specifying an empirical Bayes model, letting the data determine a compromise between, rather than choose between, the H_0 and H_1 estimates. If the true $\theta \in H_1$ is not too distant from H_0, then this approach will provide greater accuracy.

Should statisticians use empirical Bayes modeling in most multiparameter inference problems? Probably not, for there is a limiting factor. Choosing the class of prior distributions inappropriately may destroy any hoped-for advantages and produce distorted estimates. For example, if one includes with the 18 baseball players of Section 4 a 19th estimate of the success rate of a new product, or even the batting average of a pitcher, one probably will harm estimation of both groups of parameters. Thus, as always, improving estimates depends on proper use of valid new information. In this case, to be valid, the class of prior distributions Π must contain members that make the true θ likely.

The celebrated minimax result for Stein's estimator may seem to contradict the preceding statement because in Stein's case it sounds as if the sample mean can be improved upon without extra information. Not so. Stein's estimator is minimax in the equal variances case for the sum of squared errors loss function. It is not minimax for loss functions $\sum L_i(\hat{\theta}_i - \theta_i)^2$ that weight the squared errors quite differently; in fact, it is not minimax if any one L_i is even twice the average of all others. Thus, Stein's estimator provides advantages over the vector of sample means only if additional information is available. In the decision theoretic case that extra information is the correct loss function.

Most statisticians are more comfortable specifying probabilistic models for θ, as empirical Bayes modeling requires, than they are choosing among loss functions. Furthermore, the marginal density (3.4) is available to aid modeling, but choosing loss functions is pure guesswork. And minimax generalizations of Stein's estimator for the unequal variances case with equally weighted squared errors require absurd shrinking patterns: greater shrinkage for small than for large variances V_i. For practical work, then, the empirical Bayes viewpoint dominates the minimax approach.

Empirical Bayes modeling permits statisticians to incorporate additional information in problems by viewing the parameters as a second stochastic process having an unknown, but restricted, class of distributions. More work is needed to develop new applications, new computational software, methods for new situations, and to verify that the procedures have good properties with respect to the jointly distributed data and parameters. When this is done, statisticians will have powerful tools that can be used to good advantage on many multiparameter inference problems.

[*Received August 1982. Revised October 1982.*]

REFERENCES

BERGER, J. (1980), "A Robust Generalized Bayes Estimator and Confidence Region for a Multivariate Normal Mean," *Annals of Statistics*, 8, 716–761.

BERGER, J. (1983), "The Robust Bayesian Viewpoint," *Robustness in Bayesian Statistics*, ed. J. Kadane, Amsterdam: North Holland.

BOX, G.E.P. (1980), "Sampling and Bayes' Inference in Scientific Modelling and Robustness," *Journal of the Royal Statistical Society*, Ser. A, 143, 383–430.

BUHLMANN, H. (1970), *Mathematical Methods in Risk Theory*, New York: Springer-Verlag.

CARTER, G., and ROLPH, J. (1974), "Empirical Bayes Methods Applied to Estimating Fire Alarm Probabilities," *Journal of the American Statistical Association*, 69, 880–885.

COPAS, J. (1969), "Compound Decisions and Empirical Bayes," *Journal of the Royal Statistical Society*, Ser. B, 31, 397–423.

DEMPSTER, A.P., SCHATZOFF, M., and WERMUTH, N. (1977), "A Simulation Study of Alternatives to Ordinary Least Squares," *Journal of the American Statistical Association*, 72, 77–106.

EFRON, B. (1975), "Biased Versus Unbiased Estimation," in *Advances in Mathematics*, Vol. 16, No. 3, New York: Academy Press, 259–277.

EFRON, B., and MORRIS, C. (1973), "Stein's Estimation Rule and Its Competitors—An Empirical Bayes Approach," *Journal of the American Statistical Association*, 68, 117–130.

——— (1975), "Data Analysis Using Stein's Estimator and Its Generalizations," *Journal of the American Statistical Association*, 70, 311–319.

FAY, R.E. III, and HERRIOT, R.A. (1979), "Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data," *Journal of the American Statistical Association*, 74, 269–277.

FORCINA, A. (1982), "Gini's Contributions to the Theory of Inference," *International Statistical Review*, 50, 65–70.

GEISSER, S. (1975), "The Predictive Sample Reuse Method with Applications," *Journal of the American Statistical Association*, 70, 320–328.

GINI, C. (1911), "Considerazioni Sulla Probabilita a Posteriori e Ap-

plicazioni al Rapporto dei Sessi Nelle Nascite Umane," *Metron*, 15, 133–172.
GOOD, I.J. (1965), *The Estimation of Probabilities: An Essay on Modern Bayesian Methods*, Cambridge: Massachusetts Institute of Technology Press.
HARVILLE, D. (1980), "Predictions for National Football League Games via Linear-Model Methodology," *Journal of the American Statistical Association*, 75, 516–524.
HOADLEY, B. (1981), "Quality Management Plan (QMP)," *Bell System Technical Journal*, 60, 215–273.
JACKSON, D.A., DONOVAN, T.M., ZIMMER, W.J., and DEELEY, J.J. (1970), "Γ_2—Minimax Estimators in the Exponential Family," *Biometrika*, 57, 439–443.
JAMES, W., and STEIN, C. (1961), "Estimation with Quadratic Loss," *Proceedings of the Fourth Berkeley Sumposium on Mathematical Statistics and Probability* (Vol. 1), University of California Press, Berkeley, 361–379.
LINDLEY, D.V., and SMITH, A.F.M. (1972), "Bayes Estimates for the Linear Model," *Journal of the Royal Statistical Society*, Ser. B, 34, 1–41.
MORRIS, C. (1977), "Interval Estimation for Empirical Bayes Generalizations of Stein's Estimator," *Proceedings of the Twenty-Second Conference on the Design of Experiments in Army Research Development and Testing*, ARO Report 77-2.
——— (1981), "Parametric Empirical Bayes Confidence Intervals," to appear in the *Proceedings of the Conference on Scientific Inference, Data Analysis and Robustness*, University of Wisconsin.
——— (1982), "Natural Exponential Families With Quadratic Variance Functions," *The Annals of Statistics*, 10, 65–80.
——— (1983), "Natural Exponential Families With Quadratic Variance Functions: Statistical Theory," *The Annals of Statistics*, 11.
MORRIS, C., and VAN SLYKE, L. (1978), "Empirical Bayes Methods for Pricing Insurance Classes," *Proceedings of the Business and Economics Statistics Section, American Statistical Association*, 579–582.
ROBBINS, H. (1951), "Asymptotically Subminimax Solutions of Compound Statistical Decision Problems," *Proceedings of the Second Berkeley Symposium on Mathematics, Statistics and Probability*, University of California Press, 131–148.
——— (1955), "An Empirical Bayes Approach to Statistics," *Proceedings of the Third Berkeley Symposium on Mathematics, Statistics and Probability*, Vol. 1, 157–164.
——— (1964), "The Empirical Bayes Approach to Statistical Decision Problems," *The Annals of Mathematical Statistics*, 35, 1–20.
——— (1980), "An Empirical Bayes Estimation Problem," *Proceedings of the National Academy of Sciences USA*, 77, 12, 6988–6989.
RUBIN, D. (1981), "Using Empirical Bayes Techniques in the Law School Validity Studies," *Journal of the American Statistical Association*, 75, 801–827, with discussion.
STONE, M. (1973), "Discussion of the Paper by Professor Efron and Dr. Morris," *Journal of the Royal Statistical Society*, Ser. B, 35, 408–409.
——— (1974), "Cross-validatory Choice and Assessment of Statistical Predictions" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 36, 111–147.
VAN DER MERWE, A.J., GROENEWALD, P.C.N., NEL, D.G., and VAN DER MERWE, C.A. (1981), "Confidence Intervals for a Multivariate Normal Mean in the Case of Empirical Bayes Estimation Using Pearson Curves and Normal Approximations," Tech. Rept. No. 70, Department of Mathematics and Statistics, University of The Orange Free State, Bloemfontein, South Africa, November.
WHITNEY, A.W. (1918), "The Theory of Experience Rating," *Proceedings of the Casualty Actuarial Society*, Vol. IV, 274–292.

Comment

JAMES BERGER*

Morris has provided us with an excellent description of parametric empirical Bayes inference as it applies to estimation of normal means. The techniques he reviews (many developed by himself) seem eminently sound and useful for the types of problems he considers. My comments will address (a) applicability, (b) Bayesian aspects, (c) minimax aspects, and (d) robustness of the theory.

1. APPLICABILITY OF PARAMETRIC EMPIRICAL BAYES THEORY

For the small to moderately large dimensional problems most frequently encountered, I agree with Morris that it is necessary to postulate a model for the unknown means θ_i and estimate the parameters for this model, rather than try nonparametric estimation of the (assumed) common prior. There are rarely enough data for the nonparametric approach to prove successful, and a rather large degree of robustness with respect to the postulated model is present, as will be mentioned later.

* James Berger is Professor of Statistics, Department of Statistics, Purdue University, West Lafayette, IN 47907. Research was supported by the National Science Foundation, Grant MCS 81-01670A1.

It is also worth emphasizing Morris's comment that the parametric empirical Bayes approach is generally very worthwhile when one is trying to decide between a higher or lower dimensional model for the parameters of interest. A relatively common instance of this is when one is trying to decide whether to use a pooled estimate of variance or individual estimates. It can rarely be correct, when doubt exists, to use one extreme or the other; an empirical Bayes compromise is typically much better.

The development of empirical Bayes confidence sets is very important, not just as a necessary adjunct to estimation but also since tests can be done using the confidence sets. A related important aspect of the theory is its effect on "ranking problems." In the unequal variance case the empirical Bayes estimates will often be ordered differently from the original estimates Y_i, essentially because an "extreme" Y_i with large variance will be shrunk a substantial amount towards the estimated prior mean, perhaps leapfrogging a Y_i with a smaller variance. The resulting ordering of the population means is usually much more accurate and can allow substantially better selection of "extreme" θ_i.

It should be mentioned that the theory does not demand that simultaneous estimation of all θ_i be applicable. For