EMPIRICAL BAYES: A FREQUENCY-BAYES COMPROMISE

Carl N. Morris

University of Texas

Empirical Bayes research has expanded significantly since the ground-breaking paper (1956) of Herbert Robbins, and its province currently incorporates a range of methods in statistics. For example, Stein's famous estimator (James and Stein, 1961) is now best understood from the parametric empirical Bayes viewpoint. Appropriate generalizations and applications of Stein's rule in other settings (Efron and Morris, 1973, 1975; Morris, 1983b) are facilitated dramatically by the empirical Bayes viewpoint, relative to the frequentist perspective -- this will be indicated below.

Parametric empirical Bayes models differ from those considered in early empirical Bayes work, which focused on consistent estimation of Bayes rules for general prior distributions, allowing the number of parameters, k, to become asymptotically large. Rather, Stein's estimator and the generalizations developed by Efron and Morris take k fixed and possibly quite small, and ask for uniform improvement on standard estimators.

A series of examples are offered below to illustrate how empirical Bayes modeling is properly seen as a compromise between frequentist modeling and Bayesian modeling, and how the empirical Bayes model permits extension of various concepts, such as minimax properties and confidence regions, to more

*Research was supported by NSF Grant No. DMS-8407876, Department of Mathematics and Center for Statistical Sciences, University of Texas, Austin 78712.

AMS 1980 subject classifications: Primary 62C12; Secondary 62C10.

Key words: Parametric Empirical Bayes, Empirical Bayes Minimax, Empirical Bayes Confidence Intervals.

general settings.

We now consider a general model that includes the frequency, Bayes, and empirical Bayes viewpoints.

A General Model for Statistical Inference:

This model provides two families of distributions, one for observed data y, the other for unobserved parameters $\theta \in \Theta$, both y and θ possibly multivariate. The model may be specified in "descriptive" form or in "inferential" form.

I. <u>Descriptive form</u>:

(A) <u>Data</u>: Given $\theta \in \Theta$,

y has density $f(y|\theta)$, f fully known.

(B) <u>Parameters</u>:

θ has density $g_\alpha(\theta)$, $\alpha \in A$, A (possibly infinite dimensional) a known set of hyperparameters, g fully known.

Part (A) may be thought of as the likelihood function, (B) as the family of possible prior distributions. The descriptive form is usually considered when specifying a model. It is equivalent to the same model in inferential form.

II. <u>Inferential Form</u>:

(A') <u>Data</u>: Given $\alpha \in A$,

y has density $f^*_\alpha(y)$, f^* fully known.

(B') <u>Parameters</u>: Given y and $\alpha \in A$,

θ has density $g^*_\alpha(\theta|y)$, g^* fully known.

Part (B') is the possible family of posterior distributions for the parameters θ given the data, computed via Bayes theorem. Part (A'), the marginal distribution of the data, provides information on the likely values of $\alpha \in A$ via the marginal likelihood function $f_{\alpha}^{*}(y)$.

Evaluations within this model are made by integrating utility or loss functions with respect to both variables θ and y. However, when an ancillary statistic $T=t(y)$ is available, i.e., one having distribution independent of $\alpha \in A$ for the marginal distribution (A') for y, it is appropriate to calculate these integrals as θ and y vary, but with T fixed at its observed value.

This general model was proposed in (Morris, 1983b) as a framework for empirical Bayes analysis. Hill (1986) first recognized the importance in this context of requiring risk calculations to be conditional on ancillarity statistics, and has developed the model in a variety of ways.

The Bayesian framework, with known prior distribution, restricts A to have but one member. Thus the data in (A') are ancillary and only (B') is of interest. Evaluations then are conditional on all observed data, and so appropriate evaluations integrate over θ alone. Frequentists are unwilling to assume any knowledge about the prior distribution, and so A indexes all possible prior distributions on Θ, including those that assign point mass to any one θ. Thus the frequentist takes $A = \Theta$, and the posterior densities in (B') become trivial, ignoring the data. The frequentist then is only interested in (A') which is entirely equivalent to (A), in that context.

Empirical Bayes has considered a range of models intermediate between the frequentist and Bayesian models, with A having more than one element, but not all possible distributions. The empirical Bayesian, unlike the frequentist or the Bayesian, must deal with information in both (A') and (B'). Most familiar empirical Bayes models let $\theta = (\theta_1, \ldots, \theta_k)$ be a k-dimensional vector, $\Theta \subset R^k$ and let $y = (y_1, \ldots, y_k)$, y_i a one-dimensional sufficient statistic for θ_i, usually following an exponential family of distributions. The pairs (y_i, θ_i) are independent, $i = 1,2,\ldots,k$. Thus model (A) typically has taken the form

$$f(y|\theta) = \prod_1^k f_i(y_i|\theta_i).$$

Model (B) usually has provided independent identical (exchangeable) distributions

$$g_\alpha(\theta) = \prod_1^k p(\theta_i).$$

One example of the latter, e.g., Robbins (1956), chooses $A_1 = (p{:}p = \text{density on } R)$. Parametric examples might include all conjugate priors $A_2 = \{\alpha = (\alpha_1,\alpha_2){:}\ p = p_\alpha$ is a density on R known up two parameters $(\alpha_1, \alpha_2)\}$, with α_1 and α_2 the mean and variance of the conjugate prior distribution. These are "non-parametric" and "parametric" empirical Bayes assumptions on the prior distributions. Because $A_2 \subset A_1$, A_2 is more general, but both choices are very restrictive subsets of all possible distributions on $\Theta \subset R^k$ (the <u>same</u> p applies to each θ_i). When these assumptions are valid, they permit the substantial gains often provided by empirical Bayes methods relative to standard methods that do not use information from observations other than y_i when estimating θ_i.

Although parametric empirical Bayes methods are less general in this setting than nonparametric methods, they have the advantage of working well for k small (applications for k in the range 4-10 being plentiful). Parametric models are readily extendable to settings with non-exchangeable pairs (y_i, θ_i), as when θ_i follows a regression model, and to situations where the distribution of y_i differs from that of y_j because sample sizes vary. See (Morris, 1983b) for parametric examples, including references to applications.

Within the general model, various concepts can be defined such as unbiasedness, best unbiased, consistency, sufficiency, ancillarity, minimaxity, confidence sets, and so on. All properties are with respect the double integral over (y, θ) and must hold for all $\alpha \in A$. They reduce to the standard definitions of frequentist statistics when A contains <u>all</u> prior distributions. These definitions apply to empirical Bayes models in useful ways, several examples in the setting of independent normal distributions being considered below.

Suppose the descriptive model is, for $k \geq 3$,

$$y_i | \theta_i \sim N(\theta_i, V_i) \text{ independently, } i=1,\ldots, k \tag{3}$$

where V_i is assumed known and y_i represents the sample mean. Also let

$$\theta_i \sim N(0,\alpha) \qquad \text{independently, } i=1,\ldots, k \tag{4}$$

Thus $A = \{\alpha: \ \alpha \geq 0\}$, and (3), (4) form a parametric empirical Bayes model with α unknown.

The inferential model has (y_i, θ_i) independent,

$$y_i | \alpha \sim N(0, V_i + \alpha) \tag{5}$$

and

$$\theta_i | y_i, \alpha \sim N((1-B_i)y_i, V_i(1 - B_i)) \tag{6}$$

with $B_i \equiv V_i/(V_i + \alpha)$.

Stein's rule (James and Stein, 1961) for this situation is

$$\hat{\theta}_i = (1 - \hat{B})y_i, \qquad i = 1,\ldots,k \tag{7}$$

with $\hat{B} = (k - 2)/S$, $S \equiv \Sigma\, y_j^2/V_j$. It is known to dominate $(y_1,\ldots,y_k)$ as an estimate of $\theta = (\theta_1,\ldots,\theta_k)$ for every θ with respect to expected loss for the loss function $\Sigma\, (\hat{\theta}_i - \theta_i)^2/V_i$, and therefore is "minimax" in the usual (frequentist) sense. However, it is not minimax (risk less than ΣV_i) for the unweighted loss function $\Sigma(\hat{\theta}_i - \theta_i)^2$ and various alternatives have been offered, the simplest by Hudson (1974) and Berger (1976):

$$\hat{\theta}_i = (1 - \hat{B}_i)y_i, \ \hat{B}_i = \frac{(k-2)}{V_i S}, \ S = \Sigma\, y_j^2/V_j^{\,2}. \tag{8}$$

Note that the Hudson-Berger rule (8) reduces to Stein's (7) if the variances are equal, and that neither is minimax for the loss function justifying the opposite one if the variances differ substantially.

From the empirical Bayes standpoint, assuming exchangeable prior distributions (4), neither (7) nor (8) is satisfactory because shrinkage $\hat{B}_i$ should increase with V_i, not stay constant or decrease. In fact no rule can reasonably approximate Bayes rules ($\hat{B}_i$ near B_i as $k \to \infty$) and also be minimax for loss $\Sigma(\hat{\theta}_i - \theta_i)^2$. "Empirical Bayes minimax" rules do exist, however, where <u>empirical Bayes minimax</u> means (following definitions from the general model) that

$$E(\hat{\theta}_i - \theta_i)^2 \leq V_i \qquad \text{all } i=1,\ldots,k, \quad \text{all } \alpha \geq 0. \tag{9}$$

The expectation in (9) is with respect to variation in both y and θ. Thus empirical Bayes minimax requires minimaxity <u>for every component</u>. No weights need be specified for the loss function before adding components, because adding is not required. Thus, an empirical Bayes minimax rule retains its property independent of the weights w_i in the loss function $\Sigma w_i(\hat{\theta}_i - \theta_i)^2$.

A simple rule having the empirical Bayes minimax property is

$$\hat{\theta}_i = (1-\hat{B}_i)y_i = \frac{k-2}{k} \frac{V_i}{V_i^* + \hat{\alpha}} \qquad i = 1,\ldots,k \tag{10}$$

with

$$\hat{\alpha} \equiv \frac{1}{k} \ \Sigma(y_i^2 - V_i) \tag{11}$$

and
$$V_i^* = \max(\bar{V}, \ V_i + \frac{6}{k-1}(V_{max} - V_i)),$$

$$V_{max} \equiv \max(V_1,\ldots,V_k), \ \bar{V} = \Sigma\, V_i/k.$$

The choice $\hat{\alpha}$, although unbiased for α, is not the most efficient, and the regrettable increase from V_i to V_i^* is necessary in the denominator of (10) mainly to prevent the denominator from becoming negative. (Better rules could be offered, but the proof of empirical Bayes minimaxity, already tedious for (10), would be even harder.) Note that for large k, (10) behaves very well (near the Bayes rule) if $V_i = V_{max}$, and it behaves reasonably well if $V_i > \overline{V}$. But for components with $V_i < \overline{V}$, $\hat{B}_i$ in (10) is substantially too small. It is almost certain that these defects can be corrected without sacrificing empirical Bayes minimaxity. Of course (10) also reduces to Stein's rule when the variances are equal.

We have a dramatic example, using (7) and (10) showing how empirical Bayes minimax differes from frequentist minimax. In particular, for substantially unequal variances, V_{max} substantially larger than $\min(V_i)$, it can be proved that

(a) The Hudson-Berger rule (7) is minimax for unweighted loss, but is not empirical Bayes minimax.

(b) However, (7) is not minimax for other loss functions, e.g. $\Sigma(\hat{\theta}_i - \theta_i)^2/V_i$.

(c) The estimator (10) is empirical Bayes minimax, but not minimax for either of the loss functions discussed.

(d) The estimator (10) is "empirical Bayes consistent" (achieves the Bayes risk as $k \to \infty$ with respect to the model (3) - (4)) for components with $V_i > \overline{V}$, (but shrinks too little otherwise). (7) is inconsistent for all components.

Of course, rules that are empirical Bayes minimax and empirical Bayes consistent undoubtedly exist, and would be preferable to (10).

The concept of "empirical Bayes confidence intervals" also follows from the general statistical model. This requires that the probability (again, double integral over y and θ) of coverage exceed a pre-specified amount, say 0.95 for every $\alpha \in A$ (Morris, 1983a, b). In the setting of independent normal distributions (3) - (4), just considered, $C_i(y)$ is a 0.95 "empirical Bayes confidence interval" for θ_i if

(12) $$P_\alpha(\theta_i \in C_i(y)) \geq 0.95 \qquad \text{all } \alpha > 0$$

where (12) is computed with both y and θ random. For the equal variances case $V = V_i = V_j$ all i,j, sets of the form $C_i(y) = [\hat{\theta}_i - 1.96s_i, \hat{\theta}_i + 1.96s_i]$ have been shown to have property (12) with $\hat{\theta}_i$ close to Stein's rule and $s_i^2 = V(1 - \hat{B}) + v\, y_i^2$, v an estimate of the variance of $(\hat{B} - B)$, (Morris, 1983a).

Little attention has been paid to the interval estimation problem in the nonparametric empirical Bayes literature, although recently such ideas have been considered in the prediction setting (Robbins, 1977, 1983). As $k \to \infty$, of course, the entire posterior distribution can be estimated consistently, so the non-parametric approach could replace confidence intervals by posterior probability intervals, which would satisfy (12) asymptotically.

REFERENCES

Berger, J. (1976), "Minimax Estimation of a Multivariate Normal Mean Under Arbitrary Quadratic Loss," Journal of Multivariate Analysis, 6, 256-264.

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.

Efron, B. and Morris, C. (1975), "Data Analysis Using Stein's Estimator and Its Generalizations," Journal of the American Statistical Association, 70, 311-319.

Hill, J. (1986), "Empirical Bayes Statistics: A Comprehensive Theory for Data Analysis" Technical Report No. 39, University of Texas, Center for Statistical Sciences, Ph.D. Dissertation.

Hudson, H. M. (1974), "Empirical Bayes Estimation," Technical Report No. 58, Department of Statistics, Stanford University, Ph.D. Dissertation.

James, W. and Stein, C. (1961), "Estimation with Quadratic Loss," _Proceedings of the Fourth Berkeley Symposium on Mathemtical Statistics and Probability_, Vol. 1, University of California Press, Berkeley, pp. 361-379.

Morris, C. (1983a), "Parametric Empirical Bayes Confidence Intervals," in G.E.P. Box, T. Leonard, C. F. Wu (eds.), _Scientific Inference, Data Analysis, and Robustness_, Academic Press, 25-50.

Morris, C. (1983b), "Parametric Empirical Bayes Inference: Theory and Applications," _Journal of the American Statistical Association_, _78_, 47-55.

Robbins, H. (1956), "An Empirical Bayes Approach to Statistics," _Proceedings of Third Berkeley Symposium on Mathematical Statistics and Probability_, _1_, 157-163.

Robbins, H. (1983), "Some Thoughts on Empirical Bayes Estimation," _The Annals of Statistics_, _11_, 3, 713-723.