Statistical Science
2019, Vol. 34, No. 2, 206–208
https://doi.org/10.1214/19-STS705
Main article: https://doi.org/10.1214/18-STS674

Comment: Bayes, Oracle Bayes, and Empirical Bayes

Nan Laird

Efron has provided us with an interesting overview of several newer analytical developments for Empirical Bayes (EB) applications. He begins by telling us that empirical Bayes is new, but then immediately acknowledges that it is not so new. This paper makes several points that illustrate this dichotomy. First, there are new statistical methods to improve/sharpen inferences in the empirical Bayes setting. Second, both Bayesians and frequentists can benefit from using these approaches, and finally, the big-data era offers many new possibilities for their application. All of these points mean that we should see a lot more of empirical Bayes in practice. I agree with Efron than this should be true, although I do not feel as sanguine as Efron does. The ability to convert a complex data set into the simple EB framework as described by models 1 and 2 in Efron requires a lot of clever insight (for example, casting FDR as an empirical Bayes approach to multiple hypothesis testing) and we do not have good recipes for that part of the job. In addition, whether or not these techniques are widely accepted still suffers partly from the lack of a clear frequentist or Bayesian identity, partly on having reliable and readily available software, but also on our being able to convince potential users of their advantages, especially if the methods require complex computations and are not easy to explain. Efron's paper makes a lot of progress on all of these fronts.

Fred Mosteller introduced me to empirical Bayes ideas when I was a graduate student. Mosteller is not usually mentioned in the context of empirical Bayes, although his famous work with Wallace on determining the authorship of the disputed federalist papers had a decidedly empirical Bayes flavor (Mosteller and Wallace, 1964). Their method was widely characterized as Bayes, but they used data from papers of known authorship to estimate the "prior odds" for the two authors under consideration for the disputed papers. Their work is another good example of what I would characterize as "clever insight."

The first part of Efron's paper concerning Oracle Bayes has a decidedly frequentist bent and uses the ASE as an optimality criterion. I admit to being a frequentist, because it is generally the most practical, but I cannot get excited about the ASE (Average Squared Error). I can see it is possibly attractive in some settings, but with death rates, cure rates, hospital performance measures, or even gene expression, I find we are more interested in features of the ensemble, such as the extremes, thresholding, or in ordering the θ's. Thus my remarks will focus more on estimation of the mixing or prior density g, and on interval estimates for the θ's, such as those discussed in Efron (1996).

Estimating the prior, or mixing, distribution clearly arises in the EB setting, but also has broader application. Many of the real-life applications I have been involved with are more concerned with estimating g rather than the individual θ's (DerSimonian and Laird (1983)). For example, Mosteller and his collogues, Gilbert and McPeek (Gilbert, McPeek and Mosteller (1977)) were interested in how to quantitatively characterize progress in surgery and anesthesia. They sampled the surgical literature and obtained 13 randomized clinical trials (RCTs), each producing an estimate of the improvement in cure rate of an innovation over a standard therapy. Their objective was to use the estimates from these 13 RTCs to characterize the level of improvement.

As Efron notes, estimating g plays a central role in Bayes empirical Bayes inference, but using NonParametric Maximum Likelihood Estimation (NPMLE) for g is not attractive for this application because of its sparseness. I mention the progress in surgery example because it illustrates that sometimes sparseness is precisely what we want. It also illustrates that in real

Nan Laird is the Harvey V. Fineberg Research Professor of Biostatistics at Harvard School of Public Health, 677 Huntington Ave, Boston Massachusetts 02115, USA (e-mail: laird@biostat.hsph.edu).

problems, unequal variances are common. The RCTs all had different sample sizes and different underlying cure rates, and thus the observed measures of improvement all had different variances. In this setting $N = 13$; we let x_i denote the observed difference in cure rates, and let θ_i denote the "true" difference in cure rates for the ith RCT. We assumed each θ_i is drawn i.i.d. from $g(\theta)$. We made the simplifying assumption that each RCT had a sufficiently large sample size such that $x_i \approx N(\theta_i, v_i)$, where the variance of each x_i could be estimated from the data and assumed known. This assumption is at least credible, although one can argue about whether or not the difference in cure rates is the appropriate measure of progress. However, there was no obvious assumption to make about the form of g. In this context, f-modelling is not attractive, but it is straightforward to estimate g nonparametrically. F-modeling of the sort that Efron discusses is very attractive for parallel designed experiments such as microarrays, but there will always be settings, especially in descriptive work, where we are interested in inferences about g and/or where it is not feasible to estimate f.

The sparseness of the NPMLE of g should be no surprise (especially with $N = 13$). Even in the best of settings where $v_i = 0$ for all i, so that $x_i = \theta_i$, g will be a step function with mass $1/N$ at each x_i. When the v_i's are not zero a lot of information is lost and the number of support points can be far less than N. In our data set, the NPMLE had 3 support points, $(-0.0537, 0.041, 0.2096)$, with mass $(0.448, 0.496, 0.056)$. Most of the statisticians are dismayed because of the sparseness, but Mosteller and collogues were pretty happy, since it has such a straightforward and credible interpretation that is quite appropriate for the task at hand: Most of the time there is little difference between the innovation and the standard, but every now and then we get a winner.

This estimate is pretty crude, and a more careful analysis should consider transformations, such as θ_i normalized by the cure rate of the standard, or cure rate ratios, but simplicity and interpretation is important. I would also agree that if the primary interest is focused on this particular set of θ_i's, then a very discrete estimate of g is not attractive, although a bootstrap estimate of g could be (Laird and Louis (1987)).

The second part of Efron's paper deals with construction of confidence intervals in the EB setting. In the usual Bayes setting, a confidence interval can be constructed in a straightforward way for any θ given data x using the ordinary posterior, $g(\theta|x)$, assuming the forms and any incidental parameters of the sampling density, $p(x|\theta)$ and prior $g(\theta)$ are known. Frequentists can also typically construct a confidence interval for each theta based on $p(x|\theta)$, which has the advantage that it does not require knowledge of any prior for θ. A big disadvantage is its nonintuitive interpretation; nearly every "nonstatistician" intuitively gives the frequentist interval a Bayes interpretation. But, as Efron notes, to be Bayesian can require a lot more work to overcome lack of knowledge about $g(\theta)$ and can be complicated. Given the confusion that exists about confidence intervals in the simple $K = 1$ case, it is not surprising that there has been some difficulty over agreeing on what a confidence interval should be in the EB setting.

Efron characterizes the EB setting as one that offers Bayesians the opportunity to use the full sample $\mathbf{x} = (x_1, x_2, \ldots, x_K)$ to approximate the standard Bayesian approach of using hyperpriors for unknown features of $g(\theta)$. Personally, I hope that frequentists like it because I am again not so sanguine about Bayesians embracing frequentist approaches. Laird and Louis (1987) used the bootstrap to compute an approximate hyperprior; their approach was very general and computationally simple, but lacking in appeal because there was not a general theory to support their estimate of the hyperprior.

Another approach is to estimate the prior (or unknown parameters in the prior) to obtain $\widehat{g}(\theta)$, then use it as the prior to construct confidence intervals in the ordinary way, assuming $\widehat{g}(\theta)$ is the true prior (Morris (1983)). This is the naive EB approach as the intervals are not corrected for uncertainty in $\widehat{g}(\theta)$. Information in the data can be used to correct, or calibrate, these naive EB intervals. Exactly how this should be done has been the subject of much discussion.

An important piece of Efron's proposal for the finite Bayes approach is using g-modeling to obtain $\widehat{g}(\theta)$. The full sample $\mathbf{x}$ contributes to g-modeling, similar to f-modeling using the marginal density of $\mathbf{x}$. Rather than settle for parametric or strictly nonparametric forms for g, he uses splines and assesses the fit to the observed data $\mathbf{x}$. Having settled on a good representation for $g(\theta)$, $\widehat{g}(\theta)$, he uses the bootstrap to compute a 'corrected' prior, which is essentially $E(\widehat{g}(\theta))$ where $E(\cdot)$ is with respect to sampling distribution of $\widehat{g}$. Parametric Type II bootstrap samples (Laird and Louis (1987)) are used to give a sampling distribution for $\widehat{g}(\theta)$ to obtain a corrected prior, $\tilde{g}$. A confidence interval for any particular θ_0 given any x_0 is then

constructed from the usual posterior, where $\tilde{g}$ is used as the prior. Efron likens this to a full Bayesian flat hyperprior analysis. The approach is simple and intuitive, computationally straightforward even in complicated settings, and Efron provides some evidence that it does a reasonable job of approximating a full Bayesian.

All in all, Efron's work on empirical Bayes methods, both in this paper and many preceding ones, has been an important advance in a somewhat controversial and difficult field, and should be given careful consideration by both Bayesians and frequentists alike.

REFERENCES

DERSIMONIAN, R. and LAIRD, N. M. (1983). Evaluating the effect of coaching on SAT scores: A meta-analysis. *Harv. Educ. Rev.* **53** 1–16.

EFRON, B. (1996). Empirical Bayes methods for combining likelihoods. *J. Amer. Statist. Assoc.* **91** 538–565. MR1395725

GILBERT, J. P., MCPEEK, B. and MOSTELLER, F. (1977). Progress in surgery and anesthesia: An evaluation of innovative therapy. In *Costs, Benefits and Risks of Surgery* (B. A. Barnes, J. P. Bunker and F. Mosteller, eds.) Oxford Univ. Press, New York.

LAIRD, N. M. and LOUIS, T. A. (1987). Empirical Bayes confidence intervals based on bootstrap samples. *J. Amer. Statist. Assoc.* **82** 739–757. MR0909979

MORRIS, C. N. (1983). Parametric empirical Bayes inference: Theory and applications. *J. Amer. Statist. Assoc.* **78** 47–65. MR0696849

MOSTELLER, F. and WALLACE, D. L. (1964). *Inference and Disputed Authorship: The Federalist*. Addison-Wesley, Reading, MA. MR0175668