Statistical Science
2019, Vol. 34, No. 2, 234–235
https://doi.org/10.1214/19-STS674REJ
Main article: https://doi.org/10.1214/18-STS674

Rejoinder: Bayes, Oracle Bayes, and Empirical Bayes

Bradley Efron

This paper was originally a talk at the 2017 JSM, presenting a personal point of view on the current state of empirical Bayes inference. It is not surprising that the discussants, all of whom have written important papers on the subject, should have different points of view—from each other's and from mine. I don't have serious disagreements with any of these, but rejoinders have the nice property of being unchallengeable, at least in the short run, so I won't pass up my chance to get in a few free shots.

The paper and the commentaries touch on a range of related dichotomies, some of which have dogged discussions of empirical Bayes since its earliest days:

1. Hierarchical Bayes or frequentist empirical Bayes?
2. Omnibus loss functions or individual parameter inferences?
3. g-modeling or f-modeling?
4. Smooth parametic priors or the NPMLE?
5. Relevance considerations or inferences from the full data?
6. Random parameters θ or the compound decision model?
7. Finite sample performance or asymptotics?

Robbins' original work began at the apogee of statistical frequentism, with Bayesian thinking playing a decidedly minor role. Professors Greenshtein and Ritov's comments are fully frequentist (which attracts them to the compound decision framework) while Professor van der Vaart follows the hierarchical Bayes route. My paper tries to have it both ways. I never use hierarchical priors but, in Section 6, I employ Laird and Louis' Type III bootstrapping as a poor man's substitute. Van der Vaart is correct, the Dirichlet prior approach *is* pretty, but that doesn't mean it is right. Uninformative priors aren't guaranteed to produce accurate inferences—see Figure 13.7 of [3]—though here, in expert hands, it is probably fine. (Notice that in Professors Kroenker and Gu's careful calculations Dirichlet priors took more than an hour of computer time, compared to a few seconds for the bootstrap methods; van der Vaart is more optimistic about DP computation.)

Early empirical Bayes work focused on omnibus loss functions, for example, ASE in equation (5) of the paper, equation (1) in Professor Jiang's comments, and (1) in Greenshtein and Ritov. As emphasized in Section 2, omnibus loss favors the frequentist side of empirical Bayes; Section 6 redresses the balance in its "finite Bayes" calculations, where Bayesian ideas are dominant. Both van der Vaart and Greenshtein and Ritov consider empirical Bayes confidence intervals, more individual-parameter than omnibus in nature (though Jiang's confidence interval setup is omnibus). Sections 6 and 7, where individual confidence intervals are examined, were my own favorite parts of the paper.

The results are in for g-modeling vs f-modeling: none of the discussants had much good to say for f-modeling. As Professor Laird points out, f-modeling requires large numbers of perfectly parallel situations for its work, as well as a specialized problem set; g-modeling requires large numbers too, but not necessarily parallel ones. General g-modeling is discussed in [2], including an example incorporating covariate information (answering Greenshtein and Ritov's "plain vanilla" critique). Professor Louis' Section 2 nicely sums up the prosecution's case against f-modeling. And yet, many of the well-known empirical Bayes applications, from the butterflies and Robbins' formula and the baseball players up to false discovery rates, have depended on f-modeling, so it would be premature to banish it from the empirical Bayes toolkit. (I may be over-defensive on this point: my 2011 book used only f-modeling.)

Laird's 1978 paper [8] provided the key NPMLE theorem, while Koenker and Mizera's 2014 paper [7] translated the theory into a practical applied tool. Overall, the discussants' preferences tipped toward nonparametric maximum likelihood estimation (NPMLE) rather than the smooth parametric models in the paper. Koenker and Gu's commentary, a model for the

Bradley Efron is Professor, Department of Statistics, Stanford University, Sequoia Hall, 390 Serra Mall, Stanford, California 94305-4020, USA (e-mail: brad@stat.stanford.edu).

clear presentation of complicated results, shows the smooth approach doing fine when the underlying truth is smooth, and not so fine when it is not (though see the next paragraph on the two towers example). Jiang, whose 2009 paper with Zhang pioneered Oracle Bayes considerations, also demonstrates favorable NPMLE results for interval estimation. He used the new `deconvolveR` package for the smooth g-mod calculations ([4]).

The two towers example in the paper's Figure 1 was originally designed to show the limitations of low-dimensional smooth parametric g-modeling. It turned out to perform well, providing quite small EB regret. NPMLE also did well. Omnibus loss criteria such as ASE take the pressure off of g-modeling: having the estimated prior be too smooth or too rough makes little difference. Things are more delicate for individual parameter estimation, as in Figure 10, for example. Betting on smoothness is indeed a bet, but in this case a spiky three-point posterior density, looking like Koenker and Gu's Figure 2, probably wouldn't be of much use.

A conventional Bayesian prior distribution can be thought of as an infinite list of cases relevant to the problem at hand. Somehow this is easier to accept than the finite number of (perhaps) relevant cases available for an empirical Bayes analysis. I didn't succeed in worrying the discussants about the relevance questions in Section 7, but I did worry myself. The DTI example of Figures 13 and 14, and Table 8, show how critical the choice of relevance can be. (Thanks to Greenshtein and Ritov for the reminder of Fay and Herriot's paper.) The relevance rule of "all the cases that show up together on my desk" doesn't stand up to scrutiny, but formulating an alternative seems difficult. As a first try, the long paragraph beginning at equation (92) describes an algorithm for permitting an individual case to opt out of an empirical Bayes analysis if it looks sufficiently different from the others.

Nan Laird's hopeful prediction about big data and empirical Bayes has been instantly verified by Professors Wang, Miller, and Blei (WMB). I was aware of variational Bayes, thanks to Blei's earlier work, but hadn't thought about its connection to empirical Bayes. Here it appears as a 21st-century version of g-modeling. WMB's commentary is a stimulating small paper in itself, a jaunty combination of ideas from machine learning and statistics. The computer science side worries about applications to massive data sets, where computational feasibility is of more concern than statistical efficiency. The kind of nitty-gritty numerical comparisons in Jiang or Koenker and Gu would be on a statistician's wish list, but right now that might be missing the point. I hope the WMB team will continue developing this line of work, for the sake of both disciplines.

Among the discussants, only Tom Louis showed any interest in Poisson applications and the butterflies–what I've always thought of as the origin of empirical Bayes. (Greenshtein and Ritov even rule them out of empirical Bayes consideration!) Zipf's law always struck me as some sort of science joke, but no, here it performed like an overachiever. There's something wrong with Louis' NPMLE one-year trapping estimate of 25.5 new butterfly species; Good and Toulmin's nonparametric estimate was 45.2 while Fisher's gamma model gave 46.6. To answer Louis' final question: I'm more sympathetic to the Bayesian point of view than I was in 1986, and Bayesian methodology is a lot more practical than it was then. In general there seems to be less frequentist/Bayesian sniping these days. (It's a good thing not to be fighting a religious war while the AI atheists are at our gates.)

Somehow my list of seven questions didn't include the paper's main one: is empirical Bayes Bayesian or frequentist? "Both" was the paper's answer, which seemed okay with the discussants. My thanks to them and our editor, Cun-Hui Zhang, for putting together this discussion.

REFERENCES

[1] EFRON, B. (2011). Tweedie's formula and selection bias. *J. Amer. Statist. Assoc.* **106** 1602–1614. MR2896860

[2] EFRON, B. (2016). Empirical Bayes deconvolution estimates. *Biometrika* **103** 1–20. MR3465818

[3] EFRON, B. and HASTIE, T. (2016). *Computer Age Statistical Inference: Algorithms, Evidence, and Data Science. Institute of Mathematical Statistics (IMS) Monographs* **5**. Cambridge Univ. Press, New York. MR3523956

[4] EFRON, B. and NARASIMHAN, B. (2019). A g-modeling program for deconvolution and empirical Bayes estimation. *J. Stat. Softw.* To appear.

[5] GOOD, I. J. and TOULMIN, G. H. (1956). The number of new species, and the increase in population coverage, when a sample is increased. *Biometrika* **43** 45–63. MR0077039

[6] JIANG, W. and ZHANG, C.-H. (2009). General maximum likelihood empirical Bayes estimation of normal means. *Ann. Statist.* **37** 1647–1684. MR2533467

[7] KOENKER, R. and MIZERA, I. (2014). Convex optimization, shape constraints, compound decisions, and empirical Bayes rules. *J. Amer. Statist. Assoc.* **109** 674–685. MR3223742

[8] LAIRD, N. (1978). Nonparametric maximum likelihood estimation of a mixed distribution. *J. Amer. Statist. Assoc.* **73** 805–811. MR0521328