Empirical Bayes: Past, Present and Future

BY BRADLEY P. CARLIN AND THOMAS A. LOUIS

July 7, 1999

Abstract

Empirical Bayes (EB) methods have enjoyed widespread application in industry, law, medicine, and public policy settings. Their broad appeal appears to owe much to their ability to bridge both the frequentist and Bayesian philosophies: they are often justified via their links to classical James-Stein estimation, yet are implemented in a way reminiscent of objective (vague-prior) Bayesian approaches. In this vignette we provide a brief review of both parametric and nonparametric EB methodology, elucidate its (sometimes confusing) past, consider its present status, and speculate on its future as we enter the new millennium.

1 Introduction

Despite (or perhaps because of) the enormous literature on the subject, the term empirical Bayes (EB) is rather difficult to define precisely. To some statisticians, it refers to a class of models; to others, a style of analysis; to still others, a philosophy for screening statistical procedures. But perhaps any definition must begin with how EB differs from an ordinary, fully Bayesian analysis. To understand this distinction, suppose we have a distributional model \( f(y|\theta) \) for the observed data \( y = (y_1, \ldots, y_n) \) given a vector of unknown parameters \( \theta = (\theta_1, \ldots, \theta_k) \). While the classical, or frequentist, statistician would assume that \( \theta \) is an unknown but fixed parameter to be estimated from \( y \), the Bayesian statistician places a prior distribution \( \pi(\theta|\eta) \) on \( \theta \), where \( \eta \) is a vector of hyperparameters. With \( \eta \) known, the Bayesian uses Bayes’ rule to compute the posterior distribution:

\[
p(\theta|y, \eta) = \frac{f(y|\theta)\pi(\theta|\eta)}{\int f(y|u)\pi(u|\eta)du} = \frac{f(y|\theta)\pi(\theta|\eta)}{m(y|\eta)},
\]

where \( m(y|\eta) \) denotes the marginal distribution of \( y \).

---

1 Bradley P. Carlin and Thomas A. Louis are Professors, Division of Biostatistics, School of Public Health, University of Minnesota, Minneapolis, MN 55455. The first author’s work was supported in part by National Institute of Allergy and Infectious Diseases (NIAID) Grant R01-AI41966 and by National Institute of Environmental Health Sciences (NIEHS) Grant 1-R01-ES07756; the second author’s work was supported in part by NIEHS Grant 1-R01-ES07750 and NIAID Contract NO1-AI05073. The authors thank the editor and Prof. Melanie Wall for helpful comments which improved the manuscript.
If $\eta$ is unknown, information about it is captured by the marginal distribution. Moreover, if $f$ and $\pi$ form a *conjugate* pair of distributions (i.e., if $p(\theta|y, \eta)$ belongs to the same distributional family as $\pi$), then $m(y|\eta)$ will be available in closed form. An EB analysis uses this marginal distribution to estimate $\eta$ by $\tilde{\eta}(y)$ (for example, the marginal maximum likelihood estimator) and then uses $p(\theta|y, \tilde{\eta})$ as the posterior distribution. In contrast, a fully Bayesian analysis (sometimes called *Bayes empirical Bayes*, or BEB) augments (1) by a *hyperprior* distribution, $h(\eta|\lambda)$, and computes the posterior distribution as

$$p(\theta|y, \lambda) = \frac{\int \int f(y|\theta)\pi(\theta|\eta)h(\eta|\lambda) d\eta}{\int \int f(y|u)\pi(u|\eta)h(\eta|\lambda) du d\eta} = \int p(\theta|y, \eta)h(\eta|y, \lambda) d\eta.$$  \hspace{1cm} (2)

The second representation shows that the posterior is a mixture of posteriors (1) conditional on a fixed $\eta$, with mixing via the hyperprior, updated by the data $y$. Both EB and BEB use the observed data to provide information on $\eta$ and therefore “combine evidence.” This combining has proven very successful in improving statistical analyses in a broad array of fields, from the estimation of false fire alarm rates in New York City (combining across neighborhoods; Carter and Rolph, 1974) to modeling the decline in the CD4 counts in HIV-infected men (combining across men; DeGruttola et al., 1991).

The “empirical” in empirical Bayes arises from the fact that we are using the data to help determine the prior through the estimation of the hyperparameter $\eta$. Thus the basic distinction between EB and BEB is inclusion of the uppermost \(^2\) distribution $h$. The posterior computed via (2) incorporates uncertainties associated with not knowing $\eta$ and is generally to be preferred to basic EB. However, performance depends on choice of $h$, a difficult, important, and ongoing area of research.

The EB approach essentially replaces the integration in the rightmost part of (2) by a maximization (a substantial computational simplification), and bases inference on the *estimated posterior* distribution $p(\theta|y, \tilde{\eta})$. For example, any measure of the “middle” of this distribution (mean, median, mode) generally produces a suitable point estimate for $\theta$. Interval estimates arise similarly: for instance, the upper and lower $\alpha/2$-points of the marginal estimated posterior $p(\theta_1|y, \tilde{\eta})$ can be taken as a $100 \times (1 - \alpha)$% confidence interval for $\theta_1$. However, though EB point estimates generally perform quite well, *naive* EB interval estimates of this type are often too narrow, since they ignore the uncertainty in estimating $\eta$. Correcting EB intervals to have the desired (frequentist or Bayesian) coverage level has thus been the subject of significant research interest; see e.g., Laird and Louis (1987) and Carlin and Gelfand (1990, 1991).

The preceding development outlines what Morris (1983) refers to as *parametric EB* (PEB); see also Casella (1985) for a good introduction to the area. The approach assumes the distribution at the penultimate level of the hierarchy, $\pi(\theta|\eta)$, has a parametric form, so that choosing a (data-based) value for $\eta$ is

\(^2\)This uppermost distribution need not be precisely the third stage in the model; see the vignette on *hierarchical models* elsewhere in this issue.
all that is required to completely specify the estimated posterior distribution. Alternatively, one can adopt a nonparametric EB (NPEB) approach, where the distribution at the penultimate level is known to exist, but its form is specified only generically as \( \pi(\theta) \). Pioneered and championed by Robbins (1955, 1983), and further generalized and modernized by Maritz and Iwin (1989, Section 3.4), van Houwelingen (1977) and others, this method first represents the posterior mean in terms of the unknown prior, and then uses the data to estimate the Bayes rule directly. Recent advances substitute a nonparametric maximum likelihood (NPML) estimate of \( \pi(\theta) \) (see Laird, 1978).

EB has a long, colorful, and sometimes philosophically confused past, a vibrant present, and an uncertain future. We consider each of these in turn.

2 Past

Somewhat ironically, the history of EB is not particularly “Bayesian,” and certainly has little in common with the traditional, subjectivist Bayesian viewpoint. As noted above, the earliest attempts were of the nonparametric variety, and were essentially attempts by frequentist decision theorists to use Bayesian tools to produce decision rules having good frequentist (not Bayesian) properties. For example, working in the setting where \( y_i \mid \theta_i \overset{in.d.}{\sim} \text{Poisson}(\theta_i) \) and \( \theta_i \overset{i.i.d.}{\sim} G(\cdot) \), Robbins (1955) showed that the NPEB estimator

\[
\hat{\theta}_i^{\text{NPEB}} = E_G(\theta_i \mid y_i) = (y_i + 1) \frac{\#(ys \text{ equal to } y_i + 1)}{\#(ys \text{ equal to } y_i)} \tag{3}
\]

is asymptotically optimal, in that as \( k \to \infty \) its Bayes risk converges to the Bayes risk for the true Bayes rule where \( G \) is known. (Despite the name, “Bayes risk” is the expected value of a loss function, such as the average squared error loss \( ASEL(\hat{\theta}(y), \theta) = \frac{1}{k} \sum_{i=1}^{k} (\hat{\theta}(y) - \theta_i)^2 \), where the expectation is taken not conditionally given the data \( y \), but jointly over both \( \theta \) and \( y \).) Asymptotically optimal estimators can perform very poorly even for fairly large sample sizes, but (3) does demonstrate the characteristic EB “borrowing strength” from data values other than \( y_i \). Use of the NPML prior, \( G_{NPML} \), to compute the alternate estimator \( E_{G_{NPML}}(\theta_i \mid y_i) \) produces substantial benefits over the direct estimation of the Bayes rule in (3).

Parametric EB has close ties to Stein estimation, another primarily frequentist endeavor. In 1955, Stein showed that in the case where \( y_i \mid \theta_i \overset{in.d.}{\sim} N(\theta_i, \sigma^2) \), \( i = 1, \ldots, k \) and \( \sigma^2 \) is assumed known, the maximum likelihood estimator \( \hat{\theta}^{MLE}(y) = y \) is inadmissible as an estimator of \( \theta \). That is, under average squared error loss there must exist another estimator with frequentist risk no larger than \( \sigma^2 \) for every possible \( \theta \).
value. This dominating estimator was obtained by James and Stein (1961) as

$$\hat{\theta}_i^{JS}(y) = \left[ 1 - \frac{(k - 2)\sigma^2}{||y||^2} \right] y_i,$$

where $||y||^2 \equiv \sum y_i^2$. The connection to PEB was provided later in a celebrated series of papers by Efron and Morris (1971, 1972a,b, 1973a,b, 1975, 1977). Amongst many other things, these authors showed that $\hat{\theta}_i^{JS}$ is exactly the PEB point estimator obtained under the assumption that $\theta_i \sim N(0, \sigma^2)$ where the shrinkage factor $B \equiv \sigma^2/(\sigma^2 + \tau^2)$ is estimated by $\hat{B} = (k - 2)\sigma^2/||y||^2$.

The early EB authors’ consistent use of Bayesian tools to further frequentist goals while “using the data twice” (first to help determine the prior, then again in the usual Bayesian way when computing the posterior) was not highly regarded by the (primarily subjectivist) Bayesian community of the time (e.g., de Finetti, Lindley, and Savage). In his discussion of Copas (1969), Lindley noted that “there is no one less Bayesian than an empirical Bayesian;” later in a discussion of Morris (1983) he describes much of the asymptotics supporting NPEB as “technicalities out of control.” In his discussion of the same paper, Dempster borrows a metaphor of Savage (1961, p.578) in describing an empirical Bayesian as someone who “breaks the Bayesian egg but then declines to enjoy the Bayesian omelette.” Still, the early EB work of the 1950s and 60s helped further the rise of objective Bayesian thinking decades before the Gibbs sampler made such thinking routinely possible and acceptable. Further work by Morris (1983) and Hill (1990) encouraged a more unifying role for EB, in which evaluations of procedures could be made by averaging over both the data space $\mathcal{Y}$ and the parameter space $\Theta$, with the EB position emerging as a plausible compromise between the strict Bayes and frequentist positions.

3 Present

The cumulative impact of EB on statistical applications continues to be enormous. Statisticians and users of statistics, many of whom were trained to distrust Bayesian methods as overly subjective and theoretically mysterious, can nonetheless often appreciate the value of borrowing strength from similar but independent experiments. To use another metaphor, EB thus offers a way to dangle one’s legs in the Bayesian water without having to jump completely into the pool. An electronic search of the latest Current Index to Statistics on “empirical Bayes” confirms its increasing popularity, yielding a median of 2.5 hits per year during 1964–69, 11 hits per year during 1970–79, 32 hits per year during 1980–89, and 46 hits per year during 1990–1996. EB methods have for example enjoyed broad application in the analysis of longitudinal, survival, and spatially correlated data (Laird and Ware, 1982; Clayton and Kaldor, 1987).
The Bayes/EB formalism also ideally structures combining information from several published studies of the same research area, a scientific discipline commonly referred to as meta-analysis (Cooper and Hedges, 1994), though in this context primary interest is in the hyperparameters (\( \eta \)) rather than the parameters from individual studies (\( \theta \)). More generally, the hierarchical structure allows for honest assessment of heterogeneity both within and between groups such as clinical centers or census areas. The review article by Breslow (1990) and the accompanying discussion contain an excellent summary of past and potential future application areas for Bayes and empirical Bayes methods in the public health and biomedical sciences.

Important methodologic contributions also continue. For example, Efron (1996) develops an EB approach to combining likelihoods for similar but independent parameters \( \theta_i \). In keeping with the EB philosophy, little in the way of Bayesian prior information is required, and frequentist ideas (such as bias correction and robustness) are critical to the method. Other important recent EB work includes that on spatial statistics by Raghunathan (1993) and nonparametric growth curves by Altman and Casella (1995).

4 Future

Returning to the theme of our opening sentence, one's view of the future of EB is indelibly tied to one's view of its past and present, as well as one's own "upbringing." One of us (TAL) received his PhD from Columbia University and had Herb Robbins as one of his advisors, and so was exposed to an NPEB framework in which achieving good frequentist properties was paramount, but whose view then grew to include parametric approaches and a healthy dose of purely Bayesian thinking. This leads to an optimistic view of EB's future, in which more and more applied problems are tackled using EB methods which preserve a strong measure of frequentist validity for the resulting inferences. The other one of us (BPC) came along nearly two decades later, after the appearance of the marvelous PEB discussion paper by Morris (1983) but just prior to the emergence of the Gibbs sampler and other Markov chain Monte Carlo (MCMC) tools for implementing fully Bayesian analysis. This author started from a PEB shrinkage point of view, but one in which EB methods were thought of as approximations to overly-cumbersome fully Bayesian hierarchical analyses. With the widespread availability of MCMC tools such as the BUGS software (Spiegelhalter et al., 1995), this produces a much more pessimistic outlook for EB, since the need for such approximations (and the corresponding restrictions on which models can be handled) has more or less vanished.

Still, fully Bayesian solutions implemented via MCMC are by no means "plug and play." Convergence of the algorithms is notoriously difficult to diagnose, with most of the usual diagnostics having well-known flaws (Cowles and Carlin, 1996; Mengersen et al., 1999). Worse, the sheer power of MCMC methods has
led to the temptation to fit models larger than the data can readily support without a strongly informative prior structure – now something of a rarity in applied Bayesian work. In cases like this, an EB approach may well be superior to a full hierarchy with improper priors, since the computation will be better-behaved and the associated underlying theory may be better understood.

As an example of this point, consider the problem of specifying a vague hyperprior for a variance component \( \tau^2 \). The most widespread current choice seems to be the \( \text{Gamma}(\epsilon, \epsilon) \) prior (i.e., having mean 1 but variance \( 1/\epsilon \)); see e.g., Spiegelhalter et al. (1995). But recent work by Hodges and Sargent (1998) and Natarajan and Kass (1999) shows that such hyperpriors, while appearing “noninformative,” can actually have significant impact on the resulting posterior distributions. And while this hyperprior is proper, it is “nearly improper” for suitably small \( \epsilon \), potentially leading to the aforementioned MCMC convergence failure – or worse, the appearance of MCMC convergence when in fact the joint posterior is also improper. Thus it might well be that reverting to an EB approach here (replacing \( \tau^2 \) by \( \tilde{\tau}^2 \) ) will produce a estimated posterior that, while not fully defensible from a purely Bayesian point of view, produces improved estimates while at the same time is safer to use and easier to obtain.

Though EB and BEB approaches have proven very effective, they are by no means a panacea and require continued research and development. We urge progress in developing hierarchical analyses that are efficient and effective, but also robust with respect to prior and other specifications. If such robustness is not present, the sources of the sensitivity must be investigated, documented, and questioned. Broadening and deepening our understanding of the influence of the hyperprior \( h \) is an important aspect of these developments. Finally, we propose that one’s philosophical bent need not play a significant role in the decision to use EB or vague hyperprior BEB, since these generally produce procedures with good frequentist, Bayes, and EB properties; see e.g., Carlin and Louis (1996, Chapter 4).

References


6


