Statistical Science
2019, Vol. 34, No. 2, 202–205
https://doi.org/10.1214/19-STS704
Main article: https://doi.org/10.1214/18-STS674

Comment: Bayes, Oracle Bayes, and Empirical Bayes

Thomas A. Louis

1. INTRODUCTION

Brad Efron has done it again. He presents fascinating and insightful analyses that "open the box" on the properties of empirical Bayes methods. I especially like the exploratory data analysis theme, reminding us to look at the data, consider what information sources are relevant, and to conduct sensitivity analyses. These highlight the importance of computing diagnostics, and the dangers of black box modeling.

In what follows, I evaluate f-modeling (generate posterior summaries from the estimated marginal distribution of the data) and g-modeling (estimate the prior distribution and use Bayes' rule to obtain the posterior), consider Oracle Bayes, and address the choice between Bayes and empirical Bayes.

2. f- AND g-MODELING

Building on Efron (2014), Brad further compares f- and g-modeling as strategic approaches. While f-modeling is somewhat easier to implement, and the Robbins result for the Poisson (Robbins, 1983) is truly neat and showed how "empirical" can be wedded to "Bayes," g-estimation wins the day. Producing an effective g-model has its challenges, but the hard work pays off in that the (estimated) posterior distribution and generated summaries respect all constraints induced by prior to posterior mapping. There may be some models and goals for which f-modeling is competitive to g, but the situations are few and likely null when data aren't marginally i.i.d., in multivariate models, for goals such as histogram estimation and ranking (see below), or benchmarking (Bell, Datta and Ghosh, 2013). However, producing a good estimate of the X-marginal distribution is still very important; for example, it is central to assessing model fit (see Box, 1980).

2.1 The Basic Poisson Model

In Section 5, Efron presents the Robbins (1983) f-modeling approach to estimating the posterior mean, $e_g(x)$, and variance, $v_g(x)$, of the Poisson rate parameter. That $v_g(x)$ is nonnegative implies that $e_g(x)$ is nondecreasing, and a nondecreasing $e_g(x)$ requires that,

$$\frac{f(x+2)f(x)}{f^2(x+1)} \geq \frac{x+1}{x+2}.$$

Directly estimating f does not ensure satisfaction of this or other conditions imposed by the representation, $e_g(x) = \frac{\int \theta^{x+1} g(\theta)\, d\theta}{\int \theta^x g(\theta)\, d\theta}$. For example, nonnegativity of the posterior fourth central moment also imposes restrictions on f. These and other restrictions are automatically satisfied in g-modeling, but require considerable machinations to be satisfied in f-modeling.

2.2 Corbet's Butterfly Data

Efron analyzes Corbet's butterfly data, comparing versions of f- and g-modeling for the Poisson rate parameter (θ) and its logarithm (λ). For comparison, I base g-modeling on the nonparametric, maximum likelihood estimate (NPML), implemented by the EM algorithm (Laird, 1982), starting the recursion with a sequence of 24 equiprobable mass points in the interval [0.1 to 36.0]. The recursion quickly converged the the three-point distribution in Table 1. It induces an X-marginal which is graphically close to the natural spline Poisson regression fit in Figure 4 of the article, but it gives less weight to small θ-values.

The g-NPML prior generates the posterior mean plots in Figure 1. In the left panel, the g-NPML line mimics the Robbins values displayed in Efron's Figure 5, but is monotone and respects other conditions imposed by the g-modeling approach. Zipf's/g-glm are plotted as a single line, even though in Efron's Figure 5 g-glm is slightly below Zipf's for large values of X.

Table 2 gives the (estimated) Bayes risk for the Robbins, g-NPML and g-glm priors with g-NPML less optimistic than Robbins, but more optimistic than g-glm.

Thomas A. Louis is Professor Emeritus, Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, 615 North Wolfe Street, Baltimore, Maryland 21205, USA (e-mail: tlouis@jhu.edu).

TABLE 1
The NPML prior estimated from Corbet's butterfly data. (*Masses sum to* 1.005 *due to rounding*)

mass	0.552	0.256	0.197
mass point	2.175	8.200	17.204
log(mass point)	0.777	2.104	2.845

TABLE 2
Estimated Bayes risk: *"Robbins" is via f-modeling*, *"glm" and "NPML" via g-modeling*

	Bayes risk (R_g)
Robbins	$4.27 = 6.60 - 2.33$
g-NPML	$5.12 = 6.60 - 1.48$
g-glm	$6.55 = 6.60 - 0.05$

The g-glm prior produces an estimated 47.6 number of new species after one additional year of data collection, whereas g-NPML produces 23.5. Similar discrepancies occur for additional years of follow-up. The discrepancy is likely due to g-glm relative to g-NPML giving additional weight to small θ values.

In summary, g-NPML produces an X-marginal that is similar to that for g-glm, with departures that generate substantially different estimated Bayes risk and a substantially different predicted number of new species. These remind that compatibility for some features of a model doesn't imply compatibility for all features.

I am not advocating use of g-NPML, but it sets the stage for combining a smooth approach (e.g., g-glm, or a mixtures of Gamma distributions) with g-NPML. Combining can be based on the full Bayesian formalism (e.g., a Dirichlet process prior), taking a weighted average of the two prior estimates with, for a fixed N_0, relative weight N/N_0 the g-NPML, or using the EM with data augmented by values at percentiles of the smooth distribution with weights that sum to N_0.

3. EMPIRICAL ORACLES

It is no surprise that knowing the empirical distribution function (EDF) of the θs (an Oracle) provides considerable information. And, it is pleasing that EB can provide a notable portion of the Oracle advantage. An "empirical Oracle" provides a middle ground between knowing the oracle and standard EB. It proceeds by estimating the EDF of the θs that generated the current data set using the Shen and Louis (1998) approach. Specifically,

$$\theta_1, \dots, \theta_N \text{ i.i.d. } G$$

$$\begin{aligned} G_N(t|\boldsymbol{\theta}) &= \frac{1}{N}\sum I_{\{\theta_i \le t\}} \quad \text{(the Oracle)} \\ [X_i|\theta_i] &\sim f(x_i|\theta_i). \end{aligned}$$

The optimal squared error loss (SEL) estimate is the posterior mean,

$$\begin{aligned} \bar{G}_N(t|\mathbf{X}) &= E_G[G_N(t;\boldsymbol{\theta})|\mathbf{X}] \\ &= \frac{1}{K}\sum P(\theta_k \le t|\mathbf{X}). \end{aligned} \tag{3.1}$$

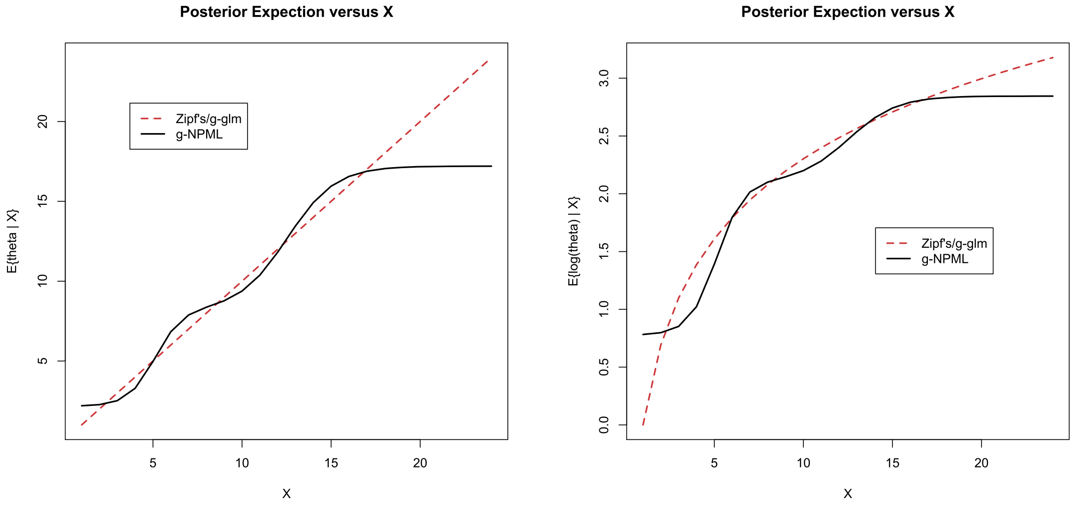

FIG. 1. *Posterior means for $E(\theta \mid X)$ (left panel) and $E(\lambda \mid X)$ (right panel). Zipf's/g-glm are plotted as a single line even though in Efron's Figure 5 g-glm is slightly below Zipf's for large values of X.*

The SEL-optimal discrete function with N mass points each with mass $1/N$ is,

$$\hat{G}_N(t \mid \mathbf{X}) : \text{mass 1/N at } \hat{U}_j = \bar{G}_N^{-1}\left(\frac{2j-1}{2N} \,\middle|\, \mathbf{X}\right),$$

an "empirical Oracle." It also produces an estimated histogram.

The empirical Oracle, $\bar{G}_N$ in equation (3.1), depends on the assumed G, but is also influenced by the observed data. An EB approach replaces the assumed prior (G) by an estimate (see Paddock et al., 2006), providing more tuning to the data. If the estimate consistently estimates G, then $\hat{G}_N$ will consistently estimate G_N which will converge to G, closing the circle.

With $\bar{G}_N^{(0)}$ an initial value for the prior, and $\bar{G}_N^{(\nu+1)}$ the result of applying equation (3.1) with $G_N^{(\nu)}$ as the prior, "roughens" the initial G toward the data (see Laird and Louis, 1991, Shen and Louis, 1999). As $\nu \to \infty$, $G_N^{(\nu)}$ converges (very slowly!) to the nonparametric, maximum likelihood (NPML) estimate of the prior, though using this approach to get the NPML most assuredly not recommended.

The SEL loss function can be replaced by posterior percentiles of $G_N(t)$ (see Paddock and Louis, 2011), which supports construction of credible intervals.

4. FREQUENTIST, BAYES, EB, OR BEB?

Making inferences solely based on information from a single study/dataset ($N = 1$) without incorporating external evidence or professional judgment is *prima facie* frequentist. However, as Brownstein et al. (2019) propose, this pure form is difficult if not impossible to achieve. All other inferential activities entail some degree of formal or informal Bayesian evaluation. If the number of relevant data sources (N) is large and they provide a large amount of information on the prior distribution, plug-in EB performs well. Many genomics examples live in this domain. For $N = 1$, if there is to be Bayes, it needs to be high-church Bayes, driven by personal/expert judgment (see O'Hagan, 2019). If N is moderate, then accommodating uncertainty in the inferred prior and posterior distributions via an expansion (Morris, 1983), the bootstrap, or hyperprior Bayes (BayesEB) is needed. Evaluating whether plug-in is sufficient usually requires comparing it an approach that captures inferred prior uncertainty, and if the latter has been implemented, then why not use it?

TABLE 3

Comparison of simulated unconditional, nominal 95% EB CI length and coverage for the exponential/inverse gamma model for $N = 5$. See Section 5.4.3 and Table 5.4 in Carlin and Louis (2009) for full details

	Method				
Feature ↓	Classical	Naive EB	Laird/Louis	h_1	h_2
Length	38.80	5.22	7.50	4.51	5.66
Coverage	0.952	0.900	0.954	0.930	0.951

4.1 Example: Exponential/Inverse-Gamma

Carlin and Louis (2009) evaluated pre-posterior CI length and coverage for the exponential/inverse gamma model with θ the rate/hazard,

$$\begin{aligned} \theta_1, \dots, \theta_N &\overset{\text{i.i.d.}}{\sim} \text{InvGamma}(\eta, 1), \\ f(y_i \mid \theta_i) &= \frac{1}{\theta_i} e^{-y_i/\theta_i}, \quad y_i > 0, \\ \hat{\theta}_i^{\text{mle}} &= y_i, \end{aligned}$$

which produces the marginal distribution and marginal mle,

$$m(y_i \mid \eta) = \eta/(y_i + 1)^{\eta+1},$$

$$\hat{\eta}^{\text{mmle}} = N \Bigg/ \sum_{i=1}^{N} \log(y_i + 1).$$

They compared the frequentist CI to naive EB (plug-in $\hat{\eta}^{\text{mmle}}$), Laird and Louis (1987) bootstrap, Carlin and Gelfand (1991) hyperprior matching using $h_1(\eta) = 1$ and $h_2(\eta) = 1/\eta$, when $N = 5$ and the true $\eta = 2$.

Table 3 gives a snippet of their results. The classical interval is well calibrated, but extremely long relative to all other methods. The naive EB interval is shortest, but at the expense of under-coverage. The Laird/Louis bootstrap and the h_2-based intervals are well calibrated, with the latter being the shorter, but the former not depending on specifying h. That the h_1-based interval under-covers shows that for small N situations, choice of hyperprior matters.

5. CLOSING

Efron's evaluations and commentary should energize similar analyses of the potentials and pitfalls of EB analysis for more complex models and goals. Some of the assessments can build on his reported mathematical analyses, others will require well-designed simulations. Assessments will be challenging, especially if activating the model requires MCMC. Careful design will be needed in a complex, multilevel, multifactor model to track information flow, and to assess which outputs are highly sensitive and which are robust to model specification (hyperprior, prior, and the data likelihood), and to model activation.

Efron's examples show that while there may be some risk in employing EB, there can be substantial rewards. EB controls the risk by reducing, but not eliminating, the need for personal opinion/judgment. Other than for the Bayesian purist, EB is an attractive approach, and the purist can find comfort in that BayesEB is "Bayes" with marginalization over the hyperparameters producing a prior for the θs that augments linkages amongst them.

Finally, in Efron (1986), Brad and discussants consider the benefits and potential drawbacks of the Bayesian approach to inference. There have been impressive advances in Bayesian and empirical Bayesian methods in the last 30+ years, and it would be great to know his updated views, possibly entitled, "Why isn't everyone an *empirical* Bayesian?"

ACKNOWLEDGMENTS

Partial support provided by the Johns Hopkins International Center of Excellence for Malaria Research cooperative agreement NIH-NIAID, U19-AI089680.

REFERENCES

BELL, W. R., DATTA, G. S. and GHOSH, M. (2013). Benchmarking small area estimators. *Biometrika* **100** 189–202. MR3034332

BOX, G. E. P. (1980). Sampling and Bayes' inference in scientific modelling and robustness. *J. Roy. Statist. Soc. Ser. A* **143** 383–430. MR0603745

BROWNSTEIN, N. C., LOUIS, T. A., O'HAGAN, A. and PENDERGAST, J. (2019). The role of expert judgment in statistical inference and evidence-based decision-making. *Amer. Statist.* **73** 56–68. MR3925709

CARLIN, B. P. and GELFAND, A. E. (1991). A sample reuse method for accurate parametric empirical Bayes confidence intervals. *J. Roy. Statist. Soc. Ser. B* **53** 189–200.

CARLIN, B. P. and LOUIS, T. A. (2009). *Bayesian Methods for Data Analysis*, 3rd ed. Chapman and Hall/CRC Press, Boca Raton, FL.

EFRON, B. (1986). Why isn't everyone a Bayesian? *Amer. Statist.* **40** 1–11. MR0828575

EFRON, B. (2014). Two modeling strategies for empirical Bayes estimation. *Statist. Sci.* **29** 285–301. MR3264543

LAIRD, N. M. (1982). Empirical Bayes Estimates Using the Nonparametric Maximum Likelihood Estimate for the Prior. *J. Stat. Comput. Simul.* **15** 211–220.

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

LAIRD, N. M. and LOUIS, T. A. (1991). Smoothing the nonparametric estimate of a prior distribution by roughening: A computational study. *Comput. Statist. Data Anal.* **12** 27–37. MR1131643

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

O'HAGAN, A. (2019). Expert knowledge elicitation: Subjective but scientific. *Amer. Statist.* **73** 69–81. MR3925710

PADDOCK, S. M. and LOUIS, T. A. (2011). Percentile-based empirical distribution function estimates for performance evaluation of healthcare providers. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **60** 575–589. MR2829191

PADDOCK, S. M., RIDGEWAY, G., LIN, R. and LOUIS, T. A. (2006). Flexible distributions of triple-goal estimates in two-stage hierarchical models. *Comput. Statist. Data Anal.* **50** 3243–3262. MR2239666

ROBBINS, H. (1983). Some thoughts on empirical Bayes estimation. *Ann. Statist.* **11** 713–723. MR0707923

SHEN, W. and LOUIS, T. A. (1998). Triple-goal estimates in two-stage hierarchical models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **60** 455–471. MR1616061

SHEN, W. and LOUIS, T. A. (1999). Empirical Bayes estimation via the smoothing by roughening approach. *J. Comput. Graph. Statist.* **8** 800–823. MR1748968