Statistical Science
2019, Vol. 34, No. 2, 214–218
https://doi.org/10.1214/19-STS707
Main article: https://doi.org/10.1214/18-STS674

Comment: Bayes, Oracle Bayes and Empirical Bayes

Aad van der Vaart

Empirical Bayes methods are intriguing, and have gained in significance by present day big data applications. Despite their early introduction, they are still not fully understood. It is a pleasure to read the review by a "statistical rock star" [13], who stood at the beginning of the methods and more recently opened our eyes to their importance for large scale inference.

Empirical Bayes combines Bayesian ways of thinking about data and what some call "frequentist" methods, often maximum likelihood. The main point of my discussion is to highlight connections to nonparametric and high-dimensional Bayesian methods, which have seen a big development in the past 20 years.

In the second paragraph of Section 6, Efron writes: "which is to say that standard Bayes is finite Bayes with $N = \infty$" and goes on to describe a fully Bayesian approach (consisting of a hyperprior $h(g)$ on the density of the parameters θ_i) as an "uncertain task". I may not be full Bayes enough to say this with absolute certainty, but would think that nowadays most Bayesians would politely disagree and consider the setting a standard one, with a Dirichlet process prior as a "default" choice [17, 18, 1, 20]. Then the setting is described by the hierarchy:

- a probability distribution $G \sim \mathrm{DP}(\alpha)$,
- latent variables $\theta_0, \dots, \theta_n | G \overset{\text{iid}}{\sim} G$,
- observations $X_0, \dots, X_n | \theta_0, \dots, \theta_n, G \overset{\text{ind}}{\sim} N(\theta_i, 1)$.

This is the model of Efron's Sections 1–4 augmented with a prior on G, and could still be preceded by extra levels to construct the parameter α (a finite distribution) of the Dirichlet process $\mathrm{DP}(\alpha)$, in particular its total mass (called "prior precision"). We restrict to the case that the observations in step three are Gaussian; it would be worth while to extend our discussion to Poisson observations, as in Efron's Section 5.

Aad van der Vaart is Professor of Stochastics, Mathematical Institute, Niels Bohrweg 3, Leiden, Netherlands (e-mail: avdvaart@math.leidenuniv.nl; URL: www.math.leidenuniv.nl/~avdvaart).

In the preceding hierarchy, the desired posterior distribution of θ_0 given data $X_0, \dots, X_n$ is the Bayesian solution to the problem posed by Efron in Section 5. It is a standard product of Bayesian inference. The standard method to compute the posterior distribution of G in the hierarchy is based on the decomposition

$$\mathrm{P}(G \in \cdot | X_0, \dots, X_n) = \int \mathrm{P}(G \in \cdot | \theta_0, \dots, \theta_n) \, d\Pi(\theta_0, \dots, \theta_n | X_0, \dots, X_n),$$

where $d\Pi(\theta_0, \dots, \theta_n | X_0, \dots, X_n)$ refers to the posterior distribution of the latent variables $\theta_0, \dots, \theta_n$, and by general theory on Dirichlet processes the integrand $\mathrm{P}(G \in \cdot | \theta_0, \dots, \theta_n)$ follows a Dirichlet process $\mathrm{DP}(\alpha + \sum_i \delta_{\theta_i})$. The usual way to exploit this is to simulate samples from $d\Pi(\theta_0, \dots, \theta_n | X_0, \dots, X_n)$. So the standard algorithms can be used also to simulate from the distribution of interest $\theta_0 | X_0, \dots, X_n$. Over the past decades, many algorithms were developed, from "exact" Gibbs samplers to fast computational shortcuts, all using the remarkable properties of the Dirichlet process [15, 42, 23, 3, 26, 61, 27] (see [20], Chapter 5, for a partial overview). Depending on the algorithm the computational burden is similar to running the bootstrap algorithm in Efron's formulas (79)–(80). In fact, because the base measure $\alpha + \sum_i \delta_{\theta_i}$ of the Dirichlet posterior of G given the latent variables θ_i is essentially the empirical distribution of the latter variables, many of the computational schemes possess a bootstrap flavour.

One advantage of this approach is that it is pretty. Another should be that, as any Bayesian approach, it automatically yields uncertainty quantification, for instance, through credible intervals obtained from the (simulated) posterior distributions of the θ_i. Now Bayesian nonparametric credible sets are not confidence sets [12, 32, 56, 58, 7], except for smooth functionals for which a Bernstein–von Mises theorem holds [31, 32, 5, 6, 50, 8, 49]. The representation

$$\mathrm{P}(\theta_0 \in B | X_0, \dots, X_n) = \int \frac{\int_B \phi(X_0 - \theta) \, dG(\theta)}{\int \phi(X_0 - \theta) \, dG(\theta)} \, d\Pi(G | X_0, \dots, X_n),$$

suggests that the posterior of the new latent variable θ_0 is *not* smooth in this sense. Furthermore, semiparametric information theory suggests that even in the class of linear functionals $\int h\,dG$ a Bernstein-von Mises theorem can hold only for the very special functions h of the form $h(z) = \int \bar{h}(x)\phi(x-z)\,dx$ (which are in the range of the adjoint score operator; see [59, 60]). Theory developed for other priors than the Dirichlet process [58, 7] further suggests that the shrinkage generated by (empirical) Bayes modelling, which is desirable and accounts for the increased efficiency, entails a complicated relationship to setting confidence intervals. In summary, although in the past decade Dirichlet–normal mixtures were shown to have remarkably good properties, both by theory and experimentation [21, 19, 54], its use for credible intervals remains to be further investigated. Preferably theory should cover the frequentist setup of independent $X_0, \dots, X_n \overset{\text{ind}}{\sim} N(\theta_i, 1)$ for arbitrary parameters θ_i, and the correct question may be to ask: "for which configurations $\theta_0, \dots, \theta_n$ is the inference satisfactory? It seems that the answer cannot include all configurations, but one might, for instance, hope for configurations that resemble a sample from a distribution.

As mentioned, the posterior distribution of G based on the "direct observations" θ_i from G is the Dirichlet process with base measure $\alpha + \sum_i \delta_{\theta_i}$. This has mean very close to the empirical distribution of these direct observations, and also (for larger n) the fluctuations of the posterior are given by a Brownian bridge process [37, 25], as for the empirical distribution around the true distribution. As the empirical distribution of the θ_i is the nonparametric maximum likelihood estimator of G, this invites to view the Dirichlet prior as a "nonparametric prior" [17]. It also suggests that the Dirichlet posterior based on the observations X_i relates in the same way to the nonparametric maximum likelihood estimator in the mixture setting: the maximiser of $G \mapsto \prod_i f_G(X_i)$ over all probability distributions G, for $f_G(x) = \int \phi(x-\theta)\,dG(\theta)$ the marginal density of the X_i, discussed in Efron's Section 6. The latter procedure also follows an age-old and proven, general principle of statistics, and is equally appealing to me. Again there is quite a bit of theory and experimentation that suggests that this procedure works excellently [30, 36, 45, 33, 28] for certain purposes. For inference on the θ_i, Efron (although he prefers parametric models for G) proposes to plug-in the maximum likelihood estimator $\hat{G}$ into

$$e_G(X_i) := \mathrm{E}(\theta_i | X_i, G) = \frac{\int \theta\phi(X_i - \theta)\,dG(\theta)}{\int \phi(X_i - \theta)\,dG(\theta)}.$$

(The notation is the same as Efron's, see his formula (12), but in our setup the expectation is conditional given G.) This may be compared to the Dirichlet approach, which would average G out over its posterior distribution:

$$\begin{aligned} e(X_i) &:= \mathrm{E}(\theta_i | X_0, \dots, X_n) \\ &= \int e_G(X_i)\,d\Pi(G | X_0, \dots, X_n). \end{aligned}$$

Although in practice one would take the average over θ_i from a sample generated from the posterior distribution rather than use this equation, the formula is useful to suggest that the two estimators are similar. Although the correspondence is not perfect [44, 55], full Bayes posterior distributions typically concentrate around the corresponding maximum likelihood estimator.

Whereas the Bayesian procedure $e(X_i)$ is the mean of a full posterior distribution, the plug-in $e_{\hat{G}}(X_i)$ is only a point estimator. Could semiparametric profile likelihood [41] based on the same likelihood function lead to valid confidence intervals? Perhaps for certain configurations of the parameters? How exactly does this relate to the full Bayesian formulation?

The Dirichlet mixture formulation fits into Efron's procedure of g-modelling, with the Dirichlet prior a nonparametric approach to G. There are plenty of other priors for G that can be used, including smooth parametric models. The smoothness of the normal density makes the marginal densities f_G of the X_i for two different mixing distributions G similar even if the G are quite different, for instance, in smoothness and number of support points [21, 52]. This diminishes the role of the prior and suggests that the gain of using a parametric prior can be small, even if the model is correct.

Such approximations, and the potential harm of misspecification, are dependent on the scale of the Gaussian kernel, here taken equal to 1 throughout, and the support of the θ_i. One other possible use of empirical Bayes methods is to set such "hyper" parameters in a data-dependent way. Maximum likelihood on the Bayesian marginal density of the data is an attractive method, and has been observed to perform similarly to a full (hierarchical) Bayesian approach that puts priors on these parameters. While Efron achieves good results using, for example, splines with 7 degrees of freedom with $N = 1500$ observations at error scale 1, some automation might be preferable, and (empirical) Bayes is an attractive way to do so.

Strong prior knowledge on the nature of the latent variables θ_i may be put to use. If these are suspected to take on a small number of different values, a finite mixture with a fixed or penalised number of support points may be preferable over a Dirichlet, as the latter, even though very sparse, still may overshoot the number of support points [39]. The sparse case, where many θ_i are (nearly) zero, is especially relevant. By its nonparametric nature, a Dirichlet process prior might work, but in the past decade attention has focused on priors that explicitly put a point mass at zero (spike-and-slab) [40, 24, 29, 10, 9] or that are continuous with a peak at zero, such as the horseshoe or two-point mixtures [4, 48, 22, 51]. Such priors naturally shrink the posterior distribution of individual parameters θ_i to zero, unless the corresponding observation X_i is clearly away from zero, and in this sense are a Bayesian competitor to the Lasso. The posterior mean $e_G(x)$ as a function of the observation is an S-curve as in Efron's Figure 2, but sparsity in the prior makes for a sharper S, distinguishing better between small and large values of x. The shrinkage effect is moderated by the size of the point mass at zero or the width of the peak at zero, which can, and should, be set based on the data. Full Bayesians will prefer to put a "hyper prior" on these parameters, but empirical Bayes (based on maximising the likelihood $\prod_i f_G(X_i)$ over the parameter in G) gives about the same behaviour [29, 57]. The marginal posterior distributions of the parameters can be used to set credible intervals for the individual parameters θ_i. Although for some configurations of parameters overshrinkage may destroy coverage, these intervals perform reasonably well [58, 7], in particular for making "discoveries", that is, filtering out nonzero parameters.

Efron discusses the estimation of the number of unseen species of butterflies as an application of G-modelling with Poisson observations. Here also there is a nonparametric Bayes connection. So-called *species sampling models* (see [20], Chapter 14, for a summary) are random discrete distributions whose point masses can serve as a prior model for the abundances (scaled to fractions) of species. Observed individuals can be viewed as a sample from such a discrete distribution, and questions about unseen species can be formulated in terms of properties of the hidden random discrete distribution and answered by posterior quantities given the observations. For instance, the probability that the next observation X_{n+1} will be a new species, given observations $X_1, \dots, X_n$, is the posterior predictive probability that this will be drawn from an atom that has not been used by $X_1, \dots, X_n$. The Dirichlet process prior is one example of such a species sampling model (it is indeed discrete [38, 2, 53]), but there are many other examples, possibly more suitable to butterflies [35, 14], with a close link to the theory of random exchangeable partitions [43, 46, 47]. Applications to estimating unseen species were developed in [34, 16, 11].

ACKNOWLEDGMENTS

The research leading to these results has received funding from the European Research Council under ERC Grant Agreement 320637.

REFERENCES

[1] ANTONIAK, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Ann. Statist.* **2** 1152–1174. MR0365969

[2] BLACKWELL, D. (1973). Discreteness of Ferguson selections. *Ann. Statist.* **1** 356–358. MR0348905

[3] BLEI, D. M. and JORDAN, M. I. (2006). Variational inference for Dirichlet process mixtures. *Bayesian Anal.* **1** 121–143. MR2227367

[4] CARVALHO, C. M., POLSON, N. G. and SCOTT, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika* **97** 465–480. MR2650751

[5] CASTILLO, I. (2012). A semiparametric Bernstein–von Mises theorem for Gaussian process priors. *Probab. Theory Related Fields* **152** 53–99. MR2875753

[6] CASTILLO, I. (2012). A semiparametric Bernstein–von Mises theorem for Gaussian process priors. *Probab. Theory Related Fields* **152** 53–99. MR2875753

[7] CASTILLO, I. and MISMER, R. (2018). Empirical Bayes analysis of spike and slab posterior distributions. *Electron. J. Stat.* **12** 3953–4001. MR3885271

[8] CASTILLO, I. and ROUSSEAU, J. (2015). A Bernstein–von Mises theorem for smooth functionals in semiparametric models. *Ann. Statist.* **43** 2353–2383. MR3405597

[9] CASTILLO, I., SCHMIDT-HIEBER, J. and VAN DER VAART, A. (2015). Bayesian linear regression with sparse priors. *Ann. Statist.* **43** 1986–2018. MR3375874

[10] CASTILLO, I. and VAN DER VAART, A. (2012). Needles and straw in a haystack: Posterior concentration for possibly sparse sequences. *Ann. Statist.* **40** 2069–2101. MR3059077

[11] CEREDA, G. (2017). Current challenges in statistical DNA evidence evaluation. Leiden Univ.

[12] COX, D. D. (1993). An analysis of Bayesian inference for nonparametric regression. *Ann. Statist.* **21** 903–923. MR1232525

[13] DAVIDE, C. (2018). Statistical 'rock star' wins coveted international prize. *Nature*. Published online: 12 November 2018; DOI:10.1038/d41586-018-07395-w.

[14] DE BLASI, P., LIJOI, A. and PRÜNSTER, I. (2013). An asymptotic analysis of a class of discrete nonparametric priors. *Statist. Sinica* **23** 1299–1321. MR3114715

[15] ESCOBAR, M. D. (1994). Estimating normal means with a Dirichlet process prior. *J. Amer. Statist. Assoc.* **89** 268–277. MR1266299

[16] FAVARO, S., LIJOI, A. and PRÜNSTER, I. (2012). Asymptotics for a Bayesian nonparametric estimator of species variety. *Bernoulli* **18** 1267–1283. MR2995795

[17] FERGUSON, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1** 209–230. MR0350949

[18] FERGUSON, T. S. (1983). Bayesian density estimation by mixtures of normal distributions. In *Recent Advances in Statistics* 287–302. Academic Press, New York. MR0736538

[19] GHOSAL, S. and VAN DER VAART, A. (2007). Posterior convergence rates of Dirichlet mixtures at smooth densities. *Ann. Statist.* **35** 697–723. MR2336864

[20] GHOSAL, S. and VAN DER VAART, A. (2017). *Fundamentals of Nonparametric Bayesian Inference. Cambridge Series in Statistical and Probabilistic Mathematics* **44**. Cambridge Univ. Press, Cambridge. MR3587782

[21] GHOSAL, S. and VAN DER VAART, A. W. (2001). Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. *Ann. Statist.* **29** 1233–1263. MR1873329

[22] GRIFFIN, J. E. and BROWN, P. J. (2010). Inference with normal-gamma prior distributions in regression problems. *Bayesian Anal.* **5** 171–188. MR2596440

[23] ISHWARAN, H. and JAMES, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *J. Amer. Statist. Assoc.* **96** 161–173. MR1952729

[24] ISHWARAN, H. and RAO, J. S. (2005). Spike and slab variable selection: Frequentist and Bayesian strategies. *Ann. Statist.* **33** 730–773. MR2163158

[25] JAMES, L. F. (2008). Large sample asymptotics for the two-parameter Poisson–Dirichlet process. In *Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanta K. Ghosh. Inst. Math. Stat.* (*IMS*) *Collect.* **3** 187–199. IMS, Beachwood, OH. MR2459225

[26] JARA, A. (2007). Applied Bayesian non-and semi-parametric inference using dppackage. *R News* **7** 17–26.

[27] JARA, A., HANSON, T., QUINTANA, F., MUELLER, P. and ROSNER, G. (2015). Package DPpackage.

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

[29] JOHNSTONE, I. M. and SILVERMAN, B. W. (2004). Needles and straw in haystacks: Empirical Bayes estimates of possibly sparse sequences. *Ann. Statist.* **32** 1594–1649. MR2089135

[30] KIEFER, J. and WOLFOWITZ, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Ann. Math. Stat.* **27** 887–906. MR0086464

[31] KIM, Y. (2006). The Bernstein–von Mises theorem for the proportional hazard model. *Ann. Statist.* **34** 1678–1700. MR2283713

[32] KNAPIK, B. T., SZABÓ, B. T., VAN DER VAART, A. W. and VAN ZANTEN, J. H. (2016). Bayes procedures for adaptive inference in inverse problems for the white noise model. *Probab. Theory Related Fields* **164** 771–813. MR3477780

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

[34] LIJOI, A., MENA, R. H. and PRÜNSTER, I. (2007). Bayesian nonparametric estimation of the probability of discovering new species. *Biometrika* **94** 769–786. MR2416792

[35] LIJOI, A., PRÜNSTER, I. and WALKER, S. G. (2008). Bayesian nonparametric estimators derived from conditional Gibbs structures. *Ann. Appl. Probab.* **18** 1519–1547. MR2434179

[36] LINDSAY, B. (1995). Mixture models: Theory, geometry and applications. In *NSF-CBMS Regional Conference Series in Probability and Statistics* i–163. IMS, Hayward, CA.

[37] LO, A. Y. (1983). Weak convergence for Dirichlet processes. *Sankhyā Ser. A* **45** 105–111. MR0749358

[38] MCCLOSKEY, J. W. T. (1965). *A Model for the Distribution of Individuals by Species in an Environment*. ProQuest LLC, Ann Arbor, MI. Thesis (Ph.D.), Michigan State Univ. MR2615013

[39] MILLER, J. W. and HARRISON, M. T. (2014). Inconsistency of Pitman–Yor process mixtures for the number of components. *J. Mach. Learn. Res.* **15** 3333–3370. MR3277163

[40] MITCHELL, T. J. and BEAUCHAMP, J. J. (1988). Bayesian variable selection in linear regression. *J. Amer. Statist. Assoc.* **83** 1023–1036. MR0997578

[41] MURPHY, S. A. and VAN DER VAART, A. W. (2000). On profile likelihood. *J. Amer. Statist. Assoc.* **95** 449–485. MR1803168

[42] NEAL, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *J. Comput. Graph. Statist.* **9** 249–265. MR1823804

[43] PERMAN, M., PITMAN, J. and YOR, M. (1992). Size-biased sampling of Poisson point processes and excursions. *Probab. Theory Related Fields* **92** 21–39. MR1156448

[44] PETRONE, S., RIZZELLI, S., ROUSSEAU, J. and SCRICCIOLO, C. (2014). Empirical Bayes methods in classical and Bayesian inference. *Metron* **72** 201–215. MR3233149

[45] PFANZAGL, J. (1988). Consistency of maximum likelihood estimators for certain nonparametric families, in particular: Mixtures. *J. Statist. Plann. Inference* **19** 137–158. MR0944202

[46] PITMAN, J. (1995). Exchangeable and partially exchangeable random partitions. *Probab. Theory Related Fields* **102** 145–158. MR1337249

[47] PITMAN, J. (1996). Some developments of the Blackwell–MacQueen urn scheme. In *Statistics, Probability and Game Theory. Institute of Mathematical Statistics Lecture Notes—Monograph Series* **30** 245–267. IMS, Hayward, CA. MR1481784

[48] POLSON, N. G. and SCOTT, J. G. (2011). Shrink globally, act locally: Sparse Bayesian regularization and prediction. In *Bayesian Statistics* 9 (J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith and M. West, eds.) 501–538. Oxford Univ. Press, Oxford. MR3204017

[49] RAY, K. (2017). Adaptive Bernstein–von Mises theorems in Gaussian white noise. *Ann. Statist.* **45** 2511–2536. MR3737900

[50] RIVOIRARD, V. and ROUSSEAU, J. (2012). Bernstein–von Mises theorem for linear functionals of the density. *Ann. Statist.* **40** 1489–1523. MR3015033

[51] ROČKOVÁ, V. (2018). Bayesian estimation of sparse signals with a continuous spike-and-slab prior. *Ann. Statist.* **46** 401–437. MR3766957

[52] SCRICCIOLO, C. (2014). Adaptive Bayesian density estimation in L^p-metrics with Pitman–Yor or normalized inverse-Gaussian process kernel mixtures. *Bayesian Anal.* **9** 475–520. MR3217004

[53] SETHURAMAN, J. (1994). A constructive definition of Dirichlet priors. *Statist. Sinica* **4** 639–650. MR1309433

[54] SHEN, W., TOKDAR, S. T. and GHOSAL, S. (2013). Adaptive Bayesian multivariate density estimation with Dirichlet mixtures. *Biometrika* **100** 623–640. MR3094441

[55] SNIEKERS, S. and VAN DER VAART, A. (2015). Adaptive Bayesian credible sets in regression with a Gaussian process prior. *Electron. J. Stat.* **9** 2475–2527. MR3425364

[56] SZABÓ, B., VAN DER VAART, A. W. and VAN ZANTEN, J. H. (2015). Frequentist coverage of adaptive nonparametric Bayesian credible sets. *Ann. Statist.* **43** 1391–1428. MR3357861

[57] VAN DER PAS, S., SZABÓ, B. and VAN DER VAART, A. (2017). Adaptive posterior contraction rates for the horseshoe. *Electron. J. Stat.* **11** 3196–3225. MR3705450

[58] VAN DER PAS, S., SZABÓ, B. and VAN DER VAART, A. (2017). Uncertainty quantification for the horseshoe (with discussion). *Bayesian Anal.* **12** 1221–1274. MR3724985

[59] VAN DER VAART, A. (1991). On differentiable functionals. *Ann. Statist.* **19** 178–204. MR1091845

[60] VAN DER VAART, A. (2002). Semiparametric statistics. In *Lectures on Probability Theory and Statistics* (*Saint-Flour*, 1999). *Lecture Notes in Math.* **1781** 331–457. Springer, Berlin. MR1915446

[61] WALKER, S. G. (2007). Sampling the Dirichlet mixture model with slices. *Comm. Statist. Simulation Comput.* **36** 45–54. MR2370888