Statistical Science
2019, Vol. 34, No. 2, 209–213
https://doi.org/10.1214/19-STS706
Main article: https://doi.org/10.1214/18-STS674

Comment: Minimalist g-Modeling

Roger Koenker and Jiaying Gu

Abstract. Efron's elegant approach to g-modeling for empirical Bayes problems is contrasted with an implementation of the Kiefer–Wolfowitz nonparametric maximum likelihood estimator for mixture models for several examples. The latter approach has the advantage that it is free of tuning parameters and consequently provides a relatively simple complementary method.

Key words and phrases: Nonparametric maximum likelihood, mixture model, convex optimization.

1. INTRODUCTION

It is a great privilege to have the opportunity to comment on this marvelous paper. Nearly 70 years ago Herbert Robbins, the oracle of empirical Bayesianism, published an abstract in the *Annals* that begins:

> Let θ be a vector random variable with distribution function $G(\theta)$ belonging to some class $\mathcal{G}$, let X be a vector random variable whose frequency function $f(x;\theta)$ depends on θ, and let $g^*(x) = \int f(x;\theta)\,dG(\theta)$ be the resulting frequency function of X. From a sample $X_1, X_2, \ldots$ it is required to estimate $G(\theta)$. The generalized method of maximum likelihood consists in using the estimates $G_n(\theta; x_1, \ldots, x_n)$ in $\mathcal{G}$ for which $\prod g^*(x_i)$ is a maximum. Under certain restrictions, this method is consistent as $n \to \infty$. (Robbins, 1950)

Of course, since this was only an abstract; no details were provided, or forthcoming, until Kiefer and Wolfowitz (1956), elaborating on Wald, provided details for the consistency claim. Some time then passed, until Laird (1978) described how the nascent EM algorithm could be deployed to compute G_n. The influential paper of Jiang and Zhang (2009) has renewed interest in the Kiefer–Wolfowitz NPMLE establishing precise risk bounds and demonstrating attractive simulation performance. In econometrics Heckman and Singer (1984) were among the first to take up the challenge of actually using EM to compute a G_n in an effort to explore frailty models for unemployment durations.

As Efron persuasively argues the time is now ripe for a major revival of interest in these methods. Data sources are much more plentiful and computational wherewithal is vastly improved. Efron's g-modeling offers an extremely flexible approach to achieving Robbins objective of effectively estimating the mixing distribution, G. This seems already astonishing in the Gaussian location mixture setting where maximum likelihood out-performs classical Fourier methods for deconvolution, but is even more astonishing when one realizes that similar methods may be applied to a much wider class of general mixture problems. The decision to model $g = G'$ as an exponential family brings many attendant advantages, not the least of which is the elegant inference apparatus laid out in Efron's paper. However, the B-spline basis expansion and the Euclidean penalization of its coefficients adds a layer of hierarchical Bayesian artistry that may frighten away some researchers. In what follows, we will try to make a case for a complementary, more minimalist approach based on the nonparametric MLE of Robbins and Kiefer–Wolfowitz.

2. A MINIMALIST G_n

The Kiefer–Wolfowitz NPMLE, like Efron's $\hat{g}$, relies upon a grid of values $t_1, \ldots, t_m$, that constitute potential support points. The number of these potential support points can be quite large; we generally take

Roger Koenker is Honorary Professor of Economics, Department of Economics, University College London, London, WC1H OAX, United Kingdom (e-mail: r.koenker@ucl.ac.uk). Jiaying Gu is Assistant Professor of Economics, Department of Economics, University of Toronto, Toronto, Ontario, M5S 3G7, Canada (e-mail: jiaying.gu@utoronto.ca).

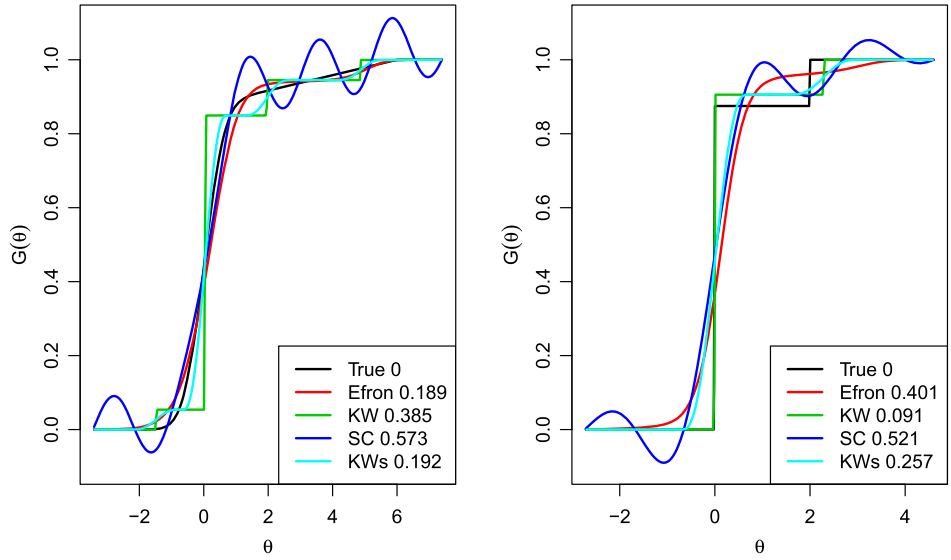

FIG. 1. *Four estimates of the mixing distributions G: In the left panel the true mixing distribution is a smooth scale mixture of Gaussians, in the right panel it is discrete with mass points at* 0 *and* 2. *The legend reports Wasserstein (L_1) distances between each of the estimates $\hat{G}_n$ and the true G.*

$m = 300$ with equally spaced t_i over the support of the observed X's, at least for Gaussian location mixtures with moderate n. The primal version of the NPMLE problem is then

$$\min_g \left\{ -\sum_{i=1}^{n} \log f(x_i) \;\middle|\; f = Ag, g \geq 0, 1_m^\top g = 1 \right\},$$

where A denotes an n by m matrix with ij element, $\varphi(x_i - t_j)$. This is a relatively simple convex optimization problem and as such admits a unique solution. As for the Breiman nonnegative garrotte, the requirement that $g \geq 0$ acts as a powerful regularization device. No more that n of the m elements of $\hat{g}$ can be strictly positive, and typically this number grows like $\mathcal{O}(\sqrt{n})$. As shown in Koenker and Mizera (2014), the corresponding dual problem

$$\max_v \left\{ \sum_{i=1}^{n} \log v_i \;\middle|\; A^\top v \leq n \right\}$$

is somewhat more convenient for computations. In either case, we obtain as a solution a discrete G_n with a small number of distinct mass points.

To illustrate the basic differences among the various methods of estimating G, we consider two variants of a simulation setting from Efron (2016). In the first of these, G is a smooth scale mixture of Gaussians: $G(\theta) = G_1(\theta) = \frac{1}{8}\Phi(\theta/6) + \frac{7}{8}\Phi(2\theta)$; in the second, we have a discrete mixing distribution: $G(\theta) = G_2(\theta) = \frac{1}{8}I(\theta \geq 0) + \frac{7}{8}I(\theta \geq 2)$, In Figure 1, we depict several estimates of G for each of these models based on a sample of size 1000. Performance, measured by Wasserstein (L_1) distance, $W_1(G, \hat{G}_n) = \int |\hat{G}_n(x) - G(x)|\, dx$, is reported in the legend for each estimator. In the smooth setting of the left panel, the Efron estimator is the clear winner, although the smoothed version of the Kiefer–Wolfowitz estimator that simply convolves the discrete estimate with a biweight kernel with scale 0.7 does almost as well. In the right panel, where the true G is discrete with only two mass points, the KW estimator is almost paranormal. In both settings, the kernel-based deconvolution estimator of Stefanski and Carroll (1990) does poorly particularly in the tails.

A small simulation experiment to compare performance of these four estimators in the two settings of Figure 1 is reported in Table 1. Mean Wasserstein errors are based on 1000 replications.

TABLE 1
Mean Wasserstein (L_1) error

	Efron	Kernel	NPMLE	NPMLEs
Smooth	0.185	0.591	0.342	0.180
Discrete	0.409	0.718	0.156	0.280

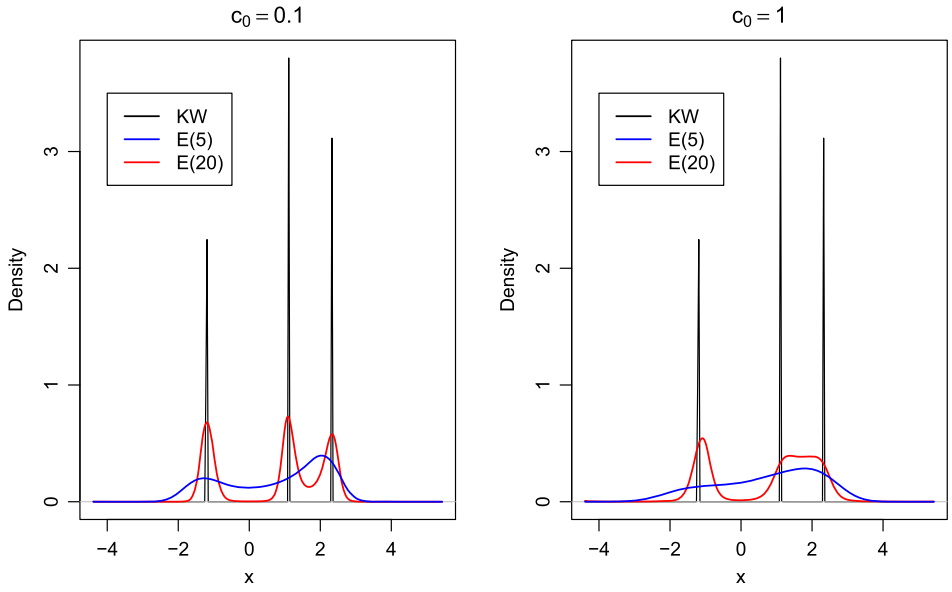

FIG. 2. *The effect of tuning on the Efron $\hat{g}$ estimator: The smooth curves depict $\hat{g}$'s computed with $K \in \{5, 20\}$ and $c_0 \in \{0.1, 1\}$ contrasted with the estimated point masses produced by the Kiefer–Wolfowitz NPMLE.*

3. TUNING

The flexibility of g-modeling as formulated in Efron (2016) arises from the opportunity to choose both the basis expansion for $\log g$ and the form and severity of the penalization of the parameters of that expansion. To explore the role of these choices in the context of Efron's "two towers" example, we compared several estimates with $g(\theta; \alpha)$ expressed as a natural spline expansion with $\alpha \in \mathbb{R}^K$ and penalty term, $c_0\|\alpha\|$. We generate data as in Efron's Figure 1, with $n = 1500$. In the left panel of Figure 2 we plot estimates of the mixing density, g, based on $K = 5$ and $K = 20$ with $c_0 = 0.1$ together with the NPMLE estimate, which has only three distinct mass points. In the right panel we do the same except that now $c_0 = 1$. It is evident that with K large and c_0 small one can obtain a $\hat{g}$ that begins to mimic the NPMLE quite well. This is somewhat similar to what happens when computing the NPMLE with the EM algorithm where early stopping of the iterations acts as a regularizing device. This is just one realization, what happens if we repeat the exercise?

To see how systematic the differences really are, we ran a small simulation experiment to compare empirical Bayes regret as defined by Efron relative to the Oracle Bayes estimator. Table 2 reports the results of this experiment based on 500 replications. The most flexible of the four Efron $\hat{g}$ estimators performs essentially the same as the NPMLE, but the other choices do not do as well, suggesting that careful tuning of the g-modeling procedure is important.

4. FREQUENTLY ALMOST BAYESIAN

It is difficult, perhaps impossible, to unravel the Bayesian and frequentist strands of the empirical Bayes tradition, and probably not terribly productive, Lindley's dictum notwithstanding. However, it does seem worthwhile at least briefly to see how the g-modeling methods already considered compare with well-established, more formal Bayes methods based on Dirichlet process methods. To this end, we reconsider the analysis of Shakespeare's vocabulary in Efron (2010).

The data consists of word counts, $\{n_1, \ldots, n_{100}\}$ where n_j denotes the number of the words in the Shakespeare canon of plays and poetry used precisely j times. Adopting the presumption that words appear as independent Poisson draws with individual intensity parameter, λ_i, this gives us a truncated Poisson mixture model,

$$\eta_j = \mathbb{E}n_j = S \int_0^\infty e^{-\lambda}\lambda^j/\big(P(\lambda)\, j!\big)\, dG(\lambda),$$

TABLE 2
Empirical Bayes Regret for "two towers" example

$c_0 = 1.0$		$c_0 = 0.1$		
df = 5	df = 20	df = 5	df = 20	NPMLE
0.03983	0.01131	0.01055	0.00805	0.00825

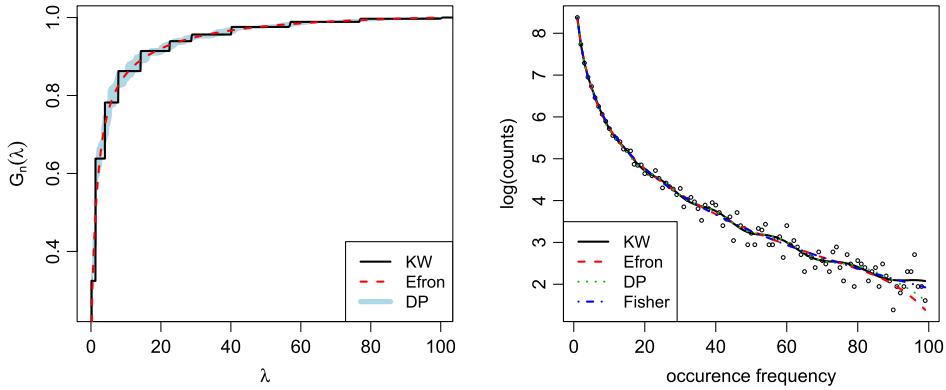

FIG. 3. *Left panel*: *Comparison of the Kiefer–Wolfowitz NPMLE and Efron g-modeling estimates with a Dirichlet process estimate of the mixing distribution G for the Shakespeare vocabulary data. The Dirichlet estimate is depicted as a* 0.95 *pointwise band based on the last* 1000 *iterations of the Metropolis–Hastings MCMC chain. Right panel*: *Comparison of the NPMLE, Efron and DP predictions of the* $\eta_j = \mathbb{E}n_j$ *values for* $j = 2, 3, \ldots, 100$ *with the parametric empirical Bayes procedure of Fisher.*

where $P(\lambda) = \Lambda(100, \lambda) - \Lambda(0, \lambda)$ with $\Lambda(x, \lambda)$ denoting the Poisson distribution function, and $S = 884{,}647$ the total number of words in the Shakespeare canon. The standard Dirichlet process formulation for Poisson mixtures would specify a prior for G as $DP(\alpha, G_0)$, where the base measure, G_0 would be gamma with some specified parameters, and α denotes the concentration of the prior belief. Truncation of the Poisson complicates things somewhat, rendering the usual closed form Gibbs MCMC infeasible. Instead we can adopt the Metropolis–Hastings strategy of Algorithm 8 of Neal (2000) as implemented in the R package `dirichletprocess` of Ross and Markwick (2018).

How formal is this more formal DP estimate from a Bayesian standpoint? We have not, we confess, chosen the parameters of the prior $DP(\alpha, G_0)$ from some deep philological understanding of English poetry and prose, instead we have given the MCMC iterations free rein to update the concentration parameter, α, and the rate parameter β of G_0. The rate parameter of G_0 is taken to be conjugate Gamma with parameters $(1, 1/2)$; the shape parameter of the G_0 is held fixed at 0.25 to avoid further complicating the estimation process. This yields a relatively weak "prior" with $\hat{\alpha} = 9.27$ and gamma rate parameter $\hat{\beta} = 0.0232$, as posterior medians based on the last 1000 of 2000 MCMC iterations. The resulting DP estimate, G_n, is illustrated in left panel of Figure 3 as a 0.95 pointwise band again based on the last 1000 MCMC iterations. For comparison, the NPMLE and Efron's G_n, with $df = 5$ and $c_0 = 2$, are overlayed in the figure. Although the three estimates appear quite similar, the computational effort they require differs considerably. The DP posterior requires about an hour and a half to compute, while the NPMLE and Efron's G_n each require less than a second; this has the unfortunate consequence of making further exploration of sensitivity of the DP procedure to the choice of hyperparameters and other tuning parameters of the MCMC process quite costly.

In the right panel of Figure 3, we compare the observed values, n_j, with the predictions from the NPMLE, Efron and DP procedures for $\eta_j : j = 2, \ldots, 100$ with the parametric MLE procedure proposed by Fisher for the Corbet butterfly data. Conditioning on n_1, and using the negative binomial representation of gamma mixtures of Poissons, we can write

$$\hat{\eta}_j = n_1 \lambda_j(\hat{a}, \hat{b}) = n_1 \frac{\Gamma(\hat{a} + j)\hat{b}^{j-1}}{j!\Gamma(\hat{a} + 1)},$$

where $(\hat{a}, \hat{b}) = \operatorname{argmax}\{\sum_{j=2}^{n} \log p(n_j, n_1 \lambda_j(a, b))\}$, and $p(n, \lambda)$ denotes the Poisson density. This too may be viewed as a parametric empirical Bayes procedure, and it delivers an astonishingly good fit to the observed counts despite the fact that its estimated "prior" with $(\hat{a}, \hat{b}) = (-0.398, 0.992)$ is improper. On the basis of visual goodness of fit, there is little to distinguish the four procedures, all perform admirably.

None of the procedures we have discussed meet a stringent Bayesian standard, as formulated for example, in Deely and Lindley (1981), so perhaps it is time to modify slightly another famous dictum of Lindley:

> We will all be [*empirical*] Bayesians in 2020, and then we can be a united profession. (Lindley and Smith, 1995)

ACKNOWLEDGEMENTS

The authors wish to express their appreciation to the Editor, Cun-Hui Zhang, for the opportunity to participate in this discussion, and to Dean Markwick for clarifying several points regarding the `dirichletprocess` R package. All the computational results reported here are available from the authors on request; they rely on the R package REBayes, Koenker and Gu (2015), which relies in turn on the convex optimization software Mosek, Andersen (2010) and its R interface, Rmosek, Friberg (2012).

REFERENCES

ANDERSEN, E. D. (2010). The Mosek Optimization Tools Manual, Version 6.0. Available from: http://www.mosek.com.

DEELY, J. J. and LINDLEY, D. V. (1981). Bayes empirical Bayes. *J. Amer. Statist. Assoc.* **76** 833–841. MR0650894

EFRON, B. (2010). *Large-Scale Inference*: *Empirical Bayes Methods for Estimation*, *Testing*, *and Prediction*. *Institute of Mathematical Statistics* (*IMS*) *Monographs* **1**. Cambridge Univ. Press, Cambridge. MR2724758

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

FRIBERG, H. A. (2012). Users Guide to the R-to-Mosek Interface. Available at http://rmosek.r-forge.r-project.org.

HECKMAN, J. and SINGER, B. (1984). A method for minimizing the impact of distributional assumptions in econometric models for duration data. *Econometrica* **52** 271–320. MR0735309

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

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

KOENKER, R. and GU, J. (2015). REBayes: An R Package for Empirical Bayes Methods. Available from https://cran.r-project.org/package=REBayes.

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

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

LINDLEY, D. V. and SMITH, A. (1995). A conversation with Dennis Lindley. *Statist. Sci.* **10** 305–319.

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

ROBBINS, H. (1950). A generalization of the method of maximum likelihood; estimating a mixing distribution (abstract). *Ann. Math. Stat.* **21** 314–315.

ROSS, G. J. and MARKWICK, D. (2018). Dirichletprocess: An R Package for Fitting Complex Bayesian Nonparametric Models. Available at https://cran.r-project.org/web/packages/dirichletprocess/vignettes/dirichletprocess.pdf.

STEFANSKI, L. and CARROLL, R. J. (1990). Deconvoluting kernel density estimators. *Statistics* **21** 169–184. MR1054861