EMPIRICAL BAYES CONFIDENCE INTERVALS: AN R VINAIGRETTE

ROGER KOENKER

Abstract. Renewed interest in Stein shrinkage and empirical Bayes methods more generally has prompted new work on confidence intervals for empirical Bayes estimates. Bruce Hansen's cogent discussion of Armstrong et al. (2020) in a recent Chamberlain Seminar aroused my curiosity about performance of posterior empirical Bayes intervals based on the nonparametric maximum likelihood methods of Robbins (1950) and Kiefer and Wolfowitz (1956). This note describes some further simulation evidence based on the implementation of this approach in Koenker and Gu (2015), and closely related methods of Efron (2019). While the "minimalist" approach based on posterior percentile intervals contructed from the NPMLE of Kiefer-Wolfowitz perform poorly, the smoother G modeling approach of Efron is found to possess good "robustness of validity" and "robustness of efficiency."

1. Introduction

Consider a family of mixture models of the form,

$$f(y) = \int \varphi(y|\theta) dG(\theta),$$

where φ denotes a known parametric "base" model and G denotes an unknown, nonparametric mixing distribution. Such models are fundemental in empirical Bayes compound decision settings where we have the (exchangeable) hierarchical structure,

$$Y_i \sim \varphi(y|\theta_i); \quad \theta_i \sim G.$$

When θ is a location parameter, so $\varphi(y|\theta_i) = \varphi(y - \theta_i)$ this is a conventional deconvolution problem usually evoking characteristic function methods, however recent work has emphasized the applicability of maximum likelihood methods. An authoritative overview of this approach is provided in Efron (2019) and the discussion thereof.

When the mixing distribution, G, is Gaussian, familiar Bayesian computations yield the James-Stein shrinkage formulas for (compound) squared error loss. This is the parametric empirical Bayes setting explored in Armstrong et al. (2020). In this note I want to very briefly explore two nonparametric approaches for confidence interval construction. We will

Version: June 16, 2020. A genre manifesto for R Vinaigrettes is available at `http://davoidofmeaning.blogspot.com/2016/12/r-vinaigrettes.html`. This is my second in a projected series of vinaigrettes about the R package REBayes, Koenker and Gu (2015). Code to reproduce the computational results presented is available from `http://www.econ.uiuc.edu/~roger/research/ebayes/ebayes.html`, along with the pdf version of this note. Thanks to Bruce Hansen for inciting my interest in this topic and sharing his simulation code. Thanks too to Jiaying Gu for helpful comments and continuing collaboration on empirical Bayes methods.

see that the mimimalist NPMLE methods of Kiefer and Wolfowitz while well suited to point estimation are not so well adapted to interval estimation. This conclusion complements the recent simulation findings and discussion of Jiang (2019), who considers several smoothed versions of the NPMLE.

2. The Kiefer-Wolfowitz NPMLE

In Koenker and Mizera (2014) we have advocated the Kiefer-Wolfowitz NPMLE approach to estimating G and constructing estimates of the θ_i's for compound decision problems. In sharp contrast to finite dimensional mixture problems with highly multimodel likelihoods, discrete formulations of the general nonparametric mixture problem are strictly convex and therefore admit unique solutions. Consider a grid $t_0, t_1, \cdots, t_m$ with associated masses $\{g \in \mathbb{R}^m | g_i \geqslant 0, \ \sum_{i=1}^m g_i \Delta t_i = 1\}$, we can approximate the log likelihood by,

$$\ell(G) = \sum_{i=1}^{n} \log f_i$$

where the n vector $f = Ag$ and A is the n by m matrix with typical element $\varphi(y_i, t_j)$. As is well known from Laird (1978) or Lindsay (1983) the NPMLE, $\hat{G}$, has $p \leqslant n$ positive mass points, while in practice this p is usually closer to $\log n$ than n. Interior point methods for solving such problems are considerably more efficient than earlier EM approaches greatly facilitating the study of their performance in simulation experiments. Unfortunately, little is known about their statistical efficiency from a theoretical perspective beyond the basic consistency results of Kiefer and Wolfowitz (1956) and Pfanzagl (1988).

Given an estimator for G, the usual Bayesian machinery can be invoked to produce point estimates with respect to various loss functions. In particular, we have Tweedie's formula, see e.g. Efron (2011), for φ standard Gaussian,

$$\hat{\theta}_i = y_i + f'(y_i)/f(y_i)$$

as a (nonlinear) shrinkage formula that can be implemented by plugging in $\hat{G}$ for G. Had we known G this is just the posterior mean, the Bayes rule with respect to (compound) quadratic loss. Other loss functions produce other point estimates, so quantile loss yields posterior quantiles. This raises the natural question: are these posterior quantile estimates good for anything. In particular, could they be used to construct confidence intervals for the $\hat{\theta}_i$'s? Before putting this to the test, let me briefly describe an alternative approach due to Efron.

3. Efron's NPMLE

Efron (2016) has proposed an alternative approach to estimating G that expresses its log derivative by a regression spline,

$$g(y|\theta) = \exp\{\sum_{j=1}^{p} \theta_j \psi_j(y) - \psi_0(\theta)\},$$

as in the pure density estimation methods of Stone (1990) and Barron and Shue (1991). We can maintain the same discretization for the support of G, and set,

$$g = (g_j) = (g(t_j|\theta)),$$

so the log likelihood can be expressed as above, except that now we are estimating a finite dimensional parameter θ of predetermined dimension. Efron suggests natural splines for the ψ_j functions and the penalization,

$$\ell_n(G_\theta) + \lambda\|\theta\|$$

by the Euclidean norm of the vector θ, thereby shrinking θ toward the origin and $\hat{G}$ toward the uniform distribution. A virtue of the Efron approach over the Kiefer-Wolfowitz form of the NPMLE is that it produces smooth $\hat{G}$'s that are better suited to inference. Their downside, as we have stressed in our contributed discussion, Koenker and Gu (2019), is the need to choose tuning parameters for the dimension of the basis expansion and the penalty parameter λ.

A striking feature of both the Efron and Kiefer-Wolfowitz proposals is that neither depend upon the mixture model being a formal convolution. Of course when θ is a location parameter so $\varphi(y|\theta) = \varphi(y - \theta)$ then classical deconvolution methods are also applicable. Efron compares the performance of his procedure with the kernel deconvolution method of Stefanski and Carroll (1990), and concludes that the latter is "too variable in the tails." This is confirmed in the comparisons reported in our contributed discussion of Efron (2019).

4. Some Simulation Evidence

We consider three simulation settings: the two proposed by Hansen, and one bonus setting proposed by an astute reader of the first draft who shall remain anonymous. All settings are special cases of the standard Gaussian sequence model,

$$Y_i = \theta_i + u_i,$$

with $u_i \sim \mathcal{N}(0, 1)$ and $\theta_i \sim G$. In the first setting G is also Gaussian with mean 0, and constant variance, V. This is obviously a favorable setting for the linear shrinkage Stein rule. There is almost no "signal" we are just observing Gaussians with variance $1+V$, when V is large there will be a few observations that are sufficient "unusual" that they would be unlikely to come from the standard Gaussian, but especially when V is small this setting is quite challenging for any NPMLE. The second setting is can be viewed as more typical of genomic Gaussian sequence models: $G = 0.90\delta_0 + 0.10\delta_a$, where δ_x is the usual Dirac δ-function with mass 1 at the point x. In the former setting there are four values of V that are considered, $\{0.1, 0.5, 1, 2\}$, while in the latter setting $a = \sqrt{10V}$ for one element of same set of V's. The procedure proposed by Armstrong et al. (2020) is now included in the simulation comparisons as implemented in their R package "ebci".

In our third bonus setting G is gamma with shape parameter $6\sqrt{V}$ and rate parameter 3. The corresponding densities for these G distributions are depicted in Figure 1.

In Table 1 I report observed coverage for posterior $[0.025, 0.975]$ intervals based on the Kiefer-Wolfowitz NPMLE on the left and the Efron NPMLE as well as the AKP-M procedure. For the Efron intervals I'm using the default 5 degrees of freedom natural spline basis expansion and $\lambda = 0.1$. It is painfully apparent that the using the Kiefer-Wolfowitz $\hat{G}$ produces severe under coverage in this setting, while the default Efron procedure is quite reliable. To explore this a bit further we report mean length of the intervals in Table 2, and root mean squared error of the posterior means in Table 3. It is evident from these tables that the under coverage of the KW intervals is due to their length; the KW posterior mean estimates are slightly more accurate than the point estimates delivered by the Efron posterior means. The discrete nature of the KW estimate, $\hat{G}$ inevitably produces intervals that are sensitive to the mass points of the $\hat{G}$, and can result in wildly optimistic intervals. The Armstrong et al (AKP-M) intervals have good coverage except for the $V = 0.1$ case where they are seriously uncovering. In contrast the Efron intervals are quite reliable throughout and at the same time considerably shorter than the AKP-M intervals except in the situations in which the latter undercover.

In Tables 4 to 6 we report parallel results for (outlier) Setting 2 of the simulation experiment. The Kiefer-Wolfowitz intervals are again consistently too short under covering the true parameters. The AKP-M intervals coverage is again quite good except when the outliers are relatively close to the null mass point at $\theta = 0$. However, the lengths of the AKM-P intervals are considerably wider than than the corresponding Efron intervals despite the fact that the Efron intervals tend to be unnecessarily conservative. KW posterior means are slightly more accurate than the Efron point estimates as in Setting 1 and

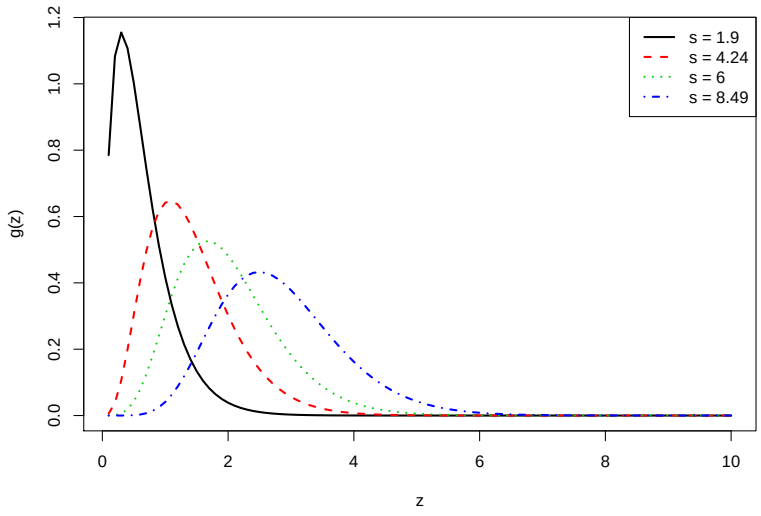

FIGURE 1. Four gamma densities for the third (bonus) simulation setting.

	Kiefer-Wolfowitz				Efron				AKP-M			
	n=100	n=200	n=500	n=1000	n=100	n=200	n=500	n=1000	n=100	n=200	n=500	n=1000
V = 0.1	0.411	0.433	0.490	0.554	0.954	0.956	0.952	0.951	0.668	0.714	0.800	0.871
V = 0.5	0.611	0.663	0.721	0.750	0.912	0.926	0.912	0.908	0.947	0.965	0.974	0.975
V = 1	0.655	0.710	0.758	0.789	0.924	0.935	0.928	0.904	0.972	0.977	0.978	0.978
V = 2	0.672	0.721	0.774	0.799	0.929	0.938	0.941	0.925	0.975	0.978	0.978	0.978

TABLE 1. Simulation Setting 1: Observed coverage proportion in 1000 trials for intended nominal coverage 0.95

	Kiefer-Wolfowitz				Efron				AKP-M			
	n=100	n=200	n=500	n=1000	n=100	n=200	n=500	n=1000	n=100	n=200	n=500	n=1000
V = 0.1	0.645	0.640	0.680	0.734	1.481	1.387	1.305	1.275	1.095	1.018	1.045	1.097
V = 0.5	1.392	1.490	1.612	1.670	2.186	2.208	2.167	2.155	2.528	2.569	2.618	2.613
V = 1	1.810	1.952	2.086	2.172	2.667	2.715	2.694	2.629	3.208	3.240	3.253	3.257
V = 2	2.156	2.301	2.463	2.542	3.084	3.130	3.158	3.113	3.720	3.730	3.735	3.738

TABLE 2. Simulation Setting 1: Observed interval length in 1000 trials for intended nominal coverage 0.95

	Kiefer-Wolfowitz				Efron				AKP-M			
	n=100	n=200	n=500	n=1000	n=100	n=200	n=500	n=1000	n=100	n=200	n=500	n=1000
V = 0.1	0.130	0.112	0.102	0.097	0.346	0.325	0.313	0.308	0.288	0.236	0.159	0.109
V = 0.5	0.385	0.362	0.349	0.343	0.608	0.594	0.592	0.589	0.347	0.339	0.337	0.335
V = 1	0.564	0.539	0.522	0.512	0.734	0.722	0.721	0.727	0.511	0.504	0.504	0.502
V = 2	0.751	0.715	0.691	0.682	0.843	0.828	0.824	0.835	0.682	0.668	0.667	0.668

TABLE 3. Simulation Setting 1: Observed root mean squared error in 1000 trials of posterior mean

	Kiefer-Wolfowitz				Efron				AKP-M			
	n=100	n=200	n=500	n=1000	n=100	n=200	n=500	n=1000	n=100	n=200	n=500	n=1000
V = 0.1	0.418	0.472	0.544	0.598	0.938	0.932	0.924	0.921	0.701	0.758	0.848	0.888
V = 0.5	0.459	0.485	0.539	0.521	0.967	0.973	0.978	0.981	0.930	0.933	0.930	0.930
V = 1	0.392	0.367	0.426	0.452	0.982	0.987	0.990	0.988	0.949	0.950	0.951	0.953
V = 2	0.340	0.338	0.371	0.381	0.992	0.995	0.998	0.997	0.964	0.964	0.967	0.967

TABLE 4. Simulation Setting 2: Observed coverage proportion in 1000 trials for intended nominal coverage 0.95

they are considerably more accurate than the point estimates delivered by linear shrinkage underlying the AKP-M procedure.

In Tables 7 to 9 we report parallel results for Setting 3 of the simulation experiment. The KW intervals are still poor in terms of coverage but now both the Efron and AKP-M intervals perform well in terms of coverage. But the AKP-M intervals are much too wide. The Efron smoothing of the MLE performs well in this asymmetric setting producing somewhat better RMSE performance than the Kiefer-Wolfowitz estimator.

	Kiefer-Wolfowitz				**Efron**				**AKP-M**			
	n=100	n=200	n=500	n=1000	n=100	n=200	n=500	n=1000	n=100	n=200	n=500	n=1000
V = 0.1	0.579	0.607	0.649	0.668	1.444	1.373	1.286	1.256	1.067	1.052	1.051	1.108
V = 0.5	1.036	1.058	1.069	1.032	1.943	1.914	1.849	1.839	2.517	2.581	2.605	2.618
V = 1	0.913	0.849	0.828	0.803	2.022	1.938	1.836	1.795	3.229	3.243	3.248	3.260
V = 2	0.564	0.521	0.480	0.433	1.881	1.749	1.607	1.522	3.734	3.732	3.739	3.740

TABLE 5. Simulation Setting 2: Observed interval length in 1000 trials for intended nominal coverage 0.95

	Kiefer-Wolfowitz				**Efron**				**AKP-M**			
	n=100	n=200	n=500	n=1000	n=100	n=200	n=500	n=1000	n=100	n=200	n=500	n=1000
V = 0.1	0.118	0.102	0.091	0.087	0.332	0.315	0.299	0.294	0.288	0.238	0.149	0.110
V = 0.5	0.285	0.261	0.242	0.236	0.534	0.517	0.504	0.498	0.348	0.341	0.336	0.335
V = 1	0.281	0.249	0.235	0.228	0.530	0.511	0.499	0.495	0.511	0.507	0.503	0.501
V = 2	0.159	0.135	0.118	0.115	0.404	0.377	0.360	0.355	0.677	0.672	0.667	0.669

TABLE 6. Simulation Setting 2: Observed root mean squared error in 1000 trials of posterior mean

	KW				KWs				Efron				AKP-M			
	n=100	n=200	n=500	n=1000	n=100	n=200	n=500	n=1000	n=100	n=200	n=500	n=1000	n=100	n=200	n=500	n=1000
V = 0.1	0.467	0.520	0.591	0.625	0.922	0.952	0.967	0.971	0.933	0.941	0.938	0.938	0.954	0.970	0.974	0.976
V = 0.5	0.581	0.620	0.689	0.716	0.926	0.948	0.962	0.967	0.913	0.927	0.919	0.934	0.977	0.979	0.978	0.979
V = 1	0.616	0.670	0.720	0.749	0.934	0.951	0.961	0.964	0.921	0.932	0.920	0.932	0.976	0.976	0.976	0.976
V = 2	0.642	0.694	0.746	0.772	0.936	0.952	0.960	0.963	0.922	0.936	0.928	0.932	0.973	0.972	0.972	0.972

TABLE 7. Simulation Setting 3: Observed coverage proportion in 1000 trials for intended nominal coverage 0.95

	KW				KWs				Efron				AKP-M			
	n=100	n=200	n=500	n=1000	n=100	n=200	n=500	n=1000	n=100	n=200	n=500	n=1000	n=100	n=200	n=500	n=1000
V = 0.1	0.858	0.928	1.023	1.077	1.762	1.850	1.915	1.928	1.667	1.633	1.585	1.582	2.725	2.777	2.802	2.810
V = 0.5	1.309	1.385	1.514	1.557	2.256	2.337	2.409	2.439	2.107	2.111	2.078	2.127	3.839	3.842	3.849	3.851
V = 1	1.518	1.640	1.747	1.808	2.515	2.610	2.666	2.679	2.355	2.385	2.345	2.376	4.076	4.078	4.079	4.079
V = 2	1.733	1.858	1.991	2.049	2.763	2.856	2.918	2.925	2.594	2.639	2.622	2.628	4.176	4.177	4.177	4.177

TABLE 8. Simulation Setting 3: Observed interval length in 1000 trials for intended nominal coverage 0.95

	KW				KWs				Efron				AKP-M			
	n=100	n=200	n=500	n=1000	n=100	n=200	n=500	n=1000	n=100	n=200	n=500	n=1000	n=100	n=200	n=500	n=1000
V = 0.1	0.212	0.194	0.182	0.177	0.215	0.200	0.190	0.186	0.202	0.188	0.179	0.176	0.393	0.386	0.382	0.380
V = 0.5	0.360	0.340	0.324	0.320	0.358	0.344	0.333	0.332	0.348	0.332	0.326	0.320	0.721	0.715	0.712	0.714
V = 1	0.438	0.422	0.406	0.397	0.433	0.424	0.417	0.410	0.424	0.412	0.410	0.400	0.827	0.826	0.825	0.823
V = 2	0.532	0.510	0.493	0.483	0.520	0.510	0.505	0.498	0.513	0.497	0.496	0.488	0.901	0.906	0.907	0.900

TABLE 9. Simulation Setting 3: Observed root mean squared error in 1000 trials of posterior mean

5. Conclusion

Viewed in retrospect it is hardly surprising that the Kiefer-Wolfowitz posterior intervals perform poorly. Their discrete character of the estimated posterior makes it almost impossible for them to adapt to a desired α level. What is more surprising about the foregoing exercise is the excellent performance of the Efron intervals. It might have been expected that since they do not account for variability in $\hat{\mathsf{G}}$, they too might under cover, but they are actually a little too conservative in all three settings of the experiment. It is tempting to suggest that the performance of the KW intervals could be ameliorated by some judicious smoothing of the KW $\hat{\mathsf{G}}$. But this would draw us back into the contested territory of tuning parameter selection and our four page limit has already been exceeded. Like his son's predictions about the length of the Iraq war, let's just concede that Jack Wolfowitz's intervals are just too short. The main take away from the experiments is that nonparametric G modeling not only produces good performance for point estimation and prediction, but it is also capable of delivering reliable estimates of the precision of such estimates. In contrast linear shrinkage and associated "least favorable" confidence interval procedures, as least as currently implemented in Armstrong et al. (2020) and the associated R package, has room for improvement on both of George Box's (1953) goals of "robustness of validity" and "robustness of efficiency."

References

Armstrong TB, Kolesár M, Plagborg-Møller M. 2020. Robust empirical Bayes confidence intervals. `http://arxiv.org/abs/2004.03448`.

Barron A, Shue C. 1991. Approximation of density functions by sequences of exponential families. *Annals of Statistics* **19**: 1347–1369.

Efron B. 2011. Tweedie's formula and selection bias. *Journal of the American Statistical Association* **106**: 1602–1614.

Efron B. 2016. Empirical Bayes deconvolution estimates. *Biometrika* **103**: 1–20.

Efron B. 2019. Bayes, oracle Bayes and empirical Bayes. *Statistical Science* **34**: 177–201.

Jiang W. 2019. Comment: Empirical Bayes interval estimation. *Statistical Science* **34**: 219–223.

Kiefer J, Wolfowitz J. 1956. Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *The Annals of Mathematical Statistics* **27**: 887–906.

Koenker R, Gu J. 2015. REBayes: An R package for empirical Bayes methods. Available from `http://cran.r-project.org`.

Koenker R, Gu J. 2019. Comment: Minimalist g-modeling. *Statistical Science* **34**: 209–213.

Koenker R, Mizera I. 2014. Convex optimization, shape constraints, compound decisions and empirical Bayes rules. *Journal of the American Statistical Association* **109**: 674–685.

Laird N. 1978. Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association* **73**: 805–811.

Lindsay B. 1983. The geometry of mixture likelihoods: A general theory. *The Annals of Statistics* **11**: 86–94.

Pfanzagl J. 1988. Consistency of maximum likelihood estimators for certain nonparametric families, in particular: mixtures. *Journal of Statistical Planning and Inference* **19**: 137–158.

Robbins H. 1950. A generalization of the method of maximum likelihood; estimating a mixing distribution (preliminary report). *The Annals of Mathematical Statistics* **21**: 314–315.

Stefanski LA, Carroll RJ. 1990. Deconvolving kernel density estimators. *Statistics* **21**: 169–184.

Stone CJ. 1990. Large-sample inference for log-spline models. *The Annals of Statistics* **18**: 717–741.