BAYESIAN DECONVOLUTION: AN R VINAIGRETTE

ROGER KOENKER

Abstract. Nonparametric maximum likelihood estimation of general mixture models pioneered by the work of Kiefer and Wolfowitz (1956) has been recently reformulated as an exponential family regression spline problem in Efron (2016). Both approaches yield a low dimensional estimate of the mixing distribution, g-modeling in the terminology of Efron. Some casual empiricism suggests that the Efron approach is preferable when the mixing distribution has a smooth density, while Kiefer-Wolfowitz is preferable for discrete mixing settings. In the classical Gaussian deconvolution problem both maximum likelihood methods appear to be preferable to (Fourier) kernel methods. Kernel smoothing of the Kiefer-Wolfowitz estimator appears to be competitive with the Efron procedure for smooth alternatives.

1. Introduction

Efron (2016) has recently introduced the phrase "Bayesian deconvolution" to describe a maximum likelihood procedure for estimating mixture models of the general 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 (iid) 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 Efron's maximum likelihood procedure recalls the NPMLE of Kiefer and Wolfowitz (1956) except rather than producing a discrete estimate of G it yields a smooth estimate.

This note contrasts the foregoing methods in a very simple, special case and argues that maximum likelihood offers considerable advantages over prior (Fourier) deconvolution methods, perhaps most significantly by extending the domain of applications beyond the location shift model.

Version: December 30, 2016. A genre manifesto for R Vinaigrettes is available at `http://davoidofmeaning.blogspot.com/2016/12/r-vinaigrettes.html`. It seemed appropriate that my first venture in this new genre should be samokritika (self-criticism) directed at 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. I would like to thank Jiaying Gu for very helpful comments on an earlier draft.

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).

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 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."

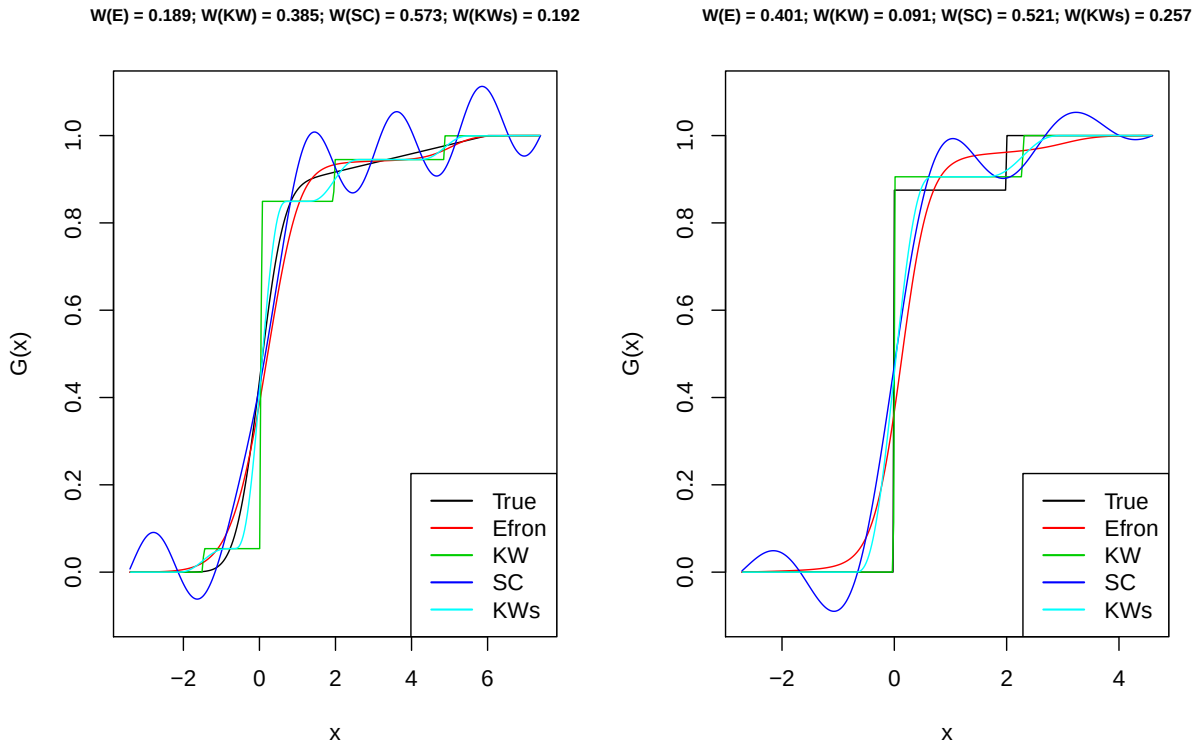

FIGURE 1. Four estimates of the mixing distributions G: In the left panel the true mixing distribution is smooth, in the left panel it is discrete as described in the text.

4. An Illustration

To compare performance of the three estimators of G described above, I have considered a slight variant of the simulation setting of Efron. The observed Y_i are $\mathcal{N}(\theta, 1)$ with θ_i's drawn iidly either a.) from the mixture of Gaussian and uniform distributions,

$$G(\theta) = (1-\epsilon)\Phi(\theta/\sigma) + \epsilon\theta I(0 \geqslant \theta < M)/M$$

with $\epsilon = 1/7$, $\sigma = 1/2$ and $M = 6$, or from b.) the discrete mixing distribution with $\epsilon = 1/7$ and,

$$G(\theta) = (1-\epsilon)I(0 \leqslant \theta) + \epsilon I(2 \leqslant \theta)$$

Figure 1 depicts typical realizations with sample size $n = 1000$. Following Efron we have set the dimension of the natural spline model to $p = 5$, and his penalty parameter to one. The scaling parameter for the Stefanski-Carroll procedure was 1/3, also following Efron's suggetion. For all three estimators the grid was equally spaced on the support of the observed Y_i with $m = 300$ distinct values. Wasserstein distance, L_1 distance between distribution functions, is reported above the figure for each of three estimates.

The performance of Efron's estimator is very impressive, while the oscillation of the kernel method in the tails confirms Efron's criticism. The Kiefer-Wolfowitz estimator is respectable, at least the estimated $\hat{\mathsf{G}}$ stays within the [0,1] bounds, but it is clearly inferior to the smoother Efron procedure. When the target G is discrete with a small number of mass points, the performance advantage, not surprisingly is reversed. One might count the lack of tuning parameters for the Kiefer-Wolfowitz NPMLE as an advantage over its competitors, or not, depending on one's outlook on minimalism.

In the spirit of competition, I couldn't resist trying to smooth the Kiefer Wolfowitz NPMLE to see whether one might be able to approach the performance of the Efron estimator, so the last (cyan) curve is a biweight kernel smooth with bandwidth equal 0.7. This does almost as well as the default Efron procedure for our test case for the smooth alternative, but spoils the auspicious performance for the discrete case. Encouraged by the former improvement I decided to wade a little further out in the water by replicating the experiment. In 1000 trials of the experiment with the smooth G_0 the mean Wasserstein error was 0.186 for the Efron estimator, 0.342 for the unsmoothed KW-NPMLE and 0.18 for the smoothed KW-NPMLE. Of course, this result proves nothing at all, except perhaps that I'm acquainted with an excellent bandwidth oracle.

References

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

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

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, 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.

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.