SHAPE CONSTRAINTS, COMPOUND DECISIONS AND EMPIRICAL BAYES RULES

ROGER KOENKER AND IVAN MIZERA

Abstract. A shape constrained version of the Brown and Greenshtein (2009) empirical Bayes Rule is very briefly described and some simulation comparisons are presented, including two variants of the generalized non-parametric maximum likelihood Bayes rule recently proposed by Jiang and Zhang (2009).

1. Introduction

Brown and Greenshtein (2009) consider the classical compound decision problem of estimating a vector $\mu = (\mu_1, \cdots, \mu_n)$ of parameters based on a conditionally Gaussian sample, $Y_i \sim \mathcal{N}(\mu_i, 1) \quad i = 1, \cdots, n$. Motivated by empirical Bayes considerations, the μ_i's are assumed to be drawn from a distribution F; so that the Y_i's have the mixture density

$$g(y) = \int \phi(y - \mu) dF(\mu).$$

They propose the decision rule,

$$\delta(y) = y + \frac{\hat{g}'(y)}{\hat{g}(y)}, \tag{1}$$

and show that kernel estimates of g can be employed to achieve attractive performance.

More recently, Brown, Greenshtein, and Ritov (2010) have reconsidered the even more classical problem of estimating a vector of Poisson intensity parameters, $\lambda = (\lambda_1, \cdots, \lambda_n)$ from a sample of $Y_i \sim Po(\lambda_i)$. Noting the connection between the two problems, they remark that monotone likelihood ratio considerations dictate that $\delta(y)$ in (1) should be monotonically increasing in y. This suggests that one may wish to consider density estimation methods for g that restrict,

$$K(y) = \tfrac{1}{2} y^2 + \log g(y) \tag{2}$$

to be convex. The initial intent of this brief note was to explore this idea and its consequences for the performance of the associated Bayes rules. The work of van Houwelingen and Stijnen (1983) provides a cogent discussion of the implication of monotonicity for Bayes rules in compound decision problems like the Gaussian case described above, and they suggest a greatest convex minorant estimator based on a preliminary histogram-type estimate of the density, g. In contrast, our approach is closely related to recent work on maximum likelihood estimation of log-concave densities.

Version: March 3, 2011. This research was partially supported by NSF grant SES-08-50060. The use of shape constrained likelihood methods for compound decision rules was originally suggested to the first author by Larry Brown. The coffee stain is intended to stress the preliminary nature of the results.

Nonparametric maximum likelihood estimation as originally proposed in Kiefer and Wolfowitz (1956) has recently been shown by Jiang and Zhang (2009) to yield excellent performance for this class of Gaussian compound decision problems. In Section 4 we briefly explore an alternative computational strategy to the EM algorithm approach of Jiang and Zhang (2009) that seems to deliver even better performance with reduced computational cost.

2. Shape Constrained Density Estimation

Recent work on shape constrained density estimation has focused primarily on imposing log concavity, see, e.g. Cule, Samworth, and Stewart (2010), Dümbgen and Rufibach (2009) and Koenker and Mizera (2010). In the present circumstance we have a somewhat different constraint, but the essential nature of the problem is very similar. Following the development in Koenker and Mizera (2010) we can consider maximizing the log likelihood,

$$\sum_{i=1}^{n} \log g(y_i),$$

over densities g satisfying the convexity constraint (2). Writing $h(y) = \log g(y)$, this task can be concisely expressed as,

$$\max_h \{\sum h(y_i) - \int e^{h(y)} dy \mid K \in \mathcal{K}\},$$

where $\mathcal{K}$ denotes the cone of convex functions on $\mathbb{R}$. As in the log-concave case, it can be shown that solutions are characterized by K being piecewise linear with knots at the observed y_i.

This characterization leads to a finite dimensional formulation, setting $\alpha_i = h(y_i)$ for the estimated function values at the knots,

$$[P] \qquad \max_\alpha \{w^\top \alpha - \sum c_i e^{\alpha_i} \mid D\alpha + 1 \geq 0\}.$$

The matrix D represents the finite difference version of the second derivative operator that appears in the variational form of the estimation problem. Here, as in Koenker and Mizera (2010), the accuracy of the Riemann approximation of the integral can be controlled by introducing pseudo observations, or "undata," on a fine grid, thereby increasing the number of estimated function values. When this is done, the vector w above becomes an evaluation operator that simply allows us to recover and sum up the contributions to the likelihood given the expanded vector of function values.

The primal problem (2) has corresponding dual problem, writing the diagonal matrix with the Riemann weights c_i as C,

$$[D] \qquad \min_\nu \{\sum c_i g_i \log g_i + 1^\top \nu \mid g = C^{-1}(w + D^\top \nu),\ \nu \geq 0\}.$$

Here the dual vector ν can be viewed as characterizing the Radon measure that determines the estimated density via the equality constraint. As with the log concave constraint the dual reveals that we are minimizing Shannon entropy, or the Kullback-Leibler distance between the estimated density and a uniform density on the support of the empirical distribution of the observations – except that the objective now contains a linear contribution not present in the log concave formulation.

3. IMPLEMENTATION

The dual form of the estimator has been implemented using two independent convex programming algorithms employing interior point methods: the PDCO algorithm of Saunders (2003) and the Mosek methods of Andersen (2010). In Figure 1 we plot a typical realization of the constrained density estimate and its corresponding K estimate. This example is based on a sample of size 100 from the model described in the introduction with the μ_i's drawn from the uniform $U[5, 15]$ distribution. The Mosek solution has been plotted in black in both figures, with the PDCO solution overplotted in red. If you look carefully in the tails of the K plot you may see a slight difference, but the estimates are for all practical purposes identical.

For those accustomed to looking at log-concave density estimates, the density plot of Figure 1 is likely to seem rather bizarre, but such are the vagaries of the convexity constraint $K \in \mathcal{K}$ that has been imposed. The fitted $\hat{K}$ is piecewise linear and consequently $\log \hat{g}$ must be piecewise quadratic. As estimates of the mixture density, g, such estimates are unlikely to win any accolades, but their implied Bayes rules nevertheless perform quite well as we shall see in Section 5.

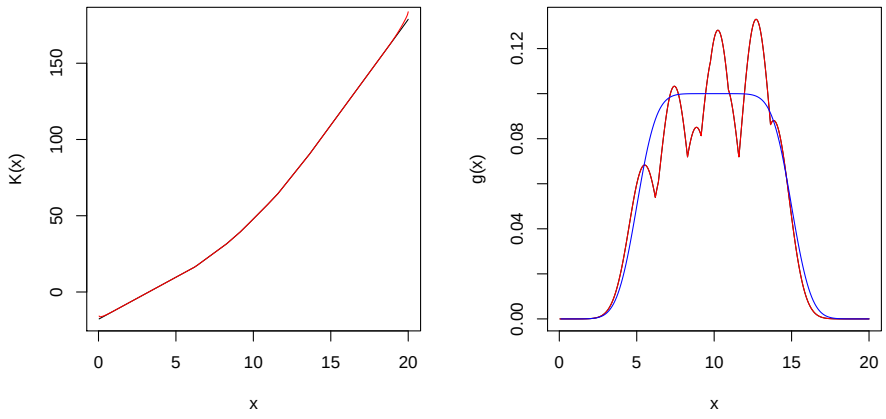

FIGURE 1. Estimated density and K function for a simple compound decision problem. Mosek and PDCO solutions essentially coincide; the target mixture density is plotted in blue.

4. NON-PARAMETRIC MAXIMUM LIKELIHOOD

Jiang and Zhang (2009) have recently suggested a variant of the Kiefer and Wolfowitz (1956) non-parametric maximum likelihood estimator as another approach to estimation of an empirical Bayes rule for the problem posed above. They suggest an EM approach to the computation and report good performance in simulations that follow the design of Johnstone and Silverman (2004). In an effort to extend the range of comparisons offered in the present note, we undertook to explore this approach a bit further.

Jiang and Zhang (2009) employ a simple fixed point EM iteration for their estimator. Given a grid $\{u_1, ...u_m\}$ containing the support of the observed sample, they

define the sequence,

$$\hat{f}_j^{(k+1)} = n^{-1} \sum_{i=1}^{n} \frac{\hat{f}_j^{(k)} \varphi(Y_i - u_j)}{\sum_{\ell=1}^{m} \hat{f}_\ell^{(k)} \varphi(Y_i - u_\ell)},$$

where $\varphi(\cdot)$ denotes the standard Gaussian density, and $\hat{g}_j^{(k)}$ denotes the value of the estimated "prior" density on the interval (u_j, u_{j+1}) at the kth iteration. In their simulations, following the design of Johnstone and Silverman (2004), the sample size is $n = 1000$ and the u_i's are equally spaced with $m = 1000$. At the conclusion of the iteration,

$$\hat{\delta}(Y_i) = \frac{\sum_{j=1}^{m} u_j \varphi(Y_i - u_j) \hat{g}_j}{\sum_{j=1}^{m} \varphi(Y_i - u_j) \hat{g}_j}.$$

This procedure performs very well as documented by Jiang and Zhang (2009) and confirmed in our independent simulation experiments reported below, but the EM iterations while simple are quite slow for moderately large sample sizes and make very lethargic progress toward their objective of maximizing the log likelihood,

$$L(g) = \sum_{i=1}^{n} \log(\sum_{j=1}^{m} \varphi(Y_i - u_j) g_j).$$

A variety of schemes to accelerate the EM iterations along the lines of Varadhan and Roland (2008) and Berlinet and Roland (2007) were investigated, which while helpful did not significantly improve the computational efficiency.

Having used interior point convex optimization methods for our shape constrained estimator, it eventually occurred to us that the task of maximizing $L(g)$ was also susceptible to such methods. For a fixed grid, $\{u_1, ..., u_m\}$ as above, denote the n by m matrix, $A = (\varphi(Y_i - u_j))$, and consider the problem,

$$\min\{-\sum_{i=1}^{n} \log(y_i) \mid Az = y,\ z \in \mathcal{S}\},$$

where $\mathcal{S}$ denotes the simplex in $\mathbb{R}^m$, i.e. $\mathcal{S} = \{s \in \mathbb{R}^m | 1^\top s = 1,\ s \geq 0\}$. The fixed grid is an important proviso, since as the grid becomes finer the solution tends to point masses, as noted by Laird (1978) and other authors. Nevertheless, for the present purpose of estimating an effective Bayes rule, a relatively fine fixed grid like that used for the EM iterations, seems entirely satisfactory.

In Figure 2 we compare the "solutions" produced by the interior point algorithm with those of the EM iteration for various limits on the number of iterations. For the test problem for this figure we have employed a structure similar to that of the simulations conducted in the following section. There is a sample of $n = 200$ observations, and a grid of $m = 300$ equally spaced points; 15 of the observations have $\mu_i = 2$, while the remainder have $\mu_i = 0$. It is obviously difficult to distinguish this mixture, yet remarkably the procedure does find, in this particular sample, a mass point near 2, as well as much more significant mass point near 0. A spurious mass point near -1 is also identified.

In Table 1 we report timing information and the values of $L(g)$ achieved for the four procedures illustrated in the figure. Although the EM procedure makes steady progress toward its goal, it leaves something to be desired even after 100,000 iterations, and nearly 10 minutes of computation. In contrast the interior point algorithm implemented in Mosek is quite quick and accurate.

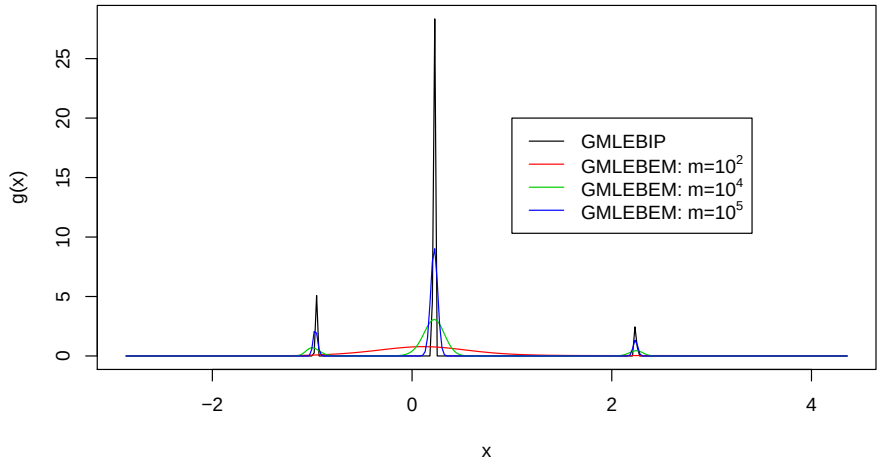

FIGURE 2. Comparison of estimates of the mixture density g: The solid black very peaked density is the interior point solution, GMLEBIP, the others as indicated by the legend represent the EM solutions with various iteration limits. See Table 1 for further details on timings and achieved likelihoods.

Estimator	EM1	EM2	EM3	IP
Iterations	100	10,000	100,000	15
Time	1	37	559	1
L(g) - 422	0.9332	1.1120	1.1204	1.1213

TABLE 1. Comparison of EM and Interior Point Solutions: Iteration counts, log likelihoods and CPU times (in seconds) for three EM variants and the interior point solver.

An alternative interior point algorithm IPOPT described in Wächter and Biegler (2006) was used to check the solutions of the Mosek implementation, but proved to be much less computationally efficient than the Mosek procedure.

5. SIMULATION PERFORMANCE

To compare performance of the shape constrained estimator with other methods we have replicated the experiment described in Johnstone and Silverman (2004), and also employed in Brown and Greenshtein (2009). In this setup the μ_i's have a simple discrete structure: there are $n = 1000$ observations, k of which have μ equal to one of the 4 values $\{3, 4, 5, 7\}$, the remaining $n - k$ have $\mu = 0$. There are three choices of k as indicated in the table. Table 2 reports results of the experiment. Each entry in the table is a sum of squared errors over the 1000 observations, averaged over the number of replications. Johnstone and Silverman (2004) evaluated 18 different procedures; the last line of the table reports the best performance achieved in their experiment for each column setting. The performance of the Brown and Greenshtein (2009) kernel based rule is given in the fourth line of the table, taken from their Table 1. Two variants of the GMLEB procedure of Jiang and Zhang (2009) appear in the second

Estimator	k = 5				k = 50				k = 500			
	μ =3	μ =4	μ =5	μ =7	μ =3	μ =4	μ =5	μ =7	μ =3	μ =4	μ =5	μ =7
$\hat{\delta}$	37	34	21	11	173	121	63	16	488	310	145	22
$\hat{\delta}_{GMLEBIP}$	33	30	16	8	153	107	51	11	454	276	127	18
$\hat{\delta}_{GMLEBEM}$	37	33	21	11	162	111	56	14	458	285	130	18
$\tilde{\delta}_{1.15}$	53	49	42	27	179	136	81	40	484	302	158	48
J-S Min	34	32	17	7	201	156	95	52	829	730	609	505

TABLE 2. Risk of Shape Constrained Rule, $\hat{\delta}$ compared to: two versions of Jaing and Zhang's GMLEB procedure, one using 100 EM iterations denoted GMLEBEM and the other, GMLEBIP, using the interior point algorithm described in the text, the kernel procedure, $\tilde{\delta}_{1.15}$ of Brown and Greenshtein, and best procedure of Johnstone and Silverman. Sum of squared errors in $n = 1000$ observations. Reported entries are based on 1000, 100, 100, 50 and 100 replications, respectively.

Estimator	n = 10,000			n = 100,000		
	k =100	k =300	k =500	k =500	k =1000	k =5000
$\hat{\delta}$	282	736	1136	1405	2659	10930
$\tilde{\delta}_{1.05}$	306	748	1134	2410	3810	10400
Oracle	295	866	1430	3335	5576	16994

TABLE 3. Risk of Shape Constrained Rule, $\hat{\delta}$ compared to kernel procedure, $\tilde{\delta}_{1.05}$ of Brown and Greenshtein, and oracle hard threshholding rule. Reported entries are sums of squared errors for the n observations averaged over 1000, 50 and 50 replications, respectively.

and third rows of the table. GMLEBEM is the original proposal as implemented by Jiang and Zhang (2009) using 100 iterations of the EM fixed point algorithm, GMLEBIP is the interior point version iterated to convergence as determined by the Mosek defaults. The shape constrained estimator described above, denoted $\hat{\delta}$ in the table, is reported in the first line. The $\hat{\delta}$ results are based on 1000 replications, the GMLEB results on 100 replications, the Brown and Greenshtein results on 50 replications, and the Johnstone and Silverman results on 100 replications.

It seems fair to say that the shape constrained estimator performs competitively in all cases, but is particularly attractive relative to the kernel rule and the Johnstone and Silverman options in the moderate k and large μ settings of the experiment. However, the GMLEB rules have a clear advantage when k is 50 and 500.

To explore performance for smoother mixing distributions F, we briefly reconsider a second example drawn from Brown and Greenshtein (2009). The mixture distribution F has a point mass at zero, and a uniform component on the interval $[-3, 3]$. Two sample sizes are considered, $n = 10,000$ and $n = 100,000$. In the former case we consider 100, 300, and 500 uniforms, in the latter case there are 500, 1000, and 5000. In Table 3 we report performance of the shape constrained estimator and compare it with the kernel estimator, now with bandwidth 1.05, and a "strong oracle" estimator both taken from Brown and Greenshtein (2009). The latter procedure is a hard-thresholding rule which takes $\delta(X)$ to be either 0 or X depending on whether $|X| > C$ for an optimal choice of C. Since the shape constrained estimator is quite quick we have done 1000 replications, while the other reported values are based on 50 replications as reported in Brown and Greenshtein (2009). As in the preceeding table

the reported values are the sum of squared errors over the n observations, averaged over the number of replications. Again, the shape constrained estimator performs quite satisfactorily, while circumventing difficult questions of bandwidth selection.

Unfortunately, given the dense form of the constraint matrix A, GMLE methods are not really feasible for sample sizes like those of the experiments reported in Table 3. Solving a single problem with $n = 10,000$ requires about one hour using the Mosek interior point algorithm. Strategies for improving the scalability of the GMLE procedure would seem to be an interesting research problem; it is known from Zhang (2009) that solutions are representable as low dimensional discrete mixtures and therefore more efficient algorithm designs seem entirely possible.

6. Provisional Conclusions

We have seen that empirical Bayes rules based on maximum likelihood estimation of Gaussian mixture densities subject to a shape constraint imposed to achieve monotonicity of the Bayes rule provide some performance improvements over unconstrained kernel based estimation methods. Likewise, Kiefer-Wolfowitz nonparametric maximum likelihood estimation of the mixing distribution offers good performance in our simulation settings. We have also seen that the computational burden of the EM implementation of the Kiefer-Wolfowitz estimator can be significantly reduced by reliance on interior point methods. However, for large sample sizes even the interior point algorithms are painfully slow. The skeptical reader may wonder at this point how it is that the shape constrained maximum likelihood estimator scales so well with sample size while the Kiefer-Wolfowitz procedure does not. How is it that we can do 1000 replications of *both* the $n = 10,000$ and $n = 100,000$ cases for Table 3 for the shape constrained estimator in about 20 minutes, while one replication of the Kiefer-Wolfowitz estimator for $n = 10,000$ takes an hour? The secret lies in the sparsity of the constraint matrix in the shape-constrained case, versus the density of the constraint matrix in the Kiefer Wolfowitz problem. Sparse linear algebra is a remarkably powerful tool when it is applicable.

References

Andersen, E. D. (2010): "The MOSEK Optimization Tools Manual, Version 6.0," Available from `http://www.mosek.com`.

Berlinet, A., and C. Roland (2007): "Acceleration Schemes with Application to the EM Algorithm," *Computational Statistics and Data Analysis*, 51, 3689–3702.

Brown, L., and E. Greenshtein (2009): "Non parametric empirical Bayes and compound decision approaches to estimation of a high dimensional vector of normal means," *The Annals of Statistics*, 37(4), 1685–1704.

Brown, L., E. Greenshtein, and Y. Ritov (2010): "The Poisson Compound Decision Problem Revisited," *Preprint arXiv:1006.4582*.

Cule, M., R. Samworth, and M. Stewart (2010): "Computing the Maximum Likelihood Estimator of a Multidimensional Log-Concave Density, with discussion," *Journal of the Royal Statistical Society, B.*, 72, 545–600.

Dümbgen, L., and K. Rufibach (2009): "Maximum Likelihood Estimation of a Log-Concave Density: Basic Properties and Uniform Consistency," *Bernoulli*, 15, 40–68.

Jiang, W., and C. Zhang (2009): "General maximum likelihood empirical Bayes estimation of normal means," *The Annals of Statistics*, 37, 1647–1684.

Johnstone, I., and B. Silverman (2004): "Needles and straw in haystacks: Empirical Bayes estimates of possibly sparse sequences," *Annals of Statistics*, pp. 1594–1649.

KIEFER, J., AND J. WOLFOWITZ (1956): "Consistency of the Maximum Likelihood Estimator in the Presence of Infinitely Many Incidental Parameters," *The Annals of Mathematical Statistics*, 27, pp. 887–906.

KOENKER, R., AND I. MIZERA (2010): "Quasi-concave density estimation," *The Annals of Statistics*, 38(5), 2998–3027.

LAIRD, N. (1978): "Nonparametric Maximum Likelihood Estimation of a Mixing Distribution," *Journal of the American Statistical Association*, 73, pp. 805–811.

SAUNDERS, M. A. (2003): "PDCO: A Primal-Dual interior solver for convex optimization," `http://www.stanford.edu/group/SOL/software/pdco.html`.

VAN HOUWELINGEN, J., AND T. STIJNEN (1983): "Monotone empirical Bayes estimators for the continuous one-parameter exponential family," *Statistica Neerlandica*, pp. 29–43.

VARADHAN, R., AND C. ROLAND (2008): "Simple and Globally Convergent Methods for Acceleratingi the Convergence of any EM Algorithm," *Scandinavian Journal of Statistics*, 35, 335–353.

WÄCHTER, A., AND L. BIEGLER (2006): "On the implementation of a primal-dual interior point filter line search algorithm for large-scale nonlinear programming," *Mathematical Programming*, 106, 25–57.

ZHANG, C.-H. (2009): "Generalized maximum likelihood estimation of normal mixture densities," *Statistica Sinica*, 19, 1297–1318.