CONVEX OPTIMIZATION, SHAPE CONSTRAINTS, COMPOUND DECISIONS, AND EMPIRICAL BAYES RULES

ROGER KOENKER AND IVAN MIZERA

Abstract. Estimation of mixture densities for the classical Gaussian compound decision problem and their associated (empirical) Bayes rules is considered from two new perspectives. The first, motivated by Brown and Greenshtein (2009), introduces a nonparametric maximum likelihood estimator of the mixture density subject to a monotonicity constraint on the resulting Bayes rule. The second, motivated by Jiang and Zhang (2009), proposes a new approach to computing the Kiefer-Wolfowitz nonparametric maximum likelihood estimator for mixtures. In contrast to prior methods for these problems, our new approaches are cast as convex optimization problems that can be efficiently solved by modern interior point methods. In particular, we show that the reformulation of the Kiefer-Wolfowitz estimator as a convex optimization problem reduces the computational effort by *several orders of magnitude for typical problems*, by comparison to prior EM-algorithm based methods, and thus greatly expands the practical applicability of the resulting methods. Our new procedures are compared with several existing empirical Bayes methods in simulations employing the well-established design of Johnstone and Silverman (2004). Some further comparisons are made based on prediction of baseball batting averages. A Bernoulli mixture application is briefly considered in the penultimate section.

1. Introduction

The recent revival of interest in empirical Bayes methods for compound decision problems, e.g. Brown (2008), Brown and Greenshtein (2009), Efron (2010, 2011), and Jiang and Zhang (2009, 2010), has renewed interest in computational methods for nonparametric mixture models. Our primary objective for this paper is to demonstrate the constructive role that modern convex optimization methods can play in this context, both from practical and theoretical viewpoints.

Version: November 20, 2013. Roger Koenker is McKinley Professor of Economics and Professor of Statistics at the University of Illinois, Urbana, IL 61801 (Email: rkoenker@uiuc.edu). Ivan Mizera is Professor of Statistics at the University of Alberta, Edmonton, Alberta T6G 2G1, Canada (Email: imizera@ualberta.ca). This research was partially supported by NSF grant SES-11-53548 and the NSERC of Canada. The authors wish to thank Larry Brown for pointing out the relevance of shape constrained density estimation to compound decision problems, and Steve Portnoy for extensive related discussions.

With this objective in mind, we focus primarily on the classical compound decision problem of estimating an n-vector $(\mu_1, \cdots, \mu_n)$ of parameters, under squared error loss, based on a conditionally Gaussian random sample, $Y_i \sim \mathcal{N}(\mu_i, 1) \quad i = 1, \cdots, n$, where the μ_i's are assumed to be drawn iid-ly from a distribution F; so that the Y_i's have the mixture density

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

For known F, and hence known g, the optimal prediction of the μ's, under squared error loss, is given by the Bayes rule

$$\delta(y) = \mathbb{E}(\mu | Y = y) = y + \frac{g'(y)}{g(y)}. \tag{1}$$

This formula is generally attributed to Robbins (1956), but Efron (2011) notes that Robbins attributes it to M.C.K. Tweedie. Efron's paper offers a broad perspective on the significance, and limitations, of its use in empirical Bayes applications; in particular, it points out that the exponential family argument of van Houwelingen and Stijnen (1983) immediately ensures that (1) is nondecreasing in y, not only for the conditional Gaussian case, but for more general one-parameter exponential families as well, regardless of the prior distribution F.

From the practical point of view, Brown and Greenshtein (2009) show that simple kernel density estimates of g can be employed to achieve attractive performance relative to some other empirical Bayes procedures. However, the monotonicity of the decision rule $\delta(y)$ suggests that also its estimated version, $\hat{\delta}(y)$, be monotone too—or equivalently, that

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

be convex. As the unconstrained kernel density estimates do not deliver this, van Houwelingen and Stijnen (1983) suggest a greatest convex minorant estimator based on a preliminary histogram-type estimate of the density, g. Rather than taking this route and starting from a kernel estimate or some form of histogram estimate we consider instead *direct* nonparametric maximum likelihood estimation of the mixture density subject to the constraint (2); this not only ensures monotonicity, but also eliminates a need to tune the bandwidth or equivalent quantity. The shape constraint itself is a sufficient regularization device. A close link can be traced here to recent work on maximum likelihood estimation of log-concave densities and other shape constrained estimation problems.

While our original intent might have been to explore this idea and its consequences for the performance of the associated empirical Bayes rules,

our second and in our view equally promising focus has been motivated by the work of Jiang and Zhang (2009), who show that good predictive performance for the class of Gaussian compound decision problems can be achieved by an implementation of the nonparametric maximum likelihood estimator originally proposed by Kiefer and Wolfowitz (1956). The implementation of Jiang and Zhang (2009) employed the EM algorithm, a strategy initially proposed by Laird (1978). In place of this, we introduce in Section 3 an alternative computational strategy based on convex optimization—which delivers better predictive performance than that observed by Jiang and Zhang (2009) based on EM, at dramatically reduced computational cost. Contrary to our first proposal that directly estimates the decision rule, the Kiefer-Wolfowitz estimator estimates rather the prior F; the implied decision rule, obtained by substituting this estimate for F, is automatically monotone—as the latter property holds regardless of the form of the prior. Moreover, our new implementation of the Kiefer-Wolfowitz estimator applies also to a broad class of mixture problems; it includes the standard exponential family problems, but extends well beyond them, thus significantly increasing its appeal.

2. Shape Constrained Density Estimation

The recent upsurge in work on shape constrained estimation has explored a wide variety of settings, but work on shape constrained density estimation has focused primarily on imposing log concavity and some weaker forms of concavity restrictions—see, e.g. Cule, Samworth, and Stewart (2010), Dümbgen and Rufibach (2009), Koenker and Mizera (2010), and Seregin and Wellner (2010). Monotonicity of the Bayes rule would seem to be a rather different constraint, but the essential nature of the problems is very similar. As in the development in Koenker and Mizera (2010), we consider maximizing the log likelihood,

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

over g; to ensure that the result is a probability density, we reparametrize the problem in terms of $h = \log(g)$, so that $g = e^h$ is automatically positive, and add a Lagrange term so that g integrates to 1. The whole task then involves maximizing

$$\sum_{i=1}^{n} h(Y_i) - \int e^{h(y)} dy$$

under the convexity constraint enforcing $\frac{1}{2}y^2 + h(y)$ to be in $\mathcal{K}$, the cone of convex functions on $\mathbb{R}$. Evidently, this constraint is equivalent to the

requirement that

$$K(y) = \log\sqrt{2\pi} + \tfrac{1}{2}y^2 + h(y)$$

is in $\mathcal{K}$. Rewriting in terms of K gives the objective function

$$\sum_{i=1}^{n} K(Y_i) + \sum_{i=1}^{n} \log\varphi(Y_i) - \int e^{K(y)}\varphi(y)dy.$$

If the optimization task is expressed in the minimization form, we obtain after omitting the constant terms the formulation,

$$\min_K\{-\sum K(Y_i) + \int e^{K(y)}d\Phi(y) \mid K \in \mathcal{K}\}, \tag{3}$$

where Φ denotes the standard Gaussian distribution function. This form differs from the prescription of Koenker and Mizera (2010) for the estimation of a log-concave density only in the sign of K, and correspondingly in the requirement that K be convex rather than concave; and in that the integration measure is $d\Phi(y) = \varphi(y)dy$ rather than dy. While it is fairly well-known in the literature that this does not necessarily imply the same form for the solutions $\hat{K}$, one might expect at least a similar one; however, the following theorem stipulates that although in both instances the solutions are piecewise linear, the knots, the breakpoints of linearity in $\hat{K}$, do not necessarily occur at the observed y_i, unlike in Theorem 2.1 of Koenker and Mizera (2010) for the log concave case. As we are looking rather for a convex interpolant with minimal integral for the concave majorant (and thus interpolant) with minimal integral, the results exhibit analogous properties to those of Groeneboom, Jongbloed, and Wellner (2001), who considered maximum likelihood estimation of a probability density (with respect to dy) that is convex and decreasing however, our problem, when recast as (3), asks for the convexity of the *logarithm* of the density (with respect to $d\Phi$); so, unfortunately, their results are not directly applicable to our case and we have furnished a brief, independent proof.

The similarity of (3) to the prescription of Koenker and Mizera (2010) is advantageous in another respect: since the objective function of (3) is convex, and minimized over the convex cone $\mathcal{K}$, the problem is convex, enabling us to provide a valuable dual formulation.

Theorem 1. *The solution,* $\hat{K}$*, of* (3) *exists and is piecewise linear. It admits a dual characterization:* $e^{\hat{K}(y)} = \hat{f}$*, where* $\hat{f}$ *is the solution of*

$$\max_f\{-\int f(y)\log f(y)d\Phi(y) \mid f = \frac{d(P_n - G)}{d\Phi}, G \in \mathcal{K}^-\}, \tag{4}$$

where $\mathcal{K}^-$ denotes the polar cone associated with $\mathcal{K}$, see e.g. Rockafellar (1970). The estimated decision rule, $\hat{\delta}$, is piecewise constant and has no jumps at $\min Y_i$ and $\max Y_i$.

Remark 1. Since $\hat{K}(y)$ is piecewise linear and convex, the resulting Bayes rule, $\hat{\delta}(y) = \hat{K}'(y)$, is piecewise constant and non-decreasing, yielding an "empirical decision rule," analogous to the empirical distribution function and determined by maximum likelihood.

Remark 2. Note that $\hat{g}(y) = \varphi(y)e^{\hat{K}(y)} = \varphi(y)\hat{f}$. In implementations we have generally found that it is more numerically stable and computationally efficient to work with the dual formulation and this is true here as well. Expressing the negative of the dual objective function in terms of g,

$$\int \left(\frac{g(y)}{\varphi(y)} \log \frac{g(y)}{\varphi(y)} \right) \varphi(y) dy = \int g(y) \log \frac{g(y)}{\varphi(y)} dy$$

reveals that the dual minimizes the Kullback-Leibler divergence between g and φ, under the constraint

$$g = \frac{d(P_n - G)}{dy}, \quad G \in \mathcal{K}^-.$$

This formulation can be further viewed as minimizing

$$\int g(y) \log g(y) - g(y) \log \varphi(y) dy = \int g(y) K(y) dy$$

and after the elimination of the constant $\log \sqrt{2\pi}$ as maximizing

$$-\int g(y) \log g(y) dy - \tfrac{1}{2} \int y^2 g(y) dy. \tag{5}$$

The latter is equivalent, in view of the independence of

$$\int y g(y) dy = \int y d(P_n - G)(y) = \int y dP_n(y) - \int y dG(y) = \frac{1}{n} \sum_{i=1}^{n} Y_i$$

on g (the last term vanishing due to the fact that $G \in \mathcal{K}^-$), to the maximization of $H(g) - \frac{1}{2} Var(g)$, where

$$H(g) = -\int g(y) \log g(y) dy$$

and

$$Var(g) = \int y^2 g(y) dy - (\int y g(y) dy)^2.$$

Remark 3. The intervals of linearity of $\hat{K}$ include $(-\infty, a)$ and $[b, +\infty)$ for some $a \geqslant \min\{Y_i\}$ and $b \leqslant \max\{Y_i\}$. Thus, unlike the case of log concave density estimators that vanish off the convex hull of the observations, solutions, $\hat{g}$ have exponential tails.

Remark 4. What do densities, g, such that $K(y) = -\log \varphi(y) + \log g(y)$ is piecewise linear and convex, look like? It is easy to show that such densities have the form,

$$g(y) = c \exp\{-\tfrac{1}{2} \sum_{j=1}^{m} (y - \mu_j)^2 I_{(a_j, a_{j+1}]}(y)\}$$

for some choice of constants, c and $\{(\mu_j, a_j) : j = 1, \cdots, m\}$. Such densities may also be expressed as mixtures of truncated normal densities, and as such do not fully comply with the features of the Gaussian mixture model. A more fully compliant estimator is introduced in the next section.

We defer further implementation details regarding the discretization of the problem to Section 4, and proceed immediately to the discussion of our second approach based on the Kiefer-Wolfowitz nonparametric MLE.

3. NON-PARAMETRIC MAXIMUM LIKELIHOOD

Jiang and Zhang (2009) have recently proposed a variant of the classical Kiefer and Wolfowitz (1956) nonparametric MLE as another promising approach to estimation of an empirical Bayes rule for the Gaussian compound decision problem.

In its original, infinite-dimensional, formulation the Kiefer and Wolfowitz (1956) nonparametric MLE solves,

$$\min \left\{ -\sum_{i=1}^{n} \log \left(\int \varphi(Y_i - \mu) dF(\mu) \right) \right\}, \tag{6}$$

where F runs over all mixing distributions. Once again, the objective function is convex, and is minimized over a convex set of F; hence again we have a convex problem, and duality theory is again applicable.

Theorem 2. *The solution, $\hat{F}$, of* (6) *exists, and is an atomic probability measure, with not more than n atoms. The locations, $\hat{\mu}_j$, and the masses, $\hat{f}_j$, at these locations can be found via the following dual characterization: the solution, $\hat{v}$, of*

$$\max\{\sum_{i=1}^{n} \log v_i \mid \sum_{i=1}^{n} v_i \varphi(Y_i - \mu) \leqslant n \text{ for all } \mu\} \tag{7}$$

satisfies the extremal equations (n *equations in less than* n *variables)*

$$\sum_j \varphi(Y_i - \hat{\mu}_j)\hat{f}_j = \frac{1}{\hat{\nu}_i}, \tag{8}$$

and $\hat{\mu}_j$ *are exactly those* μ *where the dual constraint is active—that is, the constraint function in* (7) *is equal to* n*.*

It may be interesting to compare the nonparametric MLE to the shape-constrained approach proposed the previous section; while the latter results in the piecewise constant "empirical decision rule", the nonparametric MLE rather comes with an "empirical prior distribution": a piecewise constant cumulative empirical distribution for the estimated prior.

Despite the number n appearing in the theorem, the atoms of $\hat{F}$ are not necessarily located at the datapoints; this is clear already in the example of Laird (1978) (p. 809), whose theoretical underpinning can be derived from the facts established in Theorem 2. Note that although the primal formulation is infinite-dimensional in the objective (in F), the objective of the dual formulation is finite dimensional (in ν), and infinite-dimensionality appears only in the constraint. This offers a potential for certain refinements: instead of a uniformly spaced grid supporting an atomic measure meant to approximate F, we could instead work with an adaptive (and not necessarily uniformly spaced) collection of test points where the dual constraint is enforced. In fact, if we knew the locations of maxima for the function appearing in the constraint, we could simply select these test points at these locations. Such information is typically unavailable, but practical implementations may seek to refine the solution in an iterative manner by refining the grid in regions identified by preliminary estimation.

We defer consideration of such refinements for future work since the fine grids described in the next section yield sufficient accuracy from most practical points of view. It is clear that implementations of the primal problem must discretize, thereby restricting F to a finite-dimensional approximation, as in the EM proposals of Laird (1978), and Heckman and Singer (1984). Following these earlier authors, Jiang and Zhang (2009) proposed a fixed point EM iteration that requires a grid $\{u_1, \ldots u_m\}$ containing the support of the observed sample, yielding a 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 $\hat{f}_j^{(k)}$ denotes the value of the estimated "prior" mixing density on the interval (u_j, u_{j+1}) at

the kth iteration. At the conclusion of the iteration, the decision rule is again, simply, the conditional expectation of μ_i given Y_i,

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

Jiang and Zhang (2009) report good performance in their simulations employing a design of Johnstone and Silverman (2004), with the sample size $n = 1000$ and the u_i's equally spaced with $m = 1000$. However, reproducing their results is not that straightforward: the EM iterations, while simple, make very lethargic progress toward their objective of maximizing the log likelihood,

$$L(f) = \sum_{i=1}^{n} \log(\sum_{j=1}^{m} \varphi(Y_i - u_j) f_j),$$

and the method is quite prohibitively slow even for moderately large sample sizes. To ameliorate the consequences of this slow convergence, Jiang and Zhang (2009) suggest starting the iterations with a substantial point mass at zero, rather than the perhaps more natural uniform mass points. This modification delivers improved performance for the Johnstone and Silverman (2004) simulation design by privileging $\mu_i = 0$ at the expense of a lack of equivariance.

We initially investigated a variety of schemes to accelerate the EM iterations along the lines of Varadhan and Roland (2008) and Berlinet and Roland (2007), which while somewhat helpful did not significantly improve the computational speed. The crucial insight came again only with the realization that the task of maximizing $L(f)$ is a convex problem.

4. Implementation

Both methods of the preceding section require some form of discrete implementation to render them practical for data analysis.

4.1. **Shape-constrained density estimator.** A discrete formulation of our shape-constrained MLE can be obtained by choosing a fine grid of points, $y_1 < y_2, \cdots < y_m$, setting $\alpha_i = h(y_i) \equiv \log g(y_i)$ to be the unknown function values of the mixture density and solving:

$$\max_{\alpha} \{w^\top \alpha - \sum c_i e^{\alpha_i} \mid D\alpha + 1 \geqslant 0\}, \tag{9}$$

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 is controlled by the fineness the grid, thereby

increasing the number of estimated function values. We have typically used equally spaced grids with $m \approx 300$. The vector w is an evaluation operator that simply allows us to recover and sum up the contributions to the likelihood given the expanded vector of function values. In the corresponding finite-dimensional dual problem, we write the diagonal matrix of Riemann weights, c_i, as C,

$$\min_{v}\{\sum c_i g_i \log g_i + 1^\top v \mid g = C^{-1}(w + D^\top v),\ v \geqslant 0\}. \tag{10}$$

In addition to the entropy term, the dual objective function now contains a linear contribution not present in the log-concave formulation; nevertheless, the latter is consistent with the objective function of (4), corresponding to the second term in (5).

The dual form of the estimator was implemented using two independent convex programming algorithms both 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 Bayes rule $\hat{\delta}(x) = \hat{K}'(x)$ 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 and PDCO solutions are indistinguishable in such plots. An R package REBayes, Koenker (2013), used for all the Mosek computations is available from CRAN.

For those accustomed to looking at conventional kernel density estimates, the density plot of Figure 1 is likely to appear rather bizarre, but such are the consequences 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. Careful examination of the piecewise constant Bayes rule plot illustrates that, as already remarked, its jumps do not (necessarily) occur at the observed data points represented by the "rug plot" appearing on the x-axis. As estimates of the mixture density, g, such estimates may look a bit strange, but their implied Bayes rules nevertheless conform to the monotonicity requirement and perform quite well as we shall see in Section 5.

4.2. **Nonparametric maximum likelihood.** Although the EM algorithm has dominated the literature on nonparametric maximum likelihood estimation of mixture models, others have undoubtedly recognized the advantages that convex optimization brings to such problems. Indeed, Groeneboom, Jongbloed, and Wellner (2008) have introduced an ingenious active set method they call the "support reduction algorithm" and have illustrated its performance with an application to a quantum non-locality experiment

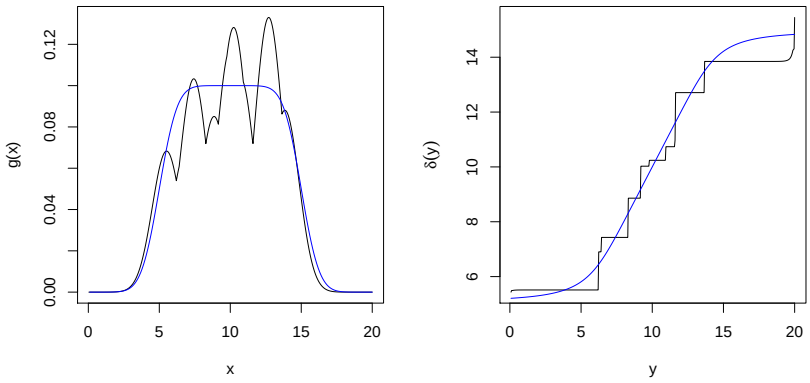

FIGURE 1. Estimated mixture density, $\hat{g}$, and corresponding Bayes rule, $\hat{\delta}$, for a simple compound decision problem. The target Bayes rule and its mixture density are plotted as smooth (blue) lines. The local maxima give y for which $\hat{\delta}(y) = y$.

that is closely related to the Gaussian mixture problem. In our experience the Mosek implementation of interior point methods for such problems has proven to be highly reliable and efficient, but a variety of other methods may eventually prove to be even better. For extremely large problems, first order gradient descent methods are likely to dominate. For further discussion of these aspects, we refer to Koenker and Mizera (2013).

The approach we chose to implement proceeds as follows. Let $\{u_1, ..., u_m\}$ be a fixed grid as above. Let A be the n by m matrix, with the elements $\varphi(Y_i - u_j)$ in the i-th row and j-th column. Consider the (primal) problem,

$$\min\{-\sum_{i=1}^{n} \log(g_i) \mid Af = g,\ f \in \mathcal{S}\},$$

where $\mathcal{S}$ denotes the unit simplex in $\mathbb{R}^m$, i.e. $\mathcal{S} = \{s \in \mathbb{R}^m | 1^\top s = 1,\ s \geqslant 0\}$. So f_j denotes the estimated mixing density estimate $\hat{f}$ evaluated at the grid point u_j, and g_i denotes the estimated mixture density estimate, $\hat{g}$, evaluated at Y_i. In this case it is again somewhat more efficient to solve the corresponding dual problem,

$$\max\{\sum_{i=1}^{n} \log v_i \mid A^\top v \leqslant n1_m,\ v \geqslant 0\},$$

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.

and subsequently recover the primal solutions. 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 3 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. By contrast, the interior point algorithm as implemented in Mosek is both quicker and more accurate. See Koenker and Mizera (2013) for more alternatives.

As another point of comparison, we illustrate in Figure 2 the Kiefer-Wolfowitz estimate of the Bayes rule and the corresponding mixture density $\hat{g}(y)$ for the example illustrated earlier in Figure 1. The estimated mixing density has four points of support in this example, but the resulting Bayes rule illustrated in the left panel of the figure is much smoother and somewhat more accurate than the piecewise constant rule produced by the shape constrained estimator as we will see in the next section.

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 both Brown and Greenshtein (2009)

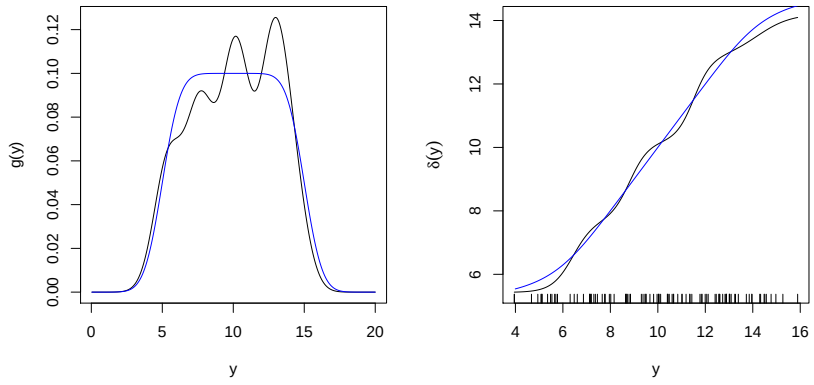

FIGURE 2. Estimated mixture density, $\hat{g}$, and corresponding Bayes rule, $\hat{\delta}$, for a simple compound decision problem. The target Bayes rule and its mixture density are again plotted in dashed blue. In contrast to the shape constrained estimator shown in Figure 1, the Kiefer-Wolfowitz MLE employed for this figure yields a much smoother and somewhat more accurate Bayes rule.

and Jiang and Zhang (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 row of the table reports the best performance, from the 18, achieved in their experiment for each column setting. The performance of the Brown and Greenshtein (2009) kernel based rule is given in the fourth row of the table, taken from their Table 1. Two variants of the GMLEB procedure of Jiang and Zhang (2009) appear in the second 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 row. The $\hat{\delta}$ and GMLEBIP results are based on 1000 replications. The GMLEB results on 100 replications, the Brown and

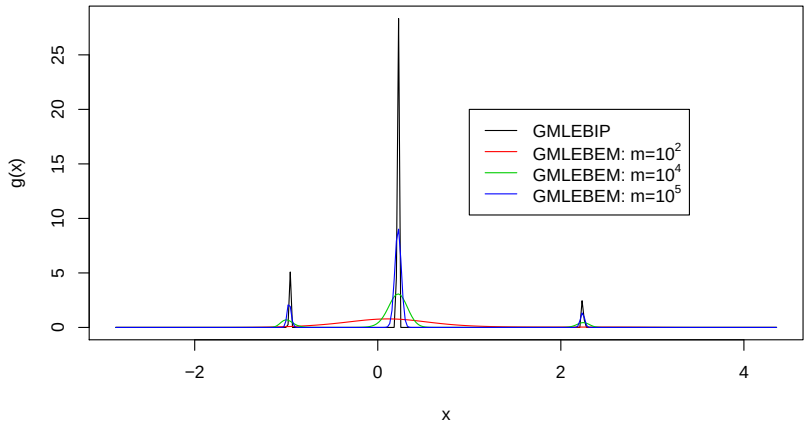

FIGURE 3. Comparison of estimates of the mixing density f: The solid black very peaked density is the interior point solution, GMLEBIP, the others as indicated by the legend represent the EM solutions with various iteration limits. After 100,000 EM iterations the peak near zero is about 8, while after 15 interior point iterations this peak is about 27. See Table 1 for further details on timings and achieved likelihoods.

Greenshtein results on 50 replications, and the Johnstone and Silverman results on 100 replications, as reported in the respective original sources.

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 procedures 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 simulation setting 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, taken from Brown and Greenshtein (2009). The row labeled "strong oracle," also taken from Brown and Greenshtein (2009), is a hard-thresholding rule which takes $\delta(X)$ to be either 0 or X depending on whether $|X| > C$ for an optimal

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
$\hat{\delta}_{\text{GMLEBIP}}$	33	30	16	8	153	107	51	11	454	276	127	18
$\hat{\delta}_{\text{GMLEBEM}}$	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
$\hat{\delta}_{\text{GMLEB}}$	268	703	1085	1365	2590	10709
$\hat{\delta}$	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. Empirical Risk of a gridded version of the GMLEB rule, the Shape Constrained Rule, $\hat{\delta}$, compared to kernel procedure, $\tilde{\delta}_{1.05}$ of Brown and Greenshtein, and an oracle hard threshholding rule described in the text. The first two rows of the table are based on 1000 replications. The last two rows are as reported in Brown and Greenshtein and based on 50 replications.

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 preceding 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.

Given the dense form of the constraint matrix A, neither the EM or IP forms of the GMLE methods are 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, so the EM implementation would be even more prohibitively time-consuming. However, it is possible to bin the observations on a fairly fine grid and employ a slight variant of the proposed interior point approach in which the likelihood terms are weighted by the relative (multinomial) bin counts. This approach, when implemented with a equally spaced grid of 600 points yields the results in the first row of Table 3. Not too unexpectedly given the earlier results, this procedure performs somewhat better than the shape constrained rule, $\hat{\delta}$. Binning the observations for the GMLEB procedure makes its computational effort comparable to the shape constrained MLE overcoming the latter's advantage due to the sparsity of its required linear algebra, and enabling us to do 1000 replications for both procedures reported in Table 3.

6. Empirical Bayesball

The ultimate test of any empirical Bayes procedure is known to be: How well does it predict second-half-of-the-season baseball batting averages? To explore this question we adopt the framework of Brown (2008) and Jiang and Zhang (2010), who use data from the 2005 Major League baseball season.

Following prior protocol, we employ midseason batting averages, $R_{i1} = H_{i1}/N_{i1}$, for $i = 1, \cdots, n_1$ to predict second half averages, $R_{i2} = H_{i2}/N_{i2}$, for $i = 1, \cdots, n_2$. All players with more than ten at bats in the first three months of the season $\mathcal{S}_1$ are used to construct predictions for their second half average, provided they also have at least 10 at bats in the second half, $\mathcal{S}_1 \cap \mathcal{S}_2$. Thus, data on $n_1 = |\mathcal{S}_1| = 567$ players in the first half is used to predict performance of $n_2 = |\mathcal{S}_1 \cap \mathcal{S}_2| = 499$ players in the second half. The transformation,

$$Y_i = \operatorname{asin}\left(\sqrt{\frac{H_{i1} + 1/4}{N_{i1} + 1/2}}\right)$$

is used to induce approximate normality and three measures of quadratic loss are employed to judge performance of the predictions:

$$\begin{aligned}
\text{TSE} &= \frac{\sum_{i=1}^{n_2}[(Y_{i2}-\hat{Y}_{i2})^2-\sigma_{i2}^2]}{\sum_{i=1}^{n_2}[(Y_{i2}-\tilde{Y}_{i2})^2-\sigma_{i2}^2]}\\
\text{TSER} &= \frac{\sum_{i=1}^{n_2}[(R_{i2}-\hat{R}_{i2})^2-R_{i2}(1-R_{i2})/N_{i2}]}{\sum_{i=1}^{n_2}[(R_{i2}-\tilde{R}_{i2})^2-R_{i2}(1-R_{i2})/N_{i2}]}\\
\text{TWSE} &= \frac{\sum_{i=1}^{n_2}[(Y_{i2}-\hat{Y}_{i2})^2-\sigma_{i2}^2]/(4\sigma_{i2}^2)}{\sum_{i=1}^{n_2}[(Y_{i2}-\tilde{Y}_{i2})^2-\sigma_{i2}^2]/(4\sigma_{i2}^2)},
\end{aligned}$$

where $\sigma_{it}^2 = 1/(4N_{it})$, $\hat{R}_{i2} = \sin^2(\hat{Y}_{i2})$ and $\tilde{Y}_{i2} = Y_{i1}$ is the naïve predictor. We depart from the earlier constant mean, constant variability context to consider linear regression models of the form:

$$Y_i = z_i^\top \beta + \xi_i + u_i,$$

where u_i is (approximately) normally distributed with variance $\sigma_i^2 = 1/(4N_{i1})$, the ξ_i's are iid from the unknown mixing density, f, and z_i denotes a vector of covariates intended to capture systematic differences in the central tendency of Y_i; these covariates are limited in the present exercise to an indicator of whether the player is a pitcher, the number of at bats taken in the first half of the season, and possible interactions thereof.

In this setting, Jiang and Zhang (2010) considered five predictors of the second half average, first two of them simply based on regression estimation,

(LSE) $$\dot{\delta}_i = z_i^\top \dot{\beta}, \quad \dot{\beta} = (Z^\top Z)^{-1} Z^\top y,$$

(WLSE) $$\ddot{\delta}_i = z_i^\top \ddot{\beta}, \quad \ddot{\beta} = (Z^\top \Omega^{-1} Z)^{-1} Z^\top \Omega^{-1} y, \quad \Omega = \text{diag}(1/\sigma_i^2).$$

The third estimator is a James-Stein version of the WLSE:

(EBJS)
$$\check{\delta}_i = \left(1 - \frac{p-2}{\sum (z_i^\top \ddot{\beta})^2/\sigma_i^2}\right) z_i^\top \ddot{\beta} + \left(1 - \frac{n-p-2}{\sum (y_i - z_i^\top \ddot{\beta})^2/\sigma_i^2}\right)(y_i - z_i^\top \ddot{\beta})$$

These three procedures were implemented independently in an effort to check the sample selection and variable definitions. The remaining two estimators are the EM implementations of the GMLEB estimator proposed by Jiang and Zhang. The first is:

(GMLEB) $$\hat{\delta}_i = z_i^\top \hat{\beta} + \frac{\int \xi \varphi((y_i - z_i^\top \hat{\beta} - \xi)/\sigma_i) d\hat{F}(\xi)}{\int \varphi((y_i - z_i^\top \hat{\beta} - \xi)/\sigma_i) d\hat{F}(\xi)}$$

where

$$\hat{\beta} = \text{argmax}_b \sum \log(\int \sigma_i^{-1} \varphi((y_i - z_i^\top b - \xi)/\sigma_i) d\hat{F}(\xi))$$

$$\hat{F} = \text{argmax}_F \sum \log(\int \sigma_i^{-1} \varphi((y_i - z_i^\top \hat{\beta} - \xi)/\sigma_i) dF(\xi))$$

The second is a reweighted version of the first:

$$\tilde{\delta}_i = z_i^\top \tilde{\beta} + \sigma_i \frac{\int \zeta \varphi(y_i/\sigma_i - z_i^\top \tilde{\beta}/\sigma_i - \zeta) d\tilde{F}(\zeta)}{\int \varphi(y_i/\sigma_i - z_i^\top \tilde{\beta}/\sigma_i - \zeta) d\tilde{F}(\zeta)} \tag{WGMLEB}$$

where

$$\tilde{\beta} = \text{argmax}_b \sum \log(\int \varphi((y_i/\sigma_i - z_i^\top b/\sigma_i - \zeta)) d\tilde{F}(\zeta)$$

$$\tilde{F} = \text{argmax}_F \sum \log(\int \varphi((y_i/\sigma_i - z_i^\top \tilde{\beta}/\sigma_i - \zeta)) dF(\zeta)$$

In both cases iteration proceeds by simply alternating between the two optimization problems, with the EM fixed point method employed for the $\hat{F}$ or $\tilde{F}$ solution; hence the acronyms GMLEBEM and WGMLEBEM.

The same strategy can be adapted to find $\hat{F}$ (or $\tilde{F}$) solutions via our interior point implementation, and thus find predictions that we designate as GMLEBIP and WGMLEBIP. We should stress that although the mixture likelihood problem is convex for each value of the regression parameter, β, the joint problem of optimizing over β and F is not convex, nevertheless conventional optimizers can be used to obtain a sequence of alternating iterations that improves the likelihood at each step.

Unfortunately, this is not the case for our shape-constrained Bayes rule, for which an analogous strategy for the regression parameter falters on the discontinuities of the $\hat{g}$ influence function. Moreover, it is unclear how the shape constrained estimator should be adapted to the variability of the variances in the observations: when $Y_i \sim \mathcal{N}(0, \sigma_i^2)$ then the Bayes rule 1 becomes

$$\delta(y_i) = y_i + \sigma_i^2 \frac{g'(y_i)}{g(y_i)}. \tag{11}$$

one must require that the corresponding K in

$$K_i(y_i) = \tfrac{1}{2} y_i^2 + \sigma_i^2 \log g(y_i) \tag{12}$$

be convex – for every σ_i^2. While it turns out that enforcing this only for the *largest* σ_i^2 ensures that it holds for all other σ_i^2, it seems that this is overly restrictive for the bulk of the players who have a larger number of at bats, and therefore smaller implied variances. In view of all the above, we have excluded the shape-constrained Bayes rule from the present competition.

Our comparison thus pits five estimators considered by Jiang and Zhang against two variants of our interior point version of the Kiefer-Wolfowitz estimator. Each of the estimators are paired with five different specifications of a regression design. In Table 4 we compare performance of five methods for each of the five regression specifications and two interior point implementations of the weighted and unweighted GMLEB procedures. The results reported for the five Jiang and Zhang methods match those reported in their Table 2. Somewhat surprisingly, however – given the results presented in the preceding section – we find that the interior point predictive performance is consistently worse than the EM results.

To explore this further, we report in Table 5 the likelihood values achieved by the EM and IP methods for the WGMLEB procedures. Of course there is nothing to assure us that estimators with better likelihoods will yield better predictions out of sample, and the hubris of expecting it to do so probably deserves the comeuppance observed in Table 4.

There are several tuning parameters that determine the stopping criteria of the EM implementation employed. The predictive advantage manifested by the EM estimates comes primarily from the effect of "early stopping" on the estimated regression parameters. It has been conjectured by some that early stopping of EM iterations can sometimes exert a beneficial regularization, or smoothing, effect. This may well be true in some circumstances, however we do not view this as very plausible in the present context. Instead, early EM stopping appears simply to produce a fortuitous location shift in the regression parameter estimates that mimics a systematic difference between first and second half-season batting.

7. A Tacky Application

In a effort to dispel the impression that the applicability of the Kiefer-Wolfowitz MLE is confined to Gaussian convolutions, we briefly reconsider the following binomial mixture problem.

> The example involves repeated rolls of a common thumbtack. A one was recorded if the tack landed point up and a zero was recorded if the tack landed point down. All tacks started point down. Each tack was flicked or hit with the fingers from where it last rested. A fixed tack was flicked 9 times. The data are recorded in Table 1. There are 320 9-tuples. These arose from 16 different tacks, 2 "flickers," and 10 surfaces. The tacks vary considerably in shape and in proportion of ones. The surfaces varied from rugs through tablecloths through bathroom floors.
>
> Beckett and Diaconis (1994) p. 108.

Estimator	TSE	TSER	TWSE
$y \sim 1$			
LSE	0.853	0.897	1.116
WLSE	1.074	1.129	0.742
EBJS	0.534	0.539	0.502
GMLEBEM	0.663	0.671	0.547
WGMLEBEM	0.306	0.298	0.427
GMLEBIP	0.686	0.700	0.593
WGMLEBIP	0.334	0.336	0.459
$y \sim AB$			
LSE	0.518	0.535	0.686
WLSE	0.537	0.527	0.545
EBJS	0.369	0.351	0.443
GMLEBEM	0.410	0.397	0.455
WGMLEBEM	0.301	0.291	0.424
GMLEBIP	0.424	0.418	0.490
WGMLEBIP	0.329	0.330	0.456
$y \sim Pitcher$			
LSE	0.272	0.283	0.559
WLSE	0.324	0.343	0.519
EBJS	0.243	0.244	0.426
GMLEBEM	0.259	0.266	0.429
WGMLEBEM	0.209	0.204	0.401
GMLEBIP	0.283	0.295	0.478
WGMLEBIP	0.231	0.234	0.426
$y \sim Pitcher + AB$			
LSE	0.242	0.246	0.477
WLSE	0.219	0.215	0.435
EBJS	0.183	0.175	0.390
GMLEBEM	0.191	0.183	0.387
WGMLEBEM	0.184	0.175	0.385
GMLEBIP	0.205	0.201	0.419
WGMLEBIP	0.203	0.203	0.412
$y \sim Pitcher * AB$			
LSE	0.240	0.244	0.476
WLSE	0.204	0.201	0.429
EBJS	0.171	0.162	0.386
GMLEBEM	0.178	0.170	0.382
WGMLEBEM	0.177	0.167	0.382
GMLEBIP	0.194	0.189	0.416
WGMLEBIP	0.206	0.206	0.414

TABLE 4. Midseason Prediction for all batters, $(|S_1|, |S_1 \cap S_2|) = (567, 499)$

Model	EM	IP
$y \sim 1$	-853.801	-851.107
$y \sim AB$	-854.733	-850.793
$y \sim Pitcher$	-844.234	-841.548
$y \sim Pitcher + AB$	-843.065	-838.970
$y \sim Pitcher * AB$	-843.258	-838.957

TABLE 5. Loglikelihood Values for the WGMLEBEM and WGMLEBIP Estimators

Liu (1996) provides a Bayesian analysis of this data employing Dirichlet process priors, and validates the resulting estimates by a comparison with a Kiefer-Wolfowitz estimate again relying on the EM algorithm as a computational device. Following Liu (1996) we focus exclusively on the binomial form of the data, so we have $n = 320$ and $n_i \equiv 9$ in what follows. The Kiefer-Wolfowitz estimator for this mixture problem has likelihood,

$$L(F) = \prod_{i=1}^{n} \int_0^1 \binom{n_i}{y_i} p^{y_i}(1-p^{n_i-y_i}) dF(p).$$

Again it proves convenient to solve the dual problem, which takes the form:

$$\max\{\sum_{i=1}^{n} \log v_i \mid \sum_{i=1}^{n} v_i \binom{n_i}{y_i} p^{y_i}(1-p^{n_i-y_i}) \leqslant n, \text{ for all } p \in [0,1]\}.$$

In effect, all that needs to be modified in our earlier formulation is the construction of the A matrix. Indeed, given that the n_i's are all identical, we can concentrate the likelihood into 10 distinct cell counts for the possible values of y_i, substantially reducing the computational effort per iteration for both the interior point and EM procedures.

In Figure 4 we compare our convex optimization solution with several EM solutions with various iteration limits. The point masses identified by our convex optimization are, to two significant figures, the same as those reported by Liu (1996), but the computational effort required to produce them is vastly reduced relative to the EM solutions.

It is evident from the figure that the EM iterations are eventually moving in the right direction, but at a glacial pace. While the convex optimization requires only 0.012 seconds in R, the fixed point EM implementation requires 599 seconds for 1,000,000 iterations and, as is clear from the figure, is still not very close to convergence by comparison with the solution provided by the interior point method after only 22 iterations. The "accelerated" EM iteration called "squarem" as implemented in Varadhan

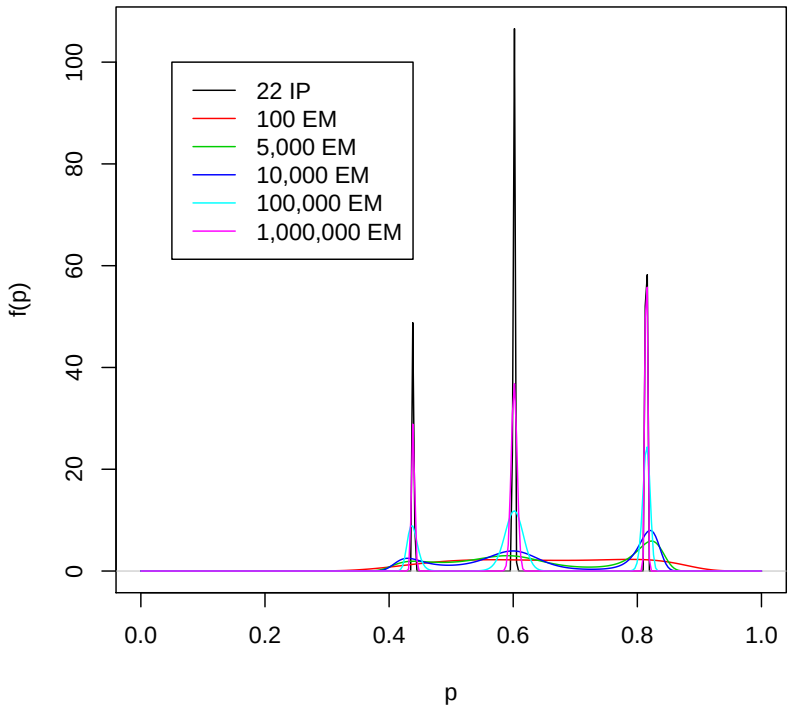

FIGURE 4. Comparison of estimates of the mixing density f: The solid black very peaked density is the interior point solution, the others as indicated by the legend represent the EM solutions with various iteration limits. Even after 1,000,000 EM iterations the peak near 0.6 is about 40, while after 22 interior point iterations this peak is about 110. One million EM iterations takes 10 minutes, but is substantially less accurate than the interior point solution that required only 0.012 seconds.

(2011) yields a somewhat more accurate solution after 1,000,000 iterations, although still not quite as precise as the interior point solution, and requires only 138 seconds to compute. However, in our experience differences in computational effort of this magnitude, more than 10,000 to 1, often make the difference between effective and ineffective statistical tools. It should be stressed in conclusion that the differences in computational

effort reported here are based on a relatively small problem with only a few observations; larger problems yield even more dramatic differences.

8. Conclusion

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 implementations of the Kiefer-Wolfowitz estimator can be dramatically reduced by reliance on interior point methods for solving convex optimization formulations of the Kiefer-Wolfowitz problem.

While the Kiefer-Wolfowitz approach seems to be typically preferable to the monotonized Bayes rule estimator introduced in Section 2, we would like to stress two advantages of the latter approach. First, because its convexity constraint is simpler, involving sparser linear algebra, computation is generally quicker. This is usually only noticeable in very large problems, but might be decisive in applications involving very large sample sizes. Second, since the monotonized Bayes rule imposes a weaker form of convexity constraint, it may have some advantages in misspecified settings where the original Gaussian location mixture assumption fails to be satisfied. In Remark 4 following Theorem 1 we have characterized the class of densities for which the piecewise constant Bayes rule is optimal and the monotonized estimator is particularly appropriate.

We would like to again stress that our convex optimization formulation of the Kiefer-Wolfowitz MLE is applicable to a wide variety of mixture problems wherever parameter heterogeneity is plausible. For longitudinal data and survival applications this is particularly germane, and we hope to pursue such applications in future work. We have focused on settings in which both empirical Bayes procedures rely on exclusively on maximum likelihood methods, but there is clearly scope for introducing further regularization, other forms of prior information, that might lend additional credibility were such information available as for example, in Bunea, Tsybakov, Wegkamp, and Barbu (2010).

Appendix A

Proof. [Theorem 1] We start by demonstrating that every candidate solution, K, can be replaced by another piecewise linear one, $\tilde{K}$, with no

larger objective function of (3). By its convexity, K has one-sided derivatives everywhere—in particular, at each Y_i; we denote the latter by $K'(Y_i-)$ and $K'(Y_i+)$. Let $\tilde{K}$ be a convex and piecewise linear function such that $\tilde{K}(Y_i) = K(Y_i)$ for all i, extending linearly left and right from every Y_i, with slopes (i) at $\min Y_i$ both equal to $K'(\min Y_i+)$; (ii) at $\max Y_i$ both equal to $K'(\max Y_i-)$; and (iii) at every other Y_i the left one equal to $K'(Y_i-)$ and the right one to $K'(Y_i+)$ (thus every Y_i is either contained in or borders a linear piece.) In the objective function of (3), such a $\tilde{K}$ yields the same summation term as K, but as $\tilde{K}$ minorizes K, that is the integral term for $\tilde{K}$ does not exceed that for K.

In view of this it is evident that solutions are restricted to a finite-dimensional class of piecewise-linear convex functions parametrized by their (finite number of) slopes and an arbitrary function value at some point, say, $K(0)$. The proof of existence then proceeds via a standard continuity/compactness argument.

First, one has to assess continuity: if $K_m(0) \to K(0)$ (hereafter, we use m for the integer index running to infinity) and the value of slopes converge as well, then uniform continuity is guaranteed on bounded intervals; but the continuity of the integral term in the objective of (3), if the integration domain is the whole real line, may be in doubt. Fortunately, the specific form of the integrand facilitates the continuity. Given an $\epsilon > 0$, we can always select a bounded closed interval that contains all the Y_i's, so that the summation part of the objective is within $\epsilon/4$ of its limit for m large enough; inside the interval, the convergence is uniform, so for m large enough, within $\epsilon/4$ of its limit again; finally, on the linear tails one can see via direct calculation that

$$\int_E e^{a_m x + b_m} d\Phi(x) = e^{\frac{1}{2}a_m^2 + b_m} \int_E e^{\frac{1}{2}(x-a_m)^2} dx, \tag{13}$$

which shows that such integrals (over various integration domains E) converge whenever $a_m \to a$ and $b_m \to b$; hence each tail integral can be made within $\epsilon/4$ of its limit for m large enough.

Once continuity of the objective function is established, it is sufficient to show the existence of nonempty compact sublevel set. Putting $K(y) \equiv 0$ yields the objective of (3) equal to 1; hence we will look now at the (nonempty) sublevel set of all parameters (that is, all slopes and $K(0)$) that yield the objective function $\leqslant 1$. We have seen that this set is nonempty; by continuity, it is closed; it remains to show that it is also bounded.

The latter will be accomplished if the following is verified for any sequence of parameter vectors: whenever at least one component of them does not stay bounded, then the objective function ultimately exceeds 1.

One can always pass to subsequences in the process, to assert that bounded sequences actually converge, and unbounded diverge either to $+\infty$ or $-\infty$. As shown in Section 2, the origin, 0, can be without any loss of generality put inside the convex hull of the Y_i's, say, in the midpoint between min and max. Suppose thus we have a sequence K_m parametrized by vectors that do not stay bounded, and such that K_m all yield values of the objective (3) not exceeding 1. As the integral term of this objective is always nonnegative, we have that $-\sum K_m(Y_i) \leqslant 1$. It follows that $K_m(0)$ has to be bounded from above: as every K_m is a convex function, it can be minorized by an affine function whose intercept is $K_m(0)$; the formula (13) then shows that the integral term (integration domain E is the whole real line now) in the objective is minorized by

$$e^{\frac{1}{2}a_m^2 + K_m(0)},$$

which, if $K_m(0) \to +\infty$, would drive the integral term to $+\infty$ even if a_m stays bounded; at the same time, the summation term stays bounded by 1. It is possible for $K_m(0)$ to diverge to $-\infty$—but not when the other parameters, slopes, stay bounded, for then one can find C such that for every i,

$$K_m(Y_i) \leqslant K_m(0) + C$$

and consequently

$$-1 \geqslant -\sum K_m(Y_i) \geqslant -nK_m(0) - nC.$$

Hence, in any case the slopes must not remain bounded; as they values are ordered in the increasing sense, this means that either the maximal slope, on the right-hand tail, diverges to $+\infty$, or the minimal slope, on the left-hand tail, diverges to $-\infty$; or both.

Now, let $Y_1 = Y_{\min}$, $Y_n = Y_{\max}$, $d = (Y_n - Y_1)/2$, and consider $K_m(Y_1)$ and $K_m(Y_n)$; one of them must remain bounded from $-\infty$, for if both of them would be $\leqslant -1/n$, then as well $K_m(Y_i) \leqslant -/n$ for all i, and then $-\sum K_m(Y_i) \geqslant 1$. Suppose that the one bounded from below is $K_m(Y_n)$, and also that it the right-hand slope that diverges,

$$\frac{K_m(Y_n) - K_m(0)}{d} \to +\infty. \tag{14}$$

Then the application of the formula (13), with integration domain $E = [Y_n, \infty)$, yields the integral term in (13) converging to 1; at the same time,

the term in the exponent in front of the integral is

$$\frac{(K_m(Y_n) - K_m(0))^2}{2d^2} + K_m(0)$$
$$= \frac{(K_m(Y_n) - K_m(0) - 2d^2)(K_m(Y_n) - K_m(0))}{2d^2} + K_m(Y_n),$$

which diverges to $+\infty$, as (14) implies that $K_m(Y_n) - K_m(Y_0) \to +\infty$, and at the same time, $K_m(Y_n)$ stays bounded away from $-\infty$.

If the latter still takes place, but the right-hand slope does not diverge to $+\infty$, then, as the line passing through $K_m(Y_n)$ minorizes the whole K_m, we obtain that $K_m(Y_1)$ must be bounded from below. At the same time, it is now the left-hand slope that must diverge to $-\infty$; the proof of the divergence of the integral is then analogous.

Finally, to backtrack in our logical tree, assume now it is not $K_m(Y_n)$ that does not stay bounded from $-\infty$, but rather $K_m(Y_1)$, then we have again that left-hand slope must diverge to $-\infty$ (otherwise $K_m(Y_n)$ would be bounded from $-\infty$ by an analogous argument as above); this situation was already encountered. Having exhausted all possibilities, we conclude that the sequence of parameter vectors cannot diverge to infinity with m when the objective function stays bounded from above by 1; hence the sublevel set is nonempty and compact, which yields the existence of the solution, $\hat{K}$, of (3).

The established form of $\hat{K}$ implies the form of the estimated density, as well as that of the decision rule. The proof of the duality result is a straightforward modification of Theorem 3.1 and Corollary 3.1 of Koenker and Mizera (2010); in particular, the change of $\mathcal{K}$ to $-\mathcal{K}$ results in the change of the sign at G in the dual constraint, and Corollary 4A from Rockafellar (1971) is used in its more general form (already formulated there). We omit the details. ∎

Proof. [Theorem 2] The statements of the theorem can be essentially established by piecing together various results from Lindsay (1983); however, some care should be exercised. First of all, Lindsay (1983) assumes that the parametric space is compact. Our parametric space is the whole real line, $\mathbb{R}$, hence not a compact; nevertheless, adding one point ∞ to it—that is, considering the one-point "circular" compactification, $\dot{\mathbb{R}}$, instead—makes the road passable.

Lindsay (1983) works with "atomic likelihood vectors" : in our case, these are defined for every $\mu \in \mathbb{R}$ as

$$(f_\mu(y_1), \ldots, f_\mu(y_n)) = (\phi(y_1 - \mu), \ldots, \phi(y_n - \mu)).$$

Compactifying $\mathbb{R}$ with ∞, we have to define the atomic likelihood vector for $\mu = \infty$; given that the limit of $\phi(y - \mu)$ for μ going to either $-\infty$ or $+\infty$ is 0, it is natural to define

$$(f_\infty(y_1), \ldots, f_\infty(y_n)) = (0, \ldots, 0). \tag{15}$$

"Mixture likelihood vectors" then correspond to terms appearing in (6); in particular, Theorem 3.1 of Lindsay (1983) asserts that the solution, $\hat{F}$, and is an atomic probability with at most n atoms. To finish the proof of this part, we have to show that ∞ will not be one of these atoms. But, if the contrary would be true, one could take the mass from ∞, which due to (15) contributes nothing to the overall likelihood, and put it instead into some other atom of $\hat{F}$ whose contribution is positive; this would increase the value of the overall likelihood, a contradiction with the claim that $\hat{F}$ is the MLE. While this cannot be done if ∞ would be the only atom of $\hat{F}$ (that is, if all the mass will be concentrated in ∞), note that in such case the value of the objective function in (6) is $+\infty$, which certainly is not an optimal value, as can be demonstrated by taking any other F, say, concentrated in some point of $\mathbb{R}$.

What Lindsay (1983) calls "mixture maximum likelihood" (and denotes by $\hat{f}$) is in our notation, for given locations of the atoms, $\hat{\mu}_j$,

$$\sum_j \varphi(Y_i - \hat{\mu}_j)\hat{f}_j.$$

Theorem 5.1 of Lindsay (1983) asserts that $\hat{v}_i$, given by (8), are solutions for his Problem 2 (page 90)—which is nothing but the dual formulation (7), albeit again with a slight difference that it is formulated for the compact parameter space. This means that the constraint in (7) is supposed to be satisfied for all $\mu \in \dot{\mathbb{R}}$; however, for $\mu = \infty$ it is satisfied automatically, due to (15), thus requiring it only for $\mu \in \mathbb{R}$ does not change the problem. Finally, our statement about the locations of the atoms is justified by Theorem 4.1.C of Lindsay (1983), in view of his formula (4.1). ∎

REFERENCES

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

BECKETT, L., AND P. DIACONIS (1994): "Spectral analysis for discrete longitudinal data," *Advances in Mathematics*, 103, 107–128.

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

BROWN, L. (2008): "In-season prediction of batting averages: A field test of empirical Bayes and Bayes methodologies," *The Annals of Applied Statistics*, 2, 113–152.

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.

Bunea, F., A. B. Tsybakov, M. H. Wegkamp, and A. Barbu (2010): "Spades and mixture models," *The Annals of Statistics*, 38(4), 2525–2558.

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.

Efron, B. (2010): *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*. Cambridge U. Press: Cambridge.

——— (2011): "Tweedie's Formula and Selection Bias," *Journal of the American Statistical Association*, 106, 1602–1614.

Groeneboom, P., G. Jongbloed, and J. A. Wellner (2001): "Estimation of a convex function: Characterization and asymptotic theory," *The Annals of Statistics*, 29(6), 1653–1698.

——— (2008): "The support reduction algorithm for computing non-parametric function estimates in mixture models," *Scandinavian Journal of Statistics*, 35, 385–399.

Heckman, J., and B. Singer (1984): "A method for minimizing the impact of distributional assumptions in econometric models for duration data," *Econometrica*, 52, 63–132.

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

Jiang, W., and C.-H. Zhang (2010): "Empirical Bayes in-season prediction of baseball batting averages," in *Borrowing Strength: Theory Powering Applications: A Festschrift for Lawrence D. Brown*, vol. 6, pp. 263–273. Institute for Mathematical Statistics.

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, 887–906.

Koenker, R. (2013): *REBayes: Empirical Bayes Estimation and Inference in R,* R package version 0.31, `http://CRAN.R-project.org/package=REBayes`.

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

——— (2013): "Convex optimization in R," preprint.

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," *Annals of Statistics*, 11, 86–94.

Liu, J. (1996): "Nonparametric hierarchical Bayes via sequential imputations," *The Annals of Statistics*, 24, 911–930.

Robbins, H. (1956): "An empirical Bayes approach to statistics," in *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, vol. I. University of California Press: Berkeley.

Rockafellar, R. T. (1970): *Convex analysis*. Princeton University Press, Princeton.

——— (1971): "Integrals which are convex functionals, II," *Pacific Journal of Mathematics*, 39, 439–469.

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

SEREGIN, A., AND J. A. WELLNER (2010): "Nonparametric estimation of multivariate convex-transformed densities," *Annals of Statistics*, 38, 3751–3781.

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. (2011): *SQUAREM: Squared extrapolation methods for accelerating fixed-point iterations*, R package version 2010.12-1, http://CRAN.R-project.org/package=SQUAREM.

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