Shape Constraints in Empirical Bayes Inference

Mu Lin[1] and Ivan Mizera[1,2,3]
[1]Department of Mathematical and Statistical Sciences
University of Alberta, Edmonton, Alberta, Canada T6G 2G1
[2]Corresponding author: Ivan Mizera, e-mail: imizera@yahoo.com

Abstract

Following the earlier work on the estimation of densities for the classical Gaussian compound decision problem and their associated (empirical) Bayes rules, the problem that has been considered from several perspectives — introducing a nonparametric maximum likelihood estimator of the mixture density subject to a monotonicity constraint on the resulting Bayes rule, and the Kiefer-Wolfowitz nonparametric maximum likelihood estimator for mixtures — we propose modifications that require a shape-constraint of log- or quasi-concave type on the mixture density. These modification exhibit superior behavior in the case when the shape assumption reasonably captures the behavior of the data, in particular, when the mixing distribution is unimodal. Like the earlier approaches, our proposals are also cast as convex optimization problems, and can be efficiently solved by modern convex optimization methods. The finite-sample properties of the procedures are compared with several existing empirical Bayes and other methods in a small simulation study.

Keywords: Empirical Bayes, Shape Constrained Inference, Nonparametric Maximum Likelihood, Shrinkage, Compound Decision Problem, Mixture Model, Random Effects, Convex Optimization.

1 Normal means

In the standard setting of normal means — see, for instance, Efron (2010) — the objective is to estimate/predict an unobserved vector $(\mu_1, \ldots, \mu_n)$, based on the observations $(Y_1, \ldots, Y_n)$. The latter are modeled as

$$Y_i|\mu_i \sim \mathcal{N}(\mu_i, \sigma^2), \tag{1}$$

with known σ^2 (which thus may be assumed to be 1). If μ_i are viewed themselves sampled from P, hereafter called mixing distribution (hereafter mixing)

$$\mu_i \sim P, \tag{2}$$

then an alternative improving, in terms of the aggregated squared error loss, upon the maximum likelihood prediction $\hat{\mu}_i = Y_i$ is given by the Bayes formula: the prediction that

[3]Supported by the NSERC of Canada

"borrows strength" from other observations is the expected value of the conditional distribution of μ_i given Y_i. This quantity can be conveniently expressed through the so-called Tweedie formula (Efron, 2011) as

$$\mathrm{E}(\mu_i|Y_i) = Y_i + \sigma^2 \ell'(Y_i), \tag{3}$$

where ℓ' is the derivative of the logarithm of the marginal, mixture distribution of Y_i,

$$\ell(y) = \log\Big(\int \varphi(y-\mu)P(d\mu)\Big), \tag{4}$$

and φ is the conditional density of Y_i given μ_i. In fact, this formula holds true for any one-parameter exponential family; for the particular φ appearing in (1) it can be verified by direct calculation.

An exponential family argument — or a direct calculation again — shows also that the derivative of (3), in Y_i, is equal to $\mathrm{Var}(\mu_i|Y_i)/\sigma^2$; this implies that the corresponding second derivative is nonnegative and thus the prediction rule is nondecreasing in Y_i (as noted by van Houwelingen and Stijnen, 1983). This agreeable property remains true if P (typically not known unless somehow elucidated from the circumstances) is replaced by an estimate; however, monotonicity may not be preserved if the marginal density of Y_i, or the subsequent prediction rule, is estimated directly — say, by a kernel density estimate as in Brown and Greenshtein (2009).

However, rather than a mere rectification of this potential deficit — the ways of monotonizing existing prediction rules, often resulting in rules uniformly better than the original ones, were investigated already by van Houwelingen (1977) and van Houwelingen and Stijnen (1983) — a more appealing opportunity that opens here is that the specific form and properties of the prediction rule may catalyze the use of the maximum likelihood method, or similar strategy, to estimate the prediction rule. In other words, for the, in general ill-posed, problem of estimating nonparametric density via maximum likelihood, the monotonicity constraint or mixture representation can act as a regularizer — and of a kind that does not call for any additional tuning.

Two instances of this general program were investigated by Koenker and Mizera (2013): maximum likelihood estimation of the marginal density under the monotonicity constraint on the prediction rule, and maximum likelihood estimation of the mixing distribution using the mixture representation of the marginal density. Here we complement those by alternatives that in addition place a shape constraint on the estimated mixture density.

2 Prediction rule with monotonicity constraint

The first approach studied by Koenker and Mizera (2013) estimates the mixture density via maximum likelihood, under the shape constraint implied by the required monotonicity of the prediction rule. Shape-constrained density estimates were considered, among others, by Dümbgen and Rufibach (2009) or Koenker and Mizera (2010); however, the typical shape constraint considered there, monotonicity or log-concavity of the estimated density f, is replaced in the present context by the requirement that the subsequent prediction rule is non-decreasing — that is, its antiderivative is convex:

$$-\sum_i g(Y_i) + \int \mathrm{e}^{g(y)}dy \;\rightsquigarrow\; \min_g !\qquad \text{subject to} \quad \frac{1}{2}y^2 + g(y) \text{ convex}. \tag{5}$$

Note that expressing the task in terms of $g = \log f$ ensures the convexity of the problem. The methodology developed in Koenker and Mizera (2007) and Koenker and Mizera (2010)

can be applied also to this case, with the implications both in theory (dual formulation) and practice (implementation).

3 Mixing density via nonparametric maximum likelihood

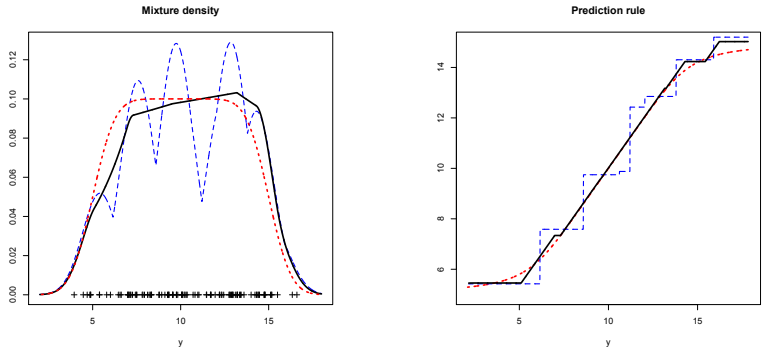

Figure 1: Estimated mixture density (left) and corresponding Bayes rule (right) for the monotone-constrained maximum likelihood (dashed blue) and the log-concave shape-constrained variant (solid). The target, "oracle" mixture density and its Bayes rule are plotted in dotted red.

It is well known that general, unrestricted estimation of a probability density via maximum likelihood is an ill-posed problem whose "solution" degenerates to a linear combination of point measures. However, the specific mixture form of the density appearing in (4) ensures that the nonparametric maximum likelihood formulation of Kiefer and Wolfowitz (1956),

$$-\sum_{i=1}^{n} \log\left(\int \varphi(Y_i - \mu) dP(\mu)\right) \rightsquigarrow \min_P ! \tag{6}$$

has a well-defined outcome — even if the problem is infinitely dimensional and P runs over all possible mixing distributions.

In practice, this solution is approximated by the solution of a closely-related finite-dimensional problem, as proposed by Jiang and Zhang (2009) and others. Their suggestions to use (variants of) the EM algorithm lead, however, to prohibitively slow implementations; a way out here is offered by the fact that the problem is again convex, with a convex objective function minimized over a convex set of putative P. This leads not only to viable practical algorithm, but also to theoretical insights, again via duality theory.

In simulation experiments of Koenker and Mizera (2013), the approach expounded in this section has a slight, but visible edge over the one expounded in the previous one; both pretty much dominate all other existing methods. The algorithm employing monotonicity constraint on the prediction rule easier scales to larger datasets; on the other hand, nonparametric maximum likelihood allows in a straightforward way for the inclusion of covariates.

4 Empirical Bayes estimation for unimodal distributions

Both methods typically yield mixture density estimates that are considerably wiggly — albeit this does not seem to influence the performance of the resulting prediction rules. The latter

when estimated via imposing the monotonicity constraint are piecewise constant; when estimated by nonparametric maximum likelihood, the mixing distribution is atomic. In certain situation, it is reasonable to believe that the mixing distribution is rather smooth; a question arises whether this knowledge cannot be utilized to obtain less rough density estimates, and subsequently better prediction rules. While it is always possible to limit the size of the atoms of the mixing distribution by a suitable upper bound, this, and other possible regularization strategies (for instance, those involving complexity penalties) would result in introducing additional tuning parameters, a potentially undesired complication.

An appealing alternative is then to impose a suitable shape constraint, for instance log-concavity, on the estimated mixing density. However, this in general leads to a non-convex problem with an uncertain implementation. The way out is offered via the observation that whenever the mixing distribution is log-concave, then so is the mixture distribution (as noted by Efron, 2011, and others). It turns out that when shape-constraint is imposed not on the mixing, but mixture distribution, the convexity of the optimization problem is preserved — for both of the approaches mentioned above. This leads to efficient implementations and theoretical insights. Moreover, the log-concavity of the mixture distribution can take place even in the case when the mixing distribution is not log-concave — therefore the shape constraint covers a more general situation when applied to the mixing, rather than mixture distribution.

The first formulation we propose here is a very slight modification of (5)

$$-\sum_i g(Y_i) + \int e^{g(y)} dy \rightsquigarrow \min_g ! \text{ subject to } \frac{1}{2}y^2 + g(y) \text{ convex, and } g \text{ concave.} \qquad (7)$$

Expressed through derivatives, the constraint can be written, in somewhat simplified manner, as $0 \geq g''(y) \geq -1$, very much resembling the dual constraints arising in density estimation regularized with total-variation penalties. The implementation of (7) is an easy and straightforward modification of that of (5), and does bring virtually no increase in computational complexity or running time of the algorithm.

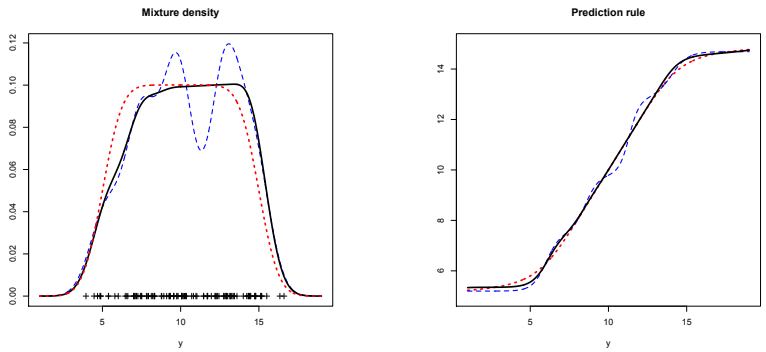

Figure 2: Estimated mixture density (left) and corresponding Bayes rule (right) for the Kiefer-Wolfowitz maximum likelihood (dashed blue) and its log-concave shape-constrained variant (solid). The target, "oracle" mixture density and its Bayes rule are again plotted in dotted red.

Imposing the shape constraint on the mixture density in (6) is a bit more tricky, and can actually create a non-convex problem — unless a slack function g is introduced and the

	$U[5,15]$	t_3	χ^2_2	$0_{95}\|2_{05}$	$0_{50}\|2_{50}$	$0_{95}\|5_{05}$	$0_{50}\|5_{50}$
br	101.5	112.4	77.8	19.7	57.3	12.6	21.1
kw	92.6	114.4	71.9	17.4	51.3	10.0	17.0
brlc	85.6	98.1	67.6	17.3	51.7	21.6	58.2
kwlc	84.9	98.2	66.8	16.5	50.4	21.2	67.6
mle	100.2	100.1	100.2	100.7	100.4	100.1	99.6
js	89.8	98.5	80.2	18.5	52.1	56.2	86.8
oracle	81.9	97.5	63.9	12.6	44.9	4.9	11.5

Table 1: The empirical risk of several estimators/prediction schemes: the MLE of the mixture density with the monotonicity constraint on the prediction rule (br); the Kiefer-Wolfowitz nonparametric MLE of the mixing distribution (kw); their versions, (brlc) and (kwlc) respectively, with mixture density constrained to be log-concave; the MLE (no shrinkage) predictor (mle); the James-Stein estimator/predictor, assuming the normal mixing distribution (js); and finally the "oracle" predictor, the Bayes rule employing the knowledge of the mixing distribution.

constraint is expressed in an epigraph, inequality form:

$$\sum_{i=1}^{n} g(Y_i) \rightarrowtail \min_{g,P}! \quad \text{subject to } g(y) \geq -\log\left(\int \varphi(y-\mu)dP(\mu)\right) \text{ and } g \text{ convex.} \quad (8)$$

The introduction of g and the need to enforce the inequality over all y from some fine grid (not only over the actually observed Y_i), considerably increases the complexity of this formulation; nevertheless, the convex character renders it still feasible, and due to the dominance of the nonparametric maximum likelihood approach over that employing the monotonicity of the prediction rule, it seems that the increased computational price may be well worth the effort. Moreover, unlike (7), this approach allows also for the incorporation of more general q-convex constraints considered in Koenker and Mizera (2010).

5 Experimental comparisons and conclusions

To compare the proposed estimators/prediction schemes in a couple of selected situations, we performed a small simulation study using the following mixing distributions: the uniform distribution on $[5,15]$ (featured in the examples shown in the figures); the t distribution with 3 degrees of freedom; the χ^2 distribution with 2 degrees of freedom; and four instances from the simulation study of Johnstone and Silverman (2004), employed also by Koenker and Mizera (2013): a mixture consisting of two numbers, k of which (we used $k = 5$ and $k = 50$) are equal to some fixed μ (we used $\mu = 2$ and $\mu = 5$) and the rest are zeros. For each distribution, we performed 1000 repetitions. The results in Table 1 show the average of the sums, computed for each repetition, of squared errors for all sampled μ_i. The sample size was in all cases $n = 100$. In addition to the methods considered above, Table 1 shows also the results of the "naïve" maximum likelihood predictor $\hat{\mu}_i = Y_i$; the predictor based on the James-Stein estimator

$$Y_i - \frac{n-3}{\sum_i (Y_i - \bar{Y})^2}(Y_i - \bar{Y}), \qquad \bar{Y} = \frac{1}{n}\sum Y_i;$$

and the "oracle" predictor, the optimal predictor assuming the knowledge of the mixing distribution.

The results indicate that the precision gain of the nonparametric maximum likelihood estimate of the mixing distribution (6) over the prediction rule with monotonicity constraint (5), as observed by Koenker and Mizera (2013), is still generally preserved for the procedures that enforce log-concavity of the mixture distribution. However, the magnitude of this effect occurs to be much smaller, hence the possible advantage of the nonparametric maximum likelihood (8) may be counterbalanced by the lower computational complexity of (7). Also, a different—reversed—behavior is observed for the heavy-tailed distribution, t_3, not only for the versions with enforced log-concavity, but also for the unrestricted ones—which are, interestingly, in this case dominated by the "naïve" predictor $\hat{\mu}_i = Y_i$ (which is however, still dominated by the oracle predictor, in accord with the theory). Finally, we can see that the versions with enforced mixture log-concavity still dominate the James-Stein predictor, based on the assumption of the normality of the mixture distribution—in the case of asymmetric mixing distribution χ^2_2, the difference in efficiency seems to be substantial.

On the other hand, it comes as no surprise that the methods enforcing mixture log-concavity yield better predictions only when the mixture distributions are truly unimodal (or close to those). In distinguished bimodal situations, the behavior is reversed in favor of the unconstrained methods, which are thus in such circumstances preferable.

References

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

DÜMBGEN, L. and RUFIBACH, K. (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 University Press, Cambridge.

EFRON, B. (2011). Tweedie's formula and selection bias. *Journal of the American Statistical Association* **106** 1602–1614.

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

JOHNSTONE, I. M. and SILVERMAN, B. W. (2004). Needles and straw in haystacks: Empirical Bayes estimates of possibly sparse sequences. *The Annals of Statistics* 1594–1649.

KIEFER, J. and 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. and MIZERA, I. (2007). Primal and dual formulations relevant for the numerical estimation of a probability density via regularization. In *Tatra Mountains Mathematical Publications* (A. Pázman, J. Volaufová and V. Witkovský, eds.) **38** Slovak Academy of Sciences Proceedings of the conference ProbaStat '06 held in Smolenice, Slovakia, June 5-9, 2006.

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

KOENKER, R. and MIZERA, I. (2013). Convex optimization, shape constraints, compound decisions, and empirical Bayes rules. Submitted.

VAN HOUWELINGEN, J. C. (1977). Monotonizing empirical Bayes estimators for a class of discrete distributions with monotone likelihood ratio. *Statistica Neerlandica* **31** 95–104.

VAN HOUWELINGEN, J. C. and STIJNEN, T. (1983). Monotone empirical Bayes estimators for the continuous one-parameter exponential family. *Statistica Neerlandica* **37** 29–43.