A GAUSSIAN COMPOUND DECISION BAKEOFF

ROGER KOENKER

Abstract. A nonparametric mixture model approach to empirical Bayes compound decisions for the Gaussian location model is compared with a parametric empirical Bayes approach recently suggested by Martin and Walker and several recent more formal Bayes procedures.

Martin and Walker (2013) have recently proposed a parametric empirical Bayes procedure for the classical Gaussian compound decision problem in which, $Y_i \sim \mathcal{N}(\theta_i, 1)$, $i = 1, 2, \cdots, n$ and we wish to estimate $\theta \in \mathbb{R}^n$ subject to squared error loss. I was curious to compare this procedure with the nonparametric empirical Bayes procedure recently introduced by Jiang and Zhang (2009) and further explored in Koenker and Mizera (2013). The latter approach is based on the nonparametric maximum likelihood estimator for mixture models of Kiefer and Wolfowitz (1956). The interior point computational approach suggested in Koenker and Mizera (2013) for the Kiefer-Wolfowitz MLE makes it feasible to study its performance much more easily.

1. Four Experiments

To this end, I initially adopted a slightly expanded variant of the simulation design of Martin and Walker: sample size is $n = 200$, with 1000 replications, $\theta = \theta_a \in \{1, 3, 5, 7\}$ for s entries, and 0 otherwise, and $s \in \{10, 20, 40, 80\}$ The default settings for the EBMW procedure including their Gibbs parameters are maintained. Table 1 reports the results of the experiment. For each replication we compute the sum of squared errors for the n observations, these are then averaged over the 1000 replications and rounded to integers, following standard practice as introduced in Johnstone and Silverman (2004).

	s = 10				**s = 20**				**s = 40**				**s = 80**			
	1	3	5	7	1	3	5	7	1	3	5	7	1	3	5	7
EBMW	10	56	22	13	20	98	36	25	40	162	62	48	79	249	106	92
EBKM	13	36	15	7	21	52	19	7	32	74	25	8	44	94	29	8

Table 1. MSE based on 1000 replications

As a second experiment I considered an expanded version of the second experiment described in Martin and Walker. This experiment employs a design used by Castillo and van der Vaart (2012). The setup is very similar to the first experiment, except that

Version: December 14, 2013. This research was partially supported by NSF grant SES-11-53548.

$n = 500$, $s \in \{25, 50, 100\}$ and θ_a takes five possible values: $\{1, 2, \cdots, 5\}$. The first 8 rows of Table 2 are taken directly from Table 1 of Castillo and van der Vaart (2012), with identifiers as described there; the ninth row is taken from Martin and Walker (2013), each based on 100 replications. In the tenth through twelfth rows we report results for a slightly expanded version of the design including $\theta_a \in \{1, 2\}$ for both the EBMW procedure and the EBKM procedure, both based on 1000 replications. In this table I have also added the performance of the monotone Bayes rule estimator, EBMR, introduced in Koenker and Mizera (2013). It performs slightly worse than EBKM, the Kiefer-Wolfowitz procedure, but considerably better than the other procedures, except in the first column of the table where the Martin and Walker estimator is the (narrow) winner. It might be initially surprising to see that the EBKM and EBMR estimators have MSE $< s$ for $s = 50$ and $s = 100$ when $\theta_a = 5$. However, when the signal is sufficiently separated from the zeros, the Kiefer-Wolfowitz estimator of the mixing distribution quite accurately estimates the two mass points and thus the Bayes rule provides not only shrinkage toward zero, but also toward this estimate of θ_a.

	s = 25					**s = 50**					**s = 100**				
	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
PM1			111	96	94			176	165	154			267	302	307
PM2			106	92	82			169	165	152			269	280	274
EBM			103	96	93			166	177	174			271	312	319
PMed1			129	83	73			205	149	130			255	279	283
PMed2			125	86	68			187	148	129			273	254	245
EBMed			110	81	72			162	148	142			255	294	300
HT			175	142	70			339	284	135			676	564	252
HTO			136	92	84			206	159	139			306	261	245
EBMW			142	98	53			240	160	93			399	262	151
EBMW	25	94	143	103	54	50	183	244	165	91	99	347	404	258	155
EBMR	30	77	89	65	35	50	123	136	92	48	79	185	193	127	62
EBKM	27	71	80	57	30	46	113	122	81	40	74	171	174	112	53

Table 2. MSE based on 1000 replications

Given the foregoing results, I was curious to see whether the improvement of the empirical Bayes (Kiefer-Wolfowitz) approach over the methods of Martin and Walker would persist if I replaced the point mass signals with more diffuse signals. So as a third experiment I maintained all the features of the second experiment except that the non-zero elements of the θ vector were now generated as standard Gaussians centered at the θ_a values, rather than as point masses at those values. Since the Kiefer-Wolfowitz MLE produces a small number of point masses my expectation was that this would make things more difficult for EBKM. The results of this experiment are reported in Table 3. As expected, this variant of the problem is more challenging, but the gap in relative performance persists.

	s = 25					**s = 50**					**s = 100**				
	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
EBMW	47	99	117	91	59	108	195	198	169	112	145	281	329	259	173
EBKM	43	69	76	64	45	79	121	123	111	78	104	165	186	165	121

TABLE 3. MSE based on 1000 replications

Finally, I thought it might be interesting to compare EBKM with the procedures suggested in Bhattacharya, Pati, Pillai, and Dunson (2012). In their simulation setting, $n = 1000$, and θ_a is 10 for the first 10 coordinates, $\theta_a \in \{2, 3, \cdots, 7\}$ for the next 90 coordinates, and zero for the remaining 900. They consider four methods: the Bayesian Lasso (BL), two variants of their Dirichlet-Laplace priors with different prior strength (DL(α)), and the horseshoe prior of Carvalho, Polson, and Scott (2009). Results in Table 4 are again substantially better than those reported by Bhattacharya et al. The first four rows of the table are taken from Bhattacharya, Pati, Pillai, and Dunson (2012) and are based on 100 replications; the last two rows are based on 1000 replications as in the previous tables. Again the empirical Bayes (Kiefer-Wolfowitz) procedure substantially outperforms the others and this advantage increases with degree of separation of the signal from the zeros.

	2	3	4	5	6	7
BL	299	386	424	450	474	493
DL(1/n)	307	354	271	205	183	169
DL(1/2)	368	679	671	374	214	160
HS	268	316	267	213	193	177
EBMW	324	439	306	175	130	123
EBKM	207	225	151	78	45	38

TABLE 4. MSE based on 100 replications for the first four rows and on 1000 replications for the last two rows

In contrast to all the other procedures considered above, the Bayes rule that emerges from the Kiefer-Wolfowitz estimator of the mixing distribution, i.e. EBKM, places no special significance on the notion of "sparsity," or at least none on the special status of the number zero. All of the other procedures purport to know that most of the θ_i's are zero, and the simulation designs justify the confidence placed on this "prior" knowledge. For Kiefer-Wolfowitz zero is just another number, indeed the Kiefer-Wolfowitz estimator of the mixing distribution is equivariant to shifts of sample location. Thus, if we were to shift our simulations from mixtures $\theta_i \sim \alpha\delta_0 + (1 - \alpha)\delta_{\theta_a}$ to $\alpha\delta_b + (1 - \alpha)\delta_{b+\theta_a}$, we would obtain the same MSE's for EBKM while the other procedures would suffer increased losses for any $b \neq 0$. Of course, zero, is a natural hypothesis in many applications and it would be entirely reasonable to consider informative priors/penalties that nudged the Kiefer-Wolfowitz MLE back

toward zero. A further advantage, however, of the original Kiefer-Wolfowitz approach is that it requires no preliminary tuning parameter elicitation/selection. It may be objected that the discretization of the convex optimization problem underlying the Kiefer-Wolfowitz method requires a choice of a grid for mass points of the mixing distribution, and the optimization itself requires a convergence tolerance. But in my experience these choices are quite benign, a few hundred equally spaced grid values and a convergence tolerance like $\epsilon = 10^{-9}$ ensure highly accurate solutions to what is, after all, a convex problem.

It is perhaps also worth noting in view of recent discussions of the tradeoff between computational difficulty and statistical efficiency that although the first of the foregoing experiments was conducted with R code made available by Ryan Martin, I found that the Gibbs procedure as provided could be significantly speeded up by a bit of vectorization. Thus, in the remaining experiments results for the EBMW estimator are based on the vectorized version. The EBKM procedure employs the default `GLmix` function from the `REBayes` package, Koenker (2013), for the R language R Core Team (2013). Code to reproduce the experimental results reported above is available from `http://www.econ.uiuc.edu/~roger/research/ebayes/ebayes.html`.

2. Some Oracle Comparisons

It would be natural to wonder at this point whether the results reported above are too good to be true. An obvious benchmark would be the performance of the Bayes rule based on the unknown mixing distribution function F. In the simulations settings we have already considered there are a fixed number of non-zero observations, but for the oracle comparisons we consider priors with a multinomial number of non-zeros with expected frequencies equal to the fixed frequencies imposed in the earlier simulations.

In Tables 5, 6 and 8 below the oracle knows exactly the value of θ_a and the *expected* number of $\theta_i \neq 0$; a smarter oracle might be assumed to know the θ_a and the *actual* number of $\theta_i \neq 0$, but this formulation yields essentially similar results. In Table 7 where, like Table 3, the non-null θ_i's are normally distributed around a θ_a, the oracle is assumed to know only the mean, θ_a.

Using this modified schema for each of our previous simulation designs we report in Tables A5-8 a comparison of performance of the EBKM estimator with the oracle estimator based on the known F distribution corresponding to each of the designs in Tables 1-4. Not surprisingly, the performance of the EBKM estimator is quite similar to its performance with fixed proportions. The oracle estimator is consistently better, but only by a relatively small margin. These results strongly support the convergence result for relative Bayes risk established in Jiang and Zhang (2009).

	s = 10				**s = 20**				**s = 40**				**s = 80**			
	1	3	5	7	1	3	5	7	1	3	5	7	1	3	5	7
Oracle	11	35	10	1	15	45	14	1	24	60	17	1	37	84	24	2
EBKM	15	40	15	6	20	51	20	7	28	66	23	7	42	90	30	9

TABLE 5. MSE based on 1000 replications

	s = 25					**s = 50**					**s = 100**				
	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
Oracle	9	25	29	19	9	16	43	46	30	14	27	66	68	43	19
EBKM	10	28	32	22	11	19	46	49	33	17	30	69	71	46	22

TABLE 6. MSE based on 1000 replications

	s = 25					**s = 50**					**s = 100**				
	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
Oracle	36	60	69	59	41	64	100	112	97	70	109	161	178	154	114
EBKM	41	66	76	66	49	70	107	120	105	78	116	169	187	163	124

TABLE 7. MSE based on 1000 replications

	2	3	4	5	6	7
Oracle	171	189	126	62	33	27
EBKM	180	198	135	71	42	37

TABLE 8. MSE 1000 replications

3. ENVOI

A reviewer has noted that the theory developed by Jiang and Zhang (2009) for the adaptive minimaxity of the Bayes risk of Kiefer-Wolfowitz procedures excludes settings for ℓ_p classes with $p = 0$. For these "nearly black" classes – to use the terminology of Johnstone and Silverman (2004) – with,

$$\ell_0(\eta) = \{\theta = (\theta_1, \cdots, \theta_n) \mid n^{-1} \sum_{i=1}^{n} I(\theta_i \neq 0) < \eta\}$$

and typifying the experimental settings employed here, little is presently known. I would like to think that the foregoing experimental results might encourage others to explore these questions further.

REFERENCES

BHATTACHARYA, A., D. PATI, N. S. PILLAI, AND D. B. DUNSON (2012): "Bayesian shrinkage," *arXiv:1212.6088*.

Carvalho, C., N. Polson, and J. Scott (2009): "Handling sparsity via the horseshoe," *Journal of Machine Learning Research*, 5, 73–80.

Castillo, I., and A. van der Vaart (2012): "Needles and Straw in a Haystack: Posterior concentration for possibly sparse sequences," *Annals of Statistics*, 40, 2069–2101.

Jiang, W., and C.-H. Zhang (2009): "General maximum likelihood empirical Bayes estimation of normal means," *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*, 32, 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 0.39, `http://www.r-project.org`.

Koenker, R., and I. Mizera (2013): "Convex Optimization, Shape Constraints, Compound Decisions, and Empirical Bayes Rules," *J of American Statistical Association*, forthcoming, available at: `http://www.econ.uiuc.edu/~roger/research/ebayes/ebayes.html`.

Martin, R., and S. G. Walker (2013): "Asymptotically minimax empirical Bayes estimation of a sparse normal mean," *arXiv:1304.7366*.

R Core Team (2013): *R: A Language and Environment for Statistical Computing,* R Foundation for Statistical Computing, Vienna, Austria, `http://www.R-project.org`.