Borrowing Strength from Experience: Empirical Bayes Methods and Convex Optimization

Ivan Mizera

University of Alberta
Edmonton, Alberta, Canada
Department of Mathematical and Statistical Sciences

("Edmonton Eulers")

Wien, January 2016

Research supported by the
Natural Sciences and Engineering Research Council of Canada

co-authors: Roger Koenker (Illinois), Mu Lin, Sile Tao (Alberta)

Compound decision problem

• Estimate (predict) a vector $\mu = (\mu_1, \cdots, \mu_n)$

Many quantities here, and not that much of sample for those

• Observing one Y_i for every μ_i, with (conditionally) known distribution

Example: $Y_i \sim \text{Binomial}(n_i, \mu_i)$, n_i known

Another example: $Y_i \sim \mathcal{N}(\mu_i, 1)$

Yet another example: $Y_i \sim \text{Poisson}(\mu_i)$

• μ_i's assumed to be sampled ($\rightarrow$ random) iid-ly from P

Thus, when the conditional density (...) of the Y_i's is $\varphi(y, \mu)$, then the marginal density of the Y_i's is

$$g(y) = \int \varphi(y, \mu) dP(\mu)$$

A sporty example

Data: known performance of individual players, typically summarized as of successes, k_i, in a number, n_i, of some repeated trials (bats, penalties) - typically, data not very extensive (start of the season, say); the objective is to predict "true" capabilities of individual players

One possibility: $Y_i = k_i \sim \text{Binomial}(n_i, \mu_i)$

Another possibility: take $Y_i = \arcsin \dfrac{k_i + 1/4}{n_i + 1/2} \mathrel{\dot\sim} N\left(\mu_i, \dfrac{1}{4n_i}\right)$

Solutions via maximum likelihood

$$\hat{\mu}_i = k_i/n_i \quad \text{or } \hat{\mu}_i = \mu_i$$

The overall mean (or marginal MLE) is often better than this

Efron and Morris (1975), Brown (2008),
Koenker and Mizera (2014): bayesball

NBA data (Agresti, 2002)

	player	n	k	prop
1	Yao	13	10	0.7692
2	Frye	10	9	0.9000
3	Camby	15	10	0.6667
4	Okur	14	9	0.6429
5	Blount	6	4	0.6667
6	Mihm	10	9	0.9000
7	Ilgauskas	10	6	0.6000
8	Brown	4	4	1.0000
9	Curry	11	6	0.5455
10	Miller	10	9	0.9000
11	Haywood	8	4	0.5000
12	Olowokandi	9	8	0.8889
13	Mourning	9	7	0.7778
14	Wallace	8	5	0.6250
15	Ostertag	6	1	0.1667

it may be better to take the overall mean!

An insurance example

Y_i - known number of accidents of individual insured motorists

Predict their expected number - rate, μ_i (in next year, say)

$Y_i \sim \text{Poisson}(\mu_i)$

Maximum likelihood: $\hat{\mu}_i = Y_i$

Nothing better?

The data of Simar (1976)

			$E_G(\theta_i \mid y_i)$		
y_i	count	$\hat{m}_G(y_i)$	Robbins	Gamma	NPML
0	7840	.82867	.168	.159	.168
1	1317	.13920	.363	.417	.372
2	239	.02526	.527	.675	.610
3	42	.00444	1.333	.933	1.001
4	14	.00148	1.429	1.191	1.952
5	4	.00042	6.000	1.449	2.836
6	4	.00042	1.750	1.707	3.123
7	1	.00011	0.000	1.965	3.142

Table 3.1 *Simar (1976) Accident Data: Observed counts and empirical Bayes posterior means for each number of claims per year for $k = 9461$ policies issued by La Royal Belge Insurance Company. The y_i are the observed frequencies, $\hat{P}_G$ is the observed relative frequency, "Robbins" is the Robbins NPEB rule, "Gamma" is the PEB posterior mean estimate based on the Poisson/gamma model, and "NPML" is the posterior mean estimate based on the EB rule for the nonparametric prior.*

So, what is better?

First, what is better?

We express it via some (expected) loss function

Most often it is averaged or aggregated squared error loss

$$\sum_i (\hat{\mu}_i - \mu_i)^2$$

But it could be also some other loss...

And then?

Well, it is sooo simple...

... if P is known!

μ_i's are sampled iid-ly from P - prior distribution

Conditionally on μ_i, the distribution of Y_i is, say, $N(\mu_i, 1)$

The optimal prediction is the posterior mean, the mean of the posterior distribution: conditional distribution of μ_i given Y_i (given that the loss function is quadratic!)

For instance, if P is $N(0, \sigma^2)$, then (homework)

the best predictor is $\hat{\mu}_i = Y_i - \frac{1}{\sigma^2 + 1} Y_i$

Borrowing strength via shrinkage

"neither will be the good that good, nor the bad that bad"

More generally, μ_i can be $N(\mu, \sigma^2)$ and Y_i then $N(\mu_i, \sigma_0^2)$,

And then $\hat{\mu}_i = Y_i - \frac{\sigma_0^2}{\sigma^2 + \sigma_0^2}(Y_i - \mu)$ (if $\sigma^2 = \sigma_0^2$, halfway to μ)

“If only all of them published posthumously...”

Thomas Bayes (1701–1761)

But do we know P (or σ^2)?

"Hierarchical model"

"Random effects"

"Smoothing"

"Empirical Bayes"

"no less Bayes than empirical Bayes"

"we know it is frequentist, but frequentists think it is Bayesian, so this is why we discuss it here"

Many inventors ...

What is mathematics?

Herbert Ellis Robbins (1915–2001)

On experience in statistical decision theory (1954)

Antonín Špaček (1911–1961)

I. J. Good (2000)

Alann Mathison Turing (1912–1954)

So, how

A. we may try to estimate the prior - "f-modeling", Efron (2014)

B. or more directly, the prediction rule - "g-modeling"

A′. Estimated normal prior (parametric)

(Nonparametric ouverture)

A. Empirical prior (nonparametric)

B. Empirical prediction rule (nonparametric)

Simulation contests

A′. Estimated normal prior

James-Stein (JS): if P is $N(0, \sigma^2)$

then the unknown part, $\frac{1}{\sigma^2 + 1}$, of the prediction rule

can be estimated by $\frac{n-2}{S}$, where $S = \sum_i Y_i^2$

A′. Estimated normal prior

James-Stein (JS): if P is $N(0, \sigma^2)$

then the unknown part, $\frac{1}{\sigma^2 + 1}$, of the prediction rule

can be estimated by $\frac{n-2}{S}$, where $S = \sum_i Y_i^2$

For general μ in place of 0, the rule is

$\hat{\mu}_i = Y_i - \frac{n-3}{S}(Y_i - \bar{Y})$, with $\bar{Y} = \frac{1}{n}\sum_i Y_i$ and $S = \sum_i (Y_i - \bar{Y})^2$

JS as empirical Bayes: Efron and Morris (1975)

Charles Stein (1920–)

Nonparametric ouverture: MLE of density

Density estimation: given the datapoints $X_1, X_2, \ldots, X_n$, solve

$$\prod_{i=1}^{n} g(X_i) \hookrightarrow \max_{g}!$$

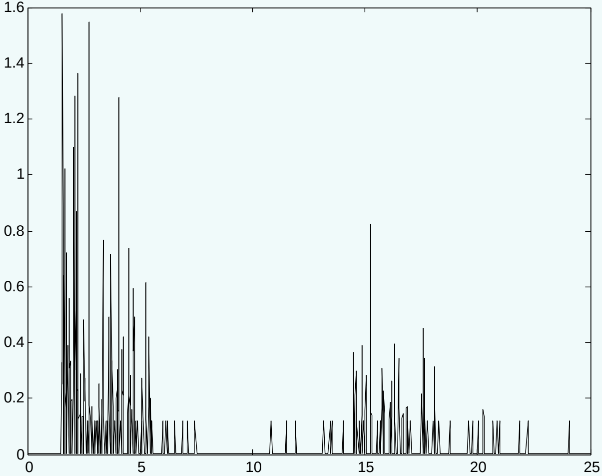
or equivalently

$$-\sum_{i=1}^{n} \log g(X_i) \hookrightarrow \min_{g}!$$

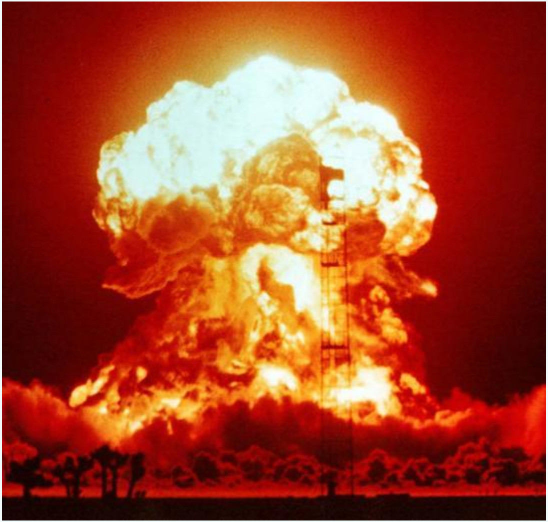
under the side conditions

$$g \geqslant 0, \quad \int g = 1$$

Doesn't work



How to prevent Dirac catastrophe?



Reference

Koenker and Mizera (2014)
... and those that cite it (Google Scholar)

"... the chance meeting on a dissecting-table of a sewing-machine and an umbrella"

See also REBayes package on CRAN

For simplicity:
$\varphi(y, \mu) = \varphi(y - \mu)$, and the latter is standard normal density

A. Empirical prior

MLE of P: Kiefer and Wolfowitz (1956)

$$-\sum_i \log\left(\int \varphi(Y_i - u)\, dP(u)\right) \leadsto \min_P !$$

The regularizer is the fact that it is a mixture

No tuning parameter needed (but "known" form of φ!)

The resulting $\hat{P}$ is atomic ("empirical prior")

However, it is an infinite-dimensional problem...

EM nonsense

Laird (1978), Jiang and Zhang (2009):

Use a grid $\{u_1, ...u_m\}$ $\quad (m = 1000)$

containing the support of the observed sample

and estimate the "prior density" via EM iterations

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

Slooooow... (original versions: 55 hours for 1000 replications)

Convex optimization!

Koenker and Mizera (2014): it is a convex problem!

$$-\sum_i \log\left(\int \varphi(Y_i - u)\, dP(u)\right) \leadsto \min_P !$$

When discretized

$$-\sum_i \log\left(\sum_m \varphi(Y_i - u_j) p_j\right) \leadsto \min_p !$$

or in a more technical form

$$-\sum_i \log y_i \leadsto \min_y ! \qquad Az = y \text{ and } z \in \mathcal{S}$$

where $A = (\varphi(Y_i - u_j))$ and $\mathcal{S} = \{s \in \mathbb{R}^m : 1^\top s = 1,\ s \geqslant 0\}$.

With a dual

The solution is an atomic probability measure, with not more than n atoms. The locations, $\hat{\mu}_j$, and the masses, $\hat{p}_j$, at these locations can be found via the following dual characterization: the solution, $\hat{\nu}$, of

$$\sum_{i=1}^{n} \log \nu_i \looparrowright \max_{\mu}! \quad \sum_{i=1}^{n} \nu_i \varphi(Y_i - \mu) \leqslant n \text{ for all } \mu$$

satisfies the extremal equations $\sum_j \varphi(Y_i - \hat{\mu}_j)\hat{p}_j = \frac{1}{\hat{\nu}_i}$,

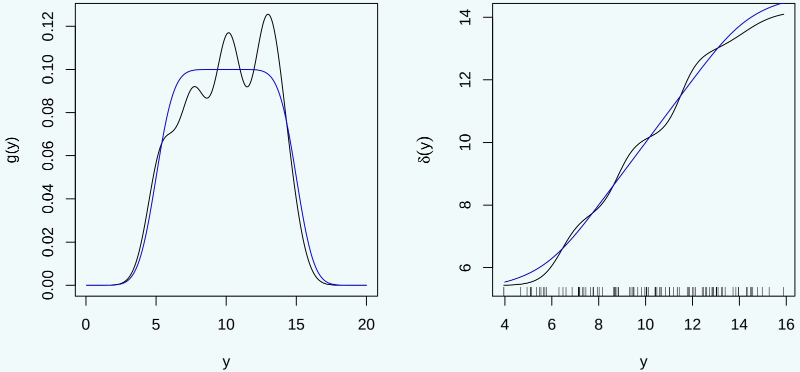
and $\hat{\mu}_j$ are exactly those μ where the dual constraint is active.

And one can use modern convex optimization methods again...

(And note: everything goes through for general $\varphi(y, \mu)$)

(And one can also handle - numerically - alternative loss functions!)

A typical result: μ_i drawn from $\mathcal{U}(5, 15)$



Left: mixture density (blue: target)

Right: decision rule (blue: target)

B. Empirical prediction rule

Lawrence Brown, personal communication

Also, looks like in Maritz and Lwin (1989)

Do not estimate P, but rather the prediction rule

Tweedie formula: for known (general) P, and hence known g, the Bayes rule is

$$\delta(y) = y + \sigma^2 \frac{g'(y)}{g(y)}$$

One may try to estimate g and plug it in - when knowing σ^2 (=1, for instance)

Brown and Greenshtein (2009)

by an exponential family argument, $\delta(y)$ is nondecreasing in y (van Houwelingen & Stijnen, 1983)

(that came automatic when the prior is estimated)

Monotone (estimate of) empirical Bayes rule

Maximum likelihood again ($h = \log g$)

- but with some shape-constraint regularization,
- like log-concavity: $(\log g)'' \leqslant 0$
- but we rather want $y + \dfrac{g'(y)}{g(y)} = y + (\log g(y))'$ nondecreasing
- that is, $\frac{1}{2}y^2 + \log g(y) = \frac{1}{2}y^2 + h(y) \quad$ convex

Monotone (estimate of) empirical Bayes rule

Maximum likelihood again ($h = \log g$)

- but with some shape-constraint regularization,
- like log-concavity: $(\log g)'' \leqslant 0$
- but we rather want $y + \dfrac{g'(y)}{g(y)} = y + (\log g(y))'$ nondecreasing
- that is, $\frac{1}{2}y^2 + \log g(y) = \frac{1}{2}y^2 + h(y)$ convex

$$-\sum_{i=1}^{n} \log g(X_i) \looparrowright \min_g ! \qquad\qquad g \geqslant 0, \quad \int g = 1$$

Monotone (estimate of) empirical Bayes rule

Maximum likelihood again ($h = \log g$)

- but with some shape-constraint regularization,
- like log-concavity: $(\log g)'' \leqslant 0$
- but we rather want $y + \dfrac{g'(y)}{g(y)} = y + (\log g(y))'$ nondecreasing
- that is, $\frac{1}{2}y^2 + \log g(y) = \frac{1}{2}y^2 + h(y)$ convex

$$-\sum_{i=1}^{n} \log g(X_i) \looparrowright \min_g ! \quad -\log g \text{ convex} \quad g \geqslant 0 \quad \int g\,dx = 1$$

Monotone (estimate of) empirical Bayes rule

Maximum likelihood again ($h = \log g$)

- but with some shape-constraint regularization,
- like log-concavity: $(\log g)'' \leqslant 0$
- but we rather want $y + \dfrac{g'(y)}{g(y)} = y + (\log g(y))'$ nondecreasing
- that is, $\frac{1}{2}y^2 + \log g(y) = \frac{1}{2}y^2 + h(y) \quad$ convex

$$-\sum_{i=1}^{n} h(X_i) \looparrowright \min_h ! \qquad -h \text{ convex} \quad e^h \geqslant 0 \quad \int e^h dx = 1$$

Monotone (estimate of) empirical Bayes rule

Maximum likelihood again ($h = \log g$)

- but with some shape-constraint regularization,
- like log-concavity: $(\log g)'' \leqslant 0$
- but we rather want $y + \dfrac{g'(y)}{g(y)} = y + (\log g(y))'$ nondecreasing
- that is, $\frac{1}{2}y^2 + \log g(y) = \frac{1}{2}y^2 + h(y) \quad$ convex

$$-\sum_{i=1}^{n} h(X_i) \looparrowright \min_h ! \qquad -h \text{ convex} \quad e^h \geqslant 0 \quad \int e^h dx = 1$$

Monotone (estimate of) empirical Bayes rule

Maximum likelihood again ($h = \log g$)

- but with some shape-constraint regularization,

- like log-concavity: $(\log g)'' \leqslant 0$

- but we rather want $y + \dfrac{g'(y)}{g(y)} = y + (\log g(y))'$ nondecreasing

- that is, $\frac{1}{2}y^2 + \log g(y) = \frac{1}{2}y^2 + h(y)$ convex

$$-\sum_{i=1}^{n} h(X_i) \looparrowright \min_{h}! \quad -h \text{ convex} \quad \int e^h dx = 1$$

Monotone (estimate of) empirical Bayes rule

Maximum likelihood again ($h = \log g$)

- but with some shape-constraint regularization,

- like log-concavity: $(\log g)'' \leqslant 0$

- but we rather want $y + \dfrac{g'(y)}{g(y)} = y + (\log g(y))'$ nondecreasing

- that is, $\frac{1}{2}y^2 + \log g(y) = \frac{1}{2}y^2 + h(y)$ convex

$$-\sum_{i=1}^{n} h(X_i) \leadsto \min_{h}! \quad - h \text{ convex} \quad \int e^{h} dx = 1$$

Monotone (estimate of) empirical Bayes rule

Maximum likelihood again ($h = \log g$)

- but with some shape-constraint regularization,
- like log-concavity: $(\log g)'' \leqslant 0$
- but we rather want $y + \dfrac{g'(y)}{g(y)} = y + (\log g(y))'$ nondecreasing
- that is, $\quad \frac{1}{2}y^2 + \log g(y) = \frac{1}{2}y^2 + h(y) \quad$ convex

$$-\sum_{i=1}^{n} h(X_i) + \int e^h dx \looparrowright \min_h ! \qquad - h \text{ convex}$$

Monotone (estimate of) empirical Bayes rule

Maximum likelihood again ($h = \log g$)

- but with some shape-constraint regularization,
- like log-concavity: $(\log g)'' \leqslant 0$
- but we rather want $y + \dfrac{g'(y)}{g(y)} = y + (\log g(y))'$ nondecreasing
- that is, $\frac{1}{2}y^2 + \log g(y) = \frac{1}{2}y^2 + h(y) \quad$ convex

$$-\sum_{i=1}^{n} h(X_i) + \int e^h dx \looparrowright \min_h ! \quad \frac{1}{2}y^2 + h(y) \text{ convex}$$

Monotone (estimate of) empirical Bayes rule

Maximum likelihood again ($h = \log g$)

- but with some shape-constraint regularization,
- like log-concavity: $(\log g)'' \leqslant 0$
- but we rather want $y + \dfrac{g'(y)}{g(y)} = y + (\log g(y))'$ nondecreasing
- that is, $\frac{1}{2}y^2 + \log g(y) = \frac{1}{2}y^2 + h(y) \quad$ convex

$$-\sum_{i=1}^{n} h(X_i) + \int e^h dx \leadsto \min_h ! \quad \frac{1}{2}y^2 + h(y) \text{ convex}$$

The regularizer is the monotonicity constraint

No tuning parameter, or knowledge of φ

- but knowing all the time that $\sigma^2 = 1$

A convex problem again

Some remarks

After reparametrization, omitting constants, etc. one can write it as a solution of an equivalent problem

$$-\frac{1}{n}\sum_{i=1}^{n} K(Y_i) + \int e^{K(y)} d\Phi_c(y) \looparrowright \min_K ! \quad K \in \mathcal{K}$$

Compare:

$$-\frac{1}{n}\sum_{i=1}^{n} h(X_i) + \int e^{h} dx \looparrowright \min_h ! \quad -h \in \mathcal{K}$$

Dual formulation

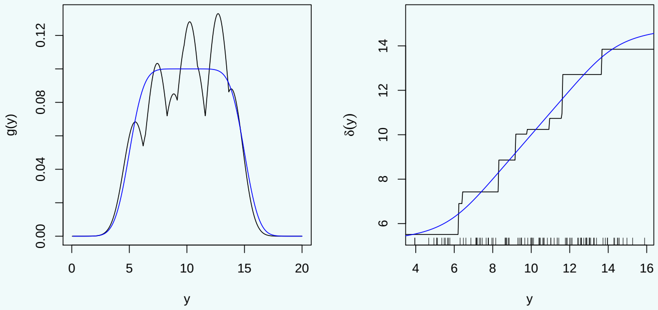
Analogous to Koenker and Mizera (2010):

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

$$-\int f(y) \log f(y) d\Phi(y) \mapsto \min_f ! \quad f = \frac{d(P_n - G)}{d\Phi}, G \in \mathcal{K}^-$$

The estimated decision rule, $\hat{\delta}$, is piecewise constant and has no jumps at $\min Y_i$ and $\max Y_i$.

A typical result: μ_i drawn from $\mathcal{U}(5, 15)$



Left: mixture density (blue: target)

Right: piecewise constant, "empirical decision rule"

Doable also for some other exponential families

However: a version of the Tweedie formula may be obtainable only for the canonical parameter (binomial!) and depends on the loss function

For the Poisson case:

- the optimal prediction with respect to the quadratic loss function is, for $x = 0, 1, 2, \dots,$

$$\hat{\mu}(x) = \frac{(x+1)g(x+1)}{g(x)} \,,$$

where g is the marginal density of the Y_i's

- for the loss function $(\mu - \hat{\mu})^2/\mu$, the optimal prediction is, for $x = 1, 2, \dots$

$$\hat{\mu}(x) = \frac{x g(x)}{g(x-1)} \,.$$

What can be done with that?

One can estimate $g(x)$ by the relative frequency, as Robbins (1956):

$$\hat{\mu}(x) = \frac{(x+1)\frac{\#\{Y_i = x+1\}}{n}}{\frac{\#\{Y_i = x\}}{n}} = \frac{(x+1)\#\{Y_i = x+1\}}{\#\{Y_i = x\}}$$

however, the predictions obtained this way are not monotone, and also erratic, especially when some denominator is 0 - the latter can be rectified by the adjustment of Maritz and Lwin (1989):

$$\hat{\mu}(x) = \frac{(x+1)\#\{Y_i = x+1\}}{1 + \#\{Y_i = x\}}$$

Better: monotonizations

The suggestion of van Houwelingen & Stijnen (1983): pool adjacent violators - also requires a grid

Or one can estimate the marginal density under the shape-restriction that the resulting prediction is monotone:

$$\frac{(x+1)\hat{g}(x+1)}{\hat{g}(x)} \leqslant \frac{(x+2)\hat{g}(x+2)}{\hat{g}(x+1)}$$

After reparametrization in terms of logarithms, the problem is almost linear: linear constraint resulting from the one above, and linear objective function - with a nonlinear Lagrange term ensuring that the result is a probability mass function. At any rate, again a convex problem - and the number of variables is the number of the x's

Why all this is feasible: interior point methods

(Leave optimization to experts)

Andersen, Christiansen, Conn, and Overton (2000)

We acknowledge using Mosek, a Danish optimization software

Mosek: E. D. Andersen (2010)

PDCO: Saunders (2003)

Nesterov and Nemirovskii (1994)

Boyd, Grant and Ye: Disciplined Convex Programming

Folk wisdom: “If it is convex, it will fly.”

Simulations - or how to be highly cited

Johnstone and Silverman (2004): empirical Bayes for sparsity

$n = 1000$ observations
k of which have μ all equal to one of the 4 values, $3, 4, 5, 7$
the remaining $n - k$ have $\mu = 0$
there are three choices of k: 5, 50, 500

Criterion: sum of squared errors, averaged over replications, and rounded

Seems like this scenario (or similar ones) became popular

The first race

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

- empirical prediction rule
- empirical prior, implementation via convex optimization
- empirical prior, implementation via EM
- Brown and Greenshtein (2009): 50 replications
  report (best?) results for bandwith-related constant 1.15
- Johnstone and Silverman (2004): 100 replications, 18 methods
  (only their winner reported here, J-S Min)

A new lineup

	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
EBB	224	243	171	92	53	45
EBKM	207	223	152	79	44	37
oracle	197	214	144	71	34	27

Bhattacharya, Pati, Pillai, Dunson (2012): "Bayesian shrinkage"

BL: "Bayesian Lasso"

DL: "Dirichlet-Laplace priors" (with different strengths)

HS: Carvalho, Polson, and Scott (2009) "horseshoe priors"

EBMW: "asympt. minimax EB" of Martin and Walker (2013)

elsewhere: Castillo & van der Vaart (2012) "posterior concentration"

Comments (Conclusions ?)

• both approaches typically outperform other methods

• Kiefer-Wolfowitz empirical prior typically outperforms monotone empirical Bayes (for the examples we considered!)

• both methods adapt to general P, in particular to those with multiple modes

• however, Kiefer-Wolfowitz empirical prior is more flexible: (much) better adapts to certain peculiarities vital in practical data analysis, like unequal σ_i, inclusion of covariates, etc

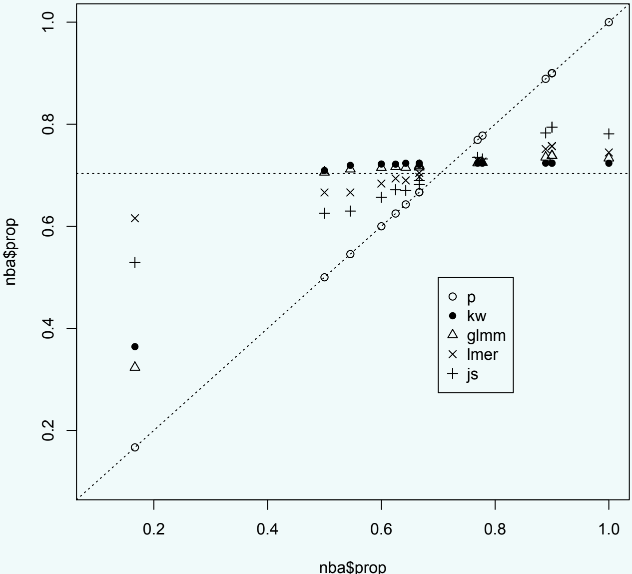
• in particular, it also exhibits certain independence of the choice of the loss function (the estimate of the prior, and hence posterior is always the same)

• but, in certain situations Kiefer-Wolfowitz (on the grid!) may be more computationally demanding

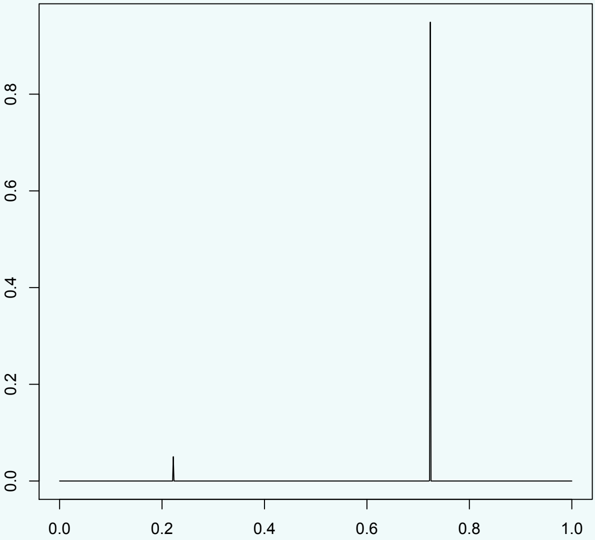
NBA data again

	player	n	prop	k	ast	sigma	ebkw	jsmm	glmm	lmer
1	Yao	13	0.769	10	1.058	0.139	0.724	0.735	0.724	0.729
2	Frye	10	0.900	9	1.219	0.158	0.724	0.794	0.738	0.757
3	Camby	15	0.667	10	0.950	0.129	0.724	0.682	0.716	0.697
4	Okur	14	0.643	9	0.925	0.134	0.724	0.670	0.715	0.690
5	Blount	6	0.667	4	0.942	0.204	0.721	0.689	0.719	0.705
6	Mihm	10	0.900	9	1.219	0.158	0.724	0.794	0.738	0.757
7	Ilgauskas	10	0.600	6	0.881	0.158	0.722	0.657	0.715	0.684
8	Brown	4	1.000	4	1.333	0.250	0.724	0.781	0.733	0.745
9	Curry	11	0.545	6	0.829	0.151	0.719	0.630	0.712	0.666
10	Miller	10	0.900	9	1.219	0.158	0.724	0.794	0.738	0.757
11	Haywood	8	0.500	4	0.785	0.177	0.709	0.626	0.706	0.666
12	Olowokandi	9	0.889	8	1.200	0.167	0.724	0.783	0.735	0.751
13	Mourning	9	0.778	7	1.063	0.167	0.724	0.732	0.725	0.727
14	Wallace	8	0.625	5	0.904	0.177	0.722	0.672	0.717	0.694
15	Ostertag	6	0.167	1	0.454	0.204	0.364	0.529	0.323	0.616

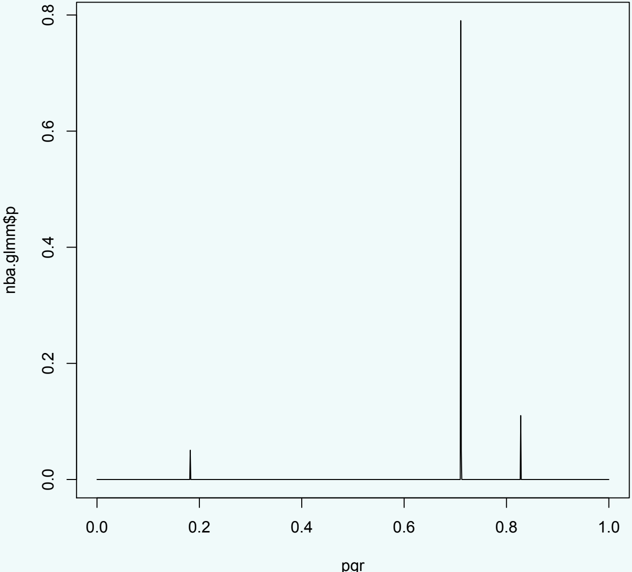
A (partial) picture



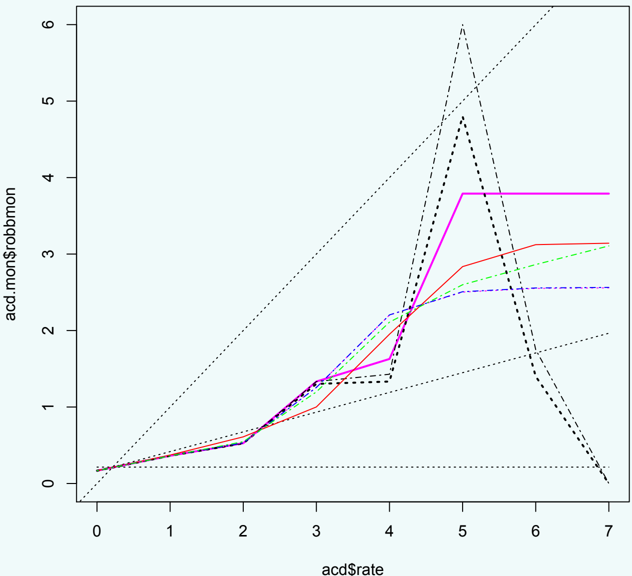
Mixing distribution ("empirical prior")



Mixing distribution for glmm



The auto insurance predictions



That's it?

What if P is unimodal? Cannot we do better in such a case?

That's it?

What if P is unimodal? Cannot we do better in such a case?

And if we can, will it be (significantly) better than James-Stein?

That's it?

What if P is unimodal? Cannot we do better in such a case?

And if we can, will it be (significantly) better than James-Stein?

Joint work with Mu Lin

OK, so just impose unimodality on P ...

... or more precisely, constrain P to be log-concave (or q-convex)
(unimodality does not work well in this context)

OK, so just impose unimodality on P ...

... or more precisely, constrain P to be log-concave (or q-convex)
(unimodality does not work well in this context)

However, the resulting problem is not convex!

OK, so just impose unimodality on P ...

... or more precisely, constrain P to be log-concave (or q-convex)
(unimodality does not work well in this context)

However, the resulting problem is not convex!

Nevertheless, given that:
log-concavity of P + that of φ implies that of the convolution

$$g(y) = \int \varphi(y - \mu) dP(\mu)$$

OK, so just impose unimodality on P ...

... or more precisely, constrain P to be log-concave (or q-convex)
(unimodality does not work well in this context)

However, the resulting problem is not convex!

Nevertheless, given that:
log-concavity of P + that of φ implies that of the convolution

$$g(y) = \int \varphi(y - \mu) dP(\mu)$$

one can impose log-concavity on the mixture!
(So that the resulting formulation then a convex problem is.)

3. “Unimodal” Kiefer-Wolfowitz

$$g \looparrowright \min_P ! \quad g = -\sum_i \log \left(\int \varphi(Y_i - u)\, dP(u) \right)$$

(Works, but needs a special version of Mosek)

May be demanding for large sample sizes

3. “Unimodal” Kiefer-Wolfowitz

$$g \looparrowright \min_P ! \quad g = -\sum_i \log\left(\int \varphi(Y_i - u)\, dP(u)\right) \quad \text{and } g \text{ convex}$$

(Works, but needs a special version of Mosek)

May be demanding for large sample sizes

3. "Unimodal" Kiefer-Wolfowitz

$$g \looparrowright \min_{g,P}! \quad g \geqslant -\sum_i \log\left(\int \varphi(Y_i - u)\,dP(u)\right) \quad \text{and } g \text{ convex}$$

(Works, but needs a special version of Mosek)

May be demanding for large sample sizes

4. “Unimodal” monotone empirical Bayes

$\frac{1}{2}y^2 + h(y)$ convex

$h(y)$ concave

$$-\sum_{i=1}^{n} h(X_i) + \int e^h dx \looparrowright \min_h ! \quad \frac{1}{2}y^2 + h(y) \text{ convex}$$

4. "Unimodal" monotone empirical Bayes

$\frac{1}{2}y^2 + h(y)$ convex

$h(y)$ concave

$$-\sum_{i=1}^{n} h(X_i) + \int e^h dx \looparrowright \min_h ! \quad 1 + h''(y) > 0$$

4. “Unimodal” monotone empirical Bayes

$\frac{1}{2}y^2 + h(y)$ convex

$h(y)$ concave

$$-\sum_{i=1}^{n} h(X_i) + \int e^h dx \hookrightarrow \min_h ! \qquad h''(y) > -1$$

4. "Unimodal" monotone empirical Bayes

$\frac{1}{2}y^2 + h(y)$ convex

$h(y)$ concave

$$-\sum_{i=1}^{n} h(X_i) + \int e^h dx \looparrowright \min_h ! \quad 0 > h''(y) > -1$$

4. "Unimodal" monotone empirical Bayes

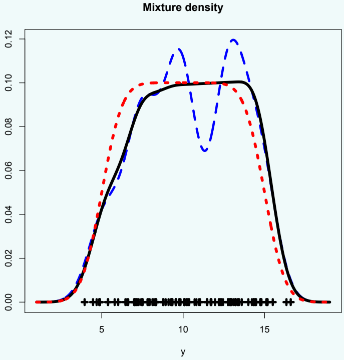
$\frac{1}{2}y^2 + h(y)$ convex

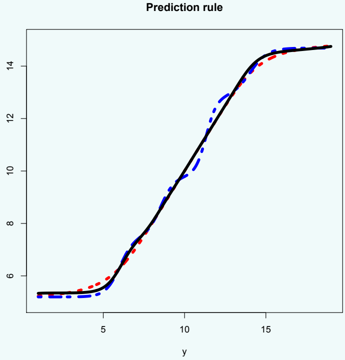
$h(y)$ concave

$$-\sum_{i=1}^{n} h(X_i) + \int e^h dx \looparrowright \min_h ! \quad 0 > h''(y) > -1$$

Very easy, very fast

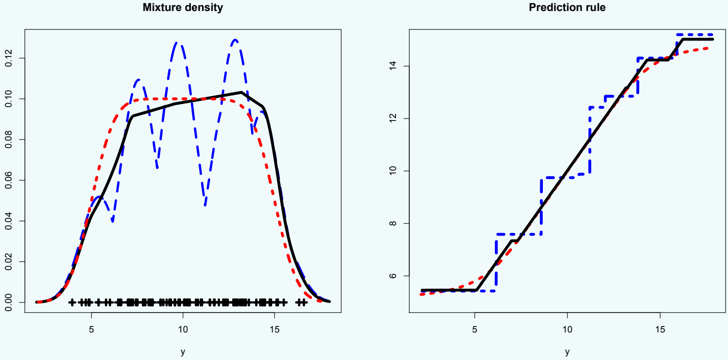
A typical result, again from $\mathcal{U}(5, 15)$





(Empirical prior, mixture unimodal)

A typical result, again from $\mathcal{U}(5, 15)$



(Empirical prediction rule, mixture unimodal)

Some simulations

Sum of squared errors, averaged over replications, rounded

	$U[5,15]$	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

Last four: the mixtures of Johnstone and Silverman (2004): $n = 1000$ observations, with 5% or 50% of μ equal to 2 or 5 and the remaining ones are 0

Conclusions II

- when the mixing (and then the mixture) distribution is unimodal, it pays to enforce this shape constraint for the estimate

Conclusions II

- when the mixing (and then the mixture) distribution is unimodal, it pays to enforce this shape constraint for the estimate

- if it is not, then it does not pay

Conclusions II

- when the mixing (and then the mixture) distribution is unimodal, it pays to enforce this shape constraint for the estimate

- if it is not, then it does not pay

- unimodal Kiefer-Wolfowitz still appears to outperform the unimodal monotonized empirical Bayes by small margin

Conclusions II

- when the mixing (and then the mixture) distribution is unimodal, it pays to enforce this shape constraint for the estimate

- if it is not, then it does not pay

- unimodal Kiefer-Wolfowitz still appears to outperform the unimodal monotonized empirical Bayes by small margin

- and both outperform James-Stein, significantly for asymmetric mixing distribution

Conclusions II

- when the mixing (and then the mixture) distribution is unimodal, it pays to enforce this shape constraint for the estimate

- if it is not, then it does not pay

- unimodal Kiefer-Wolfowitz still appears to outperform the unimodal monotonized empirical Bayes by small margin

- and both outperform James-Stein, significantly for asymmetric mixing distribution

- computationally, unimodal monotonized empirical Bayes is much more painless than unimodal Kiefer-Wolfowitz