Využitie skúsenosti v predikcii: Empirické bayesovské metódy, kvalitatívne ohraničenia a konvexná optimalizácia

Ivan Mizera

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

("Edmonton Eulers")

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

Joint work with Roger Koenker (Illinois) and Mu Lin (Alberta)

Kšaft umírající statistiky matematické

L. Breiman (1995): Reflections After Refereeing Papers for NIPS

As a result of the would-be mathematicians in statistics, it has been dominated by useless theory and fads.

- ■ Decision Theory
- → ■ Asymptotics
- → ■ Robustness
- ■ Nonparametric One and Two Sample Tests
- → ■ One-Dimensional Density Estimation
- ■ Etc.

Mikhail Lermontov: A Hero of Our Times

MLP: "Kritérium pravdy je prax."

Zložený rozhodovací problém

"(Empirical) Bayes", "Hierarchical model",
"Random effects", "Smoothing"

Estimate a vector $\mu = (\mu_1, \cdots, \mu_n)$
Conditionally normal sample, $Y_i \sim \mathcal{N}(\mu_i, 1)$, $i = 1, \cdots, n$.
μ_i's are assumed to be sampled iid-ly from P
So that the Y_i's have density (φ is the density of $\mathcal{N}(0,1)$)

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

Problem: to estimate (predict) μ_i
The MLE is $\hat{\mu}_i = Y_i$ the best we can do?
"Best": optimal w. r. t. averaged squared error loss, $(\hat{\mu} - \mu)^2$

Berme to športovo

An example:

Y_i - known performance of individual players, typically summarized as of successes, k_i, in a number, n_i, of some repeated trials (bats, penalties)

Naïve, individual MLE's: the relative frequency, k_i/n_i

predicting μ_i - the "true" capabilities of individual players, on probability scale

typically, data not very extensive (start of the season, say)

so that the overall mean is often better than the MLE's

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

Ešte jeden príklad, z NBA (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!

Technické podrobnosti

The assumption of normal distribution of Y_i typically results from an approximation of a binomial - so one can buy somewhat artificially looking assumption of unit variances

(or one can do a binomial mixture)

(or one can do something else)

An alternative to MLE: borrowing strength $\rightarrow$ shrinkage
"neither will be the good that good, nor the bad that bad"

Nič jednoduchšie

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

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

The optimal prediction is the mean of the posterior distribution: conditional distribution of μ_i given Y_i

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

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

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

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

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

Thomas Bayes (1701–1761)

Takže čo?

How do we know what is σ^2? Or why P is normal?

0. Estimated normal prior (parametric)

Nonparametric ouverture

1. Empirical prior (nonparametric)
2. Empirical prediction rule (nonparametric)

Simulation contests

A bit of data analysis

3. Empirical prior with unimodal mixture distribution
4. Empirical prediction rule with unimodal mixture distribution

A bit more simulations and conclusions

There is no less Bayes than empirical Bayes

Herbert Ellis Robbins (1915–2001)

On experience in statistical decision theory (1954)

Antonín Špaček (1911–1961)

0. Odhadované normálne apriórne rozdelenie

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 $S = \sum_i Y_i^2$

0. Odhadované normálne apriórne rozdelenie

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 ako empirický Bayes: Efron and Morris (1975)

Charles Stein (1920–)

Neparametrická predohra: maximálne vierohodný odhad hustoty

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

$$\prod_{i=1}^{n} g(X_i) \looparrowright \max_g !$$

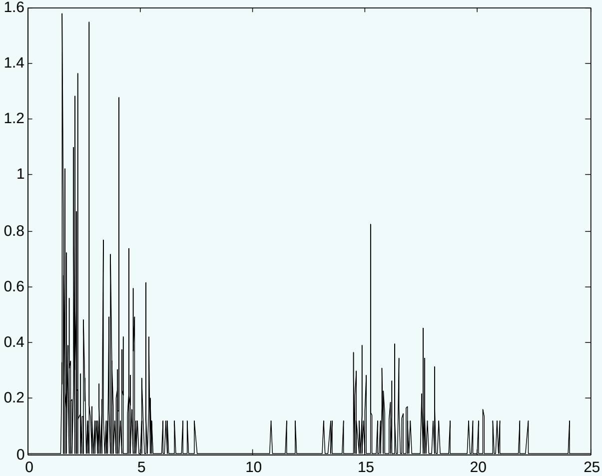
or equivalently

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

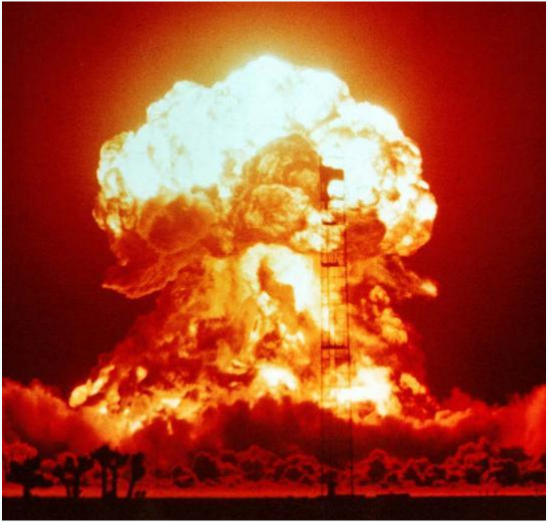
under the side conditions

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

Nejako to nefunguje ("Pr...r")



Ako zabrániť Diracovej katastrofe?



Regularizácia! Cez penalty...

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

Regularizácia! Cez penalty...

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

Regularizácia! Cez penalty...

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

$J(\cdot)$ - penalty (penalizing complexity, lack of smoothness etc.)

For instance, Koenker and Mizera (2006, 2007a)

$$J(g) = \bigvee (\log g)' = \int |(\log g)''|$$

or also $J(g) = \bigvee (\log g)'' = \int |(\log g)'''|$

Λ - regularization parameter (the extent of regularization)

Regularizácia! Cez penalty...

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

$J(\cdot)$ - penalty (penalizing complexity, lack of smoothness etc.)

For instance, Koenker and Mizera (2006, 2007a)

$J(g) = \bigvee (\log g)' = \int |(\log g)''|$

or also $J(g) = \bigvee (\log g)'' = \int |(\log g)'''|$

Λ - regularization parameter (the extent of regularization)

... a tuning parameter!

Regularizácia! Cez tvarové ohraničenia...

Monotonicity, log-concavity: $(\log g)'' \leqslant 0$
Notation: $\mathcal{K}$ is the cone of convex functions

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

Regularizácia! Cez tvarové ohraničenia...

Monotonicity, log-concavity: $(\log g)'' \leqslant 0$
Notation: $\mathcal{K}$ is the cone of convex functions

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

Regularizácia! Cez tvarové ohraničenia...

Monotonicity, log-concavity: $(\log g)'' \leqslant 0$
Notation: $\mathcal{K}$ is the cone of convex functions

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

Regularizácia! Cez tvarové ohraničenia...

Monotonicity, log-concavity: $(\log g)'' \leqslant 0$
Notation: $\mathcal{K}$ is the cone of convex functions

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

Regularizácia! Cez tvarové ohraničenia...

Monotonicity, log-concavity: $(\log g)'' \leqslant 0$
Notation: $\mathcal{K}$ is the cone of convex functions

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

Regularizácia! Cez tvarové ohraničenia...

Monotonicity, log-concavity: $(\log g)'' \leqslant 0$
Notation: $\mathcal{K}$ is the cone of convex functions

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

Regularizácia! Cez tvarové ohraničenia...

Monotonicity, log-concavity: $(\log g)'' \leqslant 0$
Notation: $\mathcal{K}$ is the cone of convex functions

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

Regularizácia! Cez tvarové ohraničenia...

Monotonicity, log-concavity: $(\log g)'' \leqslant 0$
Notation: $\mathcal{K}$ is the cone of convex functions

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

A convex problem!

Grenander (1956), Jongbloed (1998),
Groeneboom, Jongbloed, and Wellner (2001)
Eggermont and LaRiccia (2000), Walther (2000)
Rufibach and Dümbgen (2006)
Pal, Woodroofe, and Meyer (2006)

Koenker and Mizera (2007-2010): beyond log-concavity

Nie je to až tak nepodobné

The differential operator may be the same,
only the constraint is somewhat different

$$\int |(\log g)''| \leqslant \Lambda, \quad \text{in the dual } |(\log g)''| \leqslant \Lambda$$

Shape constraints: no regularization parameter to be set...
... but of course, we need to believe in the shape.

Odhadovanie hustôt na pokračovanie

Koenker and Mizera (2007)
Density estimation by total variation regularization

Koenker and Mizera (2006)
The alter egos of the regularized maximum likelihood density estimators: deregularized maximum-entropy, Shannon, Rényi, Simpson, Gini, and stretched strings

Koenker, Mizera, and Yoon (2011)
What do kernel density estimators optimize?

Koenker and Mizera (2008):
Primal and dual formulations relevant for the numerical estimation of a probability density via regularization

Koenker and Mizera (2010)
Quasi-concave density estimation

Koenker and Mizera (2014?)
`www.econ.uiuc.edu/~roger/research/ebayes/ebayes.html`

1. 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 nezmysel (“Nem EM”, “nEzMysel”)

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

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

containing the support of the observed sample

and estimate the “prior density” via EM iterations

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

where $\varphi(\cdot)$ denotes the standard normal density

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

Konvexná optimalizácia, do toho!

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

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

When discretized

$$-\sum_i \log\left(\sum_m \varphi(Y_i - u_j) f_j\right) \looparrowright \min_f !$$

or in a more technical form

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

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

Duál: Allah stvoril všetko v pároch

The solution 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, $\hat{\nu}$, of

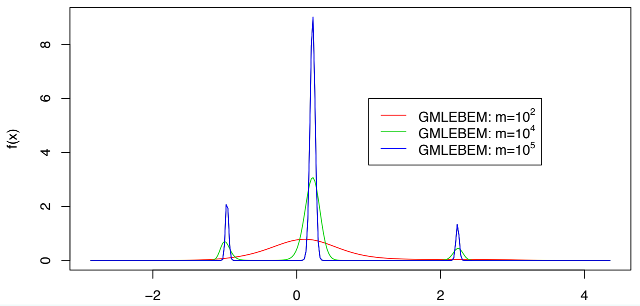
$$\sum_{i=1}^{n} \log \nu_i \leftrightarrow \max_{\mu}! \quad \sum_{i=1}^{n} \nu_i \varphi(Y_i - \mu) \leqslant n \text{ for all } \mu$$

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

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

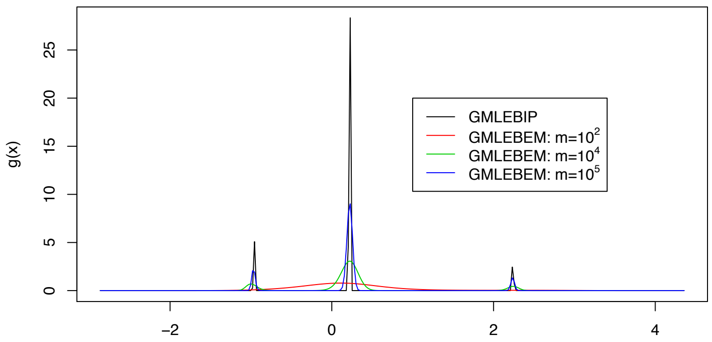
And one can use modern convex optimization methods again...

EM iterácie nemali konca...



(Original version: 55 hours for 1000 replications)

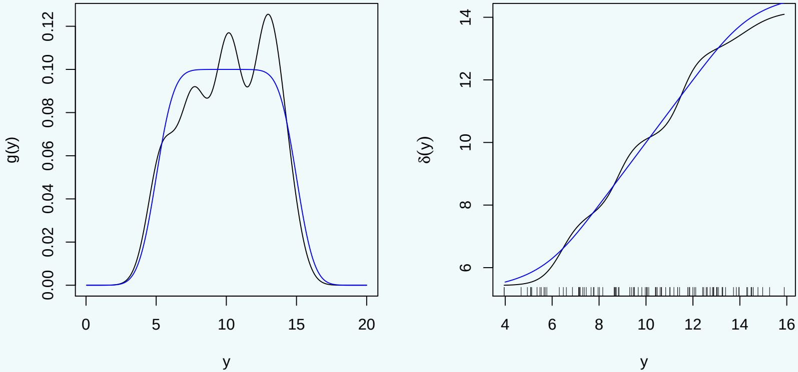
Ale konvexná optimalizácia píše!



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

$n = 200$ observations, $m = 300$ grid points

Typický výsledok keď μ_i sú z $\mathcal{U}(5, 15)$



Left: mixture density (blue: target)

Right: decision rule (blue: target)

2. Empirical prediction rule

Lawrence Brown, personal communication

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)

Monotónny odhad bayesovského rozhodovacieho pravidla

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

Monotónny odhad bayesovského rozhodovacieho pravidla

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

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

Monotónny odhad bayesovského rozhodovacieho pravidla

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

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

Monotónny odhad bayesovského rozhodovacieho pravidla

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

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

Monotónny odhad bayesovského rozhodovacieho pravidla

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

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

Monotónny odhad bayesovského rozhodovacieho pravidla

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

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

Monotónny odhad bayesovského rozhodovacieho pravidla

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

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

Monotónny odhad bayesovského rozhodovacieho pravidla

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

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

Monotónny odhad bayesovského rozhodovacieho pravidla

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

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

Monotónny odhad bayesovského rozhodovacieho pravidla

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

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

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

Poznámky

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:

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

Duálna formulácia

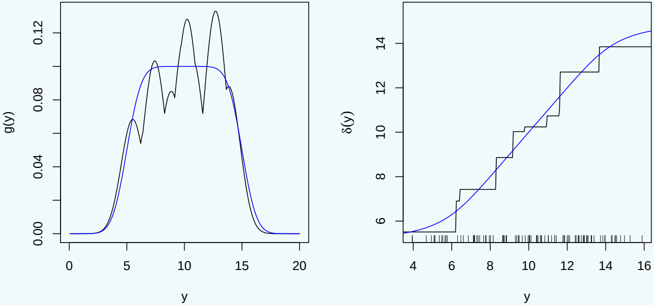
Analogous to Koenker and Mizera (2010):

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

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

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

Typický výsledok keď μ_i sú z $\mathcal{U}(5, 15)$



Left: mixture density (blue: target)

Right: piecewise constant, "empirical decision rule"

Ako to, že to funguje: metódy vnútorného bodu

(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.”

Simulácie - alebo ako byť hodne citovaný

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

Prvý turnaj

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)

Vyberaní súperi

	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"

Prvé závery

- 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
- so far, Kiefer-Wolfowitz empirical prior better adapts to some peculiarities vital in practical data analysis: unequal σ_i, inclusion of covariates,...

Znovu NBA - detaily postupu

Brown (2008)

Data: k_i successes out of n_i trials

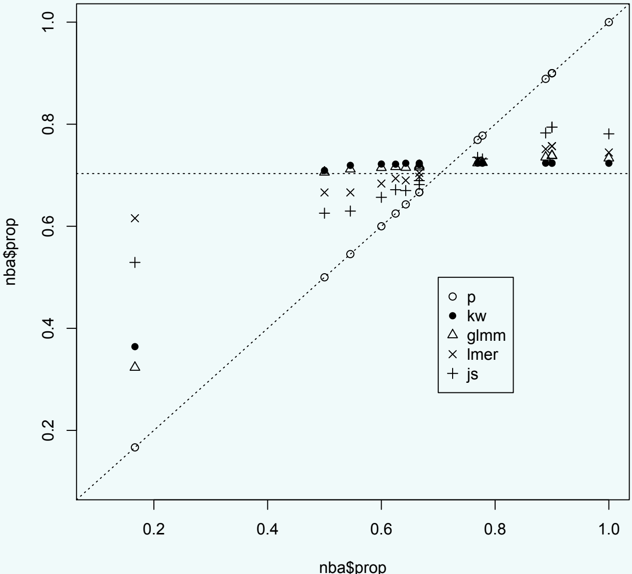
Arcsine transformation:

$$\arcsin\sqrt{\frac{k_i + 1/4}{n_i + 1/2}} \sim N\left(\arcsin\sqrt{p_i}, \frac{1}{4n_i}\right)$$

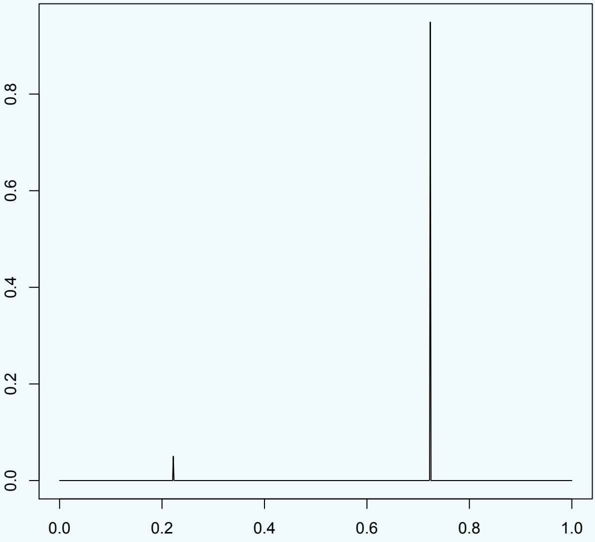
Výsledky

	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

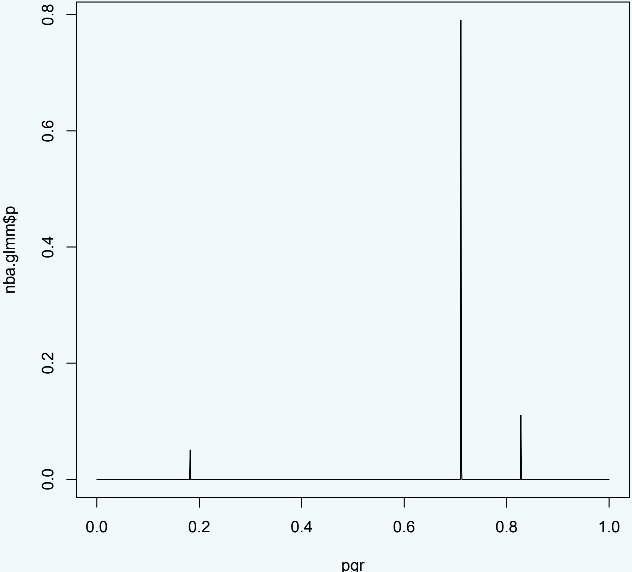
Obrázok



Zmiešavajúce rozdelenie ("empirical prior")



Zmiešavajúce rozdelenie pre glmm



To je všetko?

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

To je všetko?

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?

To je všetko?

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

Dobre, tak prikážme, aby P bola unimodálna...

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

Dobre, tak prikážme, aby P bola unimodálna...

... 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!

Dobre, tak prikážme, aby P bola unimodálna...

... 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

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

Dobre, tak prikážme, aby P bola unimodálna...

... 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. “Unimodálny” 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. "Unimodálny" 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. “Unimodálny” 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. “Unimodálny” monotónny 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. "Unimodálny" monotónny 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. "Unimodálny" monotónny empirical Bayes

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

$h(y)$ concave

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

4. “Unimodálny” monotónny 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. "Unimodálny" monotónny empirical Bayes

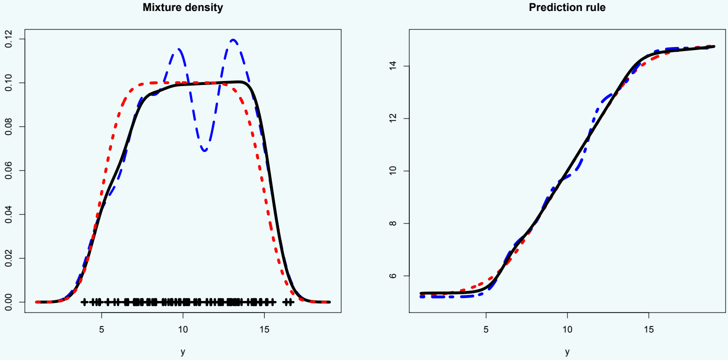
$\frac{1}{2}y^2 + h(y)$ convex

$h(y)$ concave

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

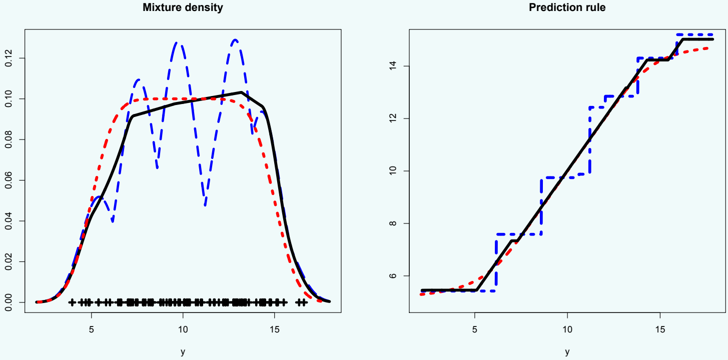
Very easy, very fast

Typický výsledok, znova pre $\mathcal{U}(5, 15)$



(Empirical prior, mixture unimodal)

Typický výsledok, znova pre $\mathcal{U}(5, 15)$



(Empirical prediction rule, mixture unimodal)

Ešte trocha simulácií

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

A nakoniec ďalšie závery

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

A nakoniec ďalšie závery

- 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

A nakoniec ďalšie závery

- 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

A nakoniec ďalšie závery

- 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

A nakoniec ďalšie závery

- 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