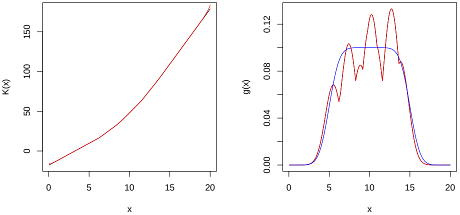
Shape Constraints, Compound Decisions and Empirical Bayes Rules

Roger Koenker

University of Illinois, Urbana-Champaign

CeMMAP Horowitz Conference: 24 June 2011

Joint work with Ivan Mizera (U. of Alberta)



An Empirical Bayes Homework Problem

Suppose you observe a sample $\{Y_1, ..., Y_n\}$ and $Y_i \sim \mathcal{N}(\mu_i, 1)$ for $i = 1, ..., n$, and would like to estimate all of the μ_i's under squared error loss. We might call this "incidental parameters with a vengence."

An Empirical Bayes Homework Problem

Suppose you observe a sample $\{Y_1, ..., Y_n\}$ and $Y_i \sim \mathcal{N}(\mu_i, 1)$ for $i = 1, ..., n$, and would like to estimate all of the μ_i's under squared error loss. We might call this "incidental parameters with a vengence."

Fact 1. If the μ_i are drawn iid-ly from a known distribution F so the Y_i have density,

$$g(y) = \int \phi(y - \mu) dF(\mu),$$

then the Bayes rule is:

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

An Empirical Bayes Homework Problem

Suppose you observe a sample $\{Y_1, ..., Y_n\}$ and $Y_i \sim \mathcal{N}(\mu_i, 1)$ for $i = 1, ..., n$, and would like to estimate all of the μ_i's under squared error loss. We might call this "incidental parameters with a vengence."

Fact 1. If the μ_i are drawn iid-ly from a known distribution F so the Y_i have density,

$$g(y) = \int \phi(y - \mu) dF(\mu),$$

then the Bayes rule is:

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

Fact 2. If F is unknown, one can try to estimate g and plug it into the Bayes rule, but exponential family considerations dictate that $\hat{\delta}(\cdot)$ should be monotone increasing.

Stein Rules!

We'd like to estimate the μ_i's with something other than the naïve decision rule, $\mu_i = Y_i$. For example, if we thought that F were $\mathcal{N}(\mu_0, \sigma_0^2)$ we would have,

$$\delta(y) = y + \frac{g'(y)}{g(y)} = \mu_0 + \frac{\sigma_0^2}{1+\sigma_0^2}(y - \mu_0).$$

Note that in this case, $Y \sim \mathcal{N}(\mu_0, 1+\sigma_0^2)$, so we can estimate (μ_0, σ_0^2) at $\sqrt{n}$ rate, and we obtain a variant of the celebrated James-Stein (1960) estimator. When the prior mean, $\mu_0 = 0$, and the prior variance, $\sigma_0^2 = 1$, then the optimal rule is "shrink by half."

$$\delta(y) = y/2$$

Unobserved Heterogeneity

More generally we can consider models of the form:

$$g(y) = \int \varphi(y, \theta) dF(\theta),$$

where φ is a known parametric likelihood, and F is again a mixing distribution over the parameter θ.

Unobserved Heterogeneity

More generally we can consider models of the form:

$$g(y) = \int \varphi(y, \theta) dF(\theta),$$

where φ is a known parametric likelihood, and F is again a mixing distribution over the parameter θ.

In survival analysis these are called ”fraility” models, or in the terminology of Heckman and Singer (1984) models of ”unobserved heterogeneity.”

The Obligatory Identification Slide

A natural question would be: When can we identify φ and F based on knowledge of the mixture distribution G. Not surprisingly, the answer is only with further assumptions. Two favorable special cases:

The Obligatory Identification Slide

A natural question would be: When can we identify φ and F based on knowledge of the mixture distribution G. Not surprisingly, the answer is only with further assumptions. Two favorable special cases:

Gaussian Location Family When $g(y) = \int \phi(y - \theta) dF(\theta)$,

$$\begin{aligned}\psi_G(t) \equiv \int e^{iyt} g(y) dy &= \int\int e^{iyt} \phi(y - \theta) dF(\theta) \\ &= e^{-t^2/2} \int e^{i\theta t} dF(\theta),\end{aligned}$$

so uniqueness of the characteristic function for G assures identifiability, of F.

The Obligatory Identification Slide

A natural question would be: When can we identify φ and F based on knowledge of the mixture distribution G. Not surprisingly, the answer is only with further assumptions. Two favorable special cases:

Gaussian Location Family When $g(y) = \int \phi(y - \theta) dF(\theta)$,

$$\begin{aligned} \psi_G(t) \equiv \int e^{iyt} g(y) dy &= \int\int e^{iyt} \phi(y - \theta) dF(\theta) \\ &= e^{-t^2/2} \int e^{i\theta t} dF(\theta), \end{aligned}$$

so uniqueness of the characteristic function for G assures identifiability, of F.

General Location Families When $g(y) = \int \varphi(y - \theta) dF(\theta)$, we have, $\psi_G(t) = \psi_\varphi(t)\psi_F(t)$, so unless $\psi_\varphi(t) = 0$ over an open interval, F is again uniquely defined.

Back to the Homework

Most of the applications of our homework problem choose φ as a one parameter exponential family with a "natural" parameter, θ, so we may write,

$$\varphi(y,\theta) = m(y)e^{y\theta}h(\theta)$$

Quadratic loss implies that the Bayes rule is a conditional mean:

$$\begin{aligned}
\delta_G(y) &= \mathbb{E}[\Theta|Y=y] \\
&= \int \theta\varphi(y,\theta)dF / \int \varphi(y,\theta)dF \\
&= \int \theta e^{y\theta}h(\theta)dF / \int e^{y\theta}h(\theta)dF \\
&= \frac{d}{dy}\log(\int e^{y\theta}h(\theta)dF \\
&= \frac{d}{dy}\log(g(y)/m(y))
\end{aligned}$$

Standard Gaussian Case

In the homework problem,

$$\varphi(y,\theta) = \phi(y-\theta) = K\exp\{-(y-\theta)^2/2\} = Ke^{-y^2/2} \cdot e^{y\theta} \cdot e^{-\theta^2/2}$$

So $m(y) = e^{-y^2/2}$ and the logarithmic derivative yields our Bayes rule:

(Fact 1) $$\delta(y) = \frac{d}{dy}\left[\frac{1}{2}y^2 + \log g(y)\right] = y + \frac{g'(y)}{g(y)}.$$

Standard Gaussian Case

In the homework problem,

$$\varphi(y,\theta) = \phi(y-\theta) = K\exp\{-(y-\theta)^2/2\} = Ke^{-y^2/2}\cdot e^{y\theta}\cdot e^{-\theta^2/2}$$

So $m(y) = e^{-y^2/2}$ and the logarithmic derivative yields our Bayes rule:

(Fact 1) $$\delta(y) = \frac{d}{dy}\left[\frac{1}{2}y^2 + \log g(y)\right] = y + \frac{g'(y)}{g(y)}.$$

For Fact 2, note that,

$$\begin{aligned}\delta_G'(y) = \frac{d}{dy}\left[\frac{\int \theta\varphi dF}{\int \varphi dF}\right] &= \frac{\int \theta^2\varphi dF}{\int \varphi dF} - \left(\frac{\int \theta\varphi dF}{\int \theta\varphi dF}\right)^2 \\ &= \mathbb{E}[\Theta^2|Y=y] - (\mathbb{E}[\Theta|Y=y])^2 \\ &= \mathbb{V}[\Theta|Y=y] \geqslant 0,\end{aligned}$$

implying that δ_G must be monotone. This is the monotone likelihood ratio property of the exponential family coming into play.

Estimating $\delta(y)$

So far we have emphasized knowing the form of the mixing distribution F as well as φ, what if F is unknown? If F is known up to a finite dimensional parameter, then there is quite a lot of literature on special cases.

Estimating $\delta(y)$

So far we have emphasized knowing the form of the mixing distribution F as well as φ, what if F is unknown? If F is known up to a finite dimensional parameter, then there is quite a lot of literature on special cases.
For example, in Johnstone and Silverman's (2004) paper "Needles and Straw in Haystacks," they consider prior densities of the form:

$$f(u) = (1 - w)\delta_0(u) + w\gamma(u)$$

so F has mass $1 - w$ at zero, and its remaining mass spread according to a density γ which is taken either to be Laplace (double exponential) or as a beta mixture of normals with Cauchy tails. They construct empirical Bayes estimators that estimate the mass w and the scale of the γ density. Estimators are then selected as the median of the posterior, or the mean, and a quite extensive simulation experiment conducted.

Johnstone and Silverman Simulation Design

Data is generated from 12 distinct models, all of the form:

$$Y_i = \mu_i + u_i, \quad u_i \sim \mathcal{N}(0, 1), \quad i = 1, ..., 1000.$$

Of the $n = 1000$ observations $n - k$ of the $\mu_i = 0$, and the remaining k take one of the four values $\{3, 4, 5, 7\}$. There are three choices of k: $\{5, 50, 500\}$. There are 50 replications for each of the 12 experimental settings and 18 different competing estimators.

Johnstone and Silverman Simulation Design

Data is generated from 12 distinct models, all of the form:

$$Y_i = \mu_i + u_i, \quad u_i \sim \mathcal{N}(0,1), \quad i = 1, ..., 1000.$$

Of the $n = 1000$ observations $n - k$ of the $\mu_i = 0$, and the remaining k take one of the four values $\{3, 4, 5, 7\}$. There are three choices of k: $\{5, 50, 500\}$. There are 50 replications for each of the 12 experimental settings and 18 different competing estimators.

Performance is measured by the mean (over replications) of the sum (over the $n = 1000$ observations) of squared errors, so a score of 500 means that the mean squared prediction error is 0.5, or half of what the naïve prediction $\hat{\mu}_i = Y_i$ would yield if the μ_i were all zero.

Johnstone and Silverman Simulation Results

Number nonzero	5				50				500			
Value nonzero	3	4	5	7	3	4	5	7	3	4	5	7
Exponential	36	32	17	8	214	156	101	73	857	873	783	658
Cauchy	37	36	18	8	271	176	103	77	922	898	829	743
Postmean	34	32	21	11	201	169	122	85	860	888	826	708
Exphard	51	43	22	11	273	189	130	91	998	998	983	817
$a = 1$	36	32	19	15	213	166	142	135	994	1099	1126	1130
$a = 0.5$	37	34	17	10	244	158	105	92	845	878	884	884
$a = 0.2$	38	37	18	7	299	188	95	69	1061	730	665	656
$a = 0.1$	38	37	18	6	339	227	102	60	1496	798	609	579
SURE	38	42	42	43	202	209	210	210	829	835	835	835
Adapt	42	63	73	76	417	620	210	210	829	835	835	835
FDR $q = 0.01$	43	51	26	5	392	299	125	55	2568	1332	656	524
FDR $q = 0.1$	40	35	19	13	280	175	113	102	1149	744	651	644
FDR $q = 0.4$	58	58	53	52	298	265	256	254	919	866	860	860
BlockThresh	46	72	72	31	444	635	600	293	1918	1276	1065	983
NeighBlock	47	64	51	26	427	543	439	227	1870	1384	1148	972
NeighCoeff	55	51	38	32	375	343	219	156	1890	1410	1032	870
Universal soft	42	63	73	76	417	620	720	746	4156	6168	7157	7413
Universal hard	39	37	18	7	370	340	163	52	3672	3355	1578	505

Non-parametric Empirical Bayes

What about nonparametric estimation of the mixture density g?
Brown and Greenshtein (Annals, 2009) propose estimating g by standard fixed bandwidth kernel methods and they compare performance of the resulting *estimated* Bayes rule with various other methods including the 18 methods investigated by Johnstone and Silverman, employing their simulation design. For these simulations they employ bandwidth $h = 1.15$.

Non-parametric Empirical Bayes

What about nonparametric estimation of the mixture density g?
Brown and Greenshtein (Annals, 2009) propose estimating g by standard fixed bandwidth kernel methods and they compare performance of the resulting *estimated* Bayes rule with various other methods including the 18 methods investigated by Johnstone and Silverman, employing their simulation design. For these simulations they employ bandwidth $h = 1.15$.

Estimator	**k = 5**				**k = 50**				**k = 500**			
	3	4	5	7	3	4	5	7	3	4	5	7
$\tilde{\delta}_{1.15}$	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

Monotone Empirical Bayes Rules

But our homework asked for a monotone Bayes rule.

Find a density estimate $\hat{g}$ for the mixture density such that

$$\hat{\delta}(y) = y + \hat{g}'(y)/\hat{g}(y)$$

is monotone increasing, or equivalently, such that,

$$K(y) = \frac{1}{2}y^2 + \log \hat{g}(y)$$

is convex. This problem is closely related to recent work on estimating log-concave densities, e.g. Cule, Samworth and Stewart (JRSSB, 2010), K and Mizera (Annals, 2010), Seregin and Wellner (2010).

Monotone Empirical Bayes Rules

We could, as van Houwelingen and Stijnen (Stat. Ned., 1983), try to make a preliminary (kernel) density estimate and then monotonize its logarithmic derivative, but why not maximum likelihood?

$$\hat{g} = \operatorname{argmax}\{\sum_{i=1}^{n} \log g(Y_i) \mid \int g\,dy = 1, \quad K(y) \in \mathcal{K}\},$$

where $\mathcal{K}$ is the convex cone of convex functions. This can be solved by standard interior point methods, or equivalently we can solve the dual problem of minimizing Shannon entropy or the Kullback-Leibler distance between the estimated density and a uniform density on the support of the empirical df.

Solutions have piecewise linear K functions, and rather funny looking $\hat{g}$'s.

Discrete Formulation

Let $h(y) = -\log g(y)$, and write the primal problem as,

$$\text{(P)} \qquad \max_{\alpha}\{w^\top \alpha - \sum c_i e^{\alpha_i} \mid D\alpha + 1 \geqslant 0\}.$$

and dual problem as,

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

Discrete Formulation

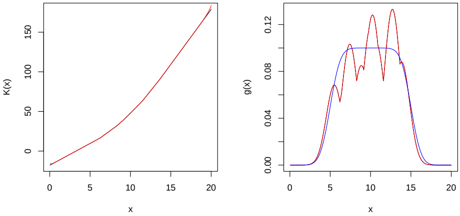
Let $h(y) = -\log g(y)$, and write the primal problem as,

$$\text{(P)} \qquad \max_{\alpha}\{w^\top \alpha - \sum c_i e^{\alpha_i} \mid D\alpha + 1 \geqslant 0\}.$$

and dual problem as,

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

For example with $F \sim U[5, 15]$ we obtain estimates like this:



Revenge of the MLE

How well do these monotone Bayes rules perform in the Johnstone and Silverman sweepstakes?

Revenge of the MLE

How well do these monotone Bayes rules perform in the Johnstone and Silverman sweepstakes?

Estimator	k = 5				k = 50				k = 500			
	3	4	5	7	3	4	5	7	3	4	5	7
$\hat{\delta}$	37	34	21	11	173	121	63	16	488	310	145	22
$\tilde{\delta}_{1.15}$	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

Revenge of the MLE

How well do these monotone Bayes rules perform in the Johnstone and Silverman sweepstakes?

Estimator	k = 5				k = 50				k = 500			
	3	4	5	7	3	4	5	7	3	4	5	7
$\hat{\delta}$	37	34	21	11	173	121	63	16	488	310	145	22
$\tilde{\delta}_{1.15}$	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

Shockingly well, actually. But as ever so, there is disappointment just around the corner.

Revenge^2 of the MLE

Kiefer and Wolfowitz (1956) reconsidering the Neyman and Scott (1948) problem showed that non-parametric maximum likelihood could be used to establish consistent estimators even when the number of incidental parameters tended to infinity. Laird (1978) and Heckman and Singer (1984) suggested that the EM algorithm could be used to compute the MLE in such cases.

Revenge^2 of the MLE

Kiefer and Wolfowitz (1956) reconsidering the Neyman and Scott (1948) problem showed that non-parametric maximum likelihood could be used to establish consistent estimators even when the number of incidental parameters tended to infinity. Laird (1978) and Heckman and Singer (1984) suggested that the EM algorithm could be used to compute the MLE in such cases.

Jiang and Zhang (Annals, 2009) adapt this approach for the empirical Bayes problem: Let $u_i : i = 1, ..., m$ denote a grid on the support of the sample Y_i's, then the prior (mixing) density f is estimated by the fixed point iteration:

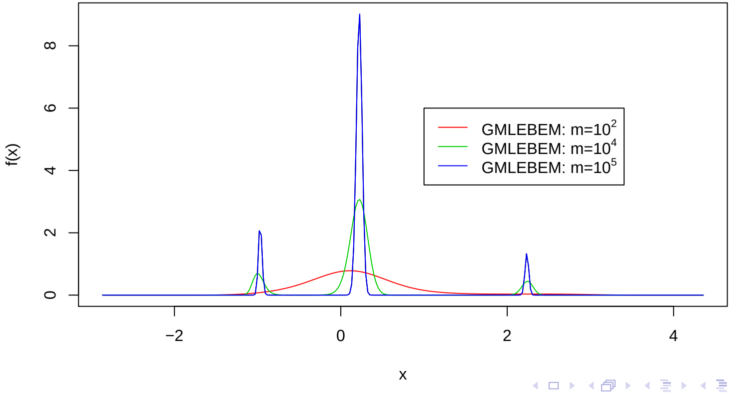
$$\hat{f}_j^{(k+1)} = n^{-1} \sum_{i=1}^{n} \frac{\hat{f}_j^{(k)} \phi(Y_i - u_j)}{\sum_{\ell=1}^{m} \hat{f}_\ell^{(k)} \phi(Y_i - u_\ell)},$$

and the implied Bayes rule becomes at convergence:

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

The Incredible Lethargy of EM-ing

Unfortunately, EM fixed point iterations are notoriously slow and this is especially apparent in the Kiefer and Wolfowitz setting. Solutions approximate discrete (point mass) distributions, but EM goes ever so slowly. (Approximation is controlled by the grid spacing of the u_i's.)



Accelerating EM

There is a large literature on accelerating EM iterations, but none of the recent developments (that I tried) seemed to help very much. Eventually it occurred to me that the problem could be reformulated as a maximum likelihood problem to exploit interior point methods for solving convex programs. Consider,

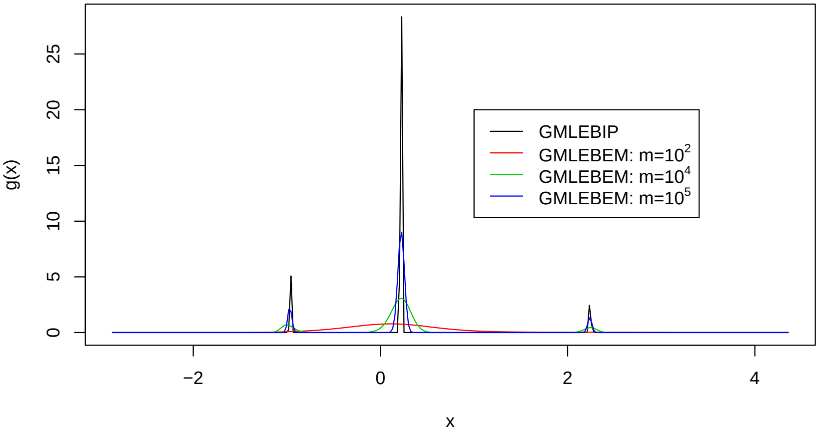
$$\max_{f \in \mathcal{F}} \sum_{i=1}^{n} \log\Big(\sum_{j=1}^{m} \phi(y_i - u_j) f_j\Big),$$

or reformulating slightly,

$$\min\{-\sum_{i=1}^{n} \log(y_i) \mid Az = y,\ z \in \mathcal{S}\},$$

where $A = (\phi(y_i - u_j))$ and $\mathcal{S} = \{s \in \mathbb{R}^m | 1^\top s = 1,\ s \geqslant 0\}$. So z_j denotes the estimated mixing density estimate $\hat{f}$ at the grid point u_j, and y_i denotes the estimated mixture density estimate, $\hat{g}$, at Y_i.

Interior Point vs. EM



Interior Point vs. EM

In the foregoing test problem we have $n = 200$ observations and $m = 300$ grid points. Timing and accuracy is summarized in this table.

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

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.

Scaling problem sizes up, the deficiency of the EM approach is even more serious.

Performance of the NP-MLE Bayes Rule

In the (now familiar) Johnstone and Silverman sweepstakes we have the following comparison of performance.

Estimator	k = 5				k = 50				k = 500			
	3	4	5	7	3	4	5	7	3	4	5	7
$\hat{\delta}_{MLE-IP}$	33	30	16	8	153	107	51	11	454	276	127	18
$\hat{\delta}_{MLE-EM}$	37	33	21	11	162	111	56	14	458	285	130	18
$\hat{\delta}$	37	34	21	11	173	121	63	16	488	310	145	22
$\tilde{\delta}_{1.15}$	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

Here MLE-EM is Jaing and Zhang's (2009) Bayes rule with their suggested 100 EM iterations. It does somewhat better than the shape constrained estimator, but the interior point version MLE-IP does even better.

... , but how does it work in theory?

The fundamental theorem of compound decisions (Robbins (1951)) asserts that the multivariate problem of estimating all the θ's can be reduced to

$$R^*(G_n) = \min_{t \in \mathcal{T}} R(t, G_n) = \min_{t \in \mathcal{T}} \mathbb{E}_{G_n}(t(Y_i) - \xi)^2$$

that is, to finding a Bayes Rule for the univariate problem:

$$Y|\xi \sim \mathcal{N}(\xi, 1), \quad \xi \sim G,$$

with $G = G_n$, the empirical df of the θ's, over the class of Borel functions.

... , but how does it work in theory?

The fundamental theorem of compound decisions (Robbins (1951)) asserts that the multivariate problem of estimating all the θ's can be reduced to

$$R^*(G_n) = \min_{t \in \mathcal{T}} R(t, G_n) = \min_{t \in \mathcal{T}} \mathbb{E}_{G_n}(t(Y_i) - \xi)^2$$

that is, to finding a Bayes Rule for the univariate problem:

$$Y|\xi \sim \mathcal{N}(\xi, 1), \quad \xi \sim G,$$

with $G = G_n$, the empirical df of the θ's, over the class of Borel functions. We can constrain the class, $\mathcal{T}$ in various ways:

- Linear $t(\cdot)$ – James-Stein estimator,
- soft thresholding $t(\cdot)$ – Stein unbiased risk estimator (SURE),
- hard thresholding $t(\cdot)$ – FDR/generalized C_p estimator,
- posterior medians – Johnstone and Silverman's EBThresh

Adaptive Minimaxity and the Oracle

Comparing performance with that of the Oracle estimator using $F = F_n$:

Estimator	k = 5				k = 50				k = 500			
	3	4	5	7	3	4	5	7	3	4	5	7
Oracle	27	22	12	1	144	93	46	3	443	273	128	8
$\hat{\delta}_{MLE-IP}$	33	30	16	8	153	107	51	11	454	276	127	18
$\hat{\delta}_{MLE-EM}$	37	33	21	11	162	111	56	14	458	285	130	18
$\hat{\delta}$	37	34	21	11	173	121	63	16	488	310	145	22
$\tilde{\delta}_{1.15}$	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

Question: How can such poor estimates of the mixing distribution produce such good performance for their associated Bayes rules?

Discrete Approximations and Entropy

The mixing density may be poor, but the mixture density is still quite good:

Lemma: (Zhang) Let $g_F(y) = \int \phi(y-u)dF(u)$, then for any F there exists a discrete F_m, with support $[-M-a, M+a]$ and at most $m = (2\lfloor 6a^2 \rfloor + 1)\lceil 2M/a + 2 \rceil + 1$ atoms such that

$$\|g_F - g_{F_m}\|_{\infty, M} \leqslant \phi(a)(1 + \phi(0)).$$

Discrete Approximations and Entropy

The mixing density may be poor, but the mixture density is still quite good:

Lemma: (Zhang) Let $g_F(y) = \int \phi(y-u)dF(u)$, then for any F there exists a discrete F_m, with support $[-M-a, M+a]$ and at most $m = (2\lfloor 6a^2 \rfloor + 1)\lceil 2M/a + 2 \rceil + 1$ atoms such that

$$\|g_F - g_{F_m}\|_{\infty,M} \leqslant \phi(a)(1+\phi(0)).$$

The existence of such parsimonious discrete approximations yield a good entropy bound (covering number) for the class of distributions and thus a large deviation inequality for the Hellinger error of the (generalized) MLE. This in turn yields strong bounds on the ”regret” for the associated Bayes rules relative to the Oracle bound.

Adaptive Minimaxity

For their approximate MLE-EM Bayes rules Jiang and Zhang prove:

Theorem: For the normal mixture problem, with a (complicated) weak pth moment restriction on Θ, the approximate non-parametric MLE, $\hat{\theta} = \hat{\delta}_{\hat{F}_n}(Y)$ is adaptively minimax, i.e.

$$\frac{\sup_{\theta} \mathbb{E}_{n,\theta} L_n(\hat{\theta}, \theta)}{\inf_{\tilde{\theta}} \sup_{\theta \in \Theta} \mathbb{E}_{n,\theta} L_n(\tilde{\theta}, \theta)} \to 1.$$

The weak pth moment condition encompasses a much broader class of both deterministic and stochastic classes Θ.

Conclusions and Extensions

- Empirical Bayes methods, employing maximum likelihood, offer some advantages over other thresholding and kernel methods,

Conclusions and Extensions

- Empirical Bayes methods, employing maximum likelihood, offer some advantages over other thresholding and kernel methods,
- Kernel based empirical Bayes rules can be improved with shape constrained MLEs and are computationally very efficient, but

Conclusions and Extensions

- Empirical Bayes methods, employing maximum likelihood, offer some advantages over other thresholding and kernel methods,
- Kernel based empirical Bayes rules can be improved with shape constrained MLEs and are computationally very efficient, but
- Kiefer-Wolfowitz type non-parametric MLEs, while computationally somewhat more demanding, perform even better, especially after replacing EM by interior point computational methods. For large sample sizes, further binning is needed to make the interior point methods practical.

Conclusions and Extensions

- Empirical Bayes methods, employing maximum likelihood, offer some advantages over other thresholding and kernel methods,
- Kernel based empirical Bayes rules can be improved with shape constrained MLEs and are computationally very efficient, but
- Kiefer-Wolfowitz type non-parametric MLEs, while computationally somewhat more demanding, perform even better, especially after replacing EM by interior point computational methods. For large sample sizes, further binning is needed to make the interior point methods practical.
- There are many opportunities for linking such methods to various semi-parametric estimation problems a la Heckman and Singer (1983) and van der Vaart (1996).