Empirical Bayes: Methodologies and Asymptotic Theorems

1. Compound and empirical Bayes decision problems
2. Estimation of normal means
3. Nonparametric regression and the white noise with drift
4. Estimation of sums of random variables

Cun-Hui Zhang
Department of Statistics, Rutgers University

Thanks to: Zhe-Da, NSF, ...

Estimation of Normal Means

Model: $X_k|\theta_k \sim N(\theta_k, 1)$, $k = 1, \ldots, n$.

Problem: estimating $\theta_{(n)} \equiv (\theta_1, \ldots, \theta_n)$ to "minimize" the compound risk

$$R_n\left(\hat{\theta}_{(n)}, \theta_{(n)}\right) \equiv n^{-1} \sum_{k=1}^{n} E_{\theta_{(n)}} \left(\hat{\theta}_k - \theta_k\right)^2.$$

- The James-Stein estimator
- Sparsity and ℓ_p balls
- Threshold estimators
- General empirical Bayes estimators
- Simulation results

The James-Stein estimator

• The MLE $\hat{\theta}_{(n)} = X_{(n)}$, with risk 1, is the UMVU, the minimum variance invariant (location group), and a minimax estimator.

• Stein (56): the MLE is inadmissible.

• The James-Stein (62) estimator

$$\hat{\theta}_{(n)} = B_n X_{(n)}, \qquad B_n \equiv 1 - \frac{n-2}{\|X_{(n)}\|^2}.$$

For $n > 2$, the risk of the JS is smaller than that of the MLE,

$$\begin{aligned} R_n\big(B_n X_{(n)}, \theta_{(n)}\big) &= 1 - (n-2)^2 n^{-1} E_{\theta_{(n)}} \|X_{(n)}\|^{-2} \\ &\leq 1 - (n-2)^2 n^{-1} \big(n + \|\theta_{(n)}\|^2\big)^{-1} \\ &< 1. \end{aligned}$$

• Empirical Bayes interpretation (Efron-Morris, 72a,b, 73a,b)

◇ Assume that θ_k are iid $N(0,\tau)$. The Bayes rule is

$$t^*_\tau(x) = B_\tau x, \quad B_\tau \equiv \frac{\tau}{1+\tau} = 1 - \frac{1}{1+\tau}.$$

◇ Since X_k are iid $N(0, 1+\tau)$, $\|X_{(n)}\|^2/n = \sum_{k=1}^n X_k^2/n$ is a "good" estimator of $1+\tau$. Thus, $B_n x$ can be viewed as an estimator of $t^*_\tau(x) = B_\tau x$.

◇ Since $\{N(0,\tau), \tau > 0\}$ is a parametric family, the JS estimator is parametric EB (Morris, 83).

◇ JS is also restricted EB (Robbins, 80, 83), since $B_n x$ can be viewed as an estimate of

$$t^*_n(x) \equiv \arg\min_{t\in\mathcal{D}} R(t, G_n) = \left(1 - \frac{n}{n + \|\theta_{(n)}\|^2}\right)x$$

with $\mathcal{D}$ being the collection of all linear functions.

- Benefits of shrinkage.
- (Proper, hierarchical) Bayes (EB) estimators (Strawderman, 71).
- Admissibility in location models (Brown, 66, 68; Berger, 75).
- Stein's (81) unbiased estimate of risk:

$$\begin{aligned} R_n\big(t_{(n)}(X_{(n)}),\theta_{(n)}\big) &= n^{-1}\textstyle\sum_{k=1}^n E_{\theta_{(n)}}\big(t_k(X_{(n)})-x_k\big)^2 \\ &+ 2n^{-1}\textstyle\sum_{k=1}^n E_{\theta_{(n)}}\big(\partial t_k/\partial x_k\big)(X_{(n)})-1 \end{aligned}$$

for any Borel $t_{(n)}:\mathbb{R}^n\to\mathbb{R}^n$ (integrating by parts).

- Calculation of the risk of the JS by Stein's (81) estimate.
- The JS estimator is NOT a "good" estimator in general for $\theta_{(n)}\in\ell_p$ for $p<2$. For example, if $\theta_1=M\sqrt{n}$ and $\theta_2=\cdots=\theta_n=0$, then $\|\theta_{(n)}\|_{p,n}=Mn^{1/2-1/p}\to 0$ for $p<2$.

Sparsity and ℓ_p balls

- Adaptation to sparse signals, avoid over-fitting, model selection.
- Sparse signals are contained in small ℓ_p balls for small $0 \le p < 2$.

$\diamond$ Strong ℓ_p balls: $\theta_{(n)} \in \Theta_{p,n}(C) \Leftrightarrow G_n \equiv \sum_{k=1}^n \delta_{\theta_k}/n \in \mathcal{G}_p(C)$,

$$\mathcal{G}_p(C) \equiv \left\{G : \int |u|^p G(du) \le C^p\right\}.$$

$\diamond$ Weak ℓ_p balls: $\theta_{(n)} \in \Theta^w_{p,n}(C) \Leftrightarrow G_n \in \mathcal{G}^w_p(C)$,

$$\mathcal{G}^w_p(C) \equiv \left\{G : t^p G(\{u : |u| > t\}) \le C^p,\ \forall t > 0\right\}.$$

$\diamond$ For $p = 0$, nearly black signals: $\theta_{(n)} \in \Theta_{0,n}(c) \Leftrightarrow G_n \in \mathcal{G}_0(c)$,

$$\mathcal{G}_0(c) \equiv \left\{G : G(\{u : u \ne 0\}) \le c\right\}.$$

$\diamond$ Example: If $G(\{a\}) = c = 1 - G(\{0\})$ for large $a > 0$, then $G \in \mathcal{G}_0(c)$, $G \in \mathcal{G}_p(ac^{1/p})$, $G \in \mathcal{G}^w_p(ac^{1/p})$, and $ac^{1/p} \le \epsilon$ for small p.

Threshold estimators

Problem: find "good" estimators for sparse signals; adaptive minimax estimation in ℓ_p balls.

- Hard threshold
- FDR
- Soft threshold
- Sure
- Posterior median

- Hard threshold:

$$\hat{\theta}_k = h_\lambda(X_k), \quad h_\lambda(x) \equiv xI\{|x| > \lambda\}, \qquad \lambda \geq 0.$$

◇ Penalty on model size: for projections Q to subsets of X_k,

$$\text{minimize } \Big\{ \|X_{(n)} - QX_{(n)}\|^2 + \lambda^2 \dim(Q) \Big\}.$$

◇ Model selection: $\lambda = \sqrt{2}$ for AIC (Akaike, 74) and C_p (Mallows, 64, 73), $\lambda = \sqrt{\log n}$ for BIC (Schwarz, 78), $\lambda = \sqrt{2 \log n}$ for RIC (Foster-George, 94), ...

◇ Universal thresholding/VisualShrink (Donoho-Johnstone, 94): $\lambda = \sqrt{2 \log n}$ to select essentially none when $\theta_k = 0$ for all k; tail of $N(0, 1)$.

• False discovery rate (FDR) (Benjamini-Hochberg, 95)

◇ Let $0 < p_k < 1, k \le n$, be p-values for hypotheses H_k and $\sigma_k \in \{0, 1\}$ be constants such that $\sigma_k = 0$ iff $p_k \sim$ unif(0,1), i.e. for true H_k.

◇ For any tests $\hat{\sigma}_k \in [0, 1]$ (power of tests), the FDR is defined as

$$EQ, \quad Q \equiv \frac{\mathrm{FD}}{\mathrm{D}} = \frac{\sum_{k=1}^{n}(1 - \sigma_k)\hat{\sigma}_k}{\sum_{k=1}^{n} \hat{\sigma}_k}$$

◇ Let $p_{(1)} \le \cdots \le p_{(n)}$. Benjamini-Hochberg procedure:

$$\hat{\sigma}_k \equiv I\{p_k \le p_{(k^*)}\}, \quad k^* \equiv \max\left\{k : p_{(k)} \le q^* k/n\right\}.$$

◇ Property: Suppose p_k are independent. Then,

$$EQ \le q^* \textstyle\sum_{k=1}^{n} \sigma_k / n \le q^*.$$

• Adaptive hard threshold via FDR (Abramovich *et al*, ABDJ 04)

◇ Let $\mathcal{R}_n\big(\Theta_n\big)$ be the minimax risk,

$$\mathcal{R}_n\big(\Theta_n\big) \equiv \inf_{t_{(n)}} \sup_{\theta_{(n)} \in \Theta_n} R_n\big(t_{(n)}, \theta_{(n)}\big).$$

◇ Let $p_k = 2\{1 - \Phi(|X_k|)\}$ and $q^* \equiv q_n^* \to 0$ with $\liminf_n q_n^* \log n > 0$. Let $\hat{\lambda}_n$ be the threshold level determined by the FDR procedure. Then, the hard threshold estimators $h_{\hat{\lambda}_n}$ are simultaneously asymptotically minimax

$$\sup_{\theta_{(n)} \in \Theta_n} R_n\big(h_{\hat{\lambda}_n}, \theta_{(n)}\big) \le (1 + o(1))\mathcal{R}_n\big(\Theta_n\big),$$

for all (strong, weak, near black) ℓ_p balls Θ_n of radii C_n, as long as $(\log n)^5/n \le C_n^p(C_n) \le 1/n^\epsilon$ for $p > 0$ $(p = 0)$, $0 < \epsilon < 1$.

◇ Problem: What happens for fixed q^*?

• Soft threshold: a smooth version of h_λ,

$$\hat{\theta}_k = s_\lambda(X_k), \quad s_\lambda(x) \equiv \text{sign}(x)(|x| - \lambda)^+$$

◇ LSE with L_1 penalty:

$$\text{minimize } \left\{ \sum_{k=1}^n (X_k - \hat{\theta}_k)^2 + 2\lambda \sum_{k=1}^n |\hat{\theta}_k| \right\}.$$

◇ Regression: LASSO (Tibshirani, 96), LARS (Efron *et al*, 03)

• Ideal selection: selecting X_k iff $|\theta_k| > 1$; the "risk" is

$$\kappa_n(\theta_{(n)}) \equiv n^{-1} \sum_{k=1}^n \min(\theta_k^2, 1).$$

◇ Compared with the ℓ_p "norm",

$$\theta_{(n)} \in \Theta_{p,n}(C) \Rightarrow \kappa_n(\theta_{(n)}) \le C^p, \quad 0 < p \le 2,$$

$$\theta_{(n)} \in \Theta^w_{p,n}(C) \Rightarrow \kappa_n(\theta_{(n)}) \le 2C^p/(2-p), \quad 0 < p < 2.$$

• Minimax risk in ℓ_p balls (DJ, 94): for $0 < p \le 2$ and small $C > 0$

$$\mathcal{R}\big(\mathcal{G}_p(C)\big) = (1+o(1))C^p\{-2\log(C^p)\}^{1-p/2}$$

and if $C^p n/(\log n)^{p/2} \to \infty$, the same for $\mathcal{R}_n(\Theta_{p,n}(C))$.

• Minimaxity of the ideal/Bayes threshold rule:

$$\hat\theta_k = s_{\lambda_n^*}(X_k), \qquad \lambda_n^* \equiv \arg\min_{\lambda\ge 0} R(s_\lambda, G_n).$$

◇ Oracle inequality: compared with the oracle $\kappa_n(\theta_{(n)})$,

$$R_n(s_\lambda, \theta_{(n)}) \le (\lambda^2+1)\kappa_n\big(\theta_{(n)}/\sqrt{\lambda^2+1}\big) + 4\lambda^{-3/2}\varphi(\lambda).$$

Consequently, for $\lambda = \sqrt{-2\log(C^p)}$ and $C \to 0+$,

$$\begin{aligned}\sup_{\theta_{(n)}\in\Theta_{p,n}(C)} R_n(s_\lambda,\theta_{(n)}) &\le (1+o(1))\lambda^{2-p}C^p + \varphi(\lambda)/\lambda \\ &= (1+o(1))\mathcal{R}\big(\mathcal{G}_p(C)\big).\end{aligned}$$

• SureShrink (DJ, 95), adaptation to sparsity, e.g. $\Theta_{p,n}(C)$.

◇ Stein's unbiased estimate of risk

$$R_n\big(s_\lambda, \theta_{(n)}\big) \approx n^{-1} \sum_{k=1}^n \Big(\min(\lambda^2, X_k^2) + 2I\{|X_k| > \lambda\} \Big) - 1.$$

◇ Use "sure" (Stein, 81) to choose $s_{\hat{\lambda}_n} \approx s_{\lambda_n^*}$, which minimizes Stein's estimate of risk; restricted EB.

◇ Oracle inequality with an error term $(\log n)^{3/2}/\sqrt{n}$.

◇ Modification: switch to universal thresholding when

$$\|X_{(n)}\|^2/n - 1 \le \gamma_n, \ \gamma_n = (\log n)^{3/2}/\sqrt{n} \ \text{ or } \ n^{-\epsilon}, \ 0 < \epsilon < 1/2.$$

◇ Adaptation to sparse signals under conditions on $\|\theta_{(n)}\|_{2,n}^2/n$.

• Posterior median (Johnstone-Silverman, 04):

$$\hat{\theta}_k = \mu_{\hat{w}_n}(X_k),$$

where for certain density g_0, $\mu_w(x)$ is the posterior median

$$\mu_w(x) \equiv \arg\min_u \left\{ w \int |u-y| e^{xy-y^2/2} g_0(y) dy + (1-w)|u| \right\}$$

with the prior $(1-w)\delta_0 + w g_0$, and w is estimated by the MLE

$$\hat{w}_n \equiv \arg\max_w \prod_{k=1}^{n} \left\{ (1-w) + w \int \exp(X_k y - y^2/2) g_0(y) dy \right\}.$$

◇ The tail of g_0 should be heavier than that of normal densities, e.g. double exponential; g_0 should be symmetric.

• Parametric EB under ℓ_1 risk.

• Threshold level:

$$\lambda(w) \equiv \sup\{x : \mu_w(x) = 0\}.$$

• Adaptation to sparse signals: rate adaptive minimax in small ℓ_p balls (Johnstone-Silverman, 04).

◇ Minor modification: use $\mu_{w_n}(X_k)$ if $\lambda(\hat{w}_n) > \sqrt{2\log n}$, where w_n is defined by $\lambda(w_n) = \sqrt{2\log n}$.

◇ Risk bounds: for $p \leq 2$

$$\sup_{\theta_{(n)} \in \Theta_{p,n}(C)} R_n\big(\mu_{\hat{w}_n}, \theta_{(n)}\big) \leq C_1 \mathcal{R}\big(\Theta_{p,n}(C)\big) + C_2 n^{-1} (\log n)^{1-p/2},$$

where C_1 and C_2 are reals depending on p and g_0.

◇ Rate adaptive minimax when $nC^p/(\log n)^{p/2} \to \infty$.

General empirical Bayes estimators

- General EB methods: approximate the Bayes

$$t_n^* \equiv \arg\min E_{\theta_{(n)}} \left\| t(X_{(n)}) - \theta_{(n)} \right\|_2^2 = x + \varphi_n'(x)/\varphi_n(x),$$

where $\varphi_n(x) \equiv \int \varphi(x-u) G_n(du)$.

- Zhang (97): $\widehat{t}_n(x) = x + \widehat{\varphi}_n'(x)/\max(\widehat{\varphi}_n(x), \rho_n)$, with

$$\widehat{\varphi}_n(x) \equiv \frac{\sqrt{2\log n}}{n} \sum_{k=1}^{n} K\big(\sqrt{2\log n}(x - X_k)\big)$$

and $\rho_n \equiv \rho_0\{2(\log n)/n\}^{1/2}$, where $K(x) \equiv \sin(x)/(\pi x)$.

- Regularized GEB:

$$\widehat{t}_n^R \equiv \arg\min_{t\uparrow} \textstyle\sum_{k=1}^{n} \big(t(X_k) - \widehat{t}_n(X_k)\big)^2.$$

- Oracle inequality: with certain $\Delta(\rho_n, G_n) \le 1 - R^*(G_n)$,

$$R_n\big(\widehat{t}_n, \theta_{(n)}\big) \le R^*(G_n) + \Delta(\rho_n, G_n) + M_0(\log n)^2/\sqrt{n}.$$

The same holds for the regularized GEB estimator $\widehat{t}_n^R$.

- Asymptotic global minimaxity:

$$\sup_{\theta_{(n)}} R_n\big(\widehat{\theta}_{(n)}, \theta_{(n)}\big) = 1 + O((\log n)^2/\sqrt{n})$$

- Asymptotic optimality & ideal adaptation: For $G_n =_D O_P(1)$,

$$R_n\big(\widehat{\theta}_{(n)}, \theta_{(n)}\big) = R^*(G_n) + o(1)$$

where $R^*(G_n) = R_n\big(t_n^*(X_{(n)}), \theta_{(n)}\big)$ is the ideal/Bayes risk.

• Adaptation in ℓ_p balls: for $p \le 2$

$$\sup_{\theta_{(n)} \in \Theta_n} R_n\left(\widehat{t}_n, \theta_{(n)}\right) \le (1 + o(1))\mathcal{R}_n(\Theta_n)$$

in strong and weak ℓ_p balls Θ_n with radii C, provided that $C^p\sqrt{n}/(\log n)^{1+p/2} \to \infty$.

◇ Good performance for dense and moderately sparse signals; possible modification for very sparse signals.

• Hybrid GEB: Given a threshold estimator with (estimated) threshold level λ_n, use the RGEB if $\lambda_n < \lambda_n^* \equiv \sqrt{-2\log\rho_n}$ and use the threshold estimator otherwise.

◇ If $\theta_k = 0$ for all k, $x + \varphi_n'(x)/\max(\varphi_n(x), \rho_n)$ has the threshold level λ_n^*.

◇ Hybrid GEB provides adaptive minimaxity for very sparse signals.

- Penalized EB:

$$\hat{\theta}_k = h_{\lambda_n}\big(\widehat{t}_n^R(X_k)\big).$$

◇ Demand model selection features via

$$h_\lambda(t_G^*) = \arg\min \int \Big[(t(x)-u)^2 + \lambda I\{|t(x)|>0\}\Big] f(x|u)G(du).$$

Since $t_n^*(x) = 0$ if $G_n =_D 0$, we choose

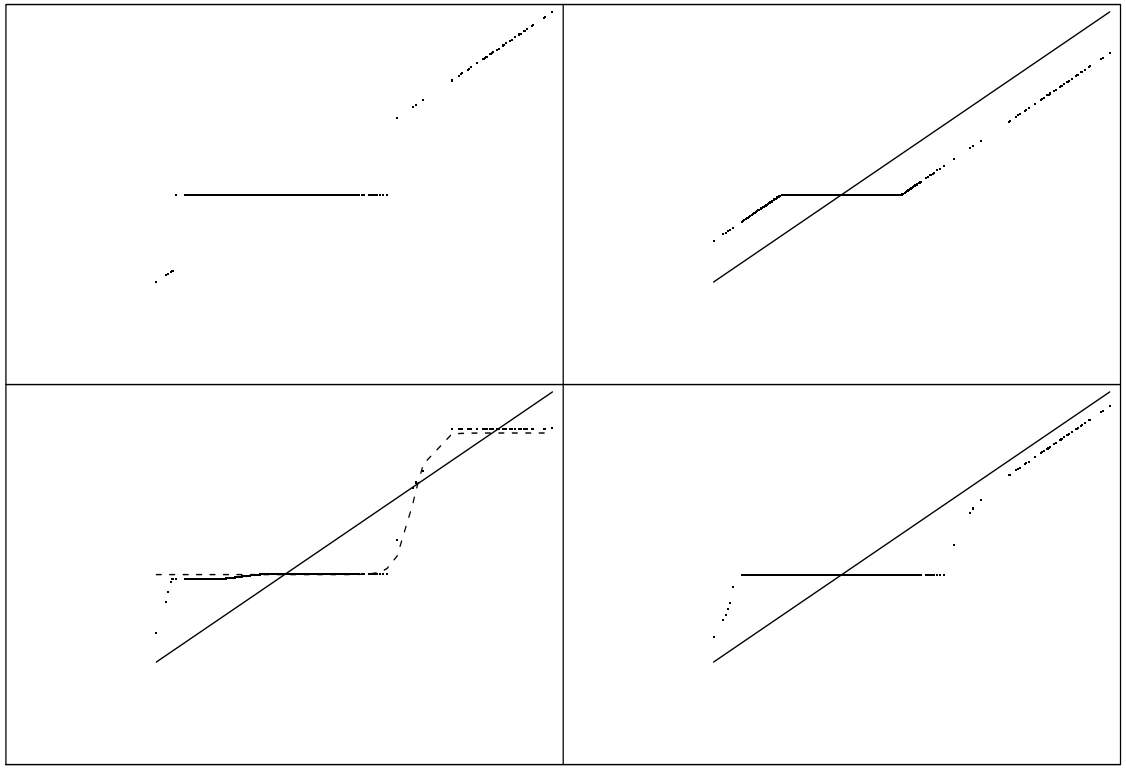
$$\lambda_n \asymp \Big\{\mathrm{Var}_0\big(\widehat{\varphi}_n(x_n)/\rho_n\big)\Big\}^{1/2}$$

where $x_n = \sqrt{-2\log\rho_n}$.

◇ Penalized GEB provides model selection features while maintaining the main optimality properties of GEB with respect to the risk.

Simulation results (normal means)

One realization: $n = 1000$, $G_n(\{5\}) = 5\% = 1 - G_n(\{0\})$



Clockwise from top left: FDR, SureShink, PostMed, BGEB

Table 1.1. 1000*(Ave MSE), $n = 1000$, θ_k = either μ or 0, 100 replications

100π	0.5	0.5	0.5	0.5	5	5	5	5	50	50	50	50
Value	3	4	5	7	3	4	5	7	3	4	5	7
JS	45	76	113	199	312	446	557	711	823	893	930	964
SURE	35	40	41	41	200	207	208	208	835	841	841	841
EBPM	35	34	20	11	237	154	103	90	849	884	890	891
FDR	39	40	22	10	268	169	120	103	1017	762	719	714
RGEB	68	63	55	52	198	152	95	70	510	344	189	85
HGEB	35	34	20	11	198	152	95	70	510	344	189	85
PGEB	48	42	31	27	235	155	83	50	510	344	189	85

Table 1.2. 1000*(Ave MSE), $n = 4000$, θ_k = either μ or 0

100π	0.5	0.5	0.5	5	5	5	10	10	10	15	15	15
Value	3	5	7	3	5	7	3	5	7	3	5	7
JS	44	112	197	311	556	711	475	716	832	575	790	881
SURE	33	37	37	198	206	206	324	333	333	422	430	430
EBPM	38	16	10	240	104	92	381	193	181	475	277	268
FDR	41	18	12	269	135	128	431	235	229	551	322	317
RGEB	46	30	27	172	70	36	257	94	40	314	107	40
HGEB	38	16	10	172	70	36	257	94	40	314	107	40
PGEB	43	23	17	218	69	29	304	94	35	347	106	35

Table 1.3. 1000*(Ave MSE), $n = 1000$, θ_k = either $N(\mu, 0.1)$ or $N(0, 0.1)$

100π	0.5	0.5	0.5	0.5	5	5	5	5	50	50	50	5
Value	3	4	5	7	3	4	5	7	3	4	5	
JS	128	154	185	258	353	472	572	716	818	887	924	95
SURE	131	135	135	135	277	283	283	284	843	847	847	84
EBPM	138	135	121	115	328	252	204	193	853	884	890	89
FDR	143	144	126	116	375	286	237	228	1054	798	761	75
RGEB	160	155	148	144	279	240	190	159	532	381	235	13
HGEB	138	135	121	115	279	240	190	159	532	381	235	13
PGEB	155	150	139	134	326	254	189	152	532	381	235	13

Table 1.4. 1000*(Ave MSE), $n = 1000$, θ_k = either $N(\mu, 2)$ or $N(0, 2)$

100π	0.5	0.5	0.5	5	5	5	50	50	50
Value	3	5	7	3	5	7	3	5	7
JS	673	682	694	702	759	813	846	931	962
SURE	761	761	761	785	785	785	945	947	947
EBPM	977	971	971	950	908	906	956	993	994
FDR	1390	1376	1376	1398	1306	1304	1354	1080	1078
RGEB	726	732	726	751	766	725	659	591	463
HGEB	726	732	726	751	766	725	659	591	463
PGEB	734	740	734	754	769	728	659	591	463

Conclusions

- Local minimax estimation for sparse signals can be achieved by threshold methods, but not linear estimators
- General empirical Bayes estimators perform well for dense and moderately sparse signals
- Empirical Bayes posterior median and FDR seem to be the best adaptive methods for very sparse signals
- It is possible to combine the general empirical Bayes and threshold estimators to achieve good performance for dense and sparse signals