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, ...

Compound and Empirical Bayes Decision Problems

- Compound decision problems
- Empirical Bayes methods
- Equivalence in minimax estimation
- Estimation of mixing distributions
- Parametric and restricted empirical Bayes

Compound decision problems

• Decision problems (Wald, 47, 50): observation X; model/density $f(x|\theta)$; parameter θ; loss function $L(a, \theta)$; decision rule $t(x)$; risk $R(t, \theta) \equiv E_\theta L(t(X), \theta)$; problem = to "minimize" the risk.

• Compound decision problems (Robbins, 51): (similar) decision problems with (conditionally independent) observations

$$X_{(n)} \equiv (X_1, \ldots, X_n), \quad X_k \sim f(x|\theta_k),$$

parameters $\theta_{(n)} \equiv (\theta_1, \ldots, \theta_n)$, and the compound risk

$$R_n\big(t_{(n)}(X_{(n)}), \theta_{(n)}\big) \equiv n^{-1} \sum_{k=1}^{n} E_{\theta_{(n)}} L(t_k(X_{(n)}), \theta_k),$$

where $t_{(n)} \equiv (t_1, \ldots, t_n)$.

• Why compound? A separable rule $t_k(X_{(n)}) = t(X_k)$ may not produce desirable results for any given t.

• Example: $X_k|\theta_k \sim N(\theta_k, 1)$, $\theta_k = \pm 1$, $L(a, \theta) = I\{a \neq \theta\}$.

◇ Separable rules $t_\lambda(x) = 2I\{x > \lambda\} - 1$ have compound risks

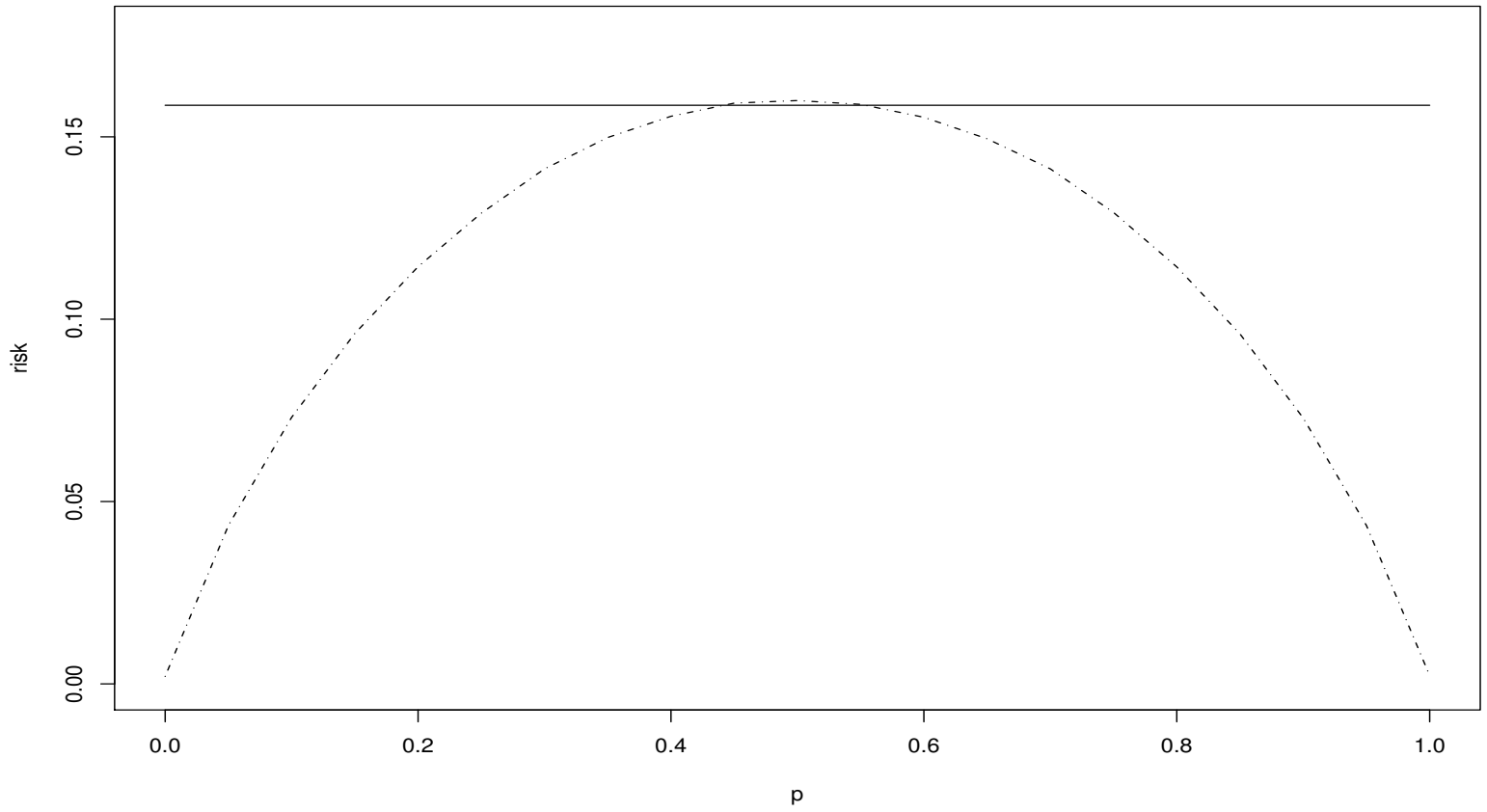
$$R_n(t_\lambda, \theta_{(n)}) = p_n \Phi(\lambda - 1) + (1 - p_n)\Phi(-\lambda - 1),$$

with $p_n \equiv \sum_{k=1}^n (\theta_k + 1)/(2n)$.

◇ The naive (minimax) choice is $\lambda = 0$.

◇ Obviously, the best choice of λ is $\lambda(p_n) \equiv 2^{-1} \log(1/p_n - 1)$, while $p_n \approx \hat{p}_n \equiv \sum_{k=1}^n (X_k + 1)/(2n)$.

◇ Simulation: solid = minimax, dash = EB, 2000 rep



Empirical Bayes methods

- Robbins (51, 56)
- For separable rules $t_k = t(X_k)$, the compound risk

$$R_n(t, \theta_{(n)}) = R(t, G_n) \equiv \int E_\theta L(t(X), \theta) G_n(d\theta)$$

is the Bayes risk with the (unknown) prior

$$G_n(A) \equiv n^{-1} \sum_{k=1}^{n} I\{\theta_k \in A\}.$$

- The ideal/Bayes rule is

$$t_n^* \equiv t_{G_n}^* \equiv \arg\min_t R(t, G_n).$$

- Empirical Bayes (EB) solutions to the compound problem

$$t_k(X_{(n)}) = \hat{t}_n(X_k), \qquad \hat{t}_n \approx t_n^*.$$

- Asymptotic optimality (ideal adaptation)

$$R_n(t_{(n)}(X_{(n)}), \theta_{(n)}) \approx R^*(G_n) \equiv R(t_n^*, G_n).$$

• Compound EB problem: a compound problem where θ_k are iid variables from an unknown G, i.e. n iid Bayes problems with an unknown prior and a compound risk.

• Sequential EB problem: only $(X_1, \ldots, X_k)$ can be used in the k-th decision problem; mathematically equivalent to the compound version.

• EB solution: use $X_{(n)}$ to estimate the Bayes rule $t^* \equiv t^*_G$ for the unknown (true) prior.

• EB methods in general: treating (a sequence of) unknown parameters as (iid) random variables with unknown prior and estimate the Bayes rule for the unknown prior based on data.

• Related work: Neyman-Scott (48), Kiefer-Wolfowitz (56), Stein (56).

Equivalence in minimax estimation

• Is the ideal risk $R^*(G_n)$ (for the ideal separate estimators) a good benchmark, i.e. "lower bound", for the compound problem?

• What lower bound? Minimaxity in many subsets of the parameter space of $\theta_{(n)}$, e.g. local minimaxity similar to LAM of Le Cam.

• Minimax equivalence between the compound and Bayes problems

$$\mathcal{R}_n(\Theta_n) \approx \mathcal{R}(\mathcal{G}_n) \geq \sup_{G \in \mathcal{G}_n} R^*(G),$$

in many subsets (balls) Θ_n, where $\mathcal{G}_n \equiv \{G_n : \theta_{(n)} \in \Theta_n\}$, and

$$\mathcal{R}_n(\Theta) \equiv \inf_{t_{(n)}} \sup_{\theta_{(n)} \in \Theta} R_n(t_{(n)}, \theta_{(n)}), \quad \mathcal{R}(\mathcal{G}) \equiv \inf_{t} \sup_{G \in \mathcal{G}} R(t, G).$$

• Ideal adaptation leads to simultaneous asymptotic local minimaxity. If the minimax equivalence holds in Θ_n, then separable decision rules are approximately minimax in them.

- Pinsker (80), Donoho-Johnstone (94, 98), ...
- Donoho-Johnstone (94): $X|\theta \sim N(\theta, 1)$, $L(a, \theta) = (a - \theta)^2$.

◇ Minimax equivalence in ℓ_p balls

$$\Theta_{p,n}(C) \equiv \left\{\theta_{(n)} : n^{-1}\textstyle\sum_{k=1}^{n} |\theta_k|^p \le C^p\right\}$$

within $o(1)\mathcal{R}_n(\Theta_{p,n}(C))$ as $nC^p/(\log n)^{p/2} \to \infty$, $0 < p \le 2$.

◇ Ideal threshold estimators are asymptotically (Bayes) minimax.

◇ Asymptotic minimaxity cannot be achieved with linear estimators.

◇ Comparison via the minimax Bayes risk in

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

Note that $\mathcal{G}_{p,n}(C) \equiv \{G_n : \theta_{(n)} \in \Theta_{p,n}(C)\} \subset \mathcal{G}_p(C)$ and $G \in \mathcal{G}_p(C) \Rightarrow \lim_n G_n = G$ (weak) for certain $G_n \in \mathcal{G}_{p,n}(C)$.

• Estimation of location vector: $X = \theta + \epsilon \in R^d$, $L(a, \theta) = \|a - \theta\|^2$.

◇ For nonnegative Borel ψ, e.g. $\psi(x) = |x/C|^p$, define ψ-balls

$$\Theta_{\psi,n} \equiv \left\{\theta_{(n)} : \sum_{k=1}^n \psi(\theta_k) \le n\right\}, \quad \mathcal{G}_\psi \equiv \left\{G : \int \psi(u) G(du) \le 1\right\}.$$

Theorem. *Suppose* $\|\epsilon\|_q \equiv (E\|\epsilon\|^q)^{1/q} < \infty$. *Then,*

$$\sup_{\theta_{(n)} \in \Theta_{\psi,n}} R^*(G_n) \le \sup_{G \in \mathcal{G}_\psi} R^*(G)$$
$$\le \mathcal{R}_n(\Theta_{\psi,n}) + \eta_n \|\epsilon\|_2^2 + 2\eta_n \left(\|\epsilon\|_2 + (\eta_n^{-1} - 1)^{1/q} \|\epsilon\|_q\right)^2,$$

where $\eta_n \approx ((\log n)/(2n))^{1/2}$ *with* $\eta_n = \exp(-2n\eta_n^2)$. *Furthermore, if* ψ *is lower semi-continuous, then* $\mathcal{R}(\mathcal{G}_\psi) = \sup_{G \in \mathcal{G}_\psi} R^*(G)$.

◇ If $\epsilon \sim N(0, 1)$, then for the best q

$$\eta_n \|\epsilon\|_2^2 + 2\eta_n \left(\|\epsilon\|_2 + (\eta_n^{-1} - 1)^{1/q} \|\epsilon\|_q\right)^2 \le (1 + o(1))(\log n)^{3/2}(2/n)^{1/2}.$$

• Outline of proof.

◇ Continuity of the Bayes risk: For $G = w_1 G_1 + w_2 G_2$,

$$w_1 R^*(G_1) \le R^*(G) \le R^*(G_1) + w_2 \big\{ \|\epsilon\|_2 + (w_1/w_2)^{1/(2q)} \|\epsilon\|_{2q} \big\}^2.$$

Problem: Degree of continuity for general/other decision problems, i.e. sensitivity to the specification of G.

◇ Large deviation: For $\int \psi dG \le 1$, let $\psi(\theta) I\{\psi(\theta) \le M\} \sim G_1$ with $\theta \sim G$ and θ_k be iid G_1 under $P_{G_1,n}$. If $P_G\{\psi(\theta) = M\} = 0$, then

$$P_{G_1,n}\Big\{ \sum_{k=1}^{n} \psi(\theta_k) > n \Big\} \le \exp\Big(-2nP_G^2\{\psi(\theta) > M\} \Big).$$

◇ Minimax theorem for lower semi-continuous ψ:

$$\mathcal{R}(\mathcal{G}_\psi) = \sup_{G \in \mathcal{G}_\psi} R^*(G) = \sup_{G \in \mathcal{G}_\psi} \inf_t R(t, G).$$

Estimation of mixing distributions

- Individual solutions to the EB problem: estimation of t^*_G or $t^*_{G_n}$.
- General solution: estimation of G or G_n and use $t^*_{\widehat{G}}$. In the EB model, X_k are iid from the mixture $f_G(x) \equiv \int f(x|u)G(du)$.
- Generalized (NP) MLE of the mixing distribution G (Kiefer-Wolfowitz, 56); EM algorithm (Dempster-Laird-Rubin, 77).
- Problem: the NPMLE of mixing distributions are consistent under very mild conditions (Phanzagl, 88), but their convergence rates are unknown in general. Examples?
- In some cases, the optimal convergence rates for the nonparametric estimation of mixing distribution are very slow, e.g. logarithmic rates.

• Asymptotic theory for (the lower bounds of) optimal minimax convergence rates in general estimation problems (Donoho-Liu, 91).

◇ Suppose we want to estimate $\tau(G)$ based on $X_{(n)}$ from $P_{G,n}$.

◇ Let ν_n be measures in $\mathcal{G}_n \equiv \{G : r_n \le d(\tau(G), \tau(G_0)) \le M r_n\}$ and $P_n \equiv \int P_{G,n} d\nu_n$. If $\liminf \|P_n - P_{G_0,n}\|_1 < 2$, then the minimax risk in $\mathcal{G}_n$ cannot be $o(r_n)$.

◇ This is a consequence of Neyman-Pearson: $(2 - \|P_n - P_{G_0,n}\|_1)/2$ is the smallest total error probabilities for testing P_n vs. $P_{G_0,n}$.

• Example (Zhang, 90): $X|\theta \sim N(\theta, 1)$

◇ Let $G_0 \sim N(0,1)$, $a_n \equiv \sqrt{2\log n}$,

$$G_n(du) \equiv \big[1 + (2a_n)^{-1}\{\cos(a_n u) - c_n\}\big] G_0(du).$$

◇ Since $\theta|x \sim N(x/2, 1/2)$, $\big|f_{G_n}(x)/f_{G_0}(x) - 1\big| \le 1/\sqrt{n}$. Thus,

$$\begin{aligned} H^2(P_{G_n,n}, P_{G,n}) &= 2\big[1 - \big\{1 - \textstyle\int(\sqrt{f_{G_n}} - \sqrt{f_{G_0}})^2/2\big\}^n\big] \\ &\le \cdots \to 2(1 - e^{1/2}) < 1, \end{aligned}$$

which implies $\liminf \|P_n - P_{G_0,n}\|_1 < 2$.

◇ Since $\|G_n - G_0\|_\infty \asymp a_n^{-2}$, $r_n = (\log n)^{-1}$ is a lower bound for the estimation of G.

◇ G_n are regular, since $(d/du)^2 G_n(u)$ are uniformly bounded.

• Estimators (not MLE) were constructed to achieve optimal rates in various models (Carroll-Hall, 88; Zhang 90, 95; Fan 91).

Parametric and restricted empirical Bayes

• Parametric EB: assume θ_k iid from $G \in \mathcal{G}$ with a parametric family $\mathcal{G}$, and approximate t^*_G for a member $G \in \mathcal{G}$, Efron-Morris (72a, b, 73a, b), Morris (83).

• Restricted EB: approximate

$$t^*_n = \arg\min_{t \in \mathcal{D}} R(t, G_n)$$

for a restricted class $\mathcal{D}$, Robbins (80, 83).

◇ Parametric EB if $\mathcal{D} = \{t^*_G, G \in \mathcal{G}\}$.

◇ General EB if $\mathcal{D} = \{$ all Borel $\}$.

Conclusions

• Empirical Bayes methods may (often) achieve great risk reduction in compound decision problems

• The ideal Bayes risk is a good benchmark for compound estimation, in the sense that the local maxima of the ideal Bayes risk provide asymptotical lower bounds for the local minimax risks

• The estimation of mixing distribution/density may have slow optimal rates of convergence, and thus may not provide a sound general solution to compound decision problems