Non-parametric Empirical Bayes and Compound Bayes Estimation of Independent Normal Means

L. Brown

Statistics Department, Wharton School
University of Pennsylvania
lbrown@wharton.upenn.edu

Joint work with E. Greenshtein

Conference:

Innovation and Inventiveness of John Hartigan

April 17, 2009

Topic Outline

1. Empirical Bayes (Concept)
2. Independent Normal Means (Setting + some theory)
3. The NP-EB Estimator (Heuristics)
4. "A Tale of Two Concepts"
   – Empirical Bayes and Compound Bayes
5. (Somewhat) Sparse Problems
6. Numerical results
7. Theorem and Proof
8. The heteroscedastic case – heuristics

Empirical Bayes:
General Background

- ***n*** Parameters to be estimated: $\theta_1,...,\theta_n$. [θ_i real-valued.]
- Observe $X_i \sim f_{\theta_i}$, independent. Let $\mathbf{X} = (X_1,.., X_n)$.
- Estimate θ_i by $\delta_i(\mathbf{X})$.
- Component-wise Loss and Overall Risk

$$L(\theta_i,\delta_i)=(\delta_i-\theta_i)^2 \text{ and}$$

$$R(\underset{\rightarrow}{\theta},\underset{\rightarrow}{\delta}) = \frac{1}{n}\sum E_{\theta_i}\left[L(\theta_i,\delta_i(\mathbf{X}))\right].$$

Bayes Estimation

- "Pretend" $\{\theta_i\}$ are iid, with a (prior) distribution G_n.
- Under G_n the Bayes Procedure would be

$$\Delta^{G_n} = \left(\delta_1^{G_n}, \ldots, \delta_n^{G_n}\right) : \delta_i^{G_n}\left(X_i\right) = E\left(\theta_i \middle| X_i\right).$$

[*Note*: $\delta_i^{G_n}$ depends only on X_i (and on G_n).]

- It would have Bayes risk

$$B_{[n]}\left(G_n\right) = B\left(G_n, \Delta^{G_n}\right) = E_{G_n}\left(R\left(\underset{\rightarrow}{\theta}, \Delta^{G_n}\right)\right).$$

[*Note*: Because of the scaling of sum-of-squared-error-loss by $1/n$ it is the case that $B\left(G_n\right)$ is also the coordinate-wise Bayes risk, ie,

$$B\left(G_n\right) = E_{G_n}\left[E_{\theta_i}\left(\theta_i - \delta_i^{G_n}\left(X_i\right)\right)^2\right].$$]

Empirical Bayes

- Introduced by Robbins:

  An empirical Bayes approach to statistics, *3rd Berk Symp*, 1956
- "Applicable when the same decision problem presents itself repeatedly and independently with a fixed but unknown a priori distribution of the parameter." Robbins, *Ann Math Stat,* 1964
- Thus: Fix G. Let $G_n = G$ for all n.
- Try to find a sequence of estimators, $\tilde{\Delta}_n\left(\mathbf{X}_n\right)$, that are asymptotically as good as Δ^G.
- *ie*, want

$$B_{[n]}\left(G,\tilde{\Delta}_n\right) - B_{[n]}\left(G\right) \to 0.$$

- Much of the subsequent literature emphasized the sequential nature of this problem.

The Fixed-Sample Empirical Goal

- Even earlier Robbins had taken a slightly different perspective.
  Asymptotically subminimax solutions of compound decision problems, *2nd Berk Symp.*, 1951. See also Zhang (2003). Robbins began,
- "When statistical problems of the same type are considered in large groups...there may exist solutions which are asymptotically ... [desirable]"
- That is, one can benefit even for fixed, but large, n (and even if G_n may change with n).
- To measure the desirability we propose

$$(1) \qquad \sup_{G_n \in \mathcal{G}_n} \frac{B_{[n]}\left(G_n, \tilde{\Delta}_n\right) - B_{[n]}\left(G_n\right)}{B_{[n]}\left(G_n\right)} \to 0.$$

- Here, $\mathcal{G}_n$ is a (very inclusive) subset of priors. [*But not all priors*].

Independent Normal Means

- Observe,

$$X_i \sim N\left(\theta_i, \sigma^2\right),\ i = 1,..,n,\ \text{indep.}$$

with σ^2 known. Let φ_{σ^2} denote normal density with Var = σ^2.

- Assume$\theta_i \sim G_n$, *iid*. Write $G = G_n$, for convenience.
- Consider the *i*-th coordinate. Write $\theta_i = \theta,\ x_i = x$ for convenience
- The Bayes estimator (for Squared error loss) is

$$\delta^G\left(x\right) = E_G\left(\theta \middle| X = x\right)$$

- Denote the marginal density of X as

$$g^*\left(x\right) = \int \varphi_{\sigma^2}\left(x-\theta\right) G\left(d\theta\right).$$

- As a general notation, let $\gamma^G\left(x\right) = \delta_G\left(x\right) - x$

- Note that

$$\gamma^G(x) = \delta^G(x) - x = \frac{\int (\theta - x)\varphi_{\sigma^2}(x-\theta)G(d\theta)}{\int \varphi_{\sigma^2}(x-\theta)G(d\theta)}$$

- Differentiate inside the integral (*always OK*), to write

(*) $$\gamma^G(x) = \sigma^2 \frac{g^{*\prime}(x)}{g^*(x)}.$$

- Moral of (*): A **really good** estimate of the marginal density $g^*(x)$ should yield a good approximation to $\gamma^G(x)$.

Validity of (*) is proved in Brown (1971).

Proposed **N**on-**P**arametric **E**mpirical-**B**ayes Estimator

- Let h be a bandwidth constant (to be chosen later).
- Estimate $g^*(x)$ by the kernel density estimator

$$\tilde{g}_h^*(x) = \sum \varphi_h(x - X_i) = \frac{1}{n}\sum \frac{1}{h}\varphi_1\left(\frac{x - X_i}{h}\right).$$

- *The normal kernel has some nice properties, to be explained later.*
- Define the NP EB estimator by $\tilde{\Delta} = (\tilde{\delta}_1, .., \tilde{\delta}_n)$ with

$$\tilde{\delta}_i(x_i) - x_i = \tilde{\gamma}_i(x_i) = \sigma^2 \frac{\tilde{g}_h^{*\prime}(x_i)}{\tilde{g}_h^*(x_i)}.$$

- A useful formula is

$$\tilde{g}_h^{*\prime}(x; \mathbf{X}) = \frac{1}{nh}\sum \frac{X_i - x}{h^2}\varphi\left(\frac{x - X_i}{h}\right).$$

A Key Lemma:

- Let $\hat{G}_n^{\mathbf{X}}$ denote the sample CDF of $\mathbf{X} = (X_1, .., X_n)$.
- Let $g_{G,v}^*$ denote the marginal density when $X_i \sim N(\theta_i, v)$.
- Let $\sigma^2 = 1$ and let $v = 1 + h^2$

Lemma 1: $E\left[\tilde{g}_h^*(x)\right] = g_{G,v}^*(x)$ and $E\left[\tilde{g}_h^{*\prime}(x)\right] = g_{G,v}^{*}{}'(x)$.

Proof:

- $\tilde{g}_h^* = \hat{G}_n^{\mathbf{X}} * \varphi_{h^2}$.
- $E\left[\hat{G}_n^{\mathbf{X}}\right] =$ CDF of $X = G * \varphi_1$.
- Hence, $E\left[\tilde{g}_h^*(x)\right] = E\left[\hat{G}_n^{\mathbf{X}} * \varphi_{h^2}\right] = G * \varphi_1 * \varphi_{h^2} = G * \varphi_v$.

  The proof for the derivatives is similar. ☺

Derivation of the Estimator

The expression for the estimator appears **in red** at the beginning and end of the following string of (approximate) equalities.

- $\gamma_1^G = \dfrac{g_{G,1}^{*\,\prime}}{g_{G,1}^{*}}$ by the fundamental equation (*).
- $\dfrac{g_{G,1}^{*\,\prime}}{g_{G,1}^{*}} \approx \dfrac{g_{G,v}^{*\,\prime}}{g_{G,v}^{*}}$ since $v = 1 + h^2 \approx 1$.
- $\dfrac{g_{G,v}^{*\,\prime}}{g_{G,v}^{*}} \approx \dfrac{\tilde{g}_h^{*\,\prime}}{\tilde{g}_h^{*}}$ from the Lemma

via plug-in Method-of-Moments in numerator and denominator.
See Jiang and Zhang (2007) for a different scheme based on a Fourier kernel.

A Tale of Two Formulations
"Compound" and "Empirical Bayes"

"Compound" Problem:

- Let $\underset{\rightarrow}{\theta}_{(\cdot)} = \left\{\theta_{(1)}, .., \theta_{(n)}\right\}$ and $\underset{\rightarrow}{X}_{(\cdot)} = \left\{X_{(1)}, .., X_{(n)}\right\}$ denote the order statistics of $\underset{\rightarrow}{\theta}$ and $\underset{\rightarrow}{X}$, resp.
- Consider estimators of the form

$$\underset{\rightarrow}{\delta} = \left\{\delta_i\right\} \ni \delta_i = \delta\left(x_i ; \underset{\rightarrow}{x}_{(\cdot)}\right)$$

These are called ***S****imple*-***S****ymmetric* est's. **SS** denotes all of them.

- Given $\underset{\rightarrow}{\theta}_{(\cdot)}$ the optimal **SS** rule is denoted as $\Delta^{\underset{\rightarrow}{\theta}_{(\cdot)}}$. It satisfies

$$R\left(\underset{\rightarrow}{\theta}, \Delta^{\underset{\rightarrow}{\theta}_{(\cdot)}}\right) = \inf_{\Delta \in \mathbf{SS}} R\left(\underset{\rightarrow}{\theta}, \Delta\right).$$

- The goal of Compound decision theory is to find rule(s) that do almost as well as $\Delta^{\underset{\rightarrow}{\theta}_{(\cdot)}}$, as judged by a criterion like (1).

The Link Between the Formulations

EB Implies CO

- Recall that $\hat{G}_n^{\underset{\to}{\theta}(\cdot)}$ denotes the sample CDF of $\underset{\to}{\theta}_{(\cdot)}$.
- Then, $\Delta \in \mathbf{SS}$ implies

$$R(\theta,\Delta) = E\left\{\frac{1}{n}\sum\left[\theta_i - \delta\left(X_i ; \underset{\to}{X}_{(\cdot)}\right)\right]^2\right\} = B\left(\hat{G}_n^{\underset{\to}{\theta}(\cdot)}, \Delta\right).$$

- Consequently: If $\tilde{\Delta}_n$ is EB [in the sense of (1)] then it is also **<u>C</u>**ompound **<u>O</u>**ptimal in the sense of: $\forall \underset{\to}{\theta} \ni \hat{G}_n^{\underset{\to}{\theta}} \in \mathcal{G}_n$.

$$(1') \qquad \frac{R_{[n]}\left(\underset{\to}{\theta}_n, \tilde{\Delta}_n\right) - \inf_{\Delta\in\mathbf{SS}} R_{[n]}\left(\underset{\to}{\theta}_n, \Delta\right)}{\inf_{\Delta\in\mathbf{SS}} R_{[n]}\left(\underset{\to}{\theta}_n, \Delta\right)} < \varepsilon_n \to 0$$

The Converse: CO $\Rightarrow$ EB

- To MOTIVATE the converse, assume $\tilde{\Delta}_n \in \mathbf{SS}$ is CO in that

$$(1') \qquad \sup_{\underset{\rightarrow}{\theta}_n \in \Theta_n} \frac{R_{[n]}\left(\underset{\rightarrow}{\theta}_n, \tilde{\Delta}_n\right) - \inf_{\Delta \in \mathbf{SS}} R_{[n]}\left(\underset{\rightarrow}{\theta}_n, \Delta\right)}{\inf_{\Delta \in \mathbf{SS}} R_{[n]}\left(\underset{\rightarrow}{\theta}_n, \Delta\right)} < \varepsilon_n \rightarrow 0.$$

- Suppose this holds when Θ_n is ALL possible vectors $\underset{\rightarrow}{\theta}_n$.
- Under a prior G_n the vector $\underset{\rightarrow}{\theta}$ has iid components, and

$$B_{[n]}\left(G_n, \Delta\right) = E\left(B_{[n]}\left(\hat{G}_n^{\underset{\rightarrow}{\theta}_{(\cdot)}}, \Delta\right) \middle| \underset{\rightarrow}{\theta}_{(\cdot)}\right).$$

- Truth of $(1')$ for all $\underset{\rightarrow}{\theta}_n$ then implies truth of its Expectation over the distribution of $\underset{\rightarrow}{\theta}_{(\cdot)}$ under G_n. This yields (1).
- In reality $(1')$ does not hold for all $\underset{\rightarrow}{\theta}_n$, but only for a very rich subset. Hence the proof requires extra details.

“Sparse” Problems

- An initial motivation for this research was to create a CO - EB method suitable for “Sparse” settings.
- The proto-typical sparse CO setting has

(sparse) $$\underset{\to(\cdot)}{\theta} = \left(\vartheta_0, \ldots\ldots\ldots\ldots, \vartheta_0, \vartheta_1, .., \vartheta_1\right)$$

with $\approx (1-\alpha)n$ values ϑ_0 and only αn values ϑ_1.

- Here, α is near 0, but ϑ_0, ϑ_1 may be either known or unknown.
- Situations with $\alpha = O(1/n)$ can be considered *extremely* sparse.
- Situations with, say, $\alpha \approx 1/n^{1-\varepsilon}$, $0 < \varepsilon < 1$ are *moderately* sparse.
- Sparse problems are of interest on their own merits (eg in genetics) – for example, as in Efron (2004+).
- And for their importance in building nonparametric regression estimators – see eg, Johnstone and Silverman (2004).

(Typical) **Comparison of NP-EB Estimator with Best Competitor**

#Signals	Est'r	Value =3	Value =4	Value =5	Value =7
5	$\tilde{\delta}_{1.15}$	53	49	42	27
5	Best other	**34**	**32**	**17**	**7**
50	$\tilde{\delta}_{1.15}$	**179**	**136**	**81**	**40**
50	Best other	201	156	95	52
500	$\tilde{\delta}_{1.15}$	**484**	**302**	**158**	**48**
500	Best other	829	730	609	505

Table: Total Expected Squared Error (via simulation; to nearest integer)
Compound Bayes setup with n=1000; 'most' means =0 and others ="Value"
"Best Other" is best performing of 18 studied in *J & S* (2004).

$\tilde{\delta}_{1.15}$ **is our NP – EB est'r with *v* = 1.15**

Statement of EB - CO Theorem

Assumptions:

- $\exists\ \varepsilon' > 0 \ni \mathcal{G}_n \subseteq \left\{ G_n : B_{[n]}\left(G_n\right) > n^{\varepsilon'} \right\}$.

  Hence, (only) moderately sparse settings are allowed.

- $\mathcal{G}_n \subseteq \left\{ G_n : G_n\left(\left[-C_n, C_n\right]\right) = 1 \right\} \ni C_n = O\left(n^{\varepsilon}\right)\ \forall\, \varepsilon > 0$.

  We believe this assumption can be relaxed, but it seems that some sort of uniformly light-tail condition on $\mathcal{G}_n$ is needed.

Theorem: Let $h_n^2 = 1/d_n$ with $d_n/\log(n) \to \infty\ \&\ d_n = o\left(n^{\varepsilon}\right)\ \forall\, \varepsilon > 0$.

Then (under above assumptions) $\tilde{\Delta}_n$ satisfies (1).

[Note: $n = 1000\ \&\ d_n = \log(n) \Rightarrow v = 1 + d^{-1} \approx 1.15$, as in Table.]

Heteroscedastic Setting

EB formulation:

- $\sigma_1^2, .., \sigma_n^2$ known
- Observe $X_i \sim N\left(\theta_i, \sigma_i^2\right)$, indep., $i = 1, .., n$.
- Assume (EB) $\theta_i \sim G_n$, indep.,
- but G_n unknown, except for $G_n \in \mathcal{G}_n$.
- Loss function, risk function, and optimality target, (1), as before.

Heuristics

- Bayes estimator on *i*-th coordinate has

$$\gamma_i^G\left(x_i\right) = \sigma_i^2\left(g_{\sigma_i^2}^{*}{}'\left(x_i\right) \Big/ g_{\sigma_i^2}^{*}\left(x_i\right)\right).$$

- Previous heuristics suggest approximating $g_{\sigma_i^2}^{*}\left(x_i\right)$ by

$$g_{\sigma_i^2}^{*}\left(x_i\right) \approx g_{\sigma_i^2\left(1+\ell^2\right)}^{*}\left(x_i\right)$$

- And then estimating $g_{\sigma_i^2\left(1+\ell^2\right)}^{*}\left(x_i\right)$ as the average of

$$g_{\sigma_i^2\left(1+\ell^2\right)}^{*}\left(x_i\right) \approx \varphi_{h^2=\sigma_i^2\left(1+\ell^2\right)-\sigma_k^2}\left(x_i - X_k\right), k = 1,..,n.$$

- To avoid impossibilities, need to use

$$h_{k,i}^2 = \left(\sigma_i^2\left(1+\ell^2\right) - \sigma_k^2\right)_+.$$

- Resulting estimator has

$$\gamma_i^G(x_i) = \sigma_i^2\left(\tilde{g}^{*\prime}(x_i)\big/\tilde{g}^*(x_i)\right)$$

with $h_{k,i}^2$ as above, and

$$\tilde{g}_i^*(X) = \frac{\sum_k I_{\left\{k:\, h_{k,i}^2>0\right\}}(k)\varphi_{h_{k,i}^2}(x_i - X_k)}{\sum_k I_{\left\{k:\, h_{k,i}^2>0\right\}}}.$$

- (With inessential modifications) this is the estimator used in Brown (AOAS, 2008).