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 sums of random variables

- The problem
- A species problem
- Data confidentiality
- Asymptotic efficiency

The problem

Let (X_j, θ_j) be random vectors such that

$$X_j|\theta_j \sim F(x|\theta_j),\ j = 1, \ldots, n,$$

for a known family $F(x|\theta)$ of distributions. Let $u(\cdot, \cdot)$ be a certain "utility" function. How do we estimate the sum

$$S_n \equiv \sum_{j=1}^{n} u(X_j, \theta_j)$$

based on observations $X_1, \ldots, X_n$?

- Zipf (1932, text analysis), Good (53, species; A.M. Turing), ...
- Robbins (77, 88), Robbins-Zhang (88, 89, 91, 00)

• Example: Given a pool of n motorists, how do we estimate the (risk) intensity of those in the pool who have a prespecified number, say a, of accidents this year? Let X_j be the number of accidents this year for the j-th motorist and θ_j the intensity. We may assume $X_j|\theta_j \sim \text{Poisson}(\theta_j)$ and set

$$S_n \equiv \sum_{j=1}^{n} \theta_j u(X_j), \quad u(x) = I\{x = a\}.$$

The Bayes estimator of S_n is

$$E\big[S_n\big|\text{data}\big] = \sum_{j=1}^{n} (a+1)\frac{P(X_j = a+1)}{P(X_j = a)} u(X_j),$$

but we don't always know the marginal distributions of X_j.

◇ A parametric method: Assume θ_j are iid exponential with an unknown mean. Then, the MLE/Bayes methods lead to

$$\widehat{S}_n = \sum_{j=1}^{n} \frac{(\alpha/n + \overline{X})u(X_j)}{(\alpha+\beta)/n + 1 + \overline{\overline{X}}},$$

since $P(X_j = a+1)/P(X_j = a) = EX/(1+EX)$.

◇ A nonparametric method: Assume that X_j are identically distributed. Then, $P(X_j = a+1)/P(X_j = a) \approx n_{a+1}/n_a$, where $n_k \equiv \sum_{j=1}^{n} I\{X_j = k\}$. This leads to

$$\widehat{S}_n = (a+1)n_{a+1}\,.$$

◇ Are these estimators asymptotically optimal?

A species problem

Suppose a random sample of size N is drawn (with replacement) from a population of d species. Let n_k be the number of species represented k times in the sample. Our problem is to estimate the total number of species d based on $\{n_k, k \geq 1\}$.

- Fisher *et al* (43), Good (53), Bunge-Fitzpatrick (93), ...
- Let X_j be the frequencies of the j-th species in the sample, so

$$(X_1, \ldots, X_d) \big| N \sim \text{multinomial}\,(N, p_1, \ldots, p_d)$$

for certain $p_j > 0$. We observe $\{n_k, k \geq 1\}$ (but not n_0), where

$$n_k \equiv \sum_{j=1}^{d} I\{X_j = k\}.$$

- The parameter d is under estimated by the observed

$$\widetilde{d} \equiv \sum_{k=1}^{\infty} n_k = \sum_{j=1}^{d} I\{X_j > 0\}.$$

• Probability models for p_j: for certain i.i.d. $\theta_j \sim G$

$$p_j = \theta_j / \textstyle\sum_{i=1}^d \theta_i, \quad N \big| \{\theta_j\} \sim \text{ Poisson}\big(c \textstyle\sum_{i=1}^d \theta_i\big).$$

Thus, $P\{X_j = k\} = \int \{e^{-cy}(cy)^k/k!\} G(dy)$ is a Poisson mixture.

• Parametric models $G \in \{G_\tau : \tau \in \mathcal{T}\}$, e.g. gamma. Assume $c = 1$. The (conditional) MLE is given by

$$\hat{d} \equiv \frac{\widetilde{d} \int_{y>0} G_{\widehat{\tau}}(dy)}{\int \big(1 - e^{-y}\big) G_{\widehat{\tau}}(dy)}, \ \widehat{\tau} \equiv \arg\max_{\tau \in \mathcal{T}} \prod_{k=1}^{\infty} \left\{ \frac{\int e^{-y} y^k G_\tau(dy)}{1 - \int e^{-y} G_\tau(dy)} \right\}^{n_k},$$

cf. Samford (55), Rao (71) and Engen (74) for Poisson/gamma.

• Nonparametric MLE: maximizing over all G, i.e. with $\{G_\tau\}$ being the collection of all distributions. The EM algorithm.

• Bias correction: Darroch-Ratcliff (80), Chao-Lee (92), Chao-Bunge (02)

• Connection to the estimation of sums of random variables

◇ Let d be treated as the number of species represented in the population out of a total of n species. Specifically, letting $p_j = 0$ if the j-th species is not represented in the population, estimating

$$d = \sum_{j=1}^{n} I\{p_j > 0\} = \sum_{j=1}^{n} I\{X_j = 0, p_j > 0\} + \sum_{k=1}^{N} n_k$$

is equivalent to estimating $S_n \equiv \sum_{j=1}^{n} u(X_j, p_j)$ with $u(x,p) = I\{p > 0\}$ or $u(x,p) = I\{x = 0, p > 0\}$, based on observations $\{X_j, j \le n\}$.

◇ For $p_j \propto \theta_j$, $(\widetilde{d}, d - \widetilde{d}, n - d)$ is trinomial, and the likelihood is

$$(1 - p_*)^d p_*^{n-d} \Big(\int e^{-y} G(dy) \Big)^{d-\widetilde{d}} \prod_{k=1}^{N} \Big(\int (e^{-y} y^k / k!) G(dy) \Big)^{n_k}$$

with $p_* = P\{p_j = 0\}$ and $\theta_j | \{\theta_j > 0\} \sim G$. In this case, $P\{\text{CMLE} = \text{MLE}\} \to 1$.

Data confidentiality

• Protection of the privacy of individuals in releasing microdata sets in the form of a high-dimensional contingency table.

If an individual belongs to a cell with small frequency, an intruder with certain knowledge about the individual may identify him and learn sensitive information about him in the data.

• Duncan and Pearson (1991), the proceedings of the joint ECE/EUROSTAT work sessions on statistical data confidentiality, e.g. Polettini and Seri (2003), Rinott (2003) and more.

- Global disclosure risk:

$$S_J \equiv \sum_{j=1}^{J} u(X_j, Y_j),$$

where X_j and Y_j are the sample and population frequencies in the j-th cell, J is the total number of cells, and $u(x, y)$ is a loss function of the form $u(x, y) = u(x)/y$, e.g. $u(x, y) = y^{-1} I\{x = 1\}$.

- Problem: estimation of S_J based on $\{X_j, j \leq J\}$.

- Model: Let $N = \sum_{j=1}^{J} Y_j$ be the population size. Suppose $N \sim \text{Poisson}(\lambda)$,

$$\{Y_j\}|N \sim \text{multinomial}\,(N, \{\pi_j\}),\ X_j|(\{Y_j\}, N) \sim \text{binomial}\,(Y_j, p_j),$$

for certain $\pi_j > 0$ with $\sum_{j=1}^{J} \pi_j = 1$, $0 \leq p_j \leq 1$ and $\lambda > 0$.

• For known $\{p_j, \pi_j, \lambda\}$, the Bayes estimator of S_J is

$$S_J^* \equiv E\Big(S_J\Big|\{X_j\}\Big) = \sum_{j=1}^{J} \bar{u}_j(X_j), \quad \bar{u}_j(x) \equiv Eu(x, Y_j - X_j + x),$$

with $Y_j - X_j \sim \text{Poisson}((1-p_j)\pi_j\lambda)$ (independent of X_j).

◇ For $u(x,y) = y^{-1}I\{x=1\}$,

$$\bar{u}_j(x) = \big\{(1-p_j)\pi_j\lambda\big\}^{-1}\Big[1 - \exp\big\{-(1-p_j)\pi_j\lambda\big\}\Big]I\{x=1\}.$$

• Connection to the species problem: for large λ

$$\begin{aligned}\sum_j \bar{u}_j(X_j) &\approx \sum_j \Big[1 - \exp\big\{-\lambda_j\big\}\Big]I\{X_j = 1, \lambda_j > 0\} \\ &\approx \sum_j I\{X_j = 0, \lambda_j > 0\}, \quad \lambda_j \equiv (1-p_j)\pi_j\lambda\end{aligned}$$

• Negative binomial models: $N \sim \mathrm{NB}(\alpha, 1/(1+\beta))$. As in Rinott (2003), $\bar{u}_j(x) = E[u(X_j, Y_j)|X_j = x]$ is

$$\bar{u}_j(x) = \frac{1 + p_j\beta_j}{(1-p_j)\beta_j} \int_{(1+p_j\beta_j)/(1+\beta_j)}^{1} t^{\alpha-1} dt I\{x = 1\}$$

for $u(x, y) = y^{-1} I\{x = 1\}$, $\beta_j \equiv \beta\pi_j$, cf. Bethlehem *et al* (1990) with $\pi_j = 1/J$ and $p_j = En/EN \approx n/N$.

◇ For $(\alpha, \beta_j) \to (0, \infty)$, $(Y_j - X_j)|\{X_j = x\} \to \mathrm{NB}(x, p_j)$ in distribution, resulting in the μ-ARGUS estimator (Benedetti and Franconi, 1998) with $\bar{u}_j(x) = p_j(1-p_j)^{-1}(-\log p_j) I\{x = 1\}$. Compared with the Poisson model in which $\lambda \approx N$, estimates of both EN and $\mathrm{Var}(N)$ are required. The μ-ARGUS model essentially assumes $\mathrm{Var}(N)/(EN)^2 \geq 1/\alpha \to \infty$.

• Parametric (regression) models

Let $\{p_j, \pi_j, \lambda\}$ be known tractable functions of an unknown vector τ and certain covariates z_j characterizing cells j, incorporating all available knowledge about the parameters, e.g. $\lambda \approx N$ and $\sum_{j=1}^{J} p_i \pi_j \approx n/N$, where $n = \sum_{j=1}^{J} X_j$ is the sample size. Consequently, $\bar{u}_j(x) = \bar{u}(x, z_j; \tau)$. This suggests

$$\widehat{S}_J \equiv \sum_{j=1}^{J} \bar{u}\big(X_j, z_j; \widehat{\tau}_J\big)$$

as an estimator of S_J, e.g. with MLE $\widehat{\tau}_J$.

◇ Example: In a two-way table with cells $j \sim (i,k)$ and known $\pi_{i,k}$ and λ, $p_{i,k} = \psi_0(\tau_1 + \tau_2' z_{i,k})$, e.g. logit or probit ψ_0. For unknown $\pi_{i,k}$, we may assume $\pi_{i,k} = \pi_{i.}\pi_{.k}$.

• How good are these estimators asymptotically?

Asymptotic efficiency

Let $(X_j, \theta_j), j \leq n$, be iid from F. We want to estimate

$$S_n \equiv S_n(F) \equiv \sum_{j=1}^{n} u(X_j, Y_j; F)$$

with certain utility function $u(x, y; F)$.

Theorem. *Under certain regularity conditions, the efficient influence function for the estimation of S_n in contiguous neighborhoods of P_{F_0} is*

$$\phi_*(x) = \psi_*(x) + \bar{u}(x; F_0) - \mu(F_0) - u_*(x)$$

where $\psi_*(x)$ is the efficient influence function for the estimation of $\mu(F) = E_F u(X, \theta)$, $\bar{u}(x; F) \equiv E_F[u(X, \theta)|X = x]$, and $u_*(x)$ is the projection of $\bar{u}(x; F_0)$ to the tangent space of all score functions based on observations.

• The estimation of $S_n(F)$ or $\mu(F)$ are closely related, but an efficient estimator of $\mu(F)$ is not necessarily efficient for the estimation of $S_n(F)/n$.

• Cramer-Rao type argument in the parametric case.

Suppose $F \equiv F_\tau$ with density f_τ and $t(x)$ is an unbiased estimator of $u(X, \theta; \tau)$. Differentiate $E_\tau t(X) = \mu_\tau \equiv E_\tau u(X, \theta; \tau)$ yields

$$E_\tau t(X)\rho_\tau(X) = E_\tau \psi_{*,\tau}(X)\rho_\tau(X)$$

where $\psi_{*,\tau}$ is the efficient influence function for the estimation of μ_τ. Under this constraint,

$$\psi_{*,\tau} + \bar{u}_\tau - u_{*,\tau} = \arg\min_{t(x)} E_\tau \big(u(X, \theta; \tau) - t(X)\big)^2,$$

where $\bar{u}_\tau(x) \equiv E_\tau \big[u(X, \theta; \tau)|X = x\big]$ and $u_{*,\tau}$ is the projection of $\bar{u}_\tau$ to $\overline{[\rho_\tau]}$.

• Implication in (regular) parametric models

Let

$$\bar{u}(x;\tau) \equiv E_\tau\big[u(X,\theta;\tau)|X=x\big].$$

Then, the "plug-in" estimator

$$\widehat{S}_n \equiv \sum_{j=1}^{n} \bar{u}(X_j;\widehat{\tau}),$$

is asymptotically efficient for $S_n \equiv \sum_{j=1}^{n} u(X_j,\theta;\tau)$, if $\widehat{\tau}$ is an efficient estimate of τ, e.g. MLE.

• Implication in nonparametric mixture models: under certain regularity conditions, the efficient influence functions $\phi_{*,F}$ at F must satisfy

$$E_F \phi_{*,F_0}(X) = E_F u(X, \theta; F)$$

for almost all F and F_0, i.e. efficient estimators are within $o(\sqrt{n})$ of "u,v" estimators of Robbins (88) of the extended form

$$\widehat{S}_n \equiv \sum_{j=1}^{n} v(X_i)$$

for certain v satisfying $E_F v(X) = E_F u(X, \theta; F)$ for all F.

Conclusions

- Estimation of sums of random variables has broad applications
- An asymptotic theorem is provided in this nonstandard estimation problem
- In parametric models, the "plug-in" estimator is asymptotically efficient
- In nonparametric mixture models, the (conditionally unbiased) "u,v" estimators are asymptotically efficient (if any)