Lec 7: Empirical Bayes

Yanjun Han

Oct. 22, 2024

A motivating example.

Given $y \sim N(\theta, I)$ with $\theta \in \mathbb{R}^p$, aim to estimate θ under quadratic loss: $\mathbb{E}_\theta \|\hat{\theta}(y) - \theta\|^2$.

Natural estimator: $\hat{\theta}^{MLE}(y) = y$

Lots of nice properties: MLE, minimax, UMVUE (uniform smallest variance unbiased estimator), MRE (minimum risk equivariant estimator), ...

Shocking observation: $\exists$ uniformly better estimator than $\hat{\theta}^{MLE}$, if $p \geq 3$.

Theorem (Stein '56, James-Stein '61)

For $p \geq 3$, the James-Stein estimator

$$\hat{\theta}^{JS} = \left(1 - \frac{p-2}{\|y\|^2}\right) y$$

is uniformly better than $\hat{\theta}^{MLE}$.

$$\mathbb{E}_\theta \|\hat{\theta}^{JS} - \theta\|^2 < \mathbb{E}_\theta \|\hat{\theta}^{MLE} - \theta\|^2, \quad \forall \theta \in \mathbb{R}^p.$$

Note: JS estimator leads to the shrinkage idea, i.e. shrink y slightly to zero (slightly increases bias, significantly reduce variance)

Pf. Let $y = \theta + \xi$ with $\xi \sim N(0, I_p)$.

Risk of MLE. $\mathbb{E}_\theta \|\hat{\theta}^{MLE} - \theta\|^2 = \mathbb{E}\|\xi\|^2 = p, \quad \forall \theta.$

Risk of JS. $$\mathbb{E}_\theta \|\hat{\theta}^{JS} - \theta\|^2 = \mathbb{E}_\theta \left[\|\xi\|^2 - 2(p-2)\left\langle \frac{y}{\|y\|^2}, \xi \right\rangle + \frac{(p-2)^2}{\|y\|^2} \right].$$

Lemma (Stein's identity).

Let $\xi \sim N(0, I_p)$ and $f: \mathbb{R}^p \to \mathbb{R}^p$. Then

$$\mathbb{E}[\langle \xi, f(\xi) \rangle] = \mathbb{E}[\nabla \cdot f(\xi)]$$

divergence: $\nabla \cdot f(\xi) = \sum_{i=1}^{p} \frac{\partial f_i}{\partial \xi_i}(\xi)$.

Pf. Suffice to prove the case $p=1$. Here

$$\mathbb{E}[f'(\xi)] = \int_{-\infty}^{+\infty} f'(\xi)\varphi(\xi)\,d\xi$$

$$= \underbrace{f(\xi)\varphi(\xi)\Big|_{-\infty}^{+\infty}}_{\substack{0 \text{ if } f \text{ has}\\ \text{sub-exponential}\\ \text{growth}}} - \int_{-\infty}^{+\infty} f(\xi)\varphi'(\xi)\,d\xi \quad \text{(integration by parts)}$$

$$= \int_{-\infty}^{+\infty} f(\xi)\,\xi\,\varphi(\xi)\,d\xi \quad (\varphi'(\xi) = -\xi\varphi(\xi))$$

$$= \mathbb{E}[\xi f(\xi)]$$

By Stein's identity,

$$\mathbb{E}_\theta \left\langle \frac{y}{\|y\|^2}, \xi \right\rangle = \mathbb{E}_\theta \left\langle \frac{\theta+\xi}{\|\theta+\xi\|^2}, \xi \right\rangle$$

$$= \mathbb{E}_\theta \left[\nabla \cdot \frac{\theta+\xi}{\|\theta+\xi\|^2} \right]$$

$$= \mathbb{E}_\theta \left[\sum_{i=1}^{p} \frac{\|\theta+\xi\|^2 - 2(\theta_i+\xi_i)^2}{\|\theta+\xi\|^4} \right]$$

$$= \mathbb{E}_\theta \left[\frac{p-2}{\|\theta+\xi\|^2} \right] = \mathbb{E}_\theta \left[\frac{p-2}{\|y\|^2} \right].$$

So

$$\mathbb{E}_\theta \|\hat\theta^{JS} - \theta\|^2 = p - 2(p-2)\,\mathbb{E}_\theta\left[\frac{p-2}{\|y\|^2}\right] + (p-2)^2\,\mathbb{E}_\theta\left[\frac{1}{\|y\|^2}\right]$$

$$= p - (p-2)^2\,\mathbb{E}_\theta\left[\frac{1}{\|y\|^2}\right]$$

$$< p$$

$$= \mathbb{E}_\theta \|\hat\theta^{MLE} - \theta\|^2 \qquad \square$$

An empirical Bayes view of JS estimator.

Consider a Bayes setting where $\theta \sim N(0, \tau^2 I_p)$.
Then

$$y \sim N(0, (1+\tau^2) I_p) \quad \text{(marginal distribution)}$$

$$\theta \mid y \sim N\left(\frac{\tau^2 y}{1+\tau^2}, \frac{\tau^2}{1+\tau^2}\right) \quad \text{(posterior)}$$

Bayes estimator:

$$\hat{\theta}^{\text{Bayes}}(y) = \mathbb{E}[\theta \mid y] = \frac{\tau^2}{1+\tau^2} y.$$

Problem: don't know how to set τ.

Empirical Bayes: estimate (functions of) τ based on marginal distribution of y!

Since $y \sim N(0, (1+\tau^2) I_p)$, then

$$\mathbb{E}\left[\frac{1}{\|y\|^2}\right] = \frac{1}{1+\tau^2} \int_0^\infty \frac{1}{2^{p/2}\Gamma(p/2)} t^{\frac{p}{2}-1} e^{-\frac{t}{2}} \cdot \frac{dt}{t}$$

$$= \frac{1}{1+\tau^2} \int_0^\infty \frac{1}{2\Gamma(p/2)} u^{\frac{p}{2}-2} e^{-u} du \quad (t = 2u)$$

$$= \frac{1}{1+\tau^2} \cdot \frac{\Gamma(\frac{p}{2}-1)}{2\Gamma(\frac{p}{2})} = \frac{1}{1+\tau^2} \cdot \frac{1}{p-2}$$

$\Longrightarrow$ $\dfrac{p-2}{\|y\|^2}$ is an unbiased estimator of $\dfrac{1}{1+\tau^2}$.

So

$$\hat{\theta}^{\text{Bayes}}(y) = \frac{\tau^2}{1+\tau^2} y$$

$$= \left(1 - \frac{1}{1+\tau^2}\right) y$$

$$\approx \left(1 - \frac{p-2}{\|y\|^2}\right) y \quad \text{(fully data-driven: JS!)}$$

Robbins' empirical Bayes model.

Given $\begin{cases} y_1 \sim P_{\theta_1} \\ y_2 \sim P_{\theta_2} \\ \vdots \\ y_k \sim P_{\theta_k} \end{cases}$, aim to estimate (functions of) $\theta \in \mathbb{R}^k$.

<u>i.i.d. model</u> (Robbins' 56) $\theta_1, \cdots, \theta_k \overset{\text{i.i.d.}}{\sim} G$ (unknown prior)

<u>compound model</u> (Robbins' 51) no distributional assumption on θ

(but usually pretend that $\theta_1, \cdots, \theta_k \overset{\text{i.i.d.}}{\sim} G$ with $G = \frac{1}{k}\sum_{i=1}^{k} \delta_{\theta_i}$)

typical steps of EB:

1. for given G, obtain the Bayes estimator $\hat{\theta}_G(y)$
2. use $y_1, \cdots, y_k$ to estimate G

   (two approaches: f-modeling and g-modeling — later)

Robbins' estimator in Poisson models

Assume $y_i \overset{\text{ind}}{\sim} \mathrm{Poi}(\theta_i)$. $i = 1, 2, \cdots, k$.

The Bayes estimator for θ:

$$\begin{aligned} \hat{\theta}_G &= \mathbb{E}_G[\theta \mid y] \\ &= \frac{\mathbb{E}_G\left[\theta \cdot e^{-\theta} \frac{\theta^y}{y!}\right]}{\mathbb{E}_G\left[e^{-\theta} \frac{\theta^y}{y!}\right]} \\ &= (y+1) \cdot \frac{\mathbb{E}_G\left[e^{-\theta} \frac{\theta^{y+1}}{(y+1)!}\right]}{\mathbb{E}_G\left[e^{-\theta} \frac{\theta^y}{y!}\right]} \\ &= (y+1) \cdot \frac{f_G(y+1)}{f_G(y)} \end{aligned}$$

(f_G: marginal distribution of y)

An unbiased estimator for $f_G(y)$:

$$\mathbb{E}\Big[\underbrace{\frac{1}{k}\sum_{i=1}^{k} 1(y_i = y)}_{=:\, N_y}\Big] = f_G(y).$$

$$\boxed{\text{Robbins' estimator: } \quad \hat{\theta}_i = (y_i+1)\frac{N_{y_i+1}}{N_{y_i}}}$$

Good-Turing estimator.

Consider a special case $y_i \sim \mathrm{Poi}(np_i)$, $i=1,\cdots,k$ with $\sum_{i=1}^{k} p_i = 1$.

(background: Poissonization of an i.i.d. sampling model.

Raw data: $x_1, \cdots, x_n \overset{i.i.d.}{\sim} (p_1, \cdots, p_k)$ (unknown)

The histogram: $y_i = \sum_{j=1}^{n} \mathbb{1}(x_j = i)$, $i=1,\cdots,k$

$(y_1,\cdots,y_k) \sim \mathrm{Multi}(n; (p_1,\cdots,p_k)) \approx \mathrm{Poi}(np_1) \times \cdots \times \mathrm{Poi}(np_k)$)

Target: estimate the portion of unseen species, i.e.

$$p^{(0)} := \sum_{i=1}^{k} \mathbb{1}(y_i = 0)\, p_i .$$

Note: MLE would be meaningless as it always outputs $p_i = 0$ if $y_i = 0$.

EB solution.

If $p_1, \cdots, p_n \overset{i.i.d}{\sim} G$, then

$$\mathbb{E}_G[p_i \mid y_i] = \frac{y_i+1}{n}\,\frac{f_G(y_i+1)}{f_G(y_i)} \approx \frac{y_i+1}{n}\,\frac{N_{y_i+1}}{N_{y_i}}$$

(Robbins' estimator)

$\Rightarrow$ a good estimator for $\sum_{i=1}^{k} \mathbb{1}(y_i=0)p_i$ is

$$\sum_{i=1}^{k} \mathbb{1}(y_i=0)\,\frac{y_i+1}{n}\cdot\frac{N_{y_i+1}}{N_{y_i}} = N_0 \cdot \frac{1}{n}\cdot\frac{N_1}{N_0} = \boxed{\frac{N_1}{n}}.$$

N_1 : # of unique species in $x_1, \cdots, x_n$

Intuition:
- if $x_1,\cdots,x_n$ are all distinct, then probably the population is large and mostly unseen, i.e. $p^{(0)} \approx 1$
- if each x appears at least twice $\Rightarrow p^{(0)} \approx 0$.

$$\boxed{\text{Good-Turing estimator.} \quad \hat{p}_i = \frac{y_i+1}{n} \cdot \frac{N_{y_i+1}}{N_{y_i}}}$$

Predicting # of new species. (Thisted & Efron '76, '87)

A related question: given a collection $x_1, \cdots, x_n$ of n observations:

- how many new species do we expect to see in a new sample of size m?
- how many species do we expect to see, in a new sample of size m, which appear exactly t times in the original sample?

The Poisson model. $y_i \overset{ind}{\sim} \mathrm{Poi}(np_i)$, $i = 1, \cdots, k$, $\sum_{i=1}^{k} p_i = 1$.

Aim to estimate

$$\sum_{i=1}^{k} \underbrace{1(y_i = t)}_{\text{species appearing exactly } t \text{ times in original sample}} \underbrace{(1 - e^{-mp_i})}_{\text{prob. of observing } i \text{ in the new sample}}, \quad k = 0, 1, \cdots$$

EB estimation of p_i^{ℓ}.

$$\mathbb{E}_G[p^{\ell} \mid y] = \frac{\mathbb{E}_G\left[p^{\ell} \cdot e^{-np} \frac{(np)^y}{y!}\right]}{\mathbb{E}_G\left[e^{-np} \frac{(np)^y}{y!}\right]} = \frac{(y+1) \cdots (y+\ell)}{n^{\ell}} \frac{f_G(y+\ell)}{f_G(y)}$$

$$\approx \frac{(y+\ell)!}{y!\, n^{\ell}} \cdot \frac{N_{y+\ell}}{N_y}.$$

Final EB estimator.

$$\sum_{i=1}^{k} 1(y_i = t)(1 - e^{-mp_i}) = \sum_{i=1}^{k} 1(y_i = t) \sum_{\ell=1}^{\infty} (-1)^{\ell+1} \frac{(mp_i)^{\ell}}{\ell!}$$

$$\approx \cancel{\sum_{i=1}^{k} 1(y_i = t)} \sum_{\ell=1}^{\infty} (-1)^{\ell+1} \binom{y_i+\ell}{\ell} \left(\frac{m}{n}\right)^{\ell} \frac{N_{y_i+\ell}}{\cancel{N_{y_i}}}$$

$$= \sum_{\ell=1}^{\infty} (-1)^{\ell+1} \binom{t+\ell}{\ell} \left(\frac{m}{n}\right)^{\ell} N_{\ell+t}.$$

Example: if $t=0$ and $m=n$, then

$\#$ of new species in the next n observations $\approx N_1 - N_2 + N_3 - \cdots$

Gaussian location model: Tweedie's formula

If $y \sim N(\theta, 1)$ with $\theta \sim G$, then

$$\mathbb{E}_G[\theta \mid y] = \frac{\mathbb{E}_G[\theta\varphi(y-\theta)]}{\mathbb{E}_G[\varphi(y-\theta)]}$$

$$= y - \frac{\mathbb{E}_G[(y-\theta)\varphi(y-\theta)]}{\mathbb{E}_G[\varphi(y-\theta)]}$$

$\mathbb{E}_G[-\varphi'(y-\theta)] = -f_G'(y)$

$f_G(y)$

$$= y + \frac{d}{dy}\log f_G(y)$$

(Tweedie's formula)

An estimator without knowledge of G: estimate the marginal distribution $\hat{f}(y)$ of y based on $y_1, \cdots, y_K$, then use $\hat{f}$ in place of f_G.

f-modeling vs. g-modeling (ongoing topic)

- All previous examples use f-modeling, where we aim to estimate the marginal distribution f_G based on $y_1, \cdots, y_K$
- f-modeling is simple, yet could have a large variance
- g-modeling: estimate G based on $y_1, \cdots, y_K$ (parametric/nonparametric)

  A popular choice: NPMLE $\hat{G} = \arg\max_G \sum_{i=1}^{k} \log \mathbb{E}_G[p_\theta(y_i)]$

  (convex, but infinite dimensional)

  $\longrightarrow$ then use $\hat{G}$ in place of G.
- some evidence that g-modeling gives better estimation performances.

EB in practice: choose hyperparameters

Model: $(x_1, y_1), \cdots, (x_n, y_n) \sim p_\theta$ for unknown θ

Bayesian inference: assume a prior on $\theta \sim \pi_\lambda$ (e.g. the conjugate prior)

Question: how to choose the hyperparameter λ?

EB approach: $\hat{\lambda} = \arg\max_{\lambda} \int p(\{x_i, y_i\}_{i=1}^{n} \mid \theta)\, \pi_\lambda(\theta)\, d\theta$

(use data to chose the hyperparameter!)