

Lec 7: Empirical Bayes

YanJun Han

Oct. 24, 2023



A motivating example.

Given $y \sim N(0, I)$ with $\theta \in \mathbb{R}^p$, aim to estimate θ under quadratic loss: $\mathbb{E}_\theta \|\hat{\theta}(y) - \theta\|^2$.

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

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

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

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

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

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

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

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

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

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

Lemma (Stein's identity).

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

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

$$\downarrow \text{divergence: } \nabla \cdot f(\mathbf{z}) = \sum_{i=1}^p \frac{\partial f_i}{\partial z_i}(\mathbf{z}).$$

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

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

$$= f(\mathbf{z}) \varphi(\mathbf{z}) \Big|_{-\infty}^{+\infty} - \int_{-\infty}^{+\infty} f(\mathbf{z}) \varphi'(\mathbf{z}) d\mathbf{z} \quad (\text{integration by parts})$$

0 if f has
sub-exponential
growth

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

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

By Stein's identity,

$$\mathbb{E}_0 \left\langle \frac{\mathbf{y}}{\|\mathbf{y}\|^2}, \mathbf{z} \right\rangle = \mathbb{E}_0 \left\langle \frac{\mathbf{0} + \mathbf{z}}{\|\mathbf{0} + \mathbf{z}\|^2}, \mathbf{z} \right\rangle$$

$$= \mathbb{E}_0 \left[\nabla \cdot \frac{\mathbf{0} + \mathbf{z}}{\|\mathbf{0} + \mathbf{z}\|^2} \right]$$

$$= \mathbb{E}_0 \left[\sum_{i=1}^p \frac{\|\mathbf{0} + \mathbf{z}\|^2 - 2(\mathbf{0}_i + \mathbf{z}_i)^2}{\|\mathbf{0} + \mathbf{z}\|^4} \right]$$

$$= \mathbb{E}_0 \left[\frac{p-2}{\|\mathbf{0} + \mathbf{z}\|^2} \right] = \mathbb{E}_0 \left[\frac{p-2}{\|\mathbf{y}\|^2} \right].$$

$$\begin{aligned} \text{So } \mathbb{E}_0 \|\hat{\theta}^{\text{JS}} - \mathbf{0}\|^2 &= p - 2(p-2) \mathbb{E}_0 \left[\frac{p-2}{\|\mathbf{y}\|^2} \right] + (p-2)^2 \mathbb{E}_0 \left[\frac{1}{\|\mathbf{y}\|^2} \right] \\ &= p - (p-2)^2 \mathbb{E}_0 \left[\frac{1}{\|\mathbf{y}\|^2} \right] \\ &< p \\ &= \mathbb{E}_0 \|\hat{\theta}^{\text{MLE}} - \mathbf{0}\|^2 \quad \square \end{aligned}$$

An empirical Bayes view of JS estimator.

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

Then

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

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

Bayes estimator:

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

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

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

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

$$\begin{aligned} \text{So } \hat{\theta}^{\text{Bayes}}(y) &= \frac{\tau^2}{1+\tau^2} y \\ &= \left(1 - \frac{1}{1+\tau^2}\right) y \\ &\approx \left(1 - \frac{p-2}{\|y\|^2}\right) y \quad (\text{fully data-driven: JS!}) \end{aligned}$$

Robbins' empirical Bayes model

$$\text{Given } \begin{cases} Y_1 \sim P_{\theta_1} \\ Y_2 \sim P_{\theta_2} \\ \vdots \\ Y_k \sim P_{\theta_k} \end{cases}, \text{ aim to estimate (functions of) } \theta \in \mathbb{R}^k.$$

i.i.d. model (Robbins' 56) $\theta_1, \dots, \theta_k \stackrel{\text{i.i.d.}}{\sim} G$ (unknown prior)

compound model (Robbins' 51) no distributional assumption on θ
(but usually pretend that $\theta_1, \dots, \theta_k \stackrel{\text{i.i.d.}}{\sim} G$ 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, \dots, y_k to estimate G

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

Robbins' estimator in Poisson models

Assume $y_i \stackrel{\text{i.i.d.}}{\sim} \text{Poi}(\theta_i)$, $i=1, 2, \dots, k$.

The Bayes estimator for θ :

$$\begin{aligned} \hat{\theta}_G &= \mathbb{E}_G[\theta | y] \\ &= \frac{\mathbb{E}_G[\theta \cdot e^{-\theta} \frac{\theta^y}{y!}]}{\mathbb{E}_G[e^{-\theta} \frac{\theta^y}{y!}]} \\ &= (y+1) \cdot \frac{\mathbb{E}_G[e^{-\theta} \frac{\theta^{y+1}}{(y+1)!}]}{\mathbb{E}_G[e^{-\theta} \frac{\theta^y}{y!}]} \\ &= (y+1) \cdot \frac{f_G(y+1)}{f_G(y)} \quad (f_G: \text{marginal distribution of } y) \end{aligned}$$

An unbiased estimator for $f_G(y)$:

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

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

Good-Turing estimator.

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

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

Raw data: $x_1, \dots, x_n \stackrel{\text{i.i.d.}}{\sim} (p_1, \dots, p_k)$ (unknown)

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

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

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

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

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

EB solution.

If $p_1, \dots, p_n \stackrel{\text{i.i.d.}}{\sim} G$, then

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

(Robbins' estimator)

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

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

N_1 : # of unique species in x_1, \dots, x_n

Intuition: - if x_1, \dots, 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 $\Rightarrow p^{(0)} \approx 0$.

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

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

A related question: given a collection x_1, \dots, 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. $y_i \stackrel{\text{ind}}{\sim} \text{Poi}(np_i)$, $i=1, \dots, k$, $\sum_{i=1}^k p_i = 1$.

Aim to estimate

$$\sum_{i=1}^k 1(y_i = t) (1 - e^{-np_i}) \quad k=0, 1, \dots$$

↑ species appearing exactly t times in original sample
↑ prob. of observing i in the new sample

EB estimation of p_i^l

$$\begin{aligned} \mathbb{E}_G[p^l | y] &= \frac{\mathbb{E}_G[p^l \cdot \frac{e^{-np} (np)^y}{y!}]}{\mathbb{E}_G[\frac{e^{-np} (np)^y}{y!}]} = \frac{(y+1) \cdots (y+l)}{n^l} \frac{f_G(y+l)}{f_G(y)} \\ &\approx \frac{(y+l)!}{y! n^l} \cdot \frac{N_{y+l}}{N_y} \end{aligned}$$

Final EB estimator.

$$\begin{aligned} \sum_{i=1}^k 1(y_i = t) (1 - e^{-np_i}) &= \sum_{i=1}^k 1(y_i = t) \sum_{\ell=1}^{\infty} (-1)^{\ell+1} \frac{(np_i)^\ell}{\ell!} \\ &\approx \sum_{i=1}^k 1(y_i = t) \sum_{\ell=1}^{\infty} (-1)^{\ell+1} \left(\frac{y_i + \ell}{\ell}\right) \left(\frac{m}{n}\right)^\ell \frac{N_{y_i + \ell}}{N_{y_i}} \\ &= \sum_{\ell=1}^{\infty} (-1)^{\ell+1} \binom{t+\ell}{\ell} \left(\frac{m}{n}\right)^\ell N_{\ell+t} \end{aligned}$$

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

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

Gaussian location model: Tweedie's formula

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

$$\begin{aligned}\mathbb{E}_G[\theta | y] &= \frac{\mathbb{E}_G[\theta \varphi(y-\theta)]}{\mathbb{E}_G[\varphi(y-\theta)]} \\ &= y - \frac{\mathbb{E}_G[(y-\theta) \varphi(y-\theta)]}{\mathbb{E}_G[\varphi(y-\theta)]} \quad \begin{array}{l} \nearrow \mathbb{E}_G[-\varphi'(y-\theta)] \\ = -f'_G(y) \end{array} \\ &= y + \frac{d}{dy} \log f_G(y) \quad \xrightarrow{\quad} f_G(y) \\ &\quad \text{(Tweedie's formula)}\end{aligned}$$

An estimator without knowledge of G : estimate the marginal distribution $\hat{f}(y)$ of y based on y_1, \dots, 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, \dots, y_k
- f-modeling is simple, yet could have a large variance
- g-modeling: estimate G based on y_1, \dots, 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)

→ then use \hat{G} in place of G .

- some evidence that g-modeling gives better estimation performances.