Empirical Bayes, James-Stein

10/3/2023

Outline

1) Empirical Bayes
2) James-Stein Paradox
3) Stein's Lemma
4) Stein's unbiased risk estimator (SURE)

Empirical Bayes

Common situation in hierarchical Bayes models:

$$\zeta \sim \lambda(\zeta)$$ ← one draw ⇒ hard to justify prior; lots of info ⇒ prior doesn't matter

$$\theta_i | \zeta \overset{iid}{\sim} \pi_\zeta(\theta)$$ ← only X_i informative ⇒ prior helps; many draws ⇒ can check fit

$$X_i | \zeta, \theta \overset{ind.}{\sim} p_{\theta_i}(x) \qquad i = 1, \ldots, d$$

Hybrid approach: treat ζ as fixed

- Estimate ζ based on observed data
- Plug in $\hat{\zeta}$ as though known

Ex.

$$\theta_i \sim N(0, \tau^2) \qquad \tau^2 \text{ fixed, unknown}$$

$$X_i | \theta \sim N(\theta_i, 1) \qquad i = 1, \ldots, d$$

Bayes estimator if we knew τ^2 is

$$\delta_i(X) = (1 - \zeta) X_i, \quad \zeta = \frac{1}{1+\tau^2}$$

To estimate ζ, use $X \sim N_d(0, \zeta^{-1} I_d) = \left(\frac{\zeta}{2\pi}\right)^{d/2} e^{-\zeta \|X\|^2/2}$ sufficient

$$\|X\|^2 \sim \zeta^{-1} \chi^2_d \Rightarrow \hat{\zeta}^{-1}_{MLE} = \|X\|^2/d,$$

Plug in: $\delta_i(X) = \left(1 - d/\|X\|^2\right) X_i$

If d large, should be near-optimal

James - Stein Estimator

James & Stein proposed instead $(d \geq 3)$:

$$\delta_{JS,i}(X) = \left(1 - \frac{d-2}{\|X\|^2}\right) x_i$$

Emp Bayes Motivation: $\frac{d-2}{\|X\|^2}$ is UMVUE of ζ

Prop: If $Y \sim \chi^2_d \overset{=}{} \text{Gamma}(\frac{d}{2}, 2)$, $n \geq 3$ then

$$\mathbb{E}\left[\frac{1}{Y}\right] = \frac{1}{d-2}$$

Proof: $\mathbb{E}\left[\frac{1}{Y}\right] = \int_0^\infty \frac{1}{y} \frac{1}{2^{d/2}\Gamma(\frac{d}{2})} \cdot y^{d/2 - 1} e^{-y/2} \, dy$

$$= \frac{2^{(d-2)/2}\Gamma(\frac{d-2}{2})}{2^{d/2}\Gamma(d/2)} \int_0^\infty \underbrace{\frac{1}{2^{(d-2)/2}\Gamma(\frac{d-2}{2})} y^{\frac{(d-2)}{2} - 1} e^{-y/2}}_{\chi^2_{d-2} \text{ density}} dy$$

Now, use $\Gamma(x) = (x-1)\Gamma(x-1) \quad \forall x > 0$

$$\ldots = \frac{1}{2} \cdot \frac{1}{(d-2)/2} = \frac{1}{d-2}$$

$\boxtimes$

$$\zeta \|X\|^2 \sim \chi^2_d \Rightarrow \zeta^{-1} E_\zeta\left[\frac{1}{\|X\|^2}\right] = \frac{1}{d-2}$$

$$\Rightarrow \hat{\zeta} = \frac{d-2}{\|X\|^2} \quad \text{UMVUE}$$

James - Stein Paradox

Back to non-Bayesian Gaussian seq. model:

$$X_i \overset{iid}{\sim} N_d(\theta, \sigma^2 I_d), \quad \theta \in \mathbb{R}^d \text{ (fixed)}, \quad \sigma^2 > 0 \text{ known}$$

$i = 1, \dots, n$

Shocking result of James & Stein (1956):

For $d \geq 3$, the sample mean $\bar{X} = \frac{1}{n} \sum X_i$

is <u>inadmissible</u> as an estimator of θ

under squared error loss:

$$\text{For } \delta_{JS}(X) = \left(1 - \frac{(d-2)\sigma^2/n}{\|\bar{X}\|^2}\right)\bar{X}$$

$$MSE(\theta, \delta_{JS}) < MSE(\theta, \bar{x}) \quad \forall \theta \in \mathbb{R}^d \quad (!!!)$$

$\bar{X}$ is UMVU, Minimax, objective Bayes, ...

Note: Might as well take $n = 1$ (Suff. reduction) $\Rightarrow \left(1 - \frac{d-2}{\|X\|^2}\right)X$

Note this result holds <u>without</u> assumption of Bayes model on θ: true for $\theta = (500, -10^{10}, 4)$

Nothing special about 0: for any $\theta_0 \in \mathbb{R}^d$

$$\delta(X) = \theta_0 + \left(1 - \frac{d-2}{\|X - \theta_0\|^2}\right)(X - \theta_0)$$

also dominates X

Deep implication: shrinkage makes sense even without Bayes justification.

Linear shrinkage w/o Bayesian assumptions

Gaussian seq. model: $X \sim N_d(\theta, I_d)$, fixed $\theta \in \mathbb{R}^d$

Let $\delta_\zeta(X) = (1-\zeta)X$, ζ is <u>tuning parameter</u>

$$R(\theta; \delta_\zeta) = \|\theta - \mathbb{E}\delta_\zeta(X)\|^2 + \sum_i \operatorname{Var}((1-\zeta)X_i)$$

(MSE)

$$= \underbrace{\zeta^2 \|\theta\|^2}_{\text{bias}^2} + \underbrace{d(1-\zeta)^2}_{\text{variance}}$$

What is optimal ζ?

$$\frac{d}{d\zeta} R(\theta; \delta_\zeta) = 2\zeta\|\theta\|^2 - 2(1-\zeta)d$$

$\Rightarrow$ minimizer $= \zeta^*(\theta) = \dfrac{d}{d + \|\theta\|^2} = \dfrac{1}{1 + \|\theta\|^2/d}$

ζ^* always > 0, but $\to 0$ as $\theta \to \infty$

What if we estimate $\zeta^*(\theta)$?

How does adaptivity of $\hat{\zeta}^*(X)$ affect MSE?

Stein's Lemma

Useful tool for computing / estimating risk in Gaussian estimation problems

Theorem (Stein's Lemma, univariate):

Suppose $X \sim N(\theta, \sigma^2)$

$h(x) : \mathbb{R} \to \mathbb{R}$ differentiable, $\mathbb{E}|\dot{h}(x)| < \infty$

Then $\underbrace{\mathbb{E}\left[(X-\theta)\, h(X)\right]}_{= \operatorname{Cov}(X,\, h(X))} = \sigma^2\, \mathbb{E}\left[\dot{h}(X)\right]$

Proof Note we can assume wlog $h(0) = 0$ (why?)

First assume $\theta = 0$, $\sigma^2 = 1$:

Note $\mathbb{E}\left[X\, h(X)\right] = \int_0^\infty x\, h(x) \phi(x)\, dx + \int_{-\infty}^0 x\, h(x) \phi(x)\, dx$

$$
\begin{aligned}
\int_0^\infty x\, h(x) \phi(x)\, dx &= \int_0^\infty x \left[\int_0^x \dot{h}(y)\, dy\right] \phi(x)\, dx \\
&= \int_0^\infty \int_0^\infty 1\{y < x\}\, x\, \dot{h}(y)\, \phi(x)\, dx\, dy \\
&= \int_0^\infty \dot{h}(y) \left[\int_y^\infty x\, \phi(x)\, dx\right] dy \\
&= \int_0^\infty \dot{h}(y)\, \phi(y)\, dy
\end{aligned}
$$

In the last step we have used:

$$\frac{d}{dx}\left[\frac{1}{\sqrt{2\pi}}\, e^{-x^2/2}\right] = x \cdot \frac{1}{\sqrt{2\pi}}\, e^{-x^2/2}$$

Similar argument shows $\int_{-\infty}^{0} x\, h(x)\, \phi(x)\, dx = \int_{-\infty}^{0} \dot{h}(x)\, \phi(x)\, dx$

$\Rightarrow$ Result holds for $\theta = 0$, $\sigma^2 = 1$

General $\theta = 0$, $\sigma^2 = 1$:

write $X = \theta + \sigma Z$, $\quad Z \sim N(0,1)$

$$\begin{aligned} \mathbb{E}\left[(X-\theta)\, h(X)\right] &= \sigma\, \mathbb{E}\left[Z\, h(\theta + \sigma Z)\right] \\ &= \sigma^2\, \mathbb{E}\left[\dot{h}(\theta + \sigma Z)\right] \\ &= \sigma^2\, \mathbb{E}\left[\dot{h}(X)\right] \end{aligned}$$

Multivariate Stein's Lemma

Def $h : \mathbb{R}^d \to \mathbb{R}^d$, $Dh \in \mathbb{R}^{d \times d}$

$$(Dh(x))_{ij} = \frac{\partial h_i}{\partial x_j}(x)$$

Def (Frobenius norm): $A \in \mathbb{R}^{d \times d}$

$$\|A\|_F = \left(\sum_{i,j} A_{ij}^2 \right)^{1/2}$$

Theorem (Stein's Lemma, Multivariate):

$$X \sim N_d(\theta, \sigma^2 I_d) \qquad \theta \in \mathbb{R}^d$$

$h : \mathbb{R}^d \to \mathbb{R}^d$ diff'able, $\mathbb{E}\|Dh(X)\|_F < \infty$

Then
$$\mathbb{E}\left[(X-\theta)' h(X)\right] = \sigma^2 \mathbb{E}\, \mathrm{tr}(Dh(X))$$
$$= \sigma^2 \sum_i \mathbb{E} \frac{\partial h_i}{\partial x_i}(X)$$

Proof

$$\mathbb{E}\left[(X_i - \theta_i) h_i(X)\right] = \mathbb{E}\left[\mathbb{E}\left[(X_i - \theta_i) h_i(X) \mid X_{-i}\right]\right]$$
$$= \mathbb{E}\left[\mathbb{E}\left[\sigma^2 \frac{\partial h_i}{\partial x_i}(X) \mid X_i\right]\right]$$
$$= \sigma^2 \mathbb{E} \frac{\partial h_i}{\partial x_i}(X) \qquad \boxtimes$$

Stein's Unbiased Risk Estimator (SURE)

Can use Stein's Lemma to get unbiased estimator of the MSE of any $\delta(X)$:

apply Stein's Lemma with $h(X) = X - \delta(X)$

Assume $\sigma^2 = 1$:

$$R(\theta;\delta) = \mathbb{E}_\theta\left[\|X - \theta - h(X)\|^2\right]$$

$$= \mathbb{E}_\theta\|X-\theta\|^2 + \mathbb{E}_\theta\|h(X)\|^2 - 2\mathbb{E}_\theta\left[(X-\theta)'h(X)\right]$$

$$= d + \mathbb{E}_\theta\|h(X)\|^2 - 2\mathbb{E}_\theta \operatorname{tr}(Dh(X))$$

$$\Rightarrow \hat{R}(X) = d + \|h(X)\|^2 - 2\operatorname{tr}(Dh(X))$$

is <u>unbiased</u> for the MSE (estimator b/c only dep. on X)

Can also compute MSE via $R = \mathbb{E}_\theta \hat{R}$

<u>Ex</u>: $\delta(X) = X \Rightarrow h(X) = 0$, $Dh'(X) = 0$

$\hat{R} = d = R(\theta;\delta) \quad \forall\theta$

<u>Ex</u>: $\delta_\zeta(X) = (1-\zeta)X$ for fixed ζ

$\Rightarrow h(X) = \zeta X$, $Dh = \zeta I_d$

$\hat{R} = d + \zeta^2\|X\|^2 - 2\zeta d = (1-2\zeta)d + \zeta^2\|X\|^2$

Risk of James-Stein

$$\delta^{JS}(X) = \left(1 - \frac{d-2}{\|X\|^2}\right) X$$

$$\Rightarrow h(X) = \frac{d-2}{\|X\|^2} X$$

$$\|h(X)\|^2 = (d-2)^2 / \|X\|^2$$

$$\frac{\partial h_i}{\partial x_i}(X) = \frac{\partial}{\partial x_i} \frac{(d-2) X_i}{\sum_j X_j^2}$$

$$= (d-2) \frac{\|X\|^2 - 2X_i^2}{\|X\|^4}$$

$$\Rightarrow tr(Dh(X)) = \frac{d-2}{\|X\|^4} \sum_i \|X\|^2 - 2X_i^2$$

$$= (d-2)^2 / \|X\|^2$$

$$\hat{R} = d + \frac{(d-2)^2}{\|X\|^2} - 2 \frac{(d-2)^2}{\|X\|^2}$$

$$= d - \frac{(d-2)^2}{\|X\|^2}$$

$$R(\theta; \delta_{JS}) = d - \overbrace{(d-2)^2 \, \mathbb{E}_\theta\left[\frac{1}{\|X\|^2}\right]}^{>0}$$

$$< d$$

$$= R(\theta; X)$$

If $\theta = 0$ then $\mathbb{E}_\theta\left[\frac{1}{\|X\|^2}\right] = d-2$

$$\Rightarrow R(\theta;\delta_{JS}) = d - (d-2) = 2$$

Possibly $\ll d$!

$\theta \to \infty$ then $\mathbb{E}_\theta\left[\frac{1}{\|X\|^2}\right] \approx \frac{1}{\|\theta\|^2}$

$$\Rightarrow R(\theta;\delta_{JS}) \approx d - \frac{(d-2)^2}{\|\theta\|^2}$$

$$\to d$$

Smaller and smaller advantage but always better.

Note $\delta_{JS}(X)$ also inadmissible:

$\delta_{JS+}(X) = \left(1 - \frac{d-2}{\|X\|^2}\right)_+ X$ is strictly better

Practically more useful version:

$$\delta_{JS,2}(X) = \bar{X} + \left(1 - \frac{d-3}{\|X-\bar{X}1_d\|^2}\right)(X - \bar{X}1_d)$$

Dominates $\delta(X) = X$ for $d \geq 4$

Taken to logical extreme, suggestion seems dumb: should everyone @ Berkeley pool their estimates?

Note $\mathbb{E}\|\cdot\|^2$ is improved, but $\mathbb{E}(X_i - \theta_i)^2$ may get worse for individual coordinates.