

Empirical Bayes, James-Stein

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:

$$\xi \sim \lambda(\xi) \quad \leftarrow \text{one draw} \Rightarrow \text{hard to justify prior}$$

lots of info \Rightarrow prior doesn't matter

$$\theta_i | \xi \stackrel{\text{iid}}{\sim} \pi_\xi(\theta) \quad \leftarrow \text{only } X_i \text{ informative} \Rightarrow \text{prior helps}$$

many draws \Rightarrow can check fit

$$X_i | \xi, \theta \stackrel{\text{ind.}}{\sim} p_{\theta_i}(x) \quad i = 1, \dots, d$$

Hybrid approach: treat ξ as fixed

- Estimate ξ based on observed data
- Plug in ξ as though known

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

$$X_i | \theta \sim N(\theta_i, 1) \quad i = 1, \dots, d$$

Bayes estimator if we knew τ^2 is

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

sufficient

To estimate ξ , use $X \sim N_d(0, \xi^{-1} \mathbf{I}_d) = \left(\frac{\xi}{2\pi}\right)^{d/2} e^{-\xi \|X\|^2/2}$

$$\|X\|^2 \sim \xi^{-1} \chi_d^2 \Rightarrow \hat{\xi}_{MLE}^{-1} = d / \|X\|^2$$

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

If d large, should be near-optimal

James - Stein Estimator

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

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

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

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

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

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

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

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

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

□

$$\begin{aligned} \zeta \|X\|^2 &\sim \chi_d^2 \Rightarrow \zeta^{-1} \mathbb{E}_\zeta \left[\frac{1}{\|X\|^2} \right] = \frac{1}{d-2} \\ \Rightarrow \hat{\zeta} &= \frac{d-2}{\|X\|^2} \quad \text{UMVUE} \end{aligned}$$

James - Stein Paradox

Back to non-Bayesian Gaussian seq. model:

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

$i = 1, \dots, n$

Shocking result of James & Stein (1956):

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

is inadmissible as an estimator of θ

under squared error loss:

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

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

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

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

Note this result holds without assumption of

Bayes model on θ : true for $\theta = (500, -10^{10}, 4)$

Nothing special about θ : 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 tuning parameter

$$\begin{aligned} R(\theta; \delta_\zeta) &= \|\theta - \mathbb{E} \delta_\zeta(X)\|^2 + \sum_i \text{Var}((1-\zeta)X_i) \\ \uparrow \\ \text{(MSE)} \quad &= \underbrace{\zeta^2 \|\theta\|^2}_{\text{bias}^2} + \underbrace{d(1-\zeta)^2}_{\text{variance}} \end{aligned}$$

What is optimal ζ ?

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

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

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

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

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

$$\text{Then } \mathbb{E}[(X - \theta)h(X)] = \sigma^2 \mathbb{E}[h'(X)]$$

$\overset{=}{\text{Cov}}(X, h(X))$

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

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

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

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

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 $\int_{-\infty}^0 x h(x) \phi(x) dx = \int_{-\infty}^0 h(x) \phi(x) dx$

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

General θ, σ^2 :

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

$$\begin{aligned} \mathbb{E}[(x - \theta) h(x)] &= \sigma \mathbb{E}[Z h(\theta + \sigma Z)] \\ &= \sigma^2 \mathbb{E}[h'(\theta + \sigma Z)] \\ &= \sigma^2 \mathbb{E}[h'(x)] \end{aligned}$$

Multivariate Stein's Lemma

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

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

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):

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

$$h: \mathbb{R}^d \rightarrow \mathbb{R}^d \quad \text{diff'able}, \quad \mathbb{E} \|Dh(x)\|_F < \infty$$

$$\begin{aligned} \text{Then } \mathbb{E} \left[(X - \theta)' h(X) \right] &= \sigma^2 \mathbb{E} \operatorname{tr}(Dh(x)) \\ &= \sigma^2 \sum_i \mathbb{E} \frac{\partial h_i}{\partial x_i}(x) \end{aligned}$$

Proof

$$\begin{aligned} \mathbb{E} \left[(X_i - \theta_i) h_i(X) \right] &= \mathbb{E} \left[\mathbb{E} \left[(X_i - \theta_i) h_i(X) \mid X_{-i} \right] \right] \\ &= \mathbb{E} \left[\mathbb{E} \left[\sigma^2 \frac{\partial h_i}{\partial x_i}(X) \mid X_i \right] \right] \\ &= \sigma^2 \mathbb{E} \frac{\partial h_i}{\partial x_i}(X) \quad \square \end{aligned}$$

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 $\sigma^2 = 1$:

$$\begin{aligned} R(\theta; \delta) &= \mathbb{E}_{\theta} [\|X - \theta - h(x)\|^2] \\ &= \mathbb{E}_{\theta} \|X - \theta\|^2 + \mathbb{E}_{\theta} \|h(x)\|^2 - 2 \mathbb{E}_{\theta} [(X - \theta)' h(x)] \\ &= d + \mathbb{E}_{\theta} \|h(x)\|^2 - 2 \mathbb{E}_{\theta} \text{tr}(Dh(x)) \end{aligned}$$

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

is unbiased for the MSE (estimator b/c only dep. on x)

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

Ex: $\delta(x) = X \Rightarrow h(x) = 0, Dh'(x) = 0$
 $\hat{R} = d = R(\theta; \delta) \quad \forall \theta$

Ex: $\delta_{\zeta}(x) = (1 - \zeta)X$ for fixed ζ

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

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

Risk of James-Stein

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

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

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

$$\frac{\partial h_i}{\partial x_i}(x) = \frac{\partial}{\partial x_i} \frac{(d-2)x_i}{\sum_j x_j^2}$$

$$= (d-2) \frac{\|x\|^2 - 2x_i^2}{\|x\|^4}$$

$$\Rightarrow \text{tr}(Dh(x)) = \frac{d-2}{\|x\|^4} \sum_i \|x\|^2 - 2x_i^2$$

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

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

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

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

$$< d$$

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

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 $\ll d$!

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

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

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:

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

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.