Statistics 210A Lecture 12 Notes

Daniel Raban

October 5, 2021

1 Analysis of the James-Stein Estimator

1.1 Recap: introduction of the James-Stein estimator

Last time, we discussed the Bayesian model with prior $\Theta_i \overset{\text{iid}}{\sim} N(0, \tau^2)$ and $X_i \mid \Theta \overset{\text{iid}}{\sim} N(\Theta_i, 1)$. This gave $\mathbb{E}[\Theta_i \mid X] = (1 - \zeta)X_i$, where $\zeta = \frac{1}{1+\tau^2}$. The **Hierarchical Bayes** approach was to put a prior on ζ, so the posterior mean is

$$\mathbb{E}[\Theta_i \mid X] = (1 - \mathbb{E}[\zeta \mid X])X_i.$$

The **Empirical Bayes** approach was to estimate ζ by an estimator $\widehat{\zeta}(X)$ to get the posterior mean

$$\widehat{E}[\Theta_0 \mid X] = (1 - \widehat{\zeta})X_i.$$

This brought us to the **James-Stein estimator**

$$\delta_i^{\text{JS}}(X) = \left(1 - \frac{d-2}{\|X\|^2}\right) X_i.$$

This estimator dominates $\delta(X) = X$, even in the Gaussian sequence model with no Bayesian assumption. In particular,

$$\text{MSE}(\theta; \delta_{\text{JS}}) < \text{MSE}(\theta; X) \qquad \forall \theta \in \mathbb{R}^d.$$

1.2 Linear shrinkage without Bayes assumptions

Suppose $X_i \overset{\text{iid}}{N}(\theta_i, 1)$ with fixed $\theta_1, \dots, \theta_d \in \mathbb{E}$. Consider the estimator $\delta_\zeta(X) = (1 - \zeta)X$ for a fixed parameter ζ. Then

$$\text{MSE}(\theta; \delta_\zeta) = \zeta^2 \|\theta\|^2 + (1 - \zeta)^2 d$$

Take the derivative over ζ to optimize:

$$0 = \frac{d}{d\zeta} \text{MSE}(\theta; \delta_\zeta) = 2\zeta \|\theta\|^2 + 2(1 - \zeta)d.$$

Solving this gives $\zeta^* = \frac{d}{d+\|\theta\|^2}$. Notice that this is always positive, so the optimal shrinkage is never 0. We can't use this value of ζ because it depends on θ. However, the James-Stein estimator is basically an adaptive ζ.

What if we try to estimate $\|\theta\|$ by using $\|X\|^2$? We have $\frac{1}{d}\|X\|^2 = \frac{1}{d}\sum_{i=1}^d X_i^2$, where each term has mean $\theta_i^2 + 1$ and variance $2 + 4\theta_i^2$. So

$$\frac{1}{d}\|X\|^2 \sim \left(\frac{d + \|\theta\|^2}{d}, \frac{2d + 4\|\theta\|^2}{d^2}\right)$$

This is nice because

$$\frac{\text{standar deviation}}{\text{mean}} = 2\frac{\sqrt{d/2 + \|\theta\|^2}}{d + \|\theta\|^2} \xrightarrow{d\to\infty} 0,$$

so this should exhibit concentration about the mean for large d.

1.3 Stein's lemma

Theorem 1.1 (Stein's lemma, univariate)**.** *Suppose $X \sim N(\theta, \sigma^2)$, and let $h : \mathbb{R} \to \mathbb{R}$ be differentiable with $\mathbb{E}[|\dot{h}(X)|] < \infty$. Then*

$$\mathbb{E}[(X - \theta)h(X)] = \sigma^2\,\mathbb{E}[\dot{h}(X)].$$

Proof. Assume without loss of generality that $h(0) = 0$. First assume $\theta = 0$ and $\sigma^2 = 1$ for simplicity. Note that

$$\mathbb{E}[Xh(X)] = \int_0^\infty xh(x)\phi(x)\,dx + \int_{-\infty}^0 xh(x)\phi(x)\,dx.$$

Dealing with these separately,

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

Using the fact that $\frac{d\phi}{dx} = \frac{d}{dx}\frac{1}{\sqrt{2\pi}}e^{-x^2/2} = -x\frac{1}{\sqrt{2\pi}}e^{-x^2/2} = -x\phi(x)$,

$$= \int_0^\infty \dot{h}(y)\phi(y)\,dy.$$

Similarly,

$$\int_{-\infty}^{0} xh(x)\phi(x)\,dx = \int_{-\infty}^{0} \dot{h}(y)\phi(y)\,dy.$$

Putting these two together gives

$$\mathbb{E}[Xh(X)] = \int_{-\infty}^{\infty} xh(x)\phi(x)\,dx = \int_{-\infty}^{\infty} \dot{h}(y)\phi(y)\,dy = \mathbb{E}[\dot{h}(X)].$$

For a general θ, σ^2, write $X = \theta + \sigma Z$, where $Z \sim N(0,1)$. Then

$$\mathbb{E}[(X-\theta)h(X)] = \sigma\,\mathbb{E}[Zh(\theta+\sigma Z)]$$

Applying the result for $g(Z) = h(\theta + \sigma Z)$ and using the chain rule,

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

□

We want to extend this to the multivariate case. Here is what we replace $\dot{h}$ with:

Definition 1.1. If $h : \mathbb{R}^d \to \mathbb{R}^d$, then the **derivative** is the matrix $Dh \in \mathbb{R}^{d\times d}$ given by

$$[Dh(x)]_{i,j} = \frac{\partial h_i}{\partial x_j}(x).$$

Definition 1.2. The **Frobenius norm** of a matrix $A \in \mathbb{R}^{d\times d}$ is

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

Theorem 1.2 (Stein's lemma, multivariate)**.** *Suppose $X \sim N_d(\theta, \sigma^2 I_d)$ with $\theta \in \mathbb{R}^d$, and let $h : \mathbb{R}^d \to \mathbb{R}^d$ be differentiable with $\mathbb{E}[\|Dh\|_F] < \infty$. Then*

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

Proof. The i-th term on the left hand side is

$$\mathbb{E}[(X_i-\theta_i)h_i(X)] = \mathbb{E}[\mathbb{E}[(X_i-\theta_i)h_i(X) \mid X_{\setminus i}]]$$

Conditionally on $X_{\setminus i}$, $X_i \sim N(\theta_i, \sigma^2)$, and $h_i(X)$ is just a function of X_i. So we can apply the univariate lemma.

$$\begin{aligned} &= \mathbb{E}\left[\sigma^2\,\mathbb{E}\left[\frac{\partial h_i}{\partial x_i}(X) \mid X_{\setminus i}\right]\right] \\ &= \sigma^2\,\mathbb{E}\left[\frac{\partial h_i}{\partial x_i}(X)\right]. \end{aligned}$$

Now sum over i on both sides to get the result. □

Remark 1.1. This differentiability condition can be relaxed somewhat.

1.4 Stein's unbiased risk estimator (SURE)

For our estimator $\delta(X)$, apply Stein's lemma on $h(X) = X - \delta(X)$. Assuming $\sigma^2 > 0$ is known,

$$\begin{aligned}\mathrm{MSE}(\theta;\delta) &= \mathbb{E}_\theta[\|X-\theta-h(X)\|^2] \\ &= \mathbb{E}_\theta[\|X-\theta\|^2] + \mathbb{E}_\theta[\|h(X)\|^2] - 2\,\mathbb{E}_\theta[(X-\theta)^\top h(X)]\end{aligned}$$

Since $\frac{1}{\sigma}(X-\theta) \sim \chi^2_d$,

$$= \sigma^2 d + \mathbb{E}_\theta[\|h(X)\|^2] - 2\sigma^2\,\mathbb{E}_\theta[\mathrm{tr}(Dh(X))].$$

So we get the estimator

$$\widehat{R} = \sigma^2 d + \|h(X)\|^2 - 2\sigma^2\,\mathrm{tr}(Dh(X)).$$

Example 1.1. If we take $\delta(X) = X$, then $h(X) = 0$, so $Dh(X) = 0$. In this case, we get

$$\widehat{R} = d\theta^2 \qquad \forall\theta.$$

Example 1.2. Now look at $\delta_\zeta(X) = (1-\zeta)X$, and let $h(X) = \zeta X$, so $Dh(X) = \zeta I_d$. Then

$$\widehat{R} = \sigma^2 d + \zeta^2\|X\|^2 - 2\sigma^2\zeta d.$$

1.5 MSE of the James-Stein estimator

We will take $\sigma^2 = 1$ for simplicity. We have

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

so

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

Then

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

and

$$\frac{\partial h_i}{\partial x_i} = \frac{\partial}{\partial x_i}\frac{d-2}{\|x\|^2}x_i = (d-2)\frac{\|X\|^2 - 2X_i^2}{\|X\|^4}.$$

Summing over i tells us that

$$\mathrm{tr}(Dh(X)) = (d-2)\frac{d\|X\|^2 - 2\|X\|^2}{\|X\|^4} = \frac{(d-2)^2}{\|X\|^2}.$$

So Stein's unbiased risk estimator is

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

The risk for the James-Stein estimator is

$$\begin{aligned}\mathrm{MSE}(\theta; \delta_{\mathrm{JS}}) &= \mathbb{E}[\widehat{R}] \\ &= d - \mathbb{E}\left[\frac{(d-2)^2}{\|X\|^2}\right] \\ &= \mathrm{MSE}(\theta; X) - \mathbb{E}\left[\frac{(d-2)^2}{\|X\|^2}\right].\end{aligned}$$

This term on the right is the improvement over X.

If $\theta = 0$,

$$\mathrm{MSE}(\theta; \delta_{\mathrm{JS}}) = d - (d-2) = 2.$$

This is a huge improvement for large d! On the other hand, if $\|\theta\| \to \infty$, then

$$\mathrm{MSE}(\theta; \delta_{\mathrm{JS}}) = d - (d-2) = 2 \to d.$$

Remark 1.2. The James-Stein estimator is inadmissible. Here is an estimator that is better: T

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

This is also inadmissible because of a "smoothed out" version of this estimator.

Remark 1.3. Here is a more practically useful estimator (when $d \geq 4$) when we have a lot of samples that estimate similar θ_i:

$$\delta^{\mathrm{JS}\,2} = \overline{X} + \left(1 - \frac{d-3}{\|X - \overline{X}\mathbf{1}_d\|}\right)(X - \overline{X}\mathbf{1}_d),$$

where $\overline{X}$ is the average value of θ.

Remark 1.4. Should we use the James-Stein estimator in practice?[1] It improves the average risk of the combined problem, but it does not improve the risk of each coordinate individually. So we may not be able to improve our estimation problem by including others' data. If we know more information about each model, it also may not be a good idea to treat them all the same.

[1] Should we go knocking on all the doors of everyone in Berkeley, asking for their samples?