STAT 460/560 Class 17: Stein's lemma, SURE, and James–Stein again

Ben Bloem-Reddy

Reading: Chapter 7.2-3, 7.4, 7.6, [Was06].

Last class, we saw the James–Stein estimator and how by shrinking estimates towards a common value (zero), it incurred some bias but traded it for lower variance, such that the overall risk was lower. Today, we'll acquire some technical tools for analyzing the JS estimator's risk. This set of tools has been used to analyze the properties of other modern statistical methods. (See https://www.stat.cmu.edu/~larry/=sml/stein.pdf for more on that.)

1. Stein's lemma

The starting point is a simple identity that has found numerous uses in probability and statistics.

Lemma 17.1 (Stein)**.** *Let $X \sim \mathcal{N}_k(\mu, \Sigma)$, and suppose that $g\colon \mathbb{R}^k \to \mathbb{R}$ is (weakly[1]) differentiable and $E[\|\nabla g(X)\|_2] < \infty$. Then*

$$E[g(X)(X - \mu)] = \Sigma E[\nabla g(X)] \,. \tag{17.1}$$

One dimension. We'll prove this for $k = 1$, in which case it follows from integration by parts. For higher dimensions, the proof is a little trickier but still surprisingly straightforward. First, assume that $\mu = 0$ and $\sigma^2 = 1$. Let ϕ denote the standard normal PDF. Note that $\phi'(z) = -z\phi(z)$. Then integration by parts yields

$$\begin{aligned} E[g'(X)] &= \int_{-\infty}^{\infty} g'(x)\phi(x) \, dx \\ &= g(x)\phi(x)\Big|_{-\infty}^{\infty} - \int_{-\infty}^{\infty} g(x)\phi'(x) \, dx \\ &= \int_{-\infty}^{\infty} xg(x)\phi(x) \, dx = E[Xg(X)] \,. \end{aligned}$$

Note that the assumption that $E[g'(X)] < \infty$ implies that $g(x)\phi(x) \to 0$ as $x \to \pm\infty$, and therefore $g(x)\phi(x)\big|_{-\infty}^{\infty} = 0$.

For $X \sim \mathcal{N}(\mu, \sigma^2)$, we get $E[(X - \mu)g(X)]/\sigma^2 = E[g'(X)]$ by applying the standard normal result to $Z = (X - \mu)/\sigma$ and to $f(z) = g(\sigma z + \mu)$. □

Note that one immediate application is an estimator of $\mathrm{Cov}(X, g(X)) = E[g'(X)]\sigma^2$ via the gradient of g (and vice versa).

2. Stein's unbiased risk estimate (SURE)

We can use the identity above to obtain an unbiased estimator of the risk for the normal means model. Recall that for the normal means model, we want to estimate a vector of parameters $\theta^n = (\theta_1, \ldots, \theta_n)$

[1] In a nutshell, this means we can apply (multivariate) integration by parts. In one dimension, this applies to $\int_a^b g(x)u'(x)dx$ for all infinitely differentiable functions u with $u(a) = u(b) = 0$.

from observations $X_i \sim \mathcal{N}(\theta_i, 1)$. Suppose we have an estimator $\hat{\theta}^n$. The risk is (using $\|\bullet\|$ to denote the Euclidean norm)

$$\begin{aligned} E[\|\theta^n - \hat{\theta}^n\|^2] &= E[\|\theta^n - X^n + X^n - \hat{\theta}^n\|^2] \\ &= n + E[\|\hat{\theta}^n - X^n\|^2] + 2E[(\theta^n - X^n)^\top (X^n - \hat{\theta}^n)] \\ &= -n + E[\|\hat{\theta}^n - X^n\|^2] + 2\sum_{i=1}^n \mathrm{Cov}(X_i, \hat{\theta}_i) \ . \end{aligned}$$

Now, using Stein's lemma on the covariance term,

$$E[\|\theta^n - \hat{\theta}^n\|^2] = -n + E[\|\hat{\theta}^n - X^n\|^2] + 2\sum_{i=1}^n E\left[\frac{\partial \hat{\theta}_i}{\partial x_i}(X^n)\right] \ . \tag{17.2}$$

Hence, an unbiased estimator for the risk is (see (7.19) of [Was06] for slightly different but equivalent expression) the SURE formula for the normal means model,

$$\hat{R}(\hat{\theta}^n) = -n + \sum_{i=1}^n (\hat{\theta}_i - X_i)^2 + 2\sum_{i=1}^n \frac{\partial \hat{\theta}_i}{\partial x_i}(X^n) \ . \tag{17.3}$$

3. Linear shrinkage estimators

Suppose we consider only **linear** estimators of the form $\hat{\theta}_i = bX_i$. Denote this class of estimators by $\mathcal{L}$. To start, suppose we knew θ^n (i.e., we were *oracles*). What is the minimum-risk value of b?

Activity 17.1. Find the value of b that minimizes the risk, and calculate its risk.

Solution: Using Eq. (17.2),

$$\begin{aligned} R(bX^n, \theta^n) &= -n + (b-1)^2 \sum_{i=1}^n E[X_i^2] + 2nb \\ &= (b-1)^2(\|\theta^n\|^2 + n) + (2b-1)n \ , \end{aligned}$$

which is minimized by

$$b_* = 1 - \frac{n}{n + \|\theta^n\|^2} = \frac{\|\theta^n\|^2}{n + \|\theta^n\|^2} \ , \tag{17.4}$$

which has risk

$$R(b_* X^n, \theta^n) = \frac{n\|\theta^n\|^2}{n + \|\theta^n\|^2} \ .$$

It is straightforward to show that the risk of the linear shrinkage estimator is a convex function of b. Hence, we know that if $\mathcal{L}$ is the class of linear shrinkage estimators,

$$\inf_{\hat{\theta}^n \in \mathcal{L}} R(\hat{\theta}^n, \theta^n) = \frac{n\|\theta^n\|^2}{n + \|\theta^n\|^2} \ ,$$

which is achievable only if we know $\|\theta^n\|^2$ (and hence is called the **oracle risk**).

Since we don't know the risk but we can estimate it, one approach is to minimize the estimated risk to obtain b. (This is conceptually analogous to empirical risk minimization.)

Activity 17.2. Write the SURE formula for the class of linear estimators. What value of b minimizes the SURE formula (call it b^*)? Does it look familiar? What is the risk of the estimator that uses b^*? Find the condition on n so that it is lower than the risk of the MLE. (Recall from last time that the risk of the MLE is n.)

Hint: Use the formula (17.2) to find the risk. You'll get a term involving $E[(\sum_i X_i^2)^{-1}]$. You don't need to evaluate this; just use the fact that it's positive.

Solution: For linear estimators in which b does not depend on the observations, the SURE formula simplifies to

$$\hat{R}(bX^n) = -n + (b-1)^2 \sum_{i=1}^n X_i^2 + 2nb$$

Minimizing this with respect to b yields

$$b^* = 1 - \frac{n}{\sum_{i=1}^n X_i^2} \, .$$

One way to view this is as a plug-in estimator of the oracle shrinkage, b_*, from the previous exercise. This is also very similar to the James–Stein shrinkage, $1 - (n-2)/\sum_{i=1}^n X_i^2$. For large n, the two will be indistinguishable.

We can calculate the risk by taking the expectation of the SURE formula. But be careful to take the derivative of $\hat{\theta}_i$ properly:

$$\frac{\partial \hat{\theta}_i}{\partial x_i}(x^n) = \frac{\partial}{\partial x_i}\left[\left(1 - \frac{n}{\sum_{j=1}^n x_j^2}\right) x_i\right] = 1 - \frac{n}{\sum_{j=1}^n x_j^2} + \frac{2nx_i^2}{\left(\sum_{j=1}^n x_j^2\right)^2} \, ,$$

so that

$$\sum_{i=1}^n \frac{\partial \hat{\theta}_i}{\partial x_i}(X^n) = n - \frac{n(n-2)}{\sum_{j=1}^n X_j^2}$$

Plugging this into the SURE formula, we find

$$\begin{aligned} \hat{R}(b^* X^n) &= -n + \frac{n^2}{\left(\sum_{j=1}^n X_j^2\right)^2} \sum_{i=1}^n X_i^2 + 2n - 2\frac{n(n-2)}{\sum_{j=1}^n X_j^2} \\ &= n - \frac{n(n-4)}{\sum_{j=1}^n X_j^2} \, . \end{aligned}$$

Since the SURE formula is unbiased for the risk, we can find the risk by taking the expectation of $\hat{R}$ (and using the inequality in the hint),

$$E[\hat{R}(b^* X^n)] = n - n(n-4)E[\|X^n\|^{-2}] \, .$$

This is less than n (the risk of the MLE) as long as $n > 4$.

Exercise 17.1. Use the same technique as Activity 17.2 to obtain the following upper bound on the

risk of the James–Stein estimator, $\hat{\theta}^{JS}$:

$$R(\hat{\theta}_n^{JS}, \theta^n) \leq n - \frac{(n-2)^2}{n-2+\|\theta^n\|^2} = 2 + \frac{(n-2)\|\theta^n\|^2}{n-2+\|\theta^n\|^2} \ .$$

To get the upper bounds, you'll need some distributional identities (see [Was06], p. 156), which result in the inequality $E[(\sum_i X_i^2)^{-1}] \geq (n-2+\|\theta^n\|^2)^{-1}$.

Solution: Following the same steps as above, except with $\hat{\theta}^{JS}$, we find that

$$R(\hat{\theta}^{JS}, \theta^n) = n - (n-2)^2 E[\|X^n\|^{-2}] \ .$$

First, we see that this risk is less than the risk of the plug-in estimator from the previous activity by $4nE[\|X^n\|^{-2}]$. Second, using the given inequality from Wasserman,

$$\begin{aligned} R(\hat{\theta}^{JS}, \theta^n) &= n - (n-2)^2 E[\|X^n\|^{-2}] \\ &\leq n - \frac{(n-2)^2}{n-2+\|\theta^n\|^2} \\ &= \frac{n(n-2+\|\theta^n\|^2) - (n-2)^2}{n-2+\|\theta^n\|^2} \\ &= \frac{2(n-2) + n\|\theta^n\|^2}{n-2+\|\theta^n\|^2} \\ &= 2 + \frac{(n-2)\|\theta^n\|^2}{n-2+\|\theta^n\|^2} \ . \end{aligned}$$

Since $\frac{(n-2)\|\theta^n\|^2}{n-2+\|\theta^n\|^2} \leq \frac{n\|\theta^n\|^2}{n+\|\theta^n\|^2}$, we can conclude (using the oracle risk above) that

$$\inf_{\hat{\theta}^n \in \mathcal{L}} R(\hat{\theta}^n, \theta^n) \leq R(\hat{\theta}_n^{JS}, \theta^n) \leq 2 + \inf_{\hat{\theta}^n \in \mathcal{L}} R(\hat{\theta}^n, \theta^n) \ .$$

References

[Was06] L. Wasserman. *All of Nonparametric Statistics*. Springer New York, 2006.