STAT 460/560 Class 18: 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 18.1 (Stein)**.** *Let $X \sim \mathcal{N}_k(\mu, \Sigma)$, and suppose that $g\colon \mathbb{R}^k \to \mathbb{R}^k$ is (weakly) differentiable and $E[g'(X)] < \infty$. Then*

$$E[g(X)(X - \mu)] = \Sigma E[\nabla g(X)] \,. \tag{18.1}$$

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 $(X - \mu)/\sigma$. □

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

2. Stein's unbiased risk estimate (SURE)

We can use the identity above to prove SURE. Recall that for the normal means model, we want to estimate a vector of parameters $\theta^n = (\theta_1, \ldots, \theta_n)$ from observations $Z_i \sim \mathcal{N}(\theta_i, 1)$. Suppose we have an estimator $\hat{\theta}^n$.

The risk is

$$\begin{aligned} E[\|\theta^n - \hat{\theta}^n\|_2^2] &= E[\|\theta^n - Z^n + Z^n - \hat{\theta}^n\|_2^2] \\ &= n + E[\|\hat{\theta}^n - Z^n\|_2^2] + 2E[(\theta^n - Z^n)^\top (Z^n - \hat{\theta}^n)] \\ &= -n + E[\|\hat{\theta}^n - Z^n\|_2^2] + 2\sum_{i=1}^n \mathrm{Cov}(Z_i, \hat{\theta}_i) \, . \end{aligned}$$

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

$$E[\|\theta^n - \hat{\theta}^n\|_2^2] = -n + E[\|\hat{\theta}^n - Z^n\|_2^2] + 2\sum_{i=1}^n E\left[\frac{\partial \hat{\theta}_i}{\partial z_i}(Z^n)\right] \, .$$

Hence, an unbiased estimate of the risk is (note the typo in (7.19) of [Was06])

$$\hat{R}(\hat{\theta}^n) = -n + \sum_{i=1}^n (\hat{\theta}_i - Z_i)^2 + 2\sum_{i=1}^n \frac{\partial \hat{\theta}_i}{\partial z_i}(Z^n) \, . \tag{18.2}$$

3. Linear shrinkage estimators

Suppose we consider only **linear** estimators of the form $\hat{\theta}_i = bZ_i$. 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 18.1. Find the value of b that minimizes the risk, and calculate its risk.

Activity 18.2. Write the SURE formula for the class of linear estimators. What value of b minimizes the SURE formula? Does it look familiar? What is its risk? Show that it is lower than the risk of the MLE. (Recall from last time that the risk of the MLE is n.)

Hint: Using the SURE formula, you'll get a term involving $E[(\sum_i Z_i^2)^{-1}]$. You don't need to evaluate this. By some distributional identities (see [Was06], p. 156), $E[(\sum_i Z_i^2)^{-1}] \geq (n - 2 + \|\theta^n\|_2^2)^{-1}$.

References

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