STAT 460/560 Class 16: The James–Stein estimator, shrinkage, and ridge regression

Ben Bloem-Reddy

Reading: Chapter 7, [EH21]; Chapter 7.1, [Was06].

Today we'll encounter a setting that departs from classical examples and is the starting point for many statistical analyses of non-parametric methods.

1. The normal means model

Suppose we have random variables $X_1, \dots, X_n$, independent and with

$$X_i \sim \mathcal{N}(\theta_i, 1) \; . \tag{16.1}$$

The goal is to estimate each θ_i. Let $\theta^n = (\theta_1, \dots, \theta_n)$.

The natural place to start is the MLE, which in this case is $\hat{\theta}_i = X_i$. Under squared error loss, the risk is

$$\mathbb{E}_{\theta^n}[\|\hat{\theta}^n - \theta^n\|^2] = \sum_{i=1}^n \mathbb{E}_{\theta^n}[(X_i - \theta_i)^2] = n \; . \tag{16.2}$$

For pedagogical purposes, let's also assume that

$$\theta_i \sim_{\text{IID}} \mathcal{N}(0, \sigma^2) \; . \tag{16.3}$$

Clearly, the MLE would miss out on this correlation between the θ_i's, so if we think such a correlation is there then our estimation procedure ought to account for it. To that end, consider the posterior distribution

$$\theta_i | X_i \sim \mathcal{N}(\tau X_i, \tau) \; , \quad \text{with} \quad \tau = \frac{\sigma^2}{1 + \sigma^2} \; . \tag{16.4}$$

Activity 16.1. What is the Bayes' estimator, $\hat{\theta}^b_i$, in this case? Show that its risk is

$$\mathbb{E}\left[\sum_{i=1}^n (\hat{\theta}^b_i - \theta_i)^2\right] = \mathbb{E}\left[\mathbb{E}\left[\sum_{i=1}^n (\hat{\theta}^b_i - \theta_i)^2 \mid \theta^n\right]\right] = n\tau \; , \tag{16.5}$$

where the expectation is taken with respect to (X^n, θ^n) according to (16.1), (16.3). How does this compare with the risk of the MLE?

2. The James–Stein estimator

The risk of $\hat{\theta}^b$ is smaller than the risk of $\hat{\theta}$ by a factor of τ. Since we don't know τ but we need it for $\hat{\theta}^b$, we might estimate it by noticing that marginally (i.e., $p(x) = \int p(\theta, x) d\theta$),

$$X_i \sim_{\text{IID}} \mathcal{N}(0, 1 + \sigma^2) \; , \tag{16.6}$$

and as long as $n > 2$, an unbiased estimate of τ is[1]

$$\hat{\tau} = 1 - \frac{n-2}{S_n}, \quad \text{with} \quad S_n = \sum_{i=1}^{n} X_i^2 . \tag{16.7}$$

This gives rise to the **James–Stein estimator**,

$$\hat{\theta}^{JS} = \hat{\tau} X^n , \tag{16.8}$$

which has risk

$$\mathbb{E}\left[\sum_{i=1}^{n} (\hat{\theta}_i^{JS} - \theta_i)^2\right] = n\tau + 2(1-\tau) , \tag{16.9}$$

which is strictly less than n (the risk of the MLE) as long as $n > 2$.

It turns out that the correlation of the θ_i's explicitly induced by the prior in (16.3) is not a necessary ingredient for this phenomenon; it only makes the exposition easier. There can still be benefit to "sharing information" between the observations.

Theorem 16.1 (James–Stein)**.** *Suppose that $X_i \sim \mathcal{N}(\theta_i, 1)$, $i = 1, \ldots, n$, are independent. Then*

$$\mathbb{E}\left[\sum_{i=1}^{n} (\hat{\theta}_i^{JS} - \theta_i)^2\right] \leq 2 + \frac{(n-2)\sum_{i=1}^{n} \theta_i^2}{n - 2 + \sum_{i=1}^{n} \theta_i^2} , \tag{16.10}$$

which is less than the risk of the MLE as long as $n > 2$.

The proof of this requires some (interesting) techniques that we'll look at next time.

The takeaway here is that when estimating many parameters, *shrinking* our estimates has the effect of introducing bias but reducing the variance, so that the overall (average) performance improves. It turns out that shrinking by $\hat{\tau}$ for the normal means problem is asymptotically optimal (in a minimax sense; see [Was06], Ch. 7.6). There is a benefit to sharing "indirect evidence" (see [EH21], Ch. 7.4) about the underlying distribution of θ_i's. Note that a Bayesian approach does this automatically; frequentist shrinkage methods are often described as biasing estimates towards zero, without the use of a prior. The approach we took in introducing the JS estimator—putting a shared prior on the θ_i's, and then using the data to estimate the parameters of the prior distribution—is an example of an **empirical Bayes** approach, which blurs the distinction between frequentist and Bayesian inference.

A potential downside of this is that the estimation of individual θ_i's can suffer for any θ_i that is "extreme."

Activity 16.2. Going back to Activity 16.1, what is the conditional risk for each θ_i, $\mathbb{E}[(\hat{\theta}_i^b - \theta_i)|\theta_i]$? Which realizations of θ_i will have the largest risk? Under what conditions will it be larger than the risk of the MLE, $\mathbb{E}[(X_i - \theta_i)^2] = 1$?

Exercise 16.1. For $n = 10$, sample θ_i and x_i according to (16.3), (16.1), with $\sigma = 5$. Set $\theta_1 = 10$ and $\theta_2 = -10$. Now simulate 1000 datasets X^n of size $n = 10$, sampled with the θ_i's fixed. Estimate the risk $\mathbb{E}_{\theta^n}[\sum_{i=1}^{n}(\hat{\theta}_i - \theta_i)^2]$ for the MLE and the JS estimators, and plot i) θ_i versus the MSE of the MLE; and ii) θ_i versus the MSE of the JS estimators. (You should get something like Fig. 1.)

3. Ridge regression

In addition to the normal means model (which can be thought of as a non-parametric version of one-way ANOVA), shrinkage estimators arise in linear regression with many covariates. Suppose we have k covariates

[1] To see that this is unbiased, observe that $S_n/(1+\sigma^2) \sim \chi^2_n$.

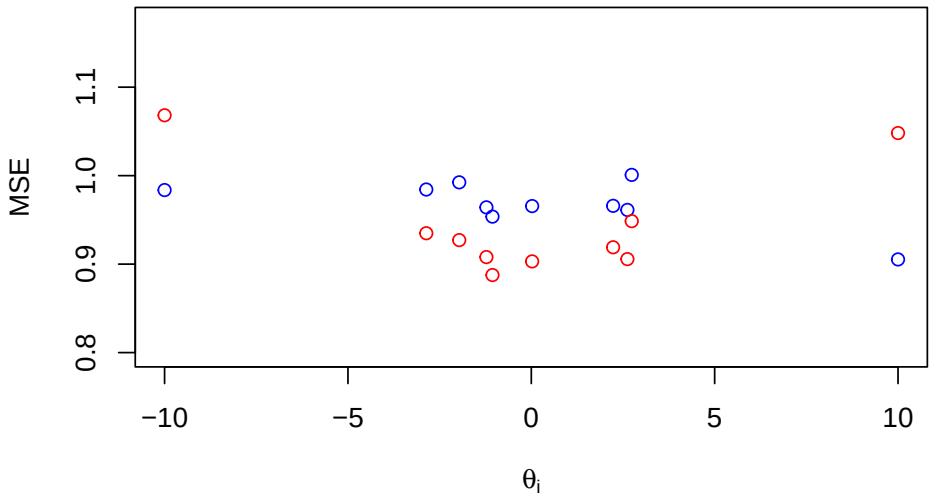

Figure 1: Estimated risk per parameter. Blue is MLE; red is JS.

and there is reason to believe that most of the coefficients are approximately zero. This can be encoded in a couple of ways. First, we might put a prior on the regression coefficients, so that

$$\beta_j \sim \mathcal{N}(0, \lambda^{-1/2}) \,, \tag{16.11}$$

with λ large. Modeling

$$Y_i = X_i\beta + \epsilon_i \,, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2) \,, \tag{16.12}$$

we know that the posterior is

$$f(\beta|y, x) \propto \exp\left(-\frac{1}{2\sigma^2}\|y - x\beta\|^2 - \frac{1}{2}\lambda\|\beta\|^2\right) \,. \tag{16.13}$$

Without going through the algebra, we know that the posterior is a normal distribution, and we can find its mean (corresponding to the Bayes estimator) by maximizing the RHS of (16.13) with respect to β. Doing so yields

$$\hat{\beta}(\lambda) = \left(X^\top X + \lambda\sigma^2\mathbb{I}\right)^{-1} X^\top Y = \left(\left(X^\top X + \lambda\sigma^2\mathbb{I}\right)^{-1} X^\top X\right)\hat{\beta}_{\text{OLS}} \,. \tag{16.14}$$

When we absorb σ^2 into λ, this is called the **ridge regression** estimator.

A second way of deriving the estimator is to ask for the sum of squared residuals to be minimized, subject to a penalty for the squared norm of the vector of coefficients,

$$\hat{\beta}(\lambda) = \arg\min_\beta \|Y - X\beta\|^2 + \frac{1}{2}\lambda\|\beta\|^2 \,. \tag{16.15}$$

This has the same solution as before, but σ^2 has been absorbed into λ. The parameter λ can be viewed as quantifying how strongly the coefficients are being penalized.

How to choose λ? Typically, via cross-validation. The `glmnet` package has an efficient implementation of ridge regression (and more).

References

[EH21] B. Efron and T. Hastie. *Computer Age Statistical Inference*. Cambridge University Press, 2021.
[Was06] L. Wasserman. *All of Nonparametric Statistics*. Springer New York, 2006.