STATS 300C: Theory of Statistics **Spring 2022**

Lecture 16 — March 19, 2022

Lecturer: Prof. Emmanuel Candès *Editor: Parth Nobel, Scribe: Asher Spector*

Warning: *These notes may contain factual and/or typographic errors. They are based on Emmanuel Candès's course from 2018 through 2022, and scribe notes written by Yuchen Hu, Alexandra Chouldechova, Albert Chiu, Gene Katsevich, Andy Tsao, and Yiguang Zhang.*

In this lecture, we discuss Stein's phenomenon and the James-Stein estimator, namely that when estimating the mean of an independent Gaussian vector with equal variance, the MLE is not admissible with respect to the squared-error risk. To prove this, we review Stein's unbiased risk estimate (SURE) and Stein's identity, and we use them to prove that the James-Stein estimator dominates the MLE. Finally, at the end we discuss improvements and extensions to the James-Stein estimator.

From a historical perspective, it is interesting to note that Charles Stein was a phenomenal statistician but also quite a political activist. For example, he was the first Stanford professor to be arrested protesting apartheid (Lawrence, 1985). By today's standards, he did not write that many papers in his lifetime—but each of his papers was very important.

16.1 Stein's phenomena

16.1.1 The James-Stein Estimator

Suppose we observe $X \sim \mathcal{N}(\mu, \sigma^2 I_p)$ and we are interested in estimating $\mu \in \mathbb{R}^p$. Equivalently, this means that we observe *independent* coordinates $X_i \overset{\text{ind}}{\sim} \mathcal{N}(\mu_i, \sigma^2)$. For now, we will assume σ^2 is known, although we will relax this assumption later.

To measure the quality of any estimator $\hat{\mu}$, we use the squared-error risk:

$$R(\hat{\mu}, \mu) = \mathbb{E}_\mu[\|\hat{\mu} - \mu\|_2^2].$$

Since the coordinates of X are independent, an intuitive estimate for μ would be $\hat{\mu} = X$, which has a risk which does not depend on μ:

$$R(\hat{\mu}, \mu) = \mathbb{E}[\|X - \mu\|_2^2] = p\sigma^2.$$

For a long time, this was thought to be the "best" estimate for a multivariate mean. Indeed, this estimator has a number of nice properties, which we list below (although for our purposes it's not crucial to define these optimality properties):

- It is the MLE.
- It is the unique minimum-variance unbiased estimator (UMVUE).
- It is minimax optimal with respect to this risk.
- It is the minimum risk equivariant estimator.

Nonetheless, it turns out that $\hat{\mu}$ is not admissible for $p \geq 3$, although it is admissible for $p = 1, 2$. A result of Stein 1956 hinted at this, and a proof was eventually provided by James and Stein (1961).

To be precise, an estimator $\hat{\mu}$ is **inadmissible** if there exists another estimator $\hat{\mu}'$ such that $R(\hat{\mu}', \mu) \leq R(\hat{\mu}, \mu)$ for all μ, where the inequality is strict for at least one $\mu \in \mathbb{R}^p$. If such a $\hat{\mu}'$ exists, it is said to **uniformly dominate** $\hat{\mu}$, since it improves upon $\hat{\mu}$ everywhere.[1] In particular, the James-Stein estimator (defined below) dominates the MLE.

Definition 1 (James-Stein estimator, from James and Stein (1961))**.** For dimensionality $p > 2$, define

$$\hat{\mu}_{\text{JS}} = \left(1 - \frac{p-2}{\|X\|_2^2}\right) X.$$

Theorem 1 (James and Stein (1961))**.** For $p \geq 3$, $\hat{\mu}_{\text{JS}}$ dominates the MLE everywhere in terms of the squared-error risk. More precisely, for all $\mu \in \mathbb{R}^p$,

$$\mathbb{E}_\mu[\|\hat{\mu}_{\text{JS}} - \mu\|_2^2] < \mathbb{E}_\mu[\|\hat{\mu} - \mu\|_2^2].$$

Note $\hat{\mu}_{\text{JS}}$ is a nonlinear, biased estimator.

This result is perhaps surprising because $\hat{\mu}_{\text{JS}}$ combines independent information. This improves overall estimation accuracy, although it may not necessarily improve accuracy for every single μ_i. It is also important to remark that $\hat{\mu}_{\text{JS}}$ is not admissible either.

16.1.2 Two intuitive interpretations

Below, we review two intuitive interpretations of the James-Stein estimator. Note there is also an empirical Bayes interpretation, which we will discuss in the next lecture.

Stein's Original Argument (1956)

In Stein's original work, he argued that a good estimate should obey $\hat{\mu}_i \approx \mu_i$, and therefore $\|\hat{\mu}\|_2^2 \approx \|\mu\|_2^2$. However, this is not true of the MLE, since

$$\mathbb{E}[\|X\|_2^2] = \sum_{i=1}^{p} \mathbb{E}[X_i^2] = \sum_{i=1}^{p} \mu_i^2 + \sigma^2 = \sigma^2 p + \|\mu\|_2^2.$$

[1] Of course, for a single μ, it is easy to improve on any estimator $\hat{\mu}$, e.g., by taking $\hat{\mu}' = 0$, since this has a risk of zero for $\mu = 0$).

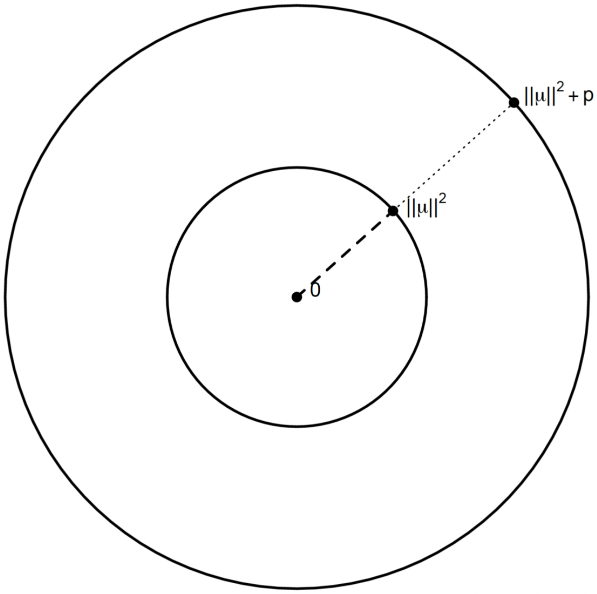

Figure 16.1. A graphical illustration of Stein's original argument. The expected squared norm of the MLE $\|\mu\|_2^2 + \sigma^2 p$ can be much greater than $\|\mu\|_2^2$, especially in high dimensions. A natural solution is to shrink the MLE towards zero.

Since the coordinates of X are independent, informally, we should expect $\|X\|_2^2$ to concentrate around $\sigma^2 p + \|\mu\|_2^2$ in high dimensions. Since the norm of $\|X\|_2^2$ is too large on average, a natural solution is to "shrink" $\|X\|_2^2$ towards zero, which is exactly what $\hat{\mu}_{\text{JS}}$ does. Indeed, in James and Stein (1961), the authors consider a family of estimators indexed by c:

$$\hat{\mu}_c = \left(1 - c\frac{\sigma^2}{\|X\|_2^2}\right) X.$$

They showed that for all $c \in (0, 2(p-2))$, $R(\hat{\mu}_c, u) < R(\hat{\mu}, \mu)$ holds uniformly.

The winner's curse

Another way to interpret Stein's phenomenon is by thinking about the *winner's curse*. In particular, consider the order statistics $X_{(1)}, \ldots, X_{(n)}$ and the order statistics $\mu_{(1)}, \ldots, \mu_{(n)}$ of μ. Jensen's inequality tells us that on average, the largest coordinate of X is larger than the largest coordinate of μ:

$$\mathbb{E}[X_{(n)}] = \mathbb{E}\left[\max_{1 \le i \le p} X_i\right] > \max_{1 \le i \le p} \mathbb{E}[X_i] = \mu_{(n)},$$

and similarly, $\mathbb{E}[X_{(1)}] < \mu_{(1)}$. As a result, it might make sense to "shrink" the order statistics (or equivalently X) towards zero, or as we shall see later, even towards the sample mean $\bar{X}$.

16.2 Stein's Identity and SURE

To prove Theorem 1, we will first review Stein's identity. Along the way, we will review Stein's unbiased risk estimate (SURE), which is of independent interest.

As setup, like before, suppose we observe $X \sim \mathcal{N}(\mu, \sigma^2 I_p)$ with σ^2 known. Given *any* estimator $\hat{\mu}$, we can always decompose $\hat{\mu} = X + g(X)$ for some function g by taking $g(X) = \hat{\mu} - X$. We would like to understand the risk $\mathbb{E}_\mu[\|\hat{\mu} - \mu\|_2^2]$. Although we will allow $\hat{\mu}$ to be very general, we will make two assumptions.

First, we assume $g : \mathbb{R}^p \to \mathbb{R}^p$ is almost differentiable. This means that there exists a function h_i such that

$$g_i(x + z) - g_i(x) = \int_0^1 \langle h_i(x + tz), z \rangle \, dt. \tag{16.1}$$

Usually, we write $h_i = \nabla g_i$, but not always.

Second, we assume that the partial derivatives of $g(X)$ are integrable, i.e.,

$$\mathbb{E}\left[\sum_{i=1}^p |\partial_i g_i(X)|\right] < \infty. \tag{16.2}$$

It is very important to note that if these two conditions are not satisfied, none of the theory developed below will hold or even make sense.

We now state Stein's identity.

Theorem 2. Under assumptions (16.1) and (16.2) and assuming $X \sim \mathcal{N}(\mu, \sigma^2 I_p)$,

$$\mathbb{E}_\mu[\|\hat{\mu} - \mu\|_2^2] = p\sigma^2 + \mathbb{E}\left[\|g(X)\|_2^2 + 2\sigma^2 \sum_{i=1}^p \partial_i g_i(X)\right].$$

Proof. Assume without loss of generality that $\sigma^2 = 1$. Then the risk of $\hat{\mu}$ is

$$\mathbb{E}[\|X + g(X) - \mu\|_2^2] = \mathbb{E}[\|X - \mu\|_2^2] + 2\mathbb{E}[(X - \mu)^T g(X)] + \mathbb{E}[\|g(X)\|_2^2].$$

As a result, we need only show that $\mathbb{E}[(X - \mu)^T g(X)] = \mathbb{E}[\mathrm{div}(g(X))]$. This follows from integration by parts. Let ϕ denote the $\mathcal{N}(0, I_p)$ pdf. Then we write

$$\mathbb{E}[(X_i - \mu_i) g_i(X)] = \int (x_i - \mu_i) g_i(x) \phi(x - \mu) dx.$$

However, note that $\partial_i \phi(x - \mu) = -(x_i - \mu_i)\phi(x - \mu)$. As a result, integrating by parts, we obtain that

$$\sum_{i=1}^p \mathbb{E}[(X_i - \mu_i) g_i(X)] = \int \sum_{i=1}^p \partial_i g_i(x) \phi(x - \mu) dx$$

where the extra term from integration by parts vanishes by the assumption that $\int \sum_{i=1}^p |\partial_i g_i(x)| \phi(x - \mu) dx < \infty$ (and the fundamental theorem of calculus). □

This result is remarkable because $\|g(X)\|_2^2 + 2\sigma^2 \sum_{i=1}^p \partial_i g_i(X)$ does not depend on μ, and thus it is an unbiased estimate of the risk. It is known as Stein's unbiased risk estimate (SURE).

Corollary 1. In the setting of Theorem 2, Stein's unbiased risk estimate (SURE), $p\sigma^2 + \|g(X)\|_2^2 + 2\sigma^2 \sum_{i=1}^p \partial_i g_i(X)$ is an unbiased estimator of the risk $R(\hat{\mu}, \mu)$.

16.3 Applying Stein's identity to James-Stein

In this section, we prove Theorem 1, which is restated for convenience.

Theorem 1. For $p \geq 3$, $\hat{\mu}_{\text{JS}}$ dominates the MLE, $\hat{\mu}$, everywhere in terms of squared-error risk. More precisely, for all $\mu \in \mathbb{R}^p$,

$$\mathbb{E}_\mu[\|\hat{\mu}_{\text{JS}} - \mu\|_2^2] < \mathbb{E}_\mu[\|\hat{\mu} - \mu\|_2^2].$$

Proof. Without loss of generality, take $\sigma^2 = 1$. To allow us to apply Stein's identity, define $g(X) = \hat{\mu}_{\text{JS}} - X = -(p-2)\frac{X}{\|X\|_2^2}$. Stein's identity then tells us that

$$\mathbb{E}[\|\hat{\mu}_{\text{JS}} - \mu\|_2^2] = \mathbb{E}[\|X - \mu\|_2^2] + \mathbb{E}[2\,\text{div}(g(X))] + \mathbb{E}[\|g(X)\|_2^2].$$

As a result, it suffices to show that sum of the last two terms is negative. To do this, observe that

$$\|g(X)\|_2^2 = \frac{(p-2)^2}{\|X\|_2^2}$$

and

$$\partial_i g_i(x) = \partial_i \left\{ -(p-2)\frac{x_i}{\|x\|_2^2} \right\} = -\frac{p-2}{\|x\|_2^2} + \frac{2(p-2)x_i^2}{\|x\|_2^4}$$

which implies that if we sum over i and scale by 2,

$$2\,\text{div}(g(x)) = -2\,\frac{p(p-2)}{\|x\|_2^2} + 2\,\frac{2(p-2)\sum_{i=1}^p x_i^2}{\|x\|_2^4} = -2\,\frac{(p-2)^2}{\|x\|_2^2}.$$

Summing these results, we obtain without any approximations:

$$\mathbb{E}[\|\hat{\mu}_{\text{JS}} - \mu\|_2^2] = p - \mathbb{E}\left[\frac{(p-2)^2}{\|X\|_2^2}\right] < p.$$

□

This result is nice because it quantifies exactly how much $\hat{\mu}_{\text{JS}}$ improves on the MLE. In particular, $\|X\|_2^2 \sim \chi^2_{p,\|\mu\|_2^2}$ follows a noncentral χ^2 distribution, and therefore Jensen's inequality yields

$$\mathbb{E}\left[\frac{1}{\|X\|_2^2}\right] \geq \frac{1}{p-2+\|\mu\|_2^2},$$

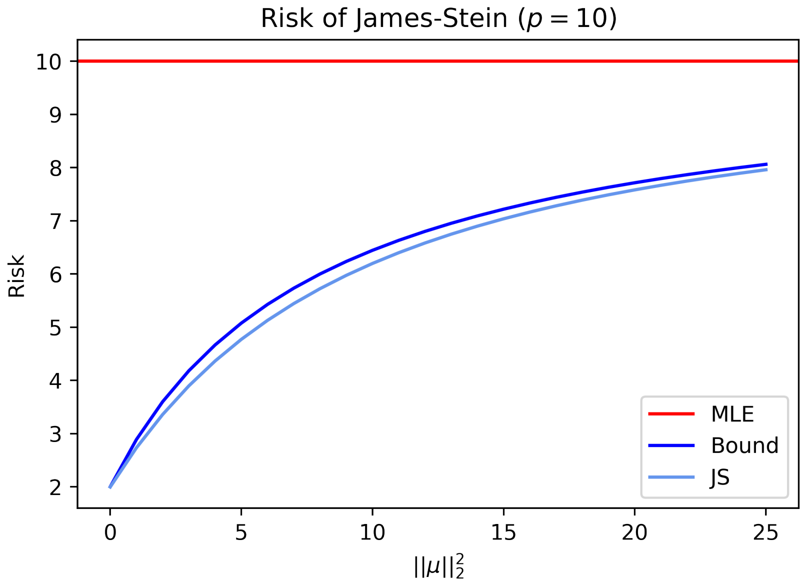

Figure 16.2. The risk of the James-Stein estimator compared to that of the MLE when $p = 10, \sigma^2 = 1$. This plot also compares the bound derived in Equation (16.3) to the exact risk.

with equality if $\mu = 0$. As a result, we can bound the risk of the James-Stein estimator by

$$\mathbb{E}_\mu[\|\hat{\mu}_{\text{JS}} - \mu\|_2^2] \leq p - \frac{p-2}{1 + \frac{\|\mu\|_2^2}{p-2}}. \tag{16.3}$$

For intuition, we consider a few special cases.

- When $\mu = 0$, $R(\hat{\mu}_{\text{JS}}, \mu) = 2$. In high dimensions, this is an *enormous* improvement over the risk of the MLE, which is p.
- If the signal-to-noise ratio is approximately 1 and $\|\mu\|_2^2 \approx p-2$, then $R(\hat{\mu}_{\text{JS}}, \mu) = p/2+1$, which is roughly half the risk of the MLE.
- As $\|\mu\|_2^2 \to \infty$, $R(\hat{\mu}_{\text{JS}}, \mu) \to p$. Thus, as μ gets larger, $\hat{\mu}_{\text{JS}}$ has a risk which approaches that of the MLE.

Figure 16.2 plots the risk of $\hat{\mu}_{\text{MLE}}, \hat{\mu}_{\text{JS}}$ for various values of $\|\mu\|_2^2$. It also shows that the bound in Equation (16.3) is fairly tight, at least in the simulated setting.

This last point motivates the next section, which is: can we improve and extend the James-Stein estimator? What if we believe that the signal size is indeed quite large? Is shrinkage towards zero a good idea in this instance?

16.4 Extensions and improvements to the James-Stein estimator

In this section, we offer a few improvements to the James-Stein estimator. Furthermore, we show that Stein's phenomenon is not unique to the special case where $X \sim \mathcal{N}(\mu, \sigma^2 I_p)$. Indeed, shrinkage is useful in *many* settings: it is a fundamental tool in high-dimensional statistics.

16.4.1 The positive James-Stein estimator

Note that when $\|X\|_2^2 < \sigma^2(p-2)$, the $\left(1 - \frac{\sigma^2(p-2)}{\|X\|_2^2}\right)$ term in $\hat{\mu}_{\text{JS}}$ becomes negative and $\hat{\mu}_{\text{JS}}$ flips the signs of X. This is counterintuitive and increases the risk of $\hat{\mu}_{\text{JS}}$. Instead, we can define the positive James-Stein estimator:

$$\hat{\mu}_{\text{js+}} = \left(1 - \frac{(p-2)\sigma^2}{\|X\|_2^2}\right)_+ X.$$

It turns out that $\hat{\mu}_{\text{js+}}$ uniformly strictly improves the JS estimator, namely $R(\hat{\mu}_{\text{js+}}, \mu) < R(\hat{\mu}_{\text{JS}}, \mu)$ for all $\mu \in \mathbb{R}^p$. However, $\hat{\mu}_{\text{js+}}$ is not admissible either.[2]

16.4.2 Shrinkage towards an arbitrary point

As discussed in the previous section, the advantage of $\hat{\mu}_{\text{JS}}$ over the MLE vanishes asymptotically as $\|\mu\|_2^2 \to \infty$. In settings where we believe a priori that $\|\mu\|_2^2$ is large, we can shrink towards an arbitrary point $\mu_0 \in \mathbb{R}^p$:

$$\hat{\mu}_{\text{JS}}(\mu_0) = \mu_0 + \left(1 - \frac{(p-2)\sigma^2}{\|X - \mu_0\|_2^2}\right)(X - \mu_0).$$

This dominates the MLE for the same reason that $\hat{\mu}_{\text{JS}}$ does when $p > 2$. In particular, note that Gaussians are a location family, so $X - \mu_0 \sim \mathcal{N}(\mu - \mu_0, \sigma^2 I_p)$, and then we can apply Theorem 1 to estimate $\mu - \mu_0$.

Of course, this raises the question: how should one pick μ_0? A good idea is to estimate it from the data, e.g., to pick $\mu_0 = \bar{X} 1_p$, where $1_p \in \mathbb{R}^p$ is the vector of ones. This estimator $\hat{\mu}_{\text{JS}}(\bar{X} 1_p)$ dominates the MLE when $p > 3$ (note that the domination requires one extra dimension because we are already using one "degree of freedom" to estimate $\bar{X}$).

[2] Of course, it is easy to come up with admissible estimators, since any unique Bayes estimator is admissible.

16.4.3 Correlated data

Suppose now that $X \sim \mathcal{N}(\mu, \Sigma)$, where Σ is an arbitrary but known covariance matrix. The MLE is still X, but the appropriate James-Stein estimate (Bock, 1975) is

$$\hat{\mu}_{\text{JS}} = \left(1 - \frac{\tilde{p} - 2}{X^T \Sigma^{-1} X}\right) X,$$

where $\tilde{p}$ is the effective dimension or "numerical rank"

$$\tilde{p} = \frac{\text{Tr}(\Sigma)}{\lambda_{\max}(\Sigma)}.$$

Just like before, if $\tilde{p} > 2$, we have that $\hat{\mu}_{JS}$ dominates the sample mean in terms of the squared-error risk.

16.4.4 Linear regression

An immediate consequence of the previous subsection is for linear regression. If we assume $y = X\beta + z$ for $z \sim \mathcal{N}(0, \sigma^2 I_p)$, then recall that $\hat{\beta}_{\text{MLE}} = (X^T X)^{-1} X^T y \sim \mathcal{N}(\beta, \sigma^2 (X^T X)^{-1})$ is a complete sufficient statistic for β, so we only need consider estimators of β depending on $\hat{\beta}_{\text{MLE}}$. Of course, this fits exactly into the setup from before! As a result, the James-Stein estimator is:

$$\tilde{\beta}_{\text{JS}} = \left(1 - \frac{(\tilde{p} - 2)\sigma^2}{\hat{\beta}_{\text{MLE}}^T X^T X \hat{\beta}_{\text{MLE}}}\right) \hat{\beta}_{\text{MLE}}.$$

By the result in the previous section, this dominates $\hat{\beta}_{\text{MLE}}$. Of course, it is perhaps slightly less interpretable because it is (e.g.) biased, but if one only cares about squared-error risk, there is no reason to use $\hat{\beta}_{\text{MLE}}$.

Bibliography

Bock, M. E. (1975). Minimax Estimators of the Mean of a Multivariate Normal Distribution. *The Annals of Statistics*, 3(1):209 – 218.

James, W. and Stein, C. (1961). Estimation with quadratic loss. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, pages 361–379.

Lawrence, M. (1985). Six arrested in apartheid sit-in. *The Stanford Daily*.