Stats 300C: Theory of Statistics	Spring 2013

Lecture 13 — April 29, 2013

Prof. Emmanuel Candes

Scribe: Alexandra Chouldechova; slighty edited by E. Candes

1 Outline

Agenda: Estimation of a Multivariate Normal Mean

1. Stein's Phenomenon
2. James-Stein Estimate
3. Stein's Unbiased Risk Estimate

We now take a break from hypothesis testing to study some results in estimation.

2 Estimation of a multivariate normal population

In this discussion, we are interested in estimating the mean μ in the multivariate normal model

$$X \sim N_p(\mu, \sigma^2 I)$$

This model can equivalently be written as

$$X_i = \mu_i + \sigma z_i \qquad z_i \overset{\text{i.i.d.}}{\sim} N(0,1) \qquad i = 1, \ldots, p$$

Our primary focus is to find an estimator $\hat{\mu}$ that performs well in terms of quadratic loss, defined as

$$\ell(\hat{\mu}, \mu) = \|\hat{\mu} - \mu\|_2^2 = \sum_{i=1}^{p} (\hat{\mu}_i - \mu_i)^2$$

The corresponding risk function, the MSE, (viewed as a function of μ) is defined as the expected loss and is given by

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

The natural estimator of μ is the MLE

$$\hat{\mu}_{\text{MLE}} = X \quad \text{[the sample mean]}$$

The MLE has risk

$$R(\hat{\mu}_{\text{MLE}}, \mu) = \mathbb{E}_\mu \|X - \mu\|_2^2 = \sigma^2 \mathbb{E}\|z\|^2 = p\sigma^2$$

For a long time, the MLE was thought to be 'the best' estimate for a multivariate mean. No estimator achieving a lower MSE for *all* values of μ was believed to exist.

Note: It is not difficult to improve on the MLE at a single point; e.g., the estimator $\hat{\mu} = 0$ outperforms the MLE at $\mu = 0$.

3 Stein's phenomenon

For $p = 1, 2$, this belief that the MLE is the best estimator is correct. However, for $p \geq 3$, it is false. A result of Stein 1956 hinted at this. A proof was eventually provided in 1961 by James & Stein.

In the 1961 paper, the authors introduce what is now referred to as the James-Stein estimator

$$\hat{\mu}_{\text{JS}} = \left[1 - \frac{p-2}{\|X\|^2}\right] X$$

This estimator is **nonlinear**, **biased**, and **shrinks the MLE towards 0**.

Theorem 1 (James, Stein 1961)**.** $\hat{\mu}_{\text{JS}}$ *dominates the MLE everywhere in terms of MSE. More precisely, for all* $\mu \in \mathbb{R}^p$,

$$\mathbb{E}_\mu \|\hat{\mu}_{\text{JS}} - \mu\|^2 < \mathbb{E}_\mu \|\hat{\mu}_{\text{MLE}} - \mu\|^2$$

In other words, this result proves the inadmissibility of the sample mean as an estimator of the mean for $p \geq 3$. It is known that the James Stein is not admissible either.

3.1 Stein's original argument (1956)

A good estimate should obey $\hat{\mu}_i \approx \mu_i$ for every i. Thus we should also have $\hat{\mu}_i^2 \approx \mu_i^2$. This further implies

$$\sum \hat{\mu}_i^2 \approx \sum \mu_i^2$$

Consider the estimator $\hat{\mu}_{\text{MLE}} = X$. For this estimator, we have

$$\begin{aligned}
\mathbb{E} \sum X_i^2 &= \mathbb{E}\left[\sum_i (\mu_i + \sigma z_i)^2\right] \\
&= \sum_i (\mu_i^2 + \sigma^2) \\
&= \|\mu\|^2 + \sigma^2 p
\end{aligned}$$

This suggests that for large p, $\|X\|^2$ is likely to be considerably larger than $\|\mu\|^2$, and hence we may be able to obtain a better estimator by shrinking the estimator toward 0. (See Figure 1 for a pictorial representation.)

In James, Stein 1961, the authors considered estimators of the form

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

They showed that for $c \in (0, 2(p-2))$,

$$R(\hat{\mu}_c, \mu) < R(\hat{\mu}_{\text{MLE}}, \mu)$$

and hence that $\hat{\mu}_{\text{JS}}$ dominates the MLE everywhere.

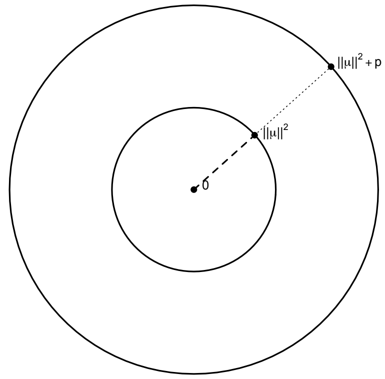

Figure 1: Pictorial version of Stein's original argument. The expected squared norm of the MLE ($\|\mu\|^2 + \sigma^2 p$) can be much greater than the 'desired' norm $\|\mu\|^2$. By the estimator shrinking toward 0, we can decrease its norm.

4 Stein's Unbiased Risk Estimate (SURE), 1981

Using tools that were developed later on, we are able to provide a simple proof of the James-Stein theorem. The proof uses Stein's unbiased risk estimate.

Suppose, as before, that $X \sim N(\mu, \sigma^2 I)$, and that we have some estimator

$$\hat{\mu} = X + g(X)$$

where g is 'almost' differentiable, and

$$\mathbb{E} \sum_{i=1}^{p} |\partial_i g_i(X)| < \infty$$

Almost differentiability means that there exists h_i so that we can write

$$g_i(x+z) - g_i(x) = \int_0^1 \langle h_i(x+tz), z \rangle dt$$

Usually, we write $h_i = \nabla g_i$.

The main result that we will use in order to compute the risk of $\hat{\mu}$ in this setup is Stein's identity.

Stein's identity (1981)

$$\mathbb{E}\|\hat{\mu} - \mu\|^2 = p\sigma^2 + \mathbb{E}\left[\|g(X)\|^2 + 2\sigma^2 \sum_i \partial_i g(X) \right]$$

An important consequence of Stein's identity is Stein's Unbiased Risk Estimate:

$$\mathrm{SURE}(\hat{\mu}) = p\sigma^2 + \|g(X)\|^2 + 2\sigma^2 \mathrm{div}\, g(X)$$

In other words, $\mathrm{SURE}(\hat{\mu})$ is an unbiased statistic for the risk.

Proof of Stein's identity. Assume without loss of generality that $\sigma = 1$. Then the risk of $\hat{\mu}$ is

$$\mathbb{E}\|X + g(X) - \mu\|^2 = \mathbb{E}\|X - \mu\|^2 + 2\mathbb{E}\left((X-\mu)^T g(X)\right) + \mathbb{E}\|g(X)\|^2$$

We just need to show that $\mathbb{E}(X-\mu)^T g(X) = \mathbb{E}\mathrm{div}\, g(X)$. This follows easily from integration by parts.

Let φ denote the $N(0, I)$ pdf. Then we can write

$$\mathbb{E}(X_i - \mu_i) g_i(X) = \int (x_i - \mu) g_i(x) \varphi(x - \mu) dx \tag{$*$}$$

Since

$$\partial_i \varphi(x - \mu) = -(x_i - \mu_i)\varphi(x - \mu)$$

$(*)$ becomes

$$(*) = \int \partial_i g_i(x) \varphi(x - \mu) dx = \mathbb{E}\partial_i g_i(X)$$

□

4.1 Applying SURE to $\hat{\mu}_{\mathrm{JS}}$ when $\sigma = 1$

We can rewrite $\hat{\mu}_{\mathrm{JS}}$ as

$$\hat{\mu}_{\mathrm{JS}} = X - \frac{p-2}{\|X\|^2} X$$

Thus $\hat{\mu}_{\mathrm{JS}}$ is of the form $X + g(X)$ where $g(x) = -(p-2)x/\|x\|^2$. This gives

$$\begin{aligned}
\|g(x)\|^2 &= \frac{(p-2)^2}{\|X\|^2} \\
\partial_i g_i(x) &= \partial_i \left\{ -(p-2)\frac{x_i}{\|x\|^2} \right\} = -\frac{p-2}{\|x\|^2} + \frac{2(p-2)x_i^2}{\|x\|^4} \\
\implies \mathrm{div}\, g(x) &= -\frac{(p-2)^2}{\|x\|^2}
\end{aligned}$$

Putting everything together gives

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

Remark: We can even be more precise. Noting that

$$\mathbb{E}\frac{1}{\|X\|^2} \geq \frac{1}{(p-2)+\|\mu\|^2}$$

we can bound the risk of the James-Stein estimator by

$$\mathbb{E}\|\hat{\mu}_{\text{JS}} - \mu\|^2 \leq p - \frac{p-2}{1+\frac{\|\mu\|^2}{p-2}}$$

It is interesting to consider a few special cases.

Under the global null, $\|\mu\|^2 = 0$, in which case

$$R(\hat{\mu}_{\text{JS}}, \mu) = 2$$

In the regime where our signal to noise ratio is 1, $\|\mu\|^2 = p$, and

$$R(\hat{\mu}_{\text{JS}}, \mu) \leq p/2$$

As $\|\mu\|^2 \to \infty$, $R(\hat{\mu}_{\text{JS}}, \mu) \to p$.

Figure 2 shows a plot of the upper bound obtained for the risk of $\hat{\mu}_{\text{JS}}$ compared to the risk of the MLE.

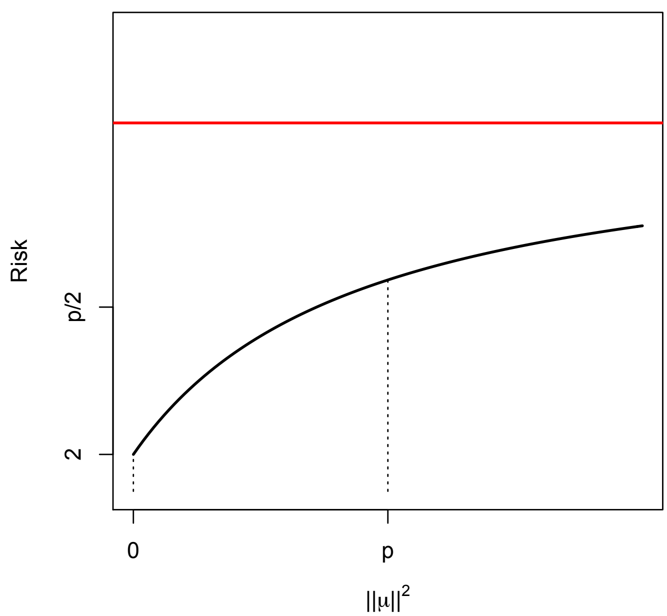

Figure 2: Comparison of the risk of the MLE (red) to the upper bound derived for risk of $\hat{\mu}_{\text{JS}}$ (black).

There's a further estimator that improves upon the JS estimate by precluding the possibility of sign reversal.

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