Proposal of Double Descent Phenomenon in SURE

Zhiyuan Chen

June 7, 2024

1 SURE

Stein's unbiased risk estimator for $y \sim \mathbb{N}\left(\theta, \sigma^2 I_n\right)$ is to minimize the following risk

$$\mathrm{SURE}(\hat{\theta}) = \|\hat{\theta} - y\|^2 - \sigma^2 n + 2\sigma^2 \sum_{j=1}^{n} \frac{\partial \hat{\theta}_j(y)}{\partial y_j}.$$

We know that Jame-Stein estimator is a shrinkage estimator. It is natural that SURE is a kind of regularization when applied on linear model.

2 SURE in Linear Regression

In linear regression $y \sim X\beta + \epsilon,\ X \in \mathbb{R}^{n\times p},\ y \sim \mathbb{N}\left(\theta, \sigma^2 I_n\right)$. Assuming σ^2 is known and X is full rank, then

$$\begin{aligned}
\hat{\theta} &= \hat{y}, \\
&= X\hat{\beta}, \\
&= X\left(X^\top X\right)^\dagger X^\top y, \\
&= Hy. \\
\sum_{j=1}^{n} \frac{\partial \hat{\theta}_j}{\partial y_j} &= \sum_{j=1}^{n} H_{jj}, \\
&= \mathrm{tr}(H).
\end{aligned}$$

2.1 Low Dimension

When $p < n,\ \mathrm{rank}(X) = p$, then

$$\begin{aligned}
H &= X\left(X^\top X\right)^{-1} X^\top. \\
\mathrm{tr}(H) &= \mathrm{tr}\left(X^\top X \left(X^\top X\right)^{-1}\right), \\
&= \mathrm{tr}(I_p), \\
&= p. \\
\Rightarrow \quad \mathrm{SURE}(\hat{\theta}) &= \|y - X\hat{\beta}\|^2 - n\sigma^2 + 2p\sigma^2.
\end{aligned}$$

We can see now n is bonus and p is penalty, which is consistent with regularization we are familiar with.(e.g. AIC)

2.2 High Dimension

If $p > n,\ \mathrm{rank}(X) = n$, we can use rank-retaining factorization to solve the pseudo inverse

$$\left(X^\top X\right)^\dagger = X^\top (XX^\top)^{-2} X.$$

Since $XX^\top \in \mathbb{R}^{n\times n}$ is full rank, its matrix inverse exists. Then we can compute the trace of hat matrix

$$\begin{aligned}
\operatorname{tr}(H) &= \operatorname{tr}\left(XX^\top (XX^\top)^{-2} XX^\top\right), \\
&= \operatorname{tr}(I_n), \\
&= n. \\
\Rightarrow \mathrm{SURE}(\hat{\theta}) &= \|\hat{\theta} - y\|^2 - n\sigma^2 + 2n\sigma^2 \\
&= \|y - X\hat{\beta}\|^2 + n\sigma^2
\end{aligned}$$

Surprisingly, n becomes penalty in the $p > n$ regime.
So when $p < n$, as $\frac{p}{n}$ increasing, the risk follows classical U-shape curve; when $p > n$, as n decreasing i.e. $\frac{p}{n}$ increasing, the risk will decrease, which is very similar to double descent phenomenon.

3 Linear System Estimation

What if just choose $\hat{\beta} = X^\dagger y$ as estimator?

$$\begin{aligned}
\hat{\theta} &= \hat{y}, \\
&= X\hat{\beta}, \\
&= XX^\dagger y, \\
&= Hy. \\
X^\dagger &= \begin{cases} (X^\top X)^{-1} X^\top & \text{if } p < n \\ X^\top (XX^\top)^{-1} & \text{if } n > p \end{cases} \\
\operatorname{tr}(H) &= \begin{cases} \operatorname{tr}(I_p) = p & \text{if } p < n \\ \operatorname{tr}(I_n) = n & \text{if } n > p \end{cases}
\end{aligned}$$

So the result is consistent with above.

4 Experimental Results

4.1 Using SURE Package

There is a rough experiment with figures of MSE $\|\hat{\theta}-y\|^2$ and SURE risk $\|\hat{\theta}-y\|^2-\sigma^2 n+2\sigma^2\sum_{j=1}^{n}\frac{\partial\hat{\theta}_j(y)}{\partial y_j}$. The SURE estimator and SURE risk are computed by R package asus. Set $p = 1000$, n from 200 to 10000.

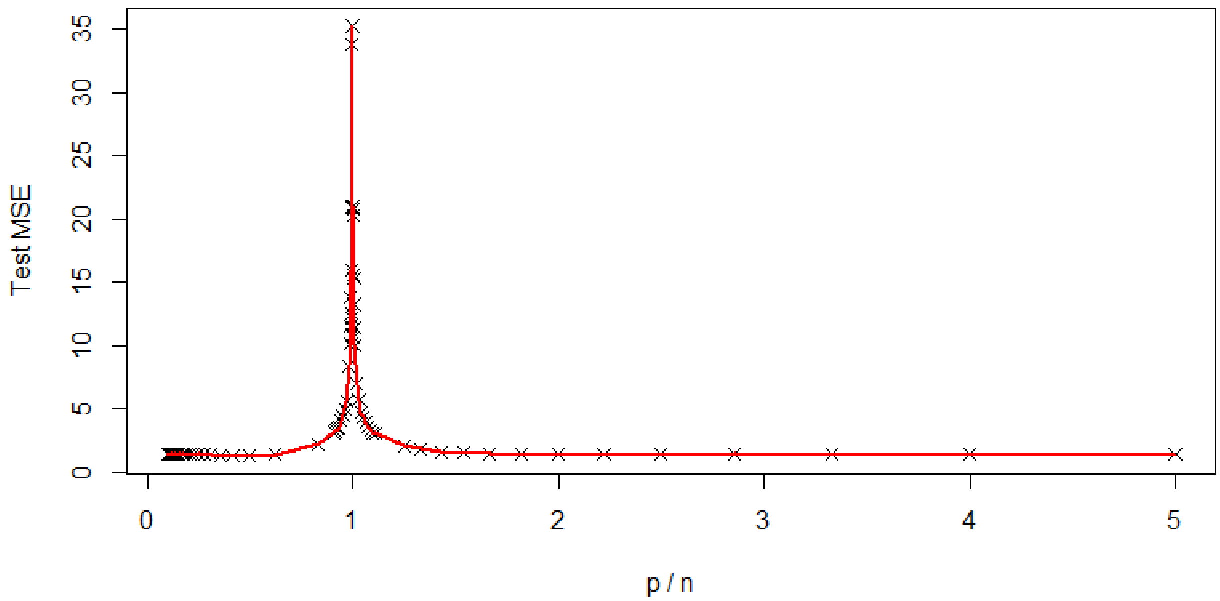

Figure 1: MSE

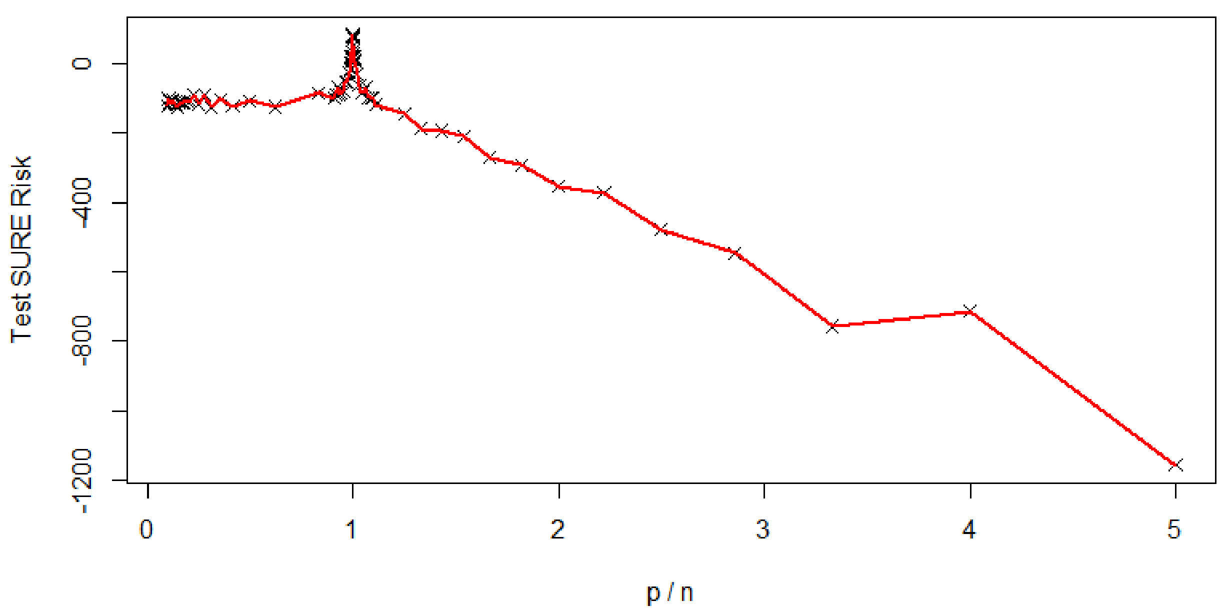

Figure 2: SURE Risk

Both figures indicate the double descent phenomenon. However, the SURE risks are negative at many points, which is unusual even considering the $-\sigma^2 n$ term in the SURE risk. Therefore, I decided to solve the optimization problem directly.

4.2 Using Optimization

Use R package nloptr to solve:

$$\min_{\hat{\beta} \in \mathbb{R}^p} \mathrm{SURE}(\hat{\beta}) = \|y - X\hat{\beta}\|^2 - n\sigma^2 + 2\sigma^2 \operatorname{tr}(H)$$

Since the computation of a large matrix (with $p = 1000$ and n ranging from 200 to $10,000$) is quite expensive, I changed the settings to $p = 100$ and n ranging from 50 to 150. The drawback is that points around $\frac{p}{n} = 1$ will be sparse.

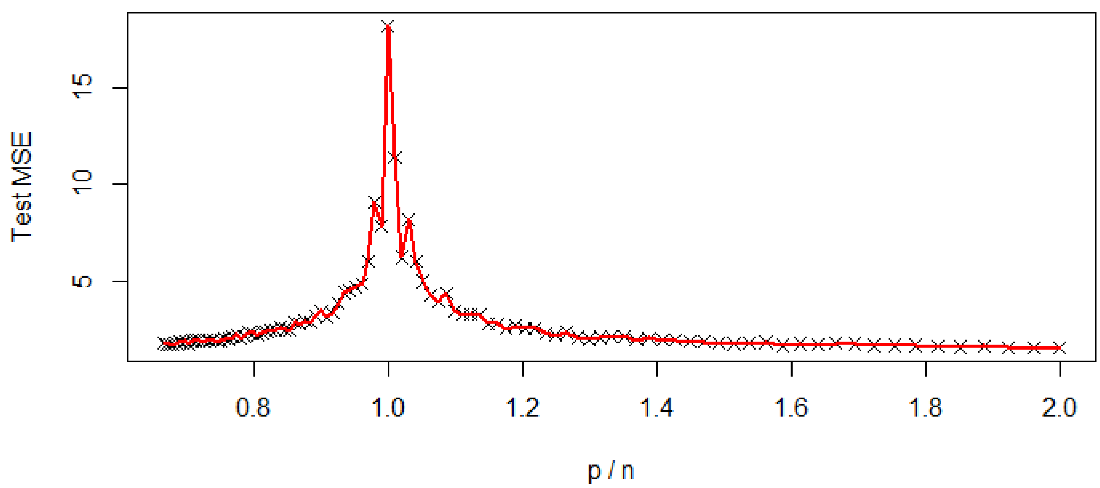

Figure 3: MSE with respect to $\frac{p}{n}$

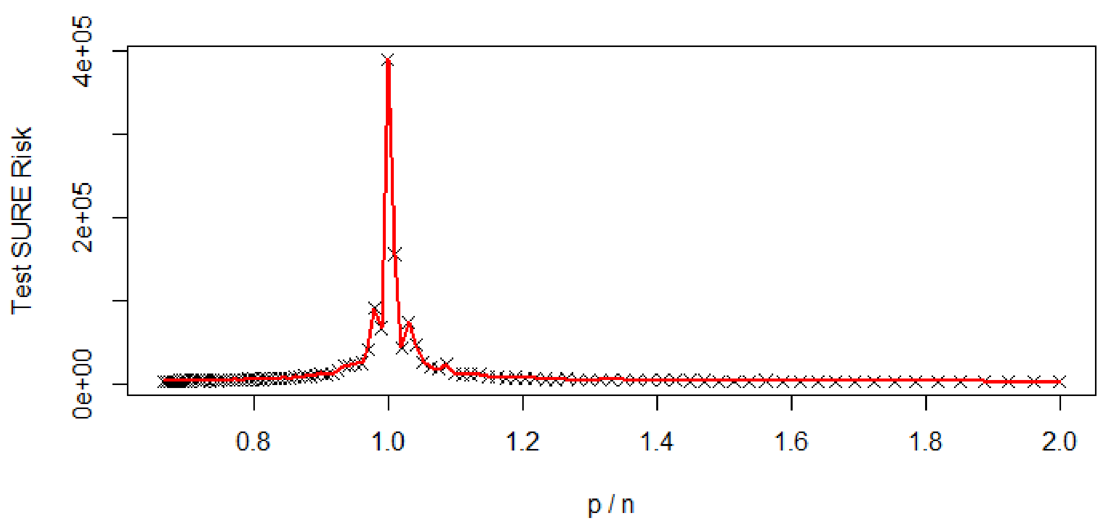

Figure 4: SURE Risk with respect to $\frac{p}{n}$

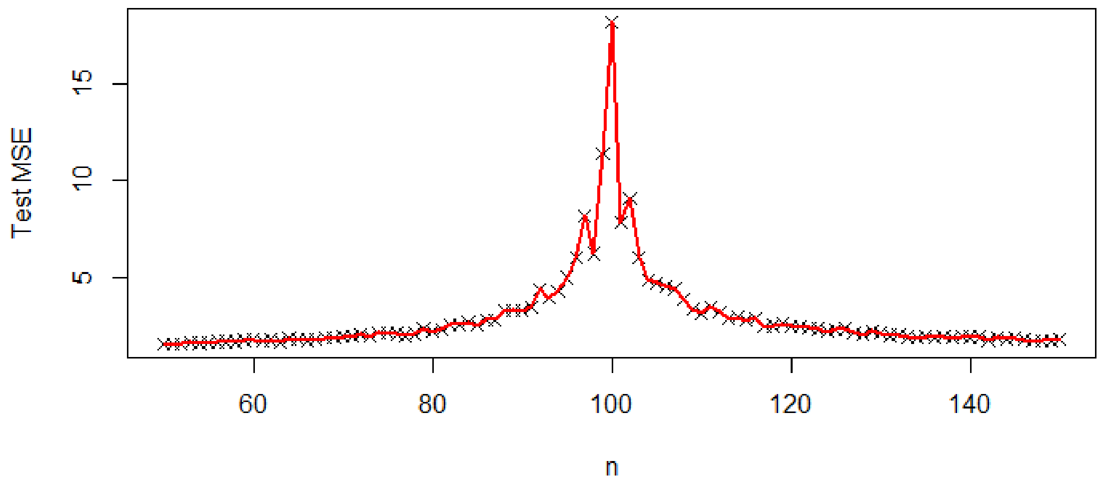

Figure 5: MSE with respect to n

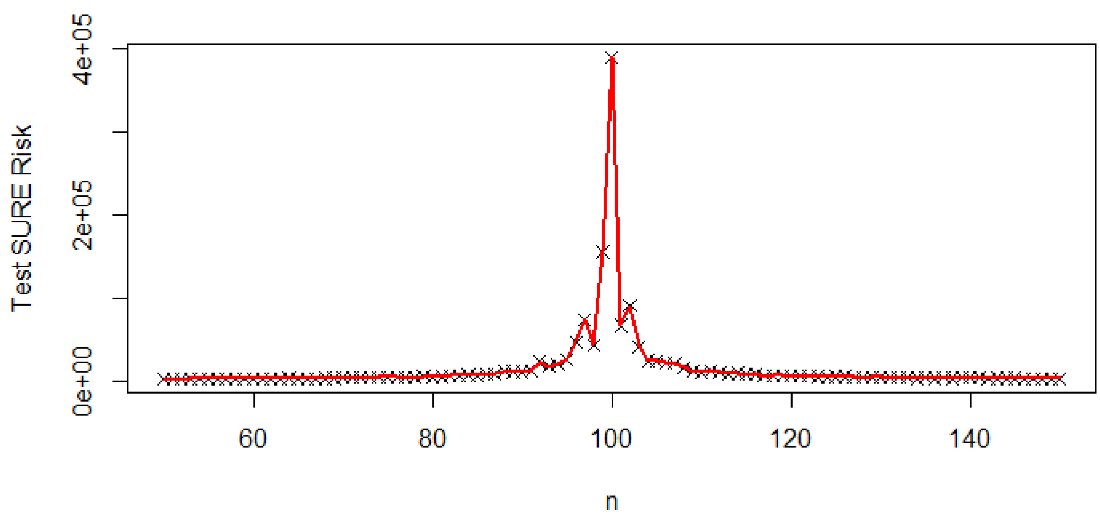

Figure 6: SURE Risk with respect to n

We can see obvious double descent phenomenon in above figures.