# High-Dimensional Estimation: Comparing Soft/Hard Thresholding and the James–Stein Estimator

**Yijin Zeng**
Imperial College London

## 1 Introduction

We first introduce our problem as follows. Suppose $Z^n = (Z_1, \ldots, Z_n)$, where

$$Z_i = \theta_i + \sigma_n \epsilon_i, i = 1, \ldots, n,$$

with $\theta^n = (\theta_1, \ldots, \theta_n) \in \mathbb{R}^n$ a vector of unknown parameters, $\epsilon_1, \ldots, \epsilon_n \overset{i.i.d}{\sim} \mathcal{N}(0, 1)$ random variables and $\sigma_n = \sigma$ a known parameter. We want to perform inference on $\theta^n$ with observed $z^n = (z_1, \ldots, z_n)$, which is a realisation of $Z^n$, as well as evaluate our inference results in a reasonable way.

However, such a problem has the main difficulty that the model has the number of parameters increasing at the same rate as the number of data points and therefore carries all the complexities and subtleties of a non parametric problem.

In this project, we will investigate this problem in more depth theoretically, as well as empirically by simulations, using three estimating methods: the soft threshold estimator, the hard threshold estimator and the James-Stein estimator. Moreover, from now on we will be assuming that $\sigma_n = \sigma = 1$ for simplicity.

## 2 Risk Function and Three Estimators

In this section, we will give a formal definition of the risk function and the three estimators that we will use, namely the soft threshold estimator, the hard threshold estimator and the James-Stein estimator.

### 2.1 Risk Function

Suppose we have an estimator $\hat{\theta}^n := \hat{\theta}^n(Z^n)$, where $Z^n$ is used to emphasize $\hat{\theta}^n$ is a function of $Z^n$. The squared error loss is given by:

$$L(\hat{\theta}^n, \theta^n) := \sum_{i=1}^{n} (\hat{\theta}_i - \theta_i)^2$$

and we will be using this loss function throughout. For any given $\theta^n$, since $\hat{\theta}^n$ is a random variable, it can be proved that $L(\hat{\theta}^n, \theta^n)$ is also a random variable. Normally, we may be more interested in the "average performance" of the estimator $\hat{\theta}^n$. This can be represented by the risk function:

$$R(\hat{\theta}^n, \theta^n) = \sum_{i=1}^{n} E_\theta (\hat{\theta}_i - \theta_i)^2.$$

Since $\theta^n$ is unknown, it would be impossible to choose $\hat{\theta}$ by minimizing the risk function directly. It is desirable that we can find a $\hat{\theta}_0^n$ that satisfies $R(\hat{\theta}_0^n, \theta^n) \leq R(\hat{\theta}^n, \theta^n)$ for any given $\theta^n \in \mathbb{R}^n$.

However, this property would be too strict and hard to satisfy. Therefore, it is common that we may turn our sight into minimizing an estimator of the risk function.

One popular unbiased risk estimator is Stein's Unbiased Risk Estimator (SURE). The definition of SURE can be given as follows: for any given $\theta \in \mathbb{R}^n$, suppose $\hat{\theta} = \hat{\theta}(Z) = (\hat{\theta}_1(Z), \hat{\theta}_2(Z), \ldots, \hat{\theta}_n(Z))^T$ is an estimator of $\theta$ and is differentiable. Let $g(Z) = \hat{\theta} - Z$ and $D_i(Z) = \frac{\partial g_i(Z)}{\partial Z_i}$. Then

$$\hat{R}(Z) = n\sigma^2 + 2\sigma^2 \sum_{i=1}^{n} D_i(Z) + \sum_{i=1}^{n} \{g_i(Z)\}^2 \tag{1}$$

is an unbiased estimator of the risk of $\hat{\theta}$. Notice that we need to make the assumption that $g$ is a weakly differentiable function. As we will see in the following sections, SURE is commonly used in order to decide the thresholding level for a given estimator.

### 2.2 Soft Threshold Estimator

One popular estimator to infer $\theta^n$ is the soft threshold estimator. It is defined as:

$$\hat{\theta}_i = \begin{cases} Z_i + \lambda & \text{if} \quad Z_i < -\lambda \\ 0 & \text{if} \quad |Z_i| \leq \lambda \\ Z_i - \lambda & \text{if} \quad Z_i > \lambda \end{cases},$$

where $\lambda$ is a constant.

An appropriate choice of $\lambda$ is essential for the soft threshold estimator to achieve a low risk value. For example, when the level of sparsity of $\theta^n$ is high, that means that we have $\theta_i = 0$ for many $i$'s, a larger $\lambda$ may be preferred since this will give $\hat{\theta}_i = 0$ more often.

In this project, we will consider two methods of setting $\lambda$. The first is a choice of $\lambda$ that minimizes SURE, as defined in equation (1), while the second is setting $\lambda = \sigma\sqrt{2\log n}$. Note that the second choice of $\lambda$ would give a very reasonable estimator when $\theta^n$ is sparse [2]. The expression for SURE with the soft threshold estimator is given by:

$$\hat{R}^{\text{SOFT}}(Z) = \sum_{i=1}^{n} \left\{1 - 2\mathbb{I}\{|Z_i| \leq \lambda\} + \min\{Z_i^2, \lambda^2\}\right\}, \tag{2}$$

where $\mathbb{I}$ is the binary indicator.

### 2.3 Hard Threshold Estimator

The hard threshold estimator is defined as follows:

$$\hat{\theta}_i = \begin{cases} Z_i & \text{if} \quad |Z_i| > \lambda \\ 0 & \text{if} \quad |Z_i| \leq \lambda \end{cases},$$

where once again $\lambda$ is the thresholding level that has to be determined.

Hard thresholding only retains the 'large' $\theta_i$'s, simply by setting everything that it considers too 'small' equal to 0. The main 'issue' with this estimator is that it does not satisfy the weak differentiability condition. Therefore, making use of SURE to determine a thresholding level $\lambda$ that will minimise the risk of a hard thresholding estimator is something that can not be done.

### 2.4 James-Stein Estimator

The James-Stein estimator [3] is a linear estimator defined by:

$$\hat{\theta}^{JS} = \left(1 - \frac{n-2}{\|Z\|^2}\right) Z,$$

where $\|\cdot\|$ denotes the Euclidean (or $L_2$) norm. Notice that the James-Stein estimator is also called a 'shrinkage' estimator, due to the fact that it shrinks $Z$ towards 0 (assuming that $n-2 < \|Z\|^2$). The expression for SURE using the James-Stein estimator is given by:

$$\hat{R}^{\mathrm{JS}}(Z) = n - \frac{(n-2)^2}{\|Z\|^2}. \tag{3}$$

An 'improvement' of the James-Stein estimator is the so-called 'positive-part James-Stein estimator', defined by:

$$\hat{\theta}_+^{JS} = \left(1 - \frac{n-2}{\|Z\|^2}\right)_+ Z,$$

where $\left(1 - \frac{n-2}{\|Z\|^2}\right)_+ = \max\left\{1 - \frac{n-2}{\|Z\|^2}, 0\right\}$. However, we will not be considering the positive-part James-Stein estimator in this report.

# 3 Simulations

## 3.1 Comparison of Estimators

We will first give a brief theoretical overview of the James-Stein and the soft thresholding estimators. It turns out that if we consider a linear shrinkage estimator $\hat{\theta}^{LIN} = cZ$, its risk will be given by the expression $(1-c)^2\|\theta^n\|^2 + nc^2$. It is straightforward to obtain the value of $c$ that minimizes this risk, which is equal to $\frac{\|\theta^n\|^2}{n+\|\theta^n\|^2}$, yielding an 'oracle risk' given by $\frac{n\|\theta^n\|^2}{n+\|\theta^n\|^2}$. The main issue with this oracle estimator is its dependence on $\theta^n$, which is unknown to us and can therefore not be used. One can show, using some simple calculations, that the risk of the James-Stein estimator is bounded above by $2 + \frac{n\|\theta^n\|^2}{n+\|\theta^n\|^2}$. Moreover, using a general result that states $\frac{1}{2}\min\{a,b\} \leq \frac{ab}{a+b} \leq \min\{a,b\}$, one easily gets to the following bounds for the oracle risk, more precisely:

$$\frac{1}{2}\min\{\|\theta^n\|^2, n\} \leq \frac{n\|\theta^n\|^2}{n+\|\theta^n\|^2} \leq \min\{\|\theta^n\|^2, n\}.$$

At this point, it is interesting to investigate whether we can get any bounds for the risk of the soft threshold estimator. If we set the threshold level to the universal value $\lambda = \sqrt{2\log n}$, it turns out that the risk of the soft threshold estimator is bounded above by $(2\log n + 1)\left(1 + \sum_{i=1}^{n}\min\{\theta_i^2, 1\}\right)$.[2] Notice that the last term, namely $\sum_{i=1}^{n}\min\{\theta_i^2, 1\}$, is the risk of an oracle estimator for highly sparse $\theta^n$. This estimator will consist of values equal to either $Z_i$ or 0, assuming that it knows when to estimate $\theta_i$ by any of the two.

Of course the soft threshold estimator can be chosen with a thresholding level that minimizes its SURE, as given by expression (2). Let us consider the expression once again:

$$\hat{R}^{\mathrm{SOFT}}(Z) = \sum_{i=1}^{n}\left\{1 - 2\mathbb{I}\{|Z_i| \leq \lambda\} + \min\{Z_i^2, \lambda^2\}\right\}$$

If we take the $|Z_i|$ values and order them in an ascending order, we obtain a permutation of the data points which we will denote by $|Z_1^*|, \ldots, |Z_n^*|$. One can see that the expression for $\hat{R}^{\mathrm{SOFT}}(Z)$ will be strictly increasing for $\lambda$ values lying between $|Z_i^*|$ and $|Z_{i+1}^*|$ (for $i = 1, \ldots, n-1$). Therefore, the 'optimal' thresholding level has to be one of the $|Z_i|$'s. The value needs to be found empirically, by calculating SURE for all $n$ possible $\lambda$ values and picking the one that minimises it. Of course, this is a much more expensive process than using the universal $\lambda$, with an effort of order $\mathcal{O}(n\log n)$, while thresholding with a fixed $\lambda$ has a cost of order $\mathcal{O}(n)$.[1]

We have implemented some simulations in order to analyse how each of the estimators behaves under certain conditions. We will be for instance interested in a comparison of the performance of each estimator as the dimensionality of our problem increases or as the sparsity level varies. The estimators that we consider are the James-Stein estimator, the soft threshold estimator with $\lambda = \sqrt{2 \log n}$ and with $\lambda$ set to minimize SURE, as well as the hard threshold estimator. The performance of each estimator is measured by calculating its empirical risk; that is achieved simply by generating the true $\theta^n$ with a specified number of zeros (which is obtained by the sparsity level that will remain constant throughout) and the remaining values are randomly drawn from a Uniform distribution on [-10, 10]. The $Z_i$ values are then generated from a multivariate Gaussian distribution with mean given by $\theta^n$ and an identity covariance matrix. We then obtain the four estimators for $\theta^n$ and calculate the sum of of their squared differences to the true parameter values. This process is repeated a total of 1000 times for each aspect that we wish to investigate. We make sure to calculate the mean of our losses to eventually obtain the empirical risk.

In Figure 1, we can see the empirical risks calculated using the four estimators mentioned above, as we increase the dimensionality $n$. We consider $n$ values ranging from $n = 3$ to $n = 100$, while the sparsity level is fixed at $90\%$. We can easily observe that generally hard thresholding yields the lowest risk. This 'keep-or-kill' estimator does a very good job but its main issue is that due to its non-differentiable form, we cannot improve much on the thresholding level, which is picked to be equal to the universal rule $\lambda = \sqrt{2 \log n}$. However, the soft threshold estimator with $\lambda$ chosen to minimise SURE is doing a similarly good job; the empirical risk values may not be as low as for hard thresholding but we can see that it is a significant improvement to the soft threshold with the much celebrated $\lambda = \sqrt{2 \log n}$. The James-Stein estimator is doing an almost similar job as soft thresholding with the universal $\lambda$, although its risk is typically larger, as can be seen. We can also observe that the empirical risk is increasing together with $n$, which is not surprising, as increasing the problem dimensionality implies that we sum over more non-negative terms when evaluating the loss.

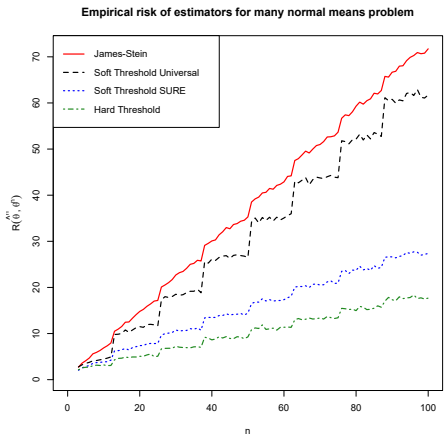


Figure 1: Empirical risks for 90% sparsity level and varying $n$

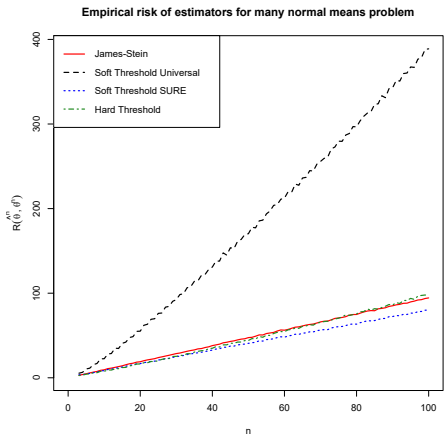


Figure 2: Empirical risks for 50% sparsity level and varying $n$

Let us now decrease the sparsity level to 50%. This means that in practice half of the values of $\theta^n$ will be set to zero, while the rest will be drawn from a uniform distribution in the range [-10, 10]. More values will be 'significantly large' now compared to when the sparsity level was 90%. Indeed, as can be seen in Figure 2, soft thresholding with $\lambda = \sqrt{2 \log n}$ has an empirical risk that increases almost linearly as more non-zero terms are introduced. The James-Stein estimator yields a much lower risk for all values of $n$ considered, while we can see that the low sparsity causes the hard threshold estimator to increase its risk as $n$ increases, with the risk of the latter even exceeding that of the former for $n \approx 70$. However, the soft threshold estimator with $\lambda$ determined by SURE yields a lower risk than all other estimators and its improvement in terms of the empirical risk gets larger with $n$. This indicates that the universal rule $\lambda = \sqrt{2 \log n}$ only works well with the soft threshold estimator

if $\theta^n$ is highly sparse. We will be looking into how the sparsity level affects the performance of soft thresholding in much more detail in the following subsection.

### 3.2 Best Thresholding Level of Soft Threshold Estimator for Different Levels of Sparsity of $\theta^n$

As we mentioned in Section 2.2, for the soft threshold estimator, we normally consider setting the threshold level $\lambda$ in two ways: the universal setting $\lambda = \sigma\sqrt{2\log n}$ and the $\lambda$ value minimizing SURE. In this section, we will compare the two candidates for $\lambda$ under different levels of sparsity of $\theta^n$. For a given vector of parameters $\theta^n = (\theta_1, \ldots, \theta_n)$, the level of sparsity of $\theta^n$ is defined as

$$s = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{\theta_i = 0\}$$

Clearly, the difficulty of the estimating task depends on the real parameter values $\theta^n$. For the following simulations, we will consider a given vector $\theta^n = (\theta_1, \ldots, \theta_n)$ $(n = 100)$, where the $\theta_i$'s are generated independently from the uniform distribution $\mathrm{U}(-10, 10)$. Before comparing the two parameter setting methods under different levels of sparsity $s$ for this $\theta^n$, we will first define the *empirical risk function* for a given level of sparsity $s$.

For any given $s$ $(s = 0.00, 0.01, \ldots, 1.00)$, we will randomly select $s * n$ numbers from the sequence $1, \ldots, 100$ and set the corresponding $\theta_i$ values to 0. The new sequence is denoted as $\theta_s^n = (\theta_{1,s}, \ldots, \theta_{n,s})$ and it satisfies the sparsity level $s$ (See Figure 3).

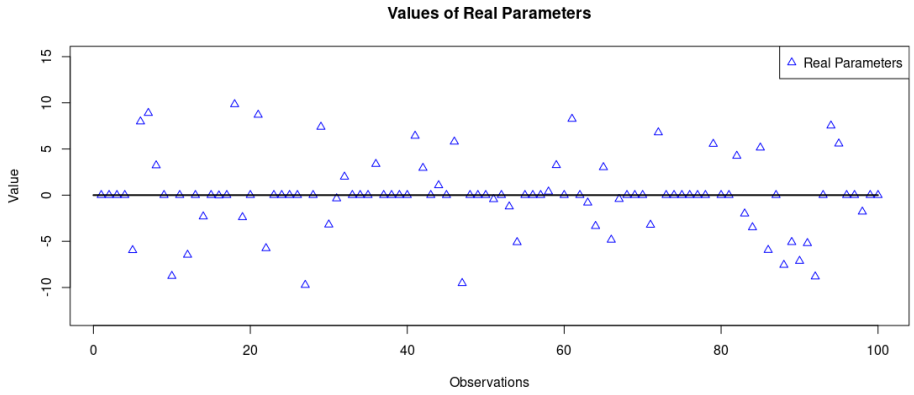


Figure 3: Example: a plot of real parameters, with level of sparsity $s = 0.5$.

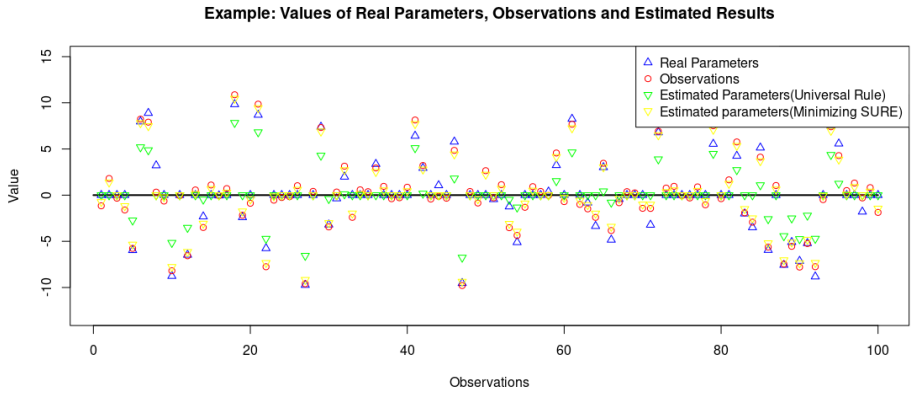


Figure 4: Example: Estimated parameters with $\lambda$ given by universal rule and minimizing SURE.

For the new sequence $\theta_s^n$, observations $z^n$ are realizations from random variable $Z_s^n = (Z_{1,s}, \ldots, Z_{n,s})$ satisfying

$$Z_{i,s} = \theta_{i,s} + \epsilon_i, \ i = 1, \ldots, n,$$

with $\epsilon_1, \ldots, \epsilon_n \overset{i.i.d}{\sim} \mathcal{N}(0,1)$. We wil now make use of the soft threshold estimator to infer the real parameters as defined in Section 2.2 (See Figure 4) and calculate the squared error loss by $L^j_{\theta_s^n}(\hat{\theta}_s^n) = \sum_{i=1}^n (\hat{\theta}_{i,s} - \theta_{i,s})^2$. This is called the $j^{\text{th}}$ *experiment* for a given $\theta_s^n$. We will repeat generating $Z_s^n$ a total of $N$ times. Therefore, $N$ loss values $L^j_{\theta_s^n}(\hat{\theta}_s^n)$ $(j = 1, \ldots, N)$ are produced for the two different parameter methods separately. Then the *empirical risk function* is defined as:

$$\hat{R}_{\theta_s^n}(\hat{\theta}_s^n) = \sum_{j=1}^{N} L^j_{\theta_s^n}(\hat{\theta}_s^n)/N.$$

The simulation results can be seen in Figure 5. What we see is that the empirical risk function value decreases when the level of sparsity $s$ increases for both parameter setting methods, while unless $s \approx 1$, the soft threshold estimator with $\lambda$ minimizing SURE will produce a much smaller empirical risk function value compared to $\lambda$ given by the universal rule.

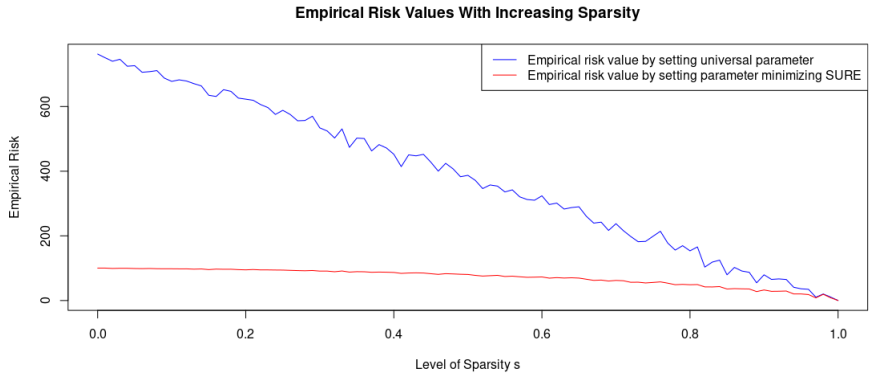


Figure 5: Empirical risk function value of setting parameter $\lambda$ using universal rule (blue line) and setting $\lambda$ by minimizing SURE (red line).

## 4 Conclusion

In this project, we have defined the sum of squares risk function and three estimators, namely the James-Stein estimator, the soft threshold estimator and the hard threshold estimator, and we have conducted two simulations. The first simulation compares the three estimators for a fixed sparsity level and a varying dimensionality. The result of it shows that under $90\%$ sparsity, the James-Stein estimator and the soft threshold estimator with the universal parameter produce some very large empirical risk values while the hard threshold estimator yields the lowest risk; under $50\%$ sparsity, the empirical risk values of the soft threshold estimator with the universal parameter are significantly larger than of all other estimators. The second simulation compares the soft threshold estimator with the universal threshold level and with the $\lambda$ value minimizing SURE. The result of this simulation shows that unless the sparsity is close to 1, the empirical risk values of the $\lambda$ value that minimizes SURE are always smaller than those we obtain when setting $\lambda$ equal to the universal value of $\sqrt{2 \log n}$.

## References


[1] D. L. Donoho and I. M. Johnstone. Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American statistical association*, 90(432):1200–1224, 1995.

[2] D. L. Donoho and J. M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *biometrika*, 81 (3):425–455, 1994.

[3] C. Stein. Inadmissibility of the usual estimator for the mean of a multivariate normal population. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Vol 1*, pages 197–206. University of California Press Berkeley, 1955.