Stats 300C: Theory of Statistics Spring 2018

Lecture 24 — May 30, 2018

Prof. Emmanuel Candes

Scribe: Martin J. Zhang, Jun Yan, Can Wang, and E. Candes

1 Outline

Agenda: High-dimensional Statistical Estimation

1. Lasso
2. $\ell_1 - \ell_0$ equivalence
3. Oracle inequalities for Lasso?

2 Canonical Selection Procedure

Consider the usual linear model

$$y = X\beta + z$$

and the canonical ℓ_0 selection procedure

$$\min \|y - X\hat{\beta}\|^2 + \lambda^2\sigma^2\|\hat{\beta}\|_0$$

We have seen several choices of the parameter λ

- AIC: $\lambda^2 = 2$
- BIC: $\lambda^2 = \log n$
- RIC: $\lambda^2 \approx 2\log p$

The critical problem shared by these methods is that finding the minimum requires an exhaustive search over all possible models. This search is completely intractable for even moderate values of p.

3 Lasso

Consider instead solving the ℓ_1 regularized problem

$$\min \frac{1}{2}\|y - X\hat{\beta}\|_2^2 + \lambda\sigma\|\hat{\beta}\|_1$$

This is a quadratic program which can be solved efficiently.

ℓ_1 regularization has a long history in statistics and applied science. This idea appeared in the early work of reflection seismology. In seismology, people try to image the earth by sound wave. With reflected sound wave, one receives a time series of signals, which is noisy and low frequency. In order to detect where the "jumps" in earth are, people came up with the idea to use ℓ_1 norm.

"In deconvolving any observed seismic trace, it is rather disappointing to discover that there is a nonzero spike at every point in time regardless of the data sampling rate. One might hope to find spikes only where real geologic discontinuities take place. Perhaps the L_1 norm can be utilized to give an output trace like Figure 13a."

Claerbout and Muir (1973)

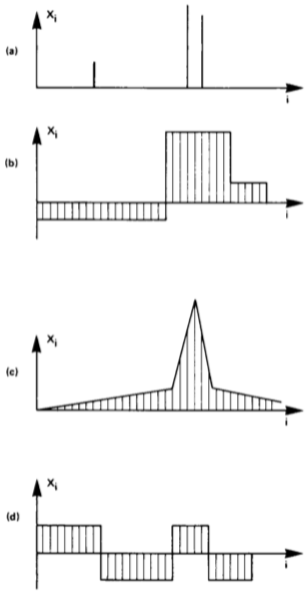


FIG. 13. Solutions to highly underdetermined asymmetric-linear norm problems where the smoothness criterion is taken to be minimization of the magnitude of (a) components of x, (b) first differences on x, (c) second differences on x, and (d) Chebyshev norm of x.

Some important references about the use of ℓ_1 norm are below.

- Logan (1956)
- Claerbout & Muir (1973)
- Taylor, Banks, McCoy (1979)
- Santosa & Synes (1983, 1986)
- Rudin, Osher, Fatemi (1992)
- Donoho et al. (1990's)
- Tibshirani (1996)

3.1 Why ℓ_1?

To motivate the ℓ_1 approach, consider first the noiseless case

$$y = X\beta,$$

which is an underdetermined system of equations in the case where we have more variables than unknowns, i.e. $p > n$.

In this setting, there is no unique solution to the least squares problem. However, by assuming that β is sparse, we can reduce our search to sparse solutions, i.e., we wish to find the sparsest solution to $y = X\beta$ by solving

$$\min \|\hat{\beta}\|_0 \quad \text{s.t.} \quad y = X\hat{\beta}$$

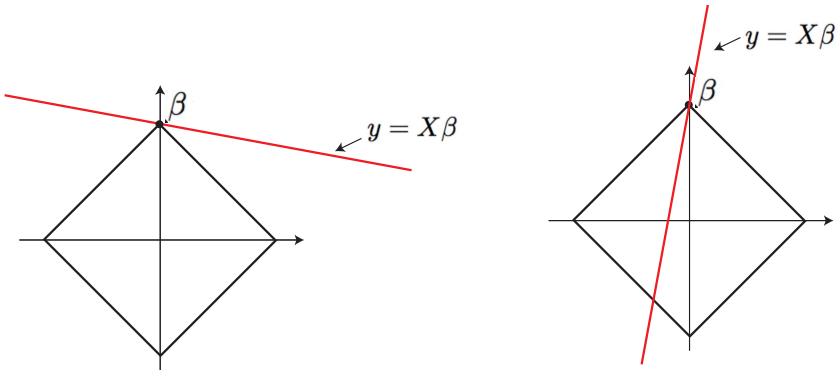
Under some conditions, the minimizer is unique, i.e., $\hat{\beta} = \beta$.

Even so, this doesn't address the issue of computational tractability. The optimization problem as stated remains intractable. This motivates the convex relaxation

$$\min \|\hat{\beta}\|_1 \quad \text{s.t.} \quad y = X\hat{\beta}$$

Under broad conditions, it can be shown that

- The minimizer is unique
- The ℓ_1 solution is **equal** to the ℓ_0 solution.

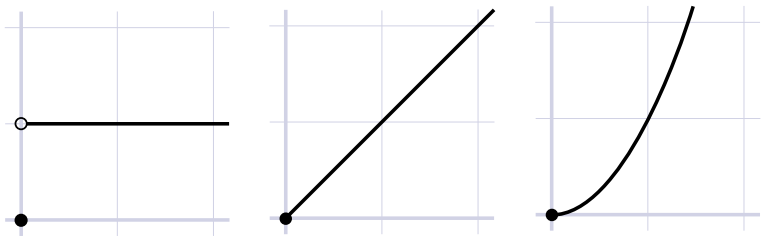


(a) $\hat{\beta} = \beta$ Good! (b) $\hat{\beta} \neq \beta$ Bad!

Figure 1: Geometry of the ℓ_1 optimization problem in the noiseless setting. To understand the picture suppose $\hat{\beta}$ is a feasible point ($X\hat{\beta} = X\beta = y$). Consider 'descent' vectors u such that $\|\hat{\beta} + u\|_1 < \|\hat{\beta}\|_1$. It is easy to see that $\hat{\beta}$ is the solution to the ℓ_1 optimization problem if $Xu \neq 0$ for all descent vectors u, i.e., if all vectors with smaller ℓ_1 norm are non-feasible. **(a)** The solution here is the true β, because all other feasible vectors have larger ℓ_1 norm. **(b)** Here, $\hat{\beta} \neq \beta$. Starting at β, we can consider the descent direction u as shown. The point $\beta + u$ remains feasible, and has ℓ_1-norm strictly smaller than β.

Rigorous results pertaining to the formal equivalence between the ℓ_0 and ℓ_1 problem can be found in

- Candes & Tao (2004)
- Donoho (2004)



(a) ℓ_0 quasi-norm (b) ℓ_1 norm (c) ℓ_2 norm

Figure 2: ℓ_1 norm is closest convex approximation to ℓ_0 quasi-norm.

3.2 ℓ_1 as a convex relaxation of ℓ_0

ℓ_1 is the tightest convex relaxation of ℓ_0. One way of seeing this is to look at the model

$$\mu = \sum_{i=1}^{p} \beta_i X_i$$

where β is sparse. The 'building blocks' of sparse solutions are vectors of the form

$$(0, \ldots, 0, \pm 1, 0, \ldots, 0)$$

and the ℓ_1 ball is the smallest convex body that contains these building blocks. It is the convex hull of these points.

Another interpretation is this: the ℓ_1 norm is the best convex lower bound to the cardinality functional over the unit cube. That is to say, if we search for a convex function f such that

$$f(\beta) \leq \|\beta\|_0 \quad \text{for all } \beta \in [-1, 1]^p,$$

then the tightest lower bound is the ℓ_1 norm, see Figure 2.

4 Lasso and Oracles?

Recall from last time that

Theorem 1. *CSP (Canonical Selection Procedures) with $\lambda_p = A \log p$, then*

$$\mathbb{E}\|X\hat{\beta} - X\beta\|^2 = O(\log p)[\sigma^2 + R^I(\mu)]$$

where $R^I(\mu)$ is the ideal risk.

As we will illustrate, we cannot hope to achieve comparable optimality results for the Lasso.

Case study 1: Take X to be

$$\begin{pmatrix} 1 & 0 & \cdots & 0 & \varepsilon \\ 0 & 1 & \cdots & 0 & \vdots \\ 0 & 0 & \ddots & 0 & \varepsilon \\ 0 & 0 & \ldots & 1 & 1 \end{pmatrix}$$

In this case, $p = n + 1$ so the system is underdetermined.

Suppose that μ is constructed with

$$\beta = \varepsilon^{-1} \begin{pmatrix} 0 \\ \vdots \\ 0 \\ -1 \\ 1 \end{pmatrix} \quad \text{so that} \quad \mu = \begin{pmatrix} 1 \\ \vdots \\ 1 \\ 1 \\ 0 \end{pmatrix}$$

In the noiseless setting:

- The ℓ_0 solution is exactly β
- The ℓ_1 solution is

$$\hat{\beta} = \begin{pmatrix} 1 \\ \vdots \\ 1 \\ 0 \\ 0 \end{pmatrix} \neq \beta$$

  in the case where $\varepsilon < 2/(n-1)$. So ℓ_1 does not recover the right solution.

Next we will consider the *noisy setting*.

With no noise, the lasso does not find the sparse solution. So with noise, there is little doubt that it will not be able to find the correct variables either. Now this may not be a problem as long as we care only about the prediction error, i.e. $\hat{\beta}$ may be a very poor estimate of β but $X\hat{\beta}$ may still be a good estimate of $X\beta$. This is, however, not the case here.

Suppose ε is small as above, and pick a noise level $\sigma = 1/10$, say. Hence, we observe

$$y_i = 1(i \neq n) + 0.1\, z_i, \quad i = 1, \ldots, n.$$

It can be shown for all reasonable values of λ, the Lasso solution is given by soft-thresholding y, i.e.

$$\hat{\beta}_i = (y_i - \lambda\sigma)_+, \quad i = 1, \ldots, n-1$$
$$\hat{\beta}_i = 0, \quad i = n, n+1$$

and that

$$\hat{\mu}_i = (y_i - \lambda\sigma)_+, \quad i = 1, \ldots, n-1$$
$$\hat{\mu}_i = 0, \quad i = n$$

Consequently, the lasso has a squared bias of roughly $(n-1)\lambda^2\sigma^2$ and a variance of $(n-1)\sigma^2$. We'd be better off by just plugging $\hat{\mu} = y$ and only pay the variance.

The empirical results under a noisy setting are shown in Figure 3.

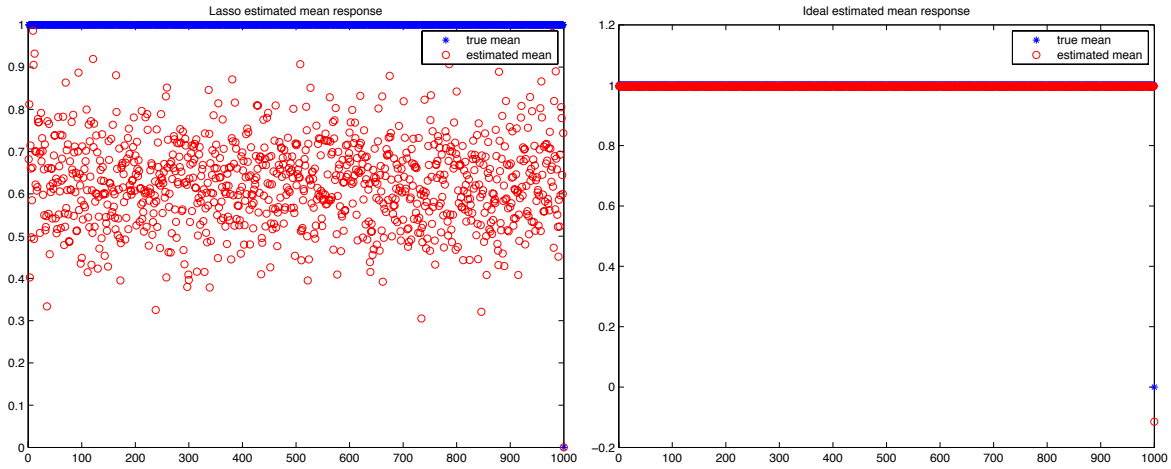


Figure 3: Graphical representation of the example above with $n = 1,000$, $p = 1,001$ and $\sigma = 0.1$. The lasso estimate fits $X\hat{\beta} = \hat{\mu} = (y - 0.1\lambda)_+$, which results in a large variance and a large bias. The ideal estimate uses only two predictors. The ratio between the lasso MSE and the ideal MSE is over $6,500 \approx 6.5n \approx 6.5p$.

Summary: This is an instance where the ℓ_1 solution is irreparably worse than the ℓ_0 solution. Indeed, with ℓ_0 we achieve

$$\mathbb{E}\|X\hat{\beta} - X\beta\|^2 \approx 2\sigma^2$$

while for the Lasso (no matter λ),

$$\mathbb{E}\|X\hat{\beta} - X\beta\|^2 \gg n\sigma^2$$

Take-away message: Even in the noiseless case, correlation between the predictors cause severe problems for the Lasso.

Following we will show another case where the Lasso does not work well even when the correlation between columns of X is very low.

Case study 2: $X = [I_n \ F_{2:n}]$, where $F_{2:n}$ is the discrete cosine transform matrix excluding the first column (the DC component). Specifically,

$$F_{2:n} = \begin{pmatrix} \varphi_1(1) & \varphi_2(1) & \cdots & \varphi_{n-1}(1) \\ \varphi_1(2) & \varphi_2(2) & \cdots & \varphi_{n-1}(2) \\ \vdots & \vdots & \cdots & \vdots \\ \varphi_1(n) & \varphi_2(n) & \cdots & \varphi_{n-1}(n) \end{pmatrix},$$

where

$$\varphi_{2k-1}(t) = \sqrt{2/n}\cos(2\pi kt/n), \quad k = 1, 2, \cdots, n/2 - 1$$
$$\varphi_{2k}(t) = \sqrt{2/n}\sin(2\pi kt/n), \quad k = 1, 2, \cdots, n/2 - 1$$
$$\varphi_{n-1}(t) = (-1)^t/\sqrt{n}.$$

In this case, $p = 2n - 1$. The maximum correlation between columns of X, $\mu(X) = \sqrt{2/n}$, which shows that the columns of X are extremely incoherent. Besides, $\|x\|^2 \leq 2$.

Then we can construct $f = X\beta$, of which each coordinate is 1, with a sparse β (shown in Figure 4).

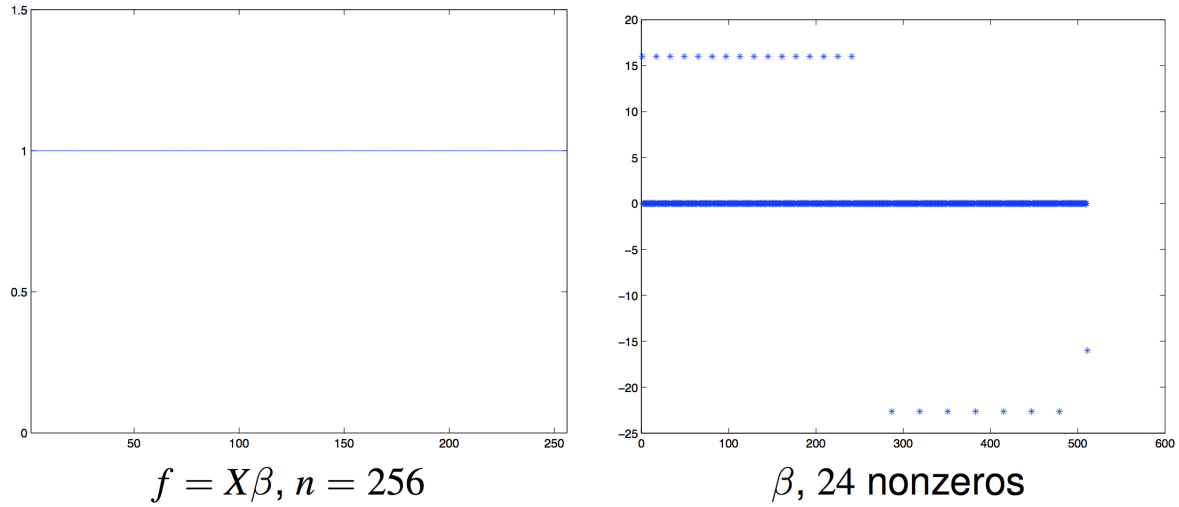


Figure 4: Graphical representation of $f = X\beta$ and β in case study 2 with n = 256, p = 511.

In noisy setting, the observed noisy $X\beta + z$, the Lasso estimate $\hat{\beta}$ and $X\hat{\beta}$ are shown in Figure 5.

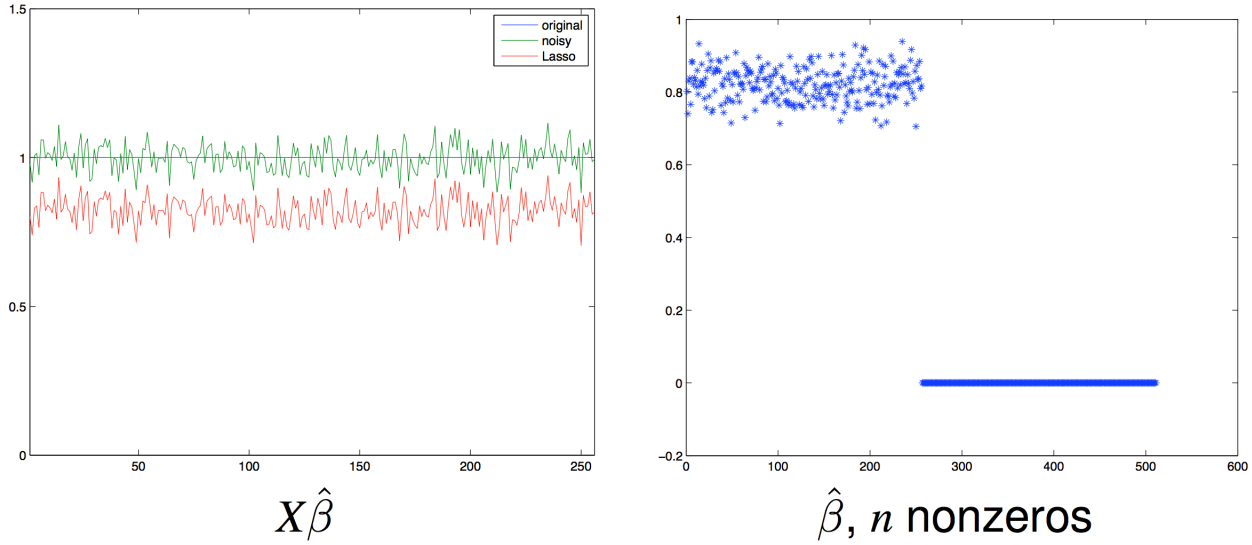


Figure 5: Graphical representation of the results in case study 2 with n = 256, p = 511. The lasso estimate $\hat{\beta}$ has far more non-zero entries than β and results in large MSE.

Empirically and provably,

$$X\hat{\beta} = y - \lambda\sigma\mathbf{1}, \quad \hat{\beta}_i = \begin{cases} y_i - \lambda\sigma, & i \in \{1, \cdots, n\}, \\ 0, & i \in \{n+1, \cdots, 2n-1\}. \end{cases}$$

The Lasso MSE, $\|\hat{f} - f\|^2 \sim n\sigma^2(1 + \lambda^2)$, is horrible. The l_0 norm of $\hat{\beta}$, $|\text{supp}(\hat{\beta})| = n$, is much higher than that of β, $|\text{supp}(\beta)| = \frac{3}{2}\sqrt{n}$.