Insights, resources, and application of "Bolasso"

BY YALIN LIAO

I spend most time reading and understanding the paper [1], which we may refer to as Bolasso for convenient. So I will expand my note based Bolasso. But first I will point out the focus of Bolasso is different from that of Terence Tao's paper [2]. They investigate the underdetermined system

$$y = \beta X + z$$

where β is the model parameter, X is the data matrix with far fewer rows than columns, $n \ll p$ (nonasymptotic), and $z \sim \mathcal{N}(0, \sigma I)$, and answer if it's possible to estimate β reliably based n observations y. The answer is yes by solving a simple convex program if the model parameter is S sparse and the data matrix X obeys the uniform uncertainty principle. However, Bolasso studies the asymptotic case, i.e., whether Lasso can recover the sparse solution (or sign pattern) as $n \to \infty$, and there is no assumption like the sparsity of β. Another difference is that Lasso is studied as a tool for selecting features (which should enter the model), not its ability to find the exact true model β.

The Bollaso first reviews the model consistency using different decay rates in Lasso. Sepically, if μ_n tends to zero slower than $n^{-1/2}$, like $\mu_n = n^{-1/3}$,[1] then the estimator $\hat{w}$ converges in probability to w[2]. It seems we get pretty good result. **Why do we care about the model consistency in terms of the sign pattern**? Imagine we have an estimator $\hat{w}$ based on n (very large) samples, satisfying $\|\hat{w} - w\| < \epsilon$ for some $0 < \epsilon < 1$.[3] But $\text{sign}(\hat{w}_i) \neq \text{sign}(w_i)$ for some component index i and the scale of correponding feature x_i of some data points is large. In this case, **although the estimated model $\hat{w}$ is close to the true w (satisfy the model consistency) but the prediction of the estimated model $\hat{w}$ can be largely deviated from the output of the true model w.**[4] This is a disaster.

So the sign pattern consistency is indeed needed. For the decay rate $\mu_n = \mu_0 n^{-1/2}$ for $\mu_0 \in (0, 1)$, they show that Lasso selects all the variables that should enter the model with probability tending to one, expentially fast, while it selects all other variables with strictly positive probability in $(0, 1)$. For convenience, we rewrite the propositions here,[5]

Proposition 1. For any sign patterns $s \in \{-1, 0, 1\}^n$ such that $s_J = \text{sign}(w_J)$, we have $\mathbb{P}(\text{sign}(\hat{w}) = s)$ tends to a limit $\rho(s, \mu_0)$. Particularly,

$$\mathbb{P}(\text{sign}(\hat{w}) = s) = \rho(s, \mu_0) + O(n^{-1/2} \log n)$$

or

$$\mathbb{P}(\text{sign}(\hat{w}_J) = \text{sign}(w_J) = s_J) = \rho(s, \mu_0) + O(n^{-1/2} \log n)$$

Proposition 2. For any sign patterns $s \in \{-1, 0, 1\}^n$ such that $s_J \neq \text{sign}(w_J)$, there exist a constant $A(\mu_0) > 0$ such that

$$\mathbb{P}(\text{sign}(\hat{w}) = s) \leq e^{-nA(\mu_0) + O(n^{-1/2})}$$

1. $n^{-1/3} \to 0$ as $n \to \infty$ and $n^{-1/3}/n^{-1/2} = \frac{n^{1/2}}{n^{1/3}} \to \infty$ as $n \to \infty$, so $\mu_n = n^{-1/3}$ tends to zero slower than $n^{-1/2}$.

2. In Bollaso, w denotes the true model, i.e., $y = Xw$ and the estimator based on finite samples is denoted by $\hat{w}$.

3. We could have this estimator based the model consistency in Lasso.

4. **Particularly, when the scale of some input points are very large in some dimensions where the sign pattern of estimated model does not match that of true model, the predictions of the estimated model are very bad for these data points.**

5. The assumptions are ignored.

or

$$\mathbb{P}(\text{sign}(\hat{w}_J) = \text{sign}(w_J) \neq s_J) \geq 1 - e^{-nA(\mu_0)+O(n^{-1/2})}$$

If don't read them carefully, you may think these two propositions describe the same thing[6] but the conclusions are contradictory. The lower bound of $\mathbb{P}(\text{sign}(\hat{w}) = s)$ for any $s_J = \text{sign}(w_J)$ tends to one while $\mathbb{P}(\text{sign}(\hat{w}) = s) \approx \rho(s, \mu_0) \in (0, 1)$ but it's impossible tends to one as $\rho(s, \mu_0)$ is indepent of n. Is any thing wrong? I think the reason is the probability estimation $\rho(s, \mu_0) + O(n^{-1/2} \log n)$ depends on s in proposition 1 while the lower bound in proposition 2 does not. How do we understand or interprete these two positions? The propostion 2 is saying, Lasso can select all relevant features X_J which should enter the model with probability tending to one, exponentially fast. This means **Lasso algorithm gennerally will not miss all relevant features** by the decay rate $\mu_n = \mu_0 n^{-1/2}$. The proposition 1 implies that for any bad sign pattern s such that $s_J = \text{sign}(w_J)$ and $s_{J^c} \neq \text{sign}(w_{J^c})$[7], if n tends to infinity, with a strictly postive probability $\rho(s, \mu_0) \in (0, 1)$ Lasso will return a model whose sign pattern is identical to s. In other word, **Lasso has a strictly positive probability to select redudant features except the relevant features**. This is the motivation to apply the bootstrap trick in the paper. The intuition is obvious: if runing Lasso once it's easy to select redudant features why not run it mutiple times and choose the intersection of these models? **Because it's much less likely to multiple models selecting redudant features simutaneously**. This is consistent with what is stated in propsition 3, the Bollaso does not exactly select the correct model, i.e., $\mathbb{P}(\hat{J} \neq J)$, has the following upper bound:

$$\mathbb{P}(\hat{J} \neq J) \leq \frac{mA_1}{e^{A_2 n}} + A_3 \frac{\log(n)}{n^{1/2}} + A_4 \frac{\log(m)}{m}$$

where A_1, A_2, A_3, A_4 are strictly positive constants. When m, n are large, $A_3 \frac{\log(n)}{n^{1/2}} + A_4 \frac{\log(m)}{m}$ tend to zero. For the first term, $e^{A_2 n}$ dominate its value unless $m \gg n$. This proposition also implies that generally Bollaso will not increase the risk to miss all relevant features though it decreases the probability to select redudant features, which cannot be seen from our intuition understanding.

Naturally, can we generalize these results to kernel Lasso regression? The question is not meaningful since the kernel trick depends on L_2 norm (induced by inner product).

Bibliography

[1] Francis R Bach. Bolasso: model consistent lasso estimation through the bootstrap. In *Proceedings of the 25th international conference on Machine learning*, pages 33–40. 2008.

[2] Emmanuel Candes and Terence Tao. The dantzig selector: statistical estimation when p is much larger than n. *The annals of Statistics*, 35(6):2313–2351, 2007.

6. If re-stating the proposition 2 as follows: for any sign patterns $s \in \{-1, 0, 1\}^n$ such that $s_J = \text{sign}(w_J)$, there exist a constant $A(\mu_0) > 0$ such that

$$\mathbb{P}(\text{sign}(\hat{w}) = s) \geq 1 - e^{-nA(\mu_0)+O(n^{-1/2})}$$

you will see they describe the same thing.

7. $\text{sign}(w_{J^c}) = 0$