
Supplementary Material for “Predicting Execution Time of Computer Programs Using Sparse Polynomial Regression”

1 Proof of Theorem 3.1

Let S be a subset of $\{1, 2, \dots, p\}$ and its complement $S^c = \{1, 2, \dots, p\} \setminus S$. Write the feature matrix X as $X = [X(S), X(S^c)]$. Let response $Y = f(X(S)) + \epsilon$, where $f(\cdot)$ is any function and ϵ is additive noise. Let n be the number of observations and s the size of S . We assume that X is deterministic, p and s are fixed, and ϵ_i i.i.d. follows the Gaussian distribution with mean 0 and variance σ^2 . Our results also hold for zero mean sub-Gaussian noise with parameter σ^2 . More general results regarding general scaling of n, p and s can also be obtained.

Recall that the LASSO is defined as

$$\hat{\beta} = \frac{1}{2} \|Y - X\beta\|_2^2 + \lambda \|\beta\|_1. \quad (1)$$

Under the following conditions, we show that Step 1 of SPORE-LASSO, the linear LASSO, selects the relevant features even if the response Y depends on predictors $X(S)$ nonlinearly:

1. The columns $(X_j, j = 1, \dots, p)$ of X are standardized: $\frac{1}{n} X_j^T X_j = 1$, for all j .
2. $\Lambda_{\min}(\frac{1}{n} X_S^T X_S) \geq c$ with a constant $c > 0$;
3. $\min |(X_S^T X_S)^{-1} X_S^T f(X_S)| > \alpha$ with a constant $\alpha > 0$;
4. $\frac{X_{S^c}^T [I - X_S (X_S^T X_S)^{-1} X_S^T] f(X_S)}{n} < \frac{\eta \alpha c}{2\sqrt{s+1}}$, for some $0 < \eta < 1$;
5. $\|X_{S^c}^T X_S (X_S^T X_S)^{-1}\|_\infty \leq 1 - \eta$;

where $\Lambda_{\min}(\cdot)$ denotes the minimum eigen value of a matrix, $\|A\|_\infty$ is defined as $\max_i \left[\sum_j |A_{ij}| \right]$ and the inequalities are defined element-wise.

By standard convex optimization theory, if $\hat{\beta} = (\hat{\beta}_S, \hat{\beta}_{S^c})$ with $\hat{\beta}_S \neq 0$ and $\hat{\beta}_{S^c} = 0$ satisfies

$$X_S^T (Y - X_S \hat{\beta}_S) = \lambda \text{sgn}(\hat{\beta}_S), \quad (2)$$

$$|X_{S^c}^T (Y - X_S \hat{\beta}_S)| < \lambda, \quad (3)$$

then it is the unique solution of the LASSO (1).

From Equation (2), we get

$$\hat{\beta}_S = (X_S^T X_S)^{-1} X_S^T f(X_S) + (X_S^T X_S)^{-1} [X_S^T \epsilon - \lambda \text{sgn}(\hat{\beta}_S)]. \quad (4)$$

Let \vec{b} be the the sign vector of $(X_S^T X_S)^{-1} X_S^T f(X_S)$. Set $\text{sgn}(\hat{\beta}_S) = \vec{b}$, substitute it into equation (4), and then we have

$$\hat{\beta}_S = (X_S^T X_S)^{-1} X_S^T f(X_S) + (X_S^T X_S)^{-1} [X_S^T \epsilon - \lambda \vec{b}]. \quad (5)$$

It can be verified that if

$$\max \left| (X_S^T X_S)^{-1} [X_S^T \epsilon - \lambda \vec{b}] \right| < \alpha, \quad (6)$$

then $\hat{\beta}_S$ defined in Equation (5) satisfies Equation (2).

Substitute $\hat{\beta}_S$ with (5) into Inequality (3), we get

$$\begin{aligned} & |X_{S^c}^T [f(X_S) - X_S (X_S^T X_S)^{-1} X_S^T f(X_S)] \\ & \quad + X_{S^c}^T [I - X_S (X_S^T X_S)^{-1} X_S^T] \epsilon \\ & \quad + \lambda X_{S^c}^T X_S (X_S^T X_S)^{-1} \vec{b} | < \lambda. \end{aligned} \quad (7)$$

By assumption,

$$|X_{S^c}^T X_S (X_S^T X_S)^{-1} \vec{b}| \leq 1 - \eta,$$

so,

$$|X_{S^c}^T [f(X_S) - X_S (X_S^T X_S)^{-1} X_S^T f(X_S)]| + |X_{S^c}^T [I - X_S (X_S^T X_S)^{-1} X_S^T] \epsilon| < \lambda \eta / 2 \quad (8)$$

is sufficient for Inequality (3).

According to the previous discussion, it suffices to prove that (6) and (8) hold with probability $\rightarrow 1$ as $n \rightarrow \infty$.

We analyze (8) first. $X_{S^c}^T [I - X_S (X_S^T X_S)^{-1} X_S^T] \epsilon$ is a Gaussian random vector with mean 0 and variance of each element at most $n\sigma^2$. So,

$$P[\max |X_{S^c}^T [I - X_S (X_S^T X_S)^{-1} X_S^T] \epsilon| > t] \leq 2(p-s) \exp \left\{ -\frac{t^2}{2n\sigma^2} \right\}.$$

Setting $t = \frac{\lambda \eta}{2} - |X_{S^c}^T [f(X_S) - X_S (X_S^T X_S)^{-1} X_S^T f(X_S)]|$, we obtain that

$$P[(8) \text{ holds}] \geq 1 - 2(p-s) \exp \left\{ -\frac{(|X_{S^c}^T [f(X_S) - X_S (X_S^T X_S)^{-1} X_S^T f(X_S)]| - \frac{\lambda \eta}{2})^2}{2n\sigma^2} \right\}.$$

Take

$$\lambda = \frac{2}{\eta} \left\{ |X_{S^c}^T [f(X_S) - X_S (X_S^T X_S)^{-1} X_S^T f(X_S)]| + \kappa \sqrt{n} \log n \right\}, \quad (9)$$

where κ is a constant. It is easy to see that the above probability goes to 1. From Condition 4., λ has the property that $\lambda/n \leq \frac{\alpha c}{\sqrt{s+1}}$ as $n \rightarrow \infty$.

Now we analyze (6). We have $|(X_S^T X_S)^{-1} [X_S^T \epsilon - \lambda \vec{b}]| \leq |(X_S^T X_S)^{-1} X_S^T \epsilon| + |(X_S^T X_S)^{-1} \lambda \vec{b}|$. Since $\|(X_S^T X_S)^{-1}\|_2 \leq \frac{1}{nc}$, we have the variance of each element of Gaussian vector $|(X_S^T X_S)^{-1} [X_S^T \epsilon]|$ at most $\frac{\sigma^2}{nc}$.

So

$$P[\max |(X_S^T X_S)^{-1} [X_S^T \epsilon]| > t] \leq 2s \exp \left\{ -\frac{nct^2}{2\sigma^2} \right\}.$$

$$|(X_S^T X_S)^{-1} \lambda \vec{b}| \leq \frac{\sqrt{s}\lambda}{nc}.$$

Set $t^2 = \frac{1}{\sqrt{n}}$ and set λ such that $\frac{\sqrt{s}\lambda}{nc} < \alpha$ (note that the previous choice of λ in Equation (9) satisfies this requirement), then (6) holds with probability greater than $1 - 2s \exp\{-\frac{c\sqrt{n}}{2\sigma^2}\} \rightarrow 1$.

2 Full version of SPORE-FoBa algorithm

Algorithm 1 SPORE-FoBa

Input: data $(x_i, y_i), i = 1, \dots, n$, the maximum degree d, ϵ

Output: polynomial terms $T^{(k)}$ and the coefficients $\beta^{(k)}$.

```
1: Let  $T^{(0)} = \emptyset, S^{(0)} = \emptyset$ 
2: let  $k = 0$  (number of terms)
3: let  $RSS^{(0)} = \sum_i y_i^2$ 
4: while True do
5:    $RSSJ = \|Y - T^{(k)}\beta^{(k)}\|_2^2$ 
6:   for  $j = 1, \dots, p$  do
7:     let  $C = \{t : t = x_j^{d_1} \prod_{l \in S} x_l^{d_l} \text{ with } d_1 > 0, d_l \geq 0, d_1 + \sum d_l \leq d\}$ 
8:     // Forward step: add terms from  $C$ 
9:     while True do
10:      let  $k = k + 1$ 
11:      let  $[t^{(k)}, \beta^{(k)}] = \arg \min_{t \in C, \beta} \|Y - [T^{(k-1)}, t]\beta\|_2^2$ 
12:      let  $RSS^{(k)} = \|Y - [T^{(k-1)}, t^{(k)}]\beta^{(k)}\|_2^2$ 
13:      let  $\delta^{(k)} = RSS^{(k-1)} - RSS^{(k)}$ 
14:       $T^{(k)} = T^{(k-1)} \cup t^{(k)}$ 
15:      if  $\delta^{(k)} \leq \epsilon$  then
16:         $k = k - 1$ 
17:        break
18:      end if
19:      // backward step: remove terms from active set  $T^{(k)}$ 
20:      while True do
21:         $RRS_{pre} = RRS^{(k)}$ 
22:        let  $[t, \beta_{now}] = \arg \min_{t \in T^k, \beta} \|Y - [T^{(k)} \setminus t]\beta\|_2^2$ 
23:        let  $RSS_{now} = \|Y - [T^{(k)} \setminus t]\beta_{now}\|_2^2$ 
24:         $\delta' = RSS_{now} - RRS_{pre}$ 
25:        if  $\delta' > 0.5\delta^{(k)}$  then
26:          break
27:        end if
28:        let  $k = k - 1$ 
29:        let  $T^{(k)} = T^{(k+1)} \setminus \{t\}$ 
30:        let  $\beta^{(k)} = \beta_{now}$ 
31:        let  $RSS^{(k)} = RSS_{now}$ 
32:      end while
33:    end while
34:    if Feature  $j$  is added into the active set  $T^{(k)}$  then
35:       $S = S \cup j$ 
36:    end if
37:  end for
38:  if  $RSS^{(k)} - RSSJ \leq \epsilon$  then
39:    break
40:  end if
41: end while
```
