On Model Selection Consistency Of Lasso

Yewon Kim

12/08/2015

Introduction

Model selection is a commonly used method to find sparsity or parsimony of statistical models, but usually involves a computationally heavy combinatorial search. Lasso (Tibshirani, 1996) is now being used as a computationally feasible alternative to model selection.

In this paper, they prove that a single condition, which they call the Irrepresentable Condition, is almost necessary and sufficient for Lasso to select the true model both in the classical fixed p setting and in the large p setting as the sample size n gets large.

Some previous results

- Knight and Fu(2000) have shown estimation consistency for Lasso for fixed p and fixed β^n
- Meinshausen and Buhlmann(2006) have shown that Lasso is consistent in estimating the dependency between Gaussian variables even when p grows faster than n
- Zhao and Yu(2006) have show model selection consistency for both fixed p and large p problems

Definition

Suppose the linear regression model:

$$Y_n = X_n \beta^n + \epsilon_n$$

Where, Y_n is a n x 1 response vector,
$X_n = (X_1^n, X_2^n, ..., X_p^n) = ((x_1)^T, (x_2)^T, ..., (x_n)^T)$ is a n x p design matrix, β^n is a p x 1 vector of model coefficients. ϵ_n is a i.i.d random error variables with mean 0 and variance σ^2

Lasso estimator is :

$$\hat{\beta}^n(\lambda) = argmin_\beta(||Y_n - X_n\beta||_2^2 + \lambda||\beta||_1)$$

with $\lambda \geq 0$

Notation

$$\beta^n = (\beta_1^n, \beta_2^n, .., \beta_q^n, \beta_{q+1}^n, ..., \beta_p^n)^T$$

Suppose, $\beta_j^n \neq 0$ for j=1,2,..,q and $\beta_j^n = 0$ for j=q+1,...,p

$$\beta_{(1)}^n = (\beta_1^n, ..., \beta_q^n), \beta_{(2)}^n = (\beta_{q+1}^n, ..., \beta_p^n)$$

$$X_n(1) = (X_1^n, ..., X_q^n), X_n(2) = (X_{q+1}^n, ..., X_p^n),$$

$$C^n = \frac{1}{n} X_n^T X_n = \begin{pmatrix} C_{11}^n & C_{12}^n \\ C_{21}^n & C_{22}^n \end{pmatrix}$$

where $C_{11}^n = \frac{1}{n} X_n(1)^T X_n(1), ..., C_{22}^n = \frac{1}{n} X_n(2)^T X_n(2)$

Definition of Consistency

- Estimation : $\hat{\beta}^n - \beta^n \to_p 0$, as n$\to \infty$
- Model selection : $P([i : \hat{\beta}_i^n \neq 0] = [i : \beta_i^n \neq 0]) \to_p 1$, as n$\to \infty$
- Sign : $P(\hat{\beta}^n =_s \beta^n) \to_p 1$, as n$\to \infty$, where $\hat{\beta}^n =_s \beta^n \Leftrightarrow \text{sign}(\hat{\beta}^n) = \text{sign}(\beta^n)$

Definition 1

Strongly Sign Consistent

Lasso is Strongly Sign Consistent if $\exists \lambda_n = f(n)$ such that
$\lim_{n \to \infty} P(\hat{\beta}^n(\lambda_n) =_s \beta^n) = 1$

General Sign Consistent

Lasso is General Sign Consistent if
$\lim_{n \to \infty} P(\exists \lambda \geq 0, \hat{\beta}^n(\lambda_n) =_s \beta^n) = 1$

Definition 2

Strong Irrepresentable Condition
$\exists \eta > 0$, such that $|C_{21}^n (C_{11}^n)^{-1} \text{sign}(\beta_{(1)}^n)| \leq 1 - \eta$

Weak Irrepresentable Condition
$|C_{21}^n (C_{11}^n)^{-1} \text{sign}(\beta_{(1)}^n)| < 1$

Result-Small p and q

Classical setting: p,q and β^n are all fixed as n$\to \infty$

Suppose the following regularity conditions:
$C^n \to C > 0$, as $n \to \infty$
$\frac{1}{n} max_{1 \le i \le n}((x_i^n)^T x_i^n) \to 0, n \to \infty$

Result-Small p and q

Theorem 1

For fixed p,q and $\beta^n = \beta$, under the previous assumptions, Lasso is strongly sign consistent if Strong Irrepresentable Condition holds. That is, when Strong Irrepresentable Condition holds, $\forall \lambda_n$ that satisfies $\frac{\lambda_n}{n} \to 0$ and $\frac{\lambda_n}{n^{\frac{1+c}{n}}} \to \infty$ with $0 \leq c < 1$, we have

$$P(\hat{\beta}^n(\lambda_n) =_s \beta^n) = 1 - o(e^{-n^c})$$

Result-Small p and q

Theorem 2

For fixed p,q and $\beta^n = \beta$, under the previous assumptions, Lasso is general sign consistent only if there exists N so that Weak Irrepresentable Condition holds for n >N

Result-Small p and q

Therefore, Strong Irrepresentable Condition implies strong sign consistency implies general sign consistency implies Weak Irrepresentable Condition. So except for the technical difference between the two conditions, Irrepresentable Condition is almost necessary and sufficient for both strong sign consistency and general sign consistency.

Result-Large p and q

Furthermore, under additional regularity conditions on the noise terms ϵ_i^n, this small p result can be extended to the large p case. That is, when p also tends to infinity not too fast as n tends to infinity, we show that Strong Irrepresentable Condition, again, implies Strong Sign Consistency for Lasso.

The dimension of the designs C^n and parameters β^n grow as n grows, then p_n and q_n are allowed to grow with n

Suppose the following conditions: $\exists\, 0 \leq c_1 < c_2 \leq 1$ and $M_1, M_2, M_3, M_4 > 0$,
$\frac{1}{n}(X_i^n)^T X_i^n \leq M_1$, for $\forall i$, $\alpha^T C_{11}^n \alpha \geq M_2$, for $\forall\, ||\alpha||_2^2 = 1$
$q_n = O(n^{c_1})$,
$n^{\frac{1-c_2}{2}}, \min_{i=1,2,..,q}|\beta_i^n| \geq M_3$

Theorem 3

Assume ϵ_i^n are i.i.d. random variable with $E(\epsilon_i^n)^{2k} < \infty$ for an integer k >0. Under the previous conditions, Strong Irrepresentable Condition implies that Lasso has strong sign consistency for $p_n = o(n^{(c_2-c_1)k})$. In particular, for $\forall \lambda_n$ that satisfies, $\frac{\lambda_n}{\sqrt{n}} = o(n^{(c_2-c_1)/2})$ and $\frac{1}{p_n}(\frac{\lambda_n}{\sqrt{n}})^{2k} \to \infty$, we have

$$P(\hat{\beta}^n(\lambda_n) =_s \beta^n) \geq 1 - O(\frac{p_n n^k}{\lambda^{2k}}) \to 1 \text{ as } n \to \infty$$

Theorem 4

Assume ϵ_i^n are i.i.d. Gaussian random variables. Under the previous conditions, if there exists $0 < c_3 < c_2 - c_1$ for which $p_n = O(e^{n^{c_3}})$, then Strong Irrepresentable Condisions implies that Lasso has strong sign consistency. In particular, for $\lambda_n \propto n^{\frac{1+c_4}{2}}$ with $c_3 < c_4 < c_2 - c_1$, $P(\hat{\beta}^n(\lambda_n) =_s \beta^n) \geq 1 - o(e^{n^{c_3}}) \to 1$ as $n \to \infty$

Discussions

In this paper, they have provided Strong and Weak Irrepresentable Conditions that are almost necessary and sufficient for model selection consistency of Lasso under both small p and large p settings. Although much of Lasso's strength lies in its finite sample performance which is not the focus here, their asymptotic results offer insights and guidance to applications of Lasso as a feature selection tool, assuming that the typical regularity conditions are satisfied on the design matrix as in Knight and Fu (2000)

References

📄 Peng Zhao, Bin Yu
On Model Selection Consistency of Lasso
Journal of Machine Learing Research 7 (2006) 2541-2563

📄 Jinhu Jia, Karl Rohe
Preconditioning To Comply With The Irrepresentable Condition
math.ST 28 Aug 2012

The End