On Model Selection Consistency of Lasso

Ganchao Wei

December 8, 2021

Overview

1. Introduction
2. Model Selection Consistency and Irrepresentable Conditions
3. Model Selection Consistency for Small q and p
4. Model Selection Consistency for Large q and p
5. Analysis and Sufficient Conditions for Strong Irrepresentable Condition
6. Simulation Studies

Introduction

The Lasso estimates $\hat{\beta}^n$ are defined by:

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

However, if an irrelevant predictor is highly correlated with the predictors in the true model, Lasso may not be able to distinguish it from the true predictors, with any amount of data and any amount of regularization. For example:

- For a fixed p and orthogonal designs, the optimal (i.t.o. parameter estimation) Lasso doesn't give consistent model selection
- When using Lasso for knot selection in spline regression, it tend to pick up knots in close proximity to one another.

Introduction

In this paper, they consider model selection consistency of the Lasso. Specifically, they focus on two problems:

- Whether there exists a deterministic amount of regularization that gives consistent selection.
- Whether there exists a correct amount of regularization that selects the true model, for each random realization.

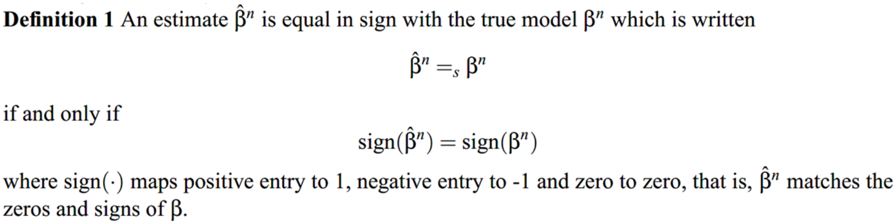
They found there exists an **Irrepresentable Condition** that is almost necessary and sufficient for both types of consistency (estimation consistency + model selection consistency)

Model Selection Consistency and Irrepresentable Conditions

Two consistencies:

- Parameter estimation consistency: $\hat{\beta}^n - \beta^n \to_p 0$, as $n \to \infty$
- Model selection consistency: $P(\{i : \hat{\beta}^n \neq 0\} = \{i : \beta^n \neq 0\}) \to 1$, as $n \to \infty$

To separate 2 consistencies, define the sign consistency that doesn't assume the estimates to be estimation consistent:



Definition 1 An estimate $\hat{\beta}^n$ is equal in sign with the true model β^n which is written

$$\hat{\beta}^n =_s \beta^n$$

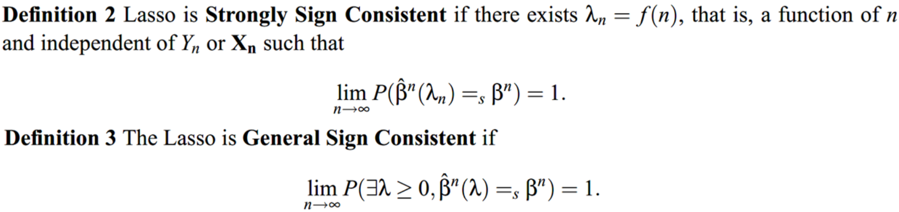
if and only if

$$\mathrm{sign}(\hat{\beta}^n) = \mathrm{sign}(\beta^n)$$

where $\mathrm{sign}(\cdot)$ maps positive entry to 1, negative entry to -1 and zero to zero, that is, $\hat{\beta}^n$ matches the zeros and signs of β.

Model Selection Consistency and Irrepresentable Conditions

2 sign consistencies for Lasso, depending on how the amount of regularization is determined:



Definition 2 Lasso is **Strongly Sign Consistent** if there exists $\lambda_n = f(n)$, that is, a function of n and independent of Y_n or $\mathbf{X_n}$ such that

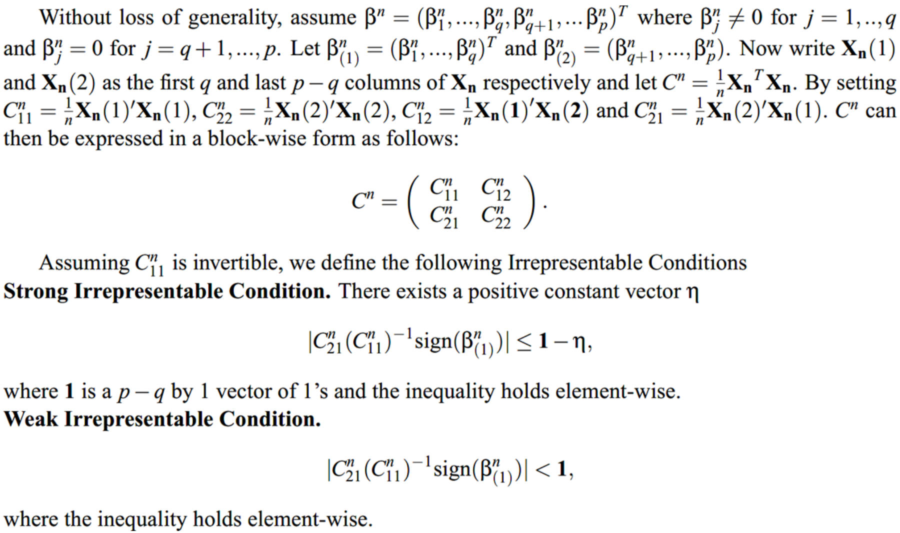
$$\lim_{n\to\infty} P(\hat{\beta}^n(\lambda_n) =_s \beta^n) = 1.$$

Definition 3 The Lasso is **General Sign Consistent** if

$$\lim_{n\to\infty} P(\exists\lambda \geq 0, \hat{\beta}^n(\lambda) =_s \beta^n) = 1.$$

We further partition the design matrix, based on the sparsity of parameters. Then we can define the strong/ weak irrepresentable condition (next page).

Model Selection Consistency and Irrepresentable Conditions



Without loss of generality, assume $\beta^n = (\beta^n_1, ..., \beta^n_q, \beta^n_{q+1}, ... \beta^n_p)^T$ where $\beta^n_j \neq 0$ for $j = 1, .., q$ and $\beta^n_j = 0$ for $j = q+1, ..., p$. Let $\beta^n_{(1)} = (\beta^n_1, ..., \beta^n_q)^T$ and $\beta^n_{(2)} = (\beta^n_{q+1}, ..., \beta^n_p)$. Now write $\mathbf{X_n}(1)$ and $\mathbf{X_n}(2)$ as the first q and last $p-q$ columns of $\mathbf{X_n}$ respectively and let $C^n = \frac{1}{n}\mathbf{X_n}^T\mathbf{X_n}$. By setting $C^n_{11} = \frac{1}{n}\mathbf{X_n}(1)'\mathbf{X_n}(1)$, $C^n_{22} = \frac{1}{n}\mathbf{X_n}(2)'\mathbf{X_n}(2)$, $C^n_{12} = \frac{1}{n}\mathbf{X_n(1)'X_n(2)}$ and $C^n_{21} = \frac{1}{n}\mathbf{X_n}(2)'\mathbf{X_n}(1)$. C^n can then be expressed in a block-wise form as follows:

$$C^n = \begin{pmatrix} C^n_{11} & C^n_{12} \\ C^n_{21} & C^n_{22} \end{pmatrix}.$$

Assuming C^n_{11} is invertible, we define the following Irrepresentable Conditions

Strong Irrepresentable Condition. There exists a positive constant vector η

$$|C^n_{21}(C^n_{11})^{-1}\text{sign}(\beta^n_{(1)})| \leq \mathbf{1} - \eta,$$

where $\mathbf{1}$ is a $p-q$ by 1 vector of 1's and the inequality holds element-wise.

Weak Irrepresentable Condition.

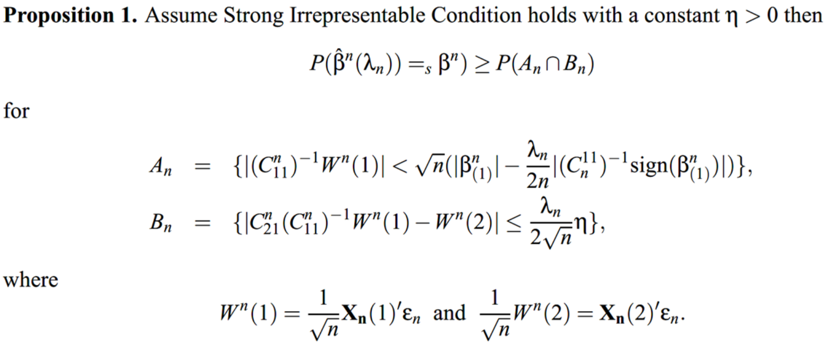
$$|C^n_{21}(C^n_{11})^{-1}\text{sign}(\beta^n_{(1)})| < \mathbf{1},$$

where the inequality holds element-wise.

Model Selection Consistency and Irrepresentable Conditions

We can put a lower bound on the probability of choosing the true model. This is quantitatively relates to:

- The probability of Lasso selecting the correct model.
- How well Strong Irrepresentable Condition holds.



Proposition 1. Assume Strong Irrepresentable Condition holds with a constant $\eta > 0$ then

$$P(\hat{\beta}^n(\lambda_n)) =_s \beta^n) \geq P(A_n \cap B_n)$$

for

$$\begin{aligned} A_n &= \{|(C_{11}^n)^{-1}W^n(1)| < \sqrt{n}(|\beta_{(1)}^n| - \frac{\lambda_n}{2n}|(C_n^{11})^{-1}\text{sign}(\beta_{(1)}^n)|)\}, \\ B_n &= \{|C_{21}^n(C_{11}^n)^{-1}W^n(1) - W^n(2)| \leq \frac{\lambda_n}{2\sqrt{n}}\eta\}, \end{aligned}$$

where

$$W^n(1) = \frac{1}{\sqrt{n}}\mathbf{X_n}(1)'\varepsilon_n \text{ and } \frac{1}{\sqrt{n}}W^n(2) = \mathbf{X_n}(2)'\varepsilon_n.$$

Model Selection Consistency and Irrepresentable Conditions

Some interpretations:

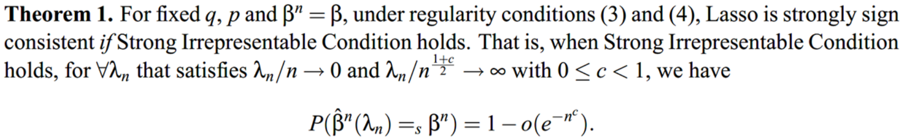
- A_n: the signs of those of $\beta^n_{(1)}$ are estimated correctly.
- Given A_n, B_n further imply $\hat{\beta}^n_{(2)}$ are shrunk to zero.
- λ_n trades off the size of these 2 events: smaller leads to larger A_n but smaller $B_n \Rightarrow$ more likely to have Lasso pick more irrelevant variables.
- larger $\eta \Rightarrow$ easier for Lasso to pick up true model (larger B_n but not impact on A_n)

Model Selection Consistency for Small q and p

We first consider q, p and β^n are fixed as $n \to \infty$. As usual, the regularity conditions:

- $C^n \to C$, as $n \to \infty$, where C is p.d.
- $\frac{1}{n} \max_{1 \leq i \leq n}((x_i^n)^T x_i^n) \to 0$, as $n \to \infty$

Then,



Theorem 1. For fixed q, p and $\beta^n = \beta$, under regularity conditions (3) and (4), Lasso is strongly sign consistent *if* Strong Irrepresentable Condition holds. That is, when Strong Irrepresentable Condition holds, for $\forall \lambda_n$ that satisfies $\lambda_n/n \to 0$ and $\lambda_n/n^{\frac{1+c}{2}} \to \infty$ with $0 \leq c < 1$, we have

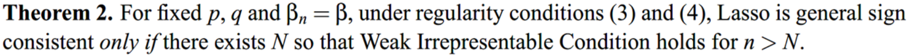
$$P(\hat{\beta}^n(\lambda_n) =_s \beta^n) = 1 - o(e^{-n^c}).$$

Model Selection Consistency for Small q and p

Theorem 1 shows: Strong Irrepresentable Condition + finite second moment of the noise $\Rightarrow$ strong sign consistency at exponentail rate

Moreover, by Knight and Fu(2000): $\lambda_n = o(n) \Rightarrow$ LASSO has consistent estimation and asymptotic normality.

They further show that Weak Irrepresentable Condition is also necessary for the weaker general sign consistency.



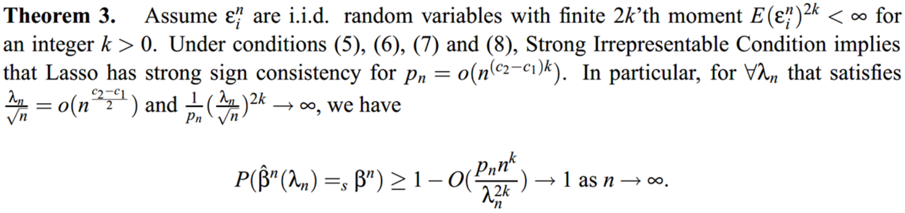
Theorem 2. For fixed p, q and $\beta_n = \beta$, under regularity conditions (3) and (4), Lasso is general sign consistent *only if* there exists N so that Weak Irrepresentable Condition holds for $n > N$.

Model Selection Consistency for Large q and p

Then ,we allow the dimension of the designs C^n and model parameters β_n grow as n grows (i.e. $p = p_n$ and $q = q_n$ are allowed to grow with n). Therefore, we need to modify the regularity conditions a bit:

- $\frac{1}{n}(X_i^n)'X_i^n \leq M_1$ for $\forall i$
- Bound the eigenvalues of C_{11}^n: $\alpha' C_{11}^n \alpha \geq M_2$, for $\forall ||\alpha||_2^2 = 1$
- Sparsity assumption: $q_n = O(n^{c_1})$
- Control the smallest entry of $\beta_{(1)}^n$: $n^{\frac{1-c_2}{2}} \min_{i=1,\dots,q} |\beta_i^n| \geq M_3$

Then we can give the following result:

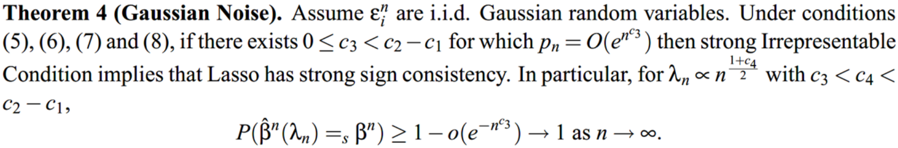


Theorem 3. Assume ε_i^n are i.i.d. random variables with finite $2k$'th moment $E(\varepsilon_i^n)^{2k} < \infty$ for an integer $k > 0$. Under conditions (5), (6), (7) and (8), Strong Irrepresentable Condition implies that Lasso has strong sign consistency for $p_n = o(n^{(c_2 - c_1)k})$. In particular, for $\forall \lambda_n$ that satisfies $\frac{\lambda_n}{\sqrt{n}} = o(n^{\frac{c_2 - c_1}{2}})$ and $\frac{1}{p_n}(\frac{\lambda_n}{\sqrt{n}})^{2k} \to \infty$, we have

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

Model Selection Consistency for Large q and p

In particular, for Gaussian noise:



Theorem 4 (Gaussian Noise). Assume ε_i^n are i.i.d. Gaussian random variables. Under conditions (5), (6), (7) and (8), if there exists $0 \leq c_3 < c_2 - c_1$ for which $p_n = O(e^{n^{c_3}})$ then strong Irrepresentable Condition implies that Lasso has strong sign consistency. In particular, for $\lambda_n \propto n^{\frac{1+c_4}{2}}$ with $c_3 < c_4 < c_2 - c_1$,

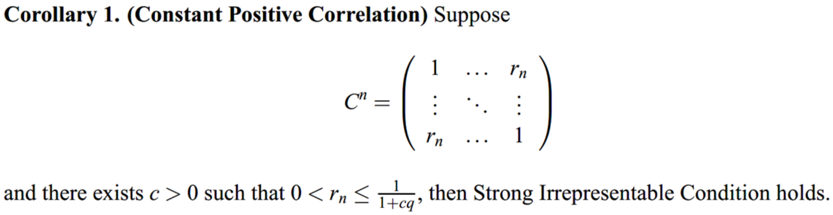
$$P(\hat{\beta}^n(\lambda_n) =_s \beta^n) \geq 1 - o(e^{-n^{c_3}}) \to 1 \text{ as } n \to \infty.$$

This is interesting: for Gaussian noise, p can grow faster than n (up to exponentially), while still allow for fast convergence of the probability of correct model selection to 1.

This is not the case for all noise distributions.

Analysis and Sufficient Conditions for Strong Irrepresentable Condition

In general, the Irrepresentable Condition is non-trivial, when number of zeros and nonzeros are of moderate sizes. Therefore, they give 5 sufficient conditions, such that Strong Irrepresentable Condition is guaranteed.



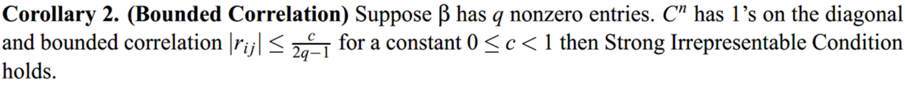
Corollary 1. (Constant Positive Correlation) Suppose

$$C^n = \begin{pmatrix} 1 & \dots & r_n \\ \vdots & \ddots & \vdots \\ r_n & \dots & 1 \end{pmatrix}$$

and there exists $c > 0$ such that $0 < r_n \leq \frac{1}{1+cq}$, then Strong Irrepresentable Condition holds.

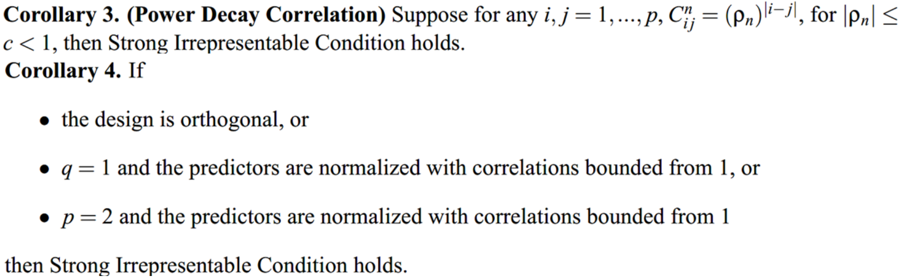
The design is symmetric, so that the covariates share a constant.

Analysis and Sufficient Conditions for Strong Irrepresentable Condition



Corollary 2. (Bounded Correlation) Suppose β has q nonzero entries. C^n has 1's on the diagonal and bounded correlation $|r_{ij}| \leq \frac{c}{2q-1}$ for a constant $0 \leq c < 1$ then Strong Irrepresentable Condition holds.

When the design matrix is slightly correlated, Lasso works consistently.



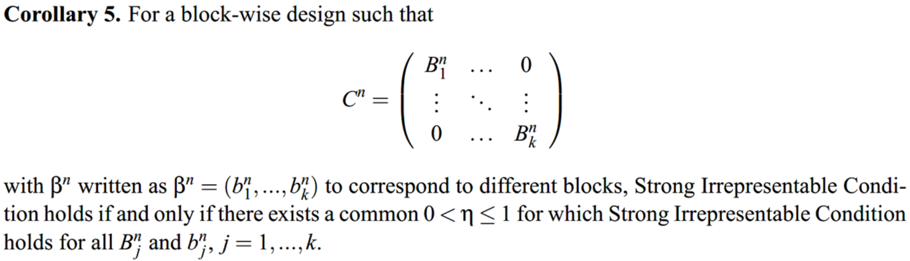
Corollary 3. (Power Decay Correlation) Suppose for any $i, j = 1, \ldots, p$, $C_{ij}^n = (\rho_n)^{|i-j|}$, for $|\rho_n| \leq c < 1$, then Strong Irrepresentable Condition holds.

Corollary 4. If

- the design is orthogonal, or
- $q = 1$ and the predictors are normalized with correlations bounded from 1, or
- $p = 2$ and the predictors are normalized with correlations bounded from 1

then Strong Irrepresentable Condition holds.

Analysis and Sufficient Conditions for Strong Irrepresentable Condition



Corollary 5. For a block-wise design such that

$$C^n = \begin{pmatrix} B_1^n & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & B_k^n \end{pmatrix}$$

with β^n written as $\beta^n = (b_1^n, \dots, b_k^n)$ to correspond to different blocks, Strong Irrepresentable Condition holds if and only if there exists a common $0 < \eta \leq 1$ for which Strong Irrepresentable Condition holds for all B_j^n and b_j^n, $j = 1, \dots, k$.

Simulations

Simulation 1: Consistency and Inconsistency with 3 Variables
Generate i.i.d. random variables x_{i1}, x_{i2}, e_i and ϵ_i, with variance 1 and mean 0, for $i = 1, \ldots, n = 1000$. x_{i3} is correlated with x_{i1} and x_{i2}:

$$x_{i3} = \frac{2}{3}x_{i1} + \frac{2}{3}x_{i2} + \frac{1}{3}e_i$$

The response:

$$Y_i = x_{i1}\beta_1 + x_{i2}\beta_2 + \epsilon_i$$

Then two settings:

- Strong Irrepresentable Condition fails: $\beta_1 = 2, \beta_2 = 3$
- Strong Irrepresentable Condition holds: $\beta_1 = -2, \beta_2 = 3$

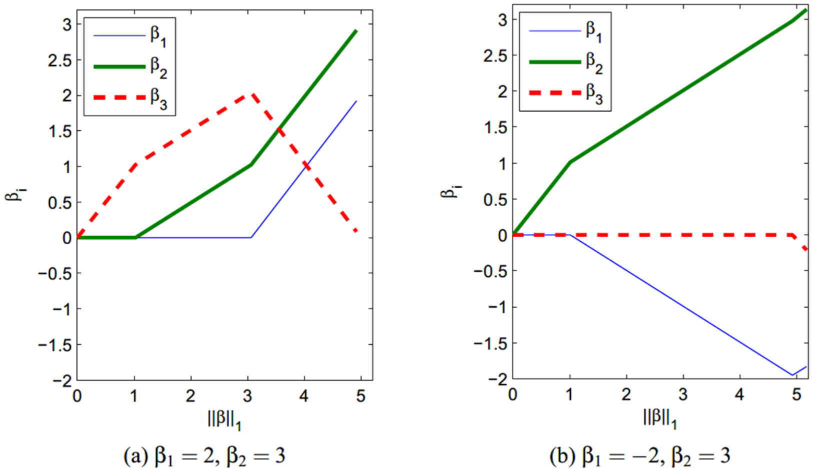
Simulations

Another view:
If we define $Y^*(\lambda) = Y - X\hat{\beta}(\lambda)$, then:

$$\begin{aligned} |X_3'Y^*| &= |(\frac{2}{3}X_1 + \frac{2}{3}X_2 + \frac{1}{3}e)'Y^*| \\ &\geq \frac{4}{3}\min(|X_1'Y^*|, |X_2'Y^*|)(\frac{\text{sign}(X_1'Y^*) + \text{sign}(X_2'Y^*)}{2}) - \frac{1}{3}|e'Y^*|. \end{aligned}$$

If $\hat{\beta}_3 = 0 \Rightarrow$ signs of X_1's and X_2's inner products with Y agree with the signs of $\hat{\beta}_1$ and $\hat{\beta}_2$ $\Rightarrow$ For Lasso to be sign consistent, the signs of β_1 and β_2 has to disagree. ("agree to" Strong Irrepresentalbe Condition)

Simulations



(a) $\beta_1 = 2, \beta_2 = 3$

(b) $\beta_1 = -2, \beta_2 = 3$

Figure 1: An example to illustrate Lasso's (in)consistency in Model Selection. The Lasso paths for settings (a) and (b) are plotted in the left and right panel respectively.

Simulations

Simulation 2: Quantitative Evaluation of Impact of Strong Irrepresentable Condition on Model Selection
Take $n = 100, p = 32, q = 5, \beta_1 = (7, 4, 2, 1, 1)^T$ and choose a small $\sigma^2 = 0.1$. Calculate $\eta_\infty = 1 - ||C_{21}^n(C_{11}^n)^{-1} sign(\beta_{(1)}^n)||_\infty$.

$\eta_\infty > 0$ means the Strong Irrepresentable Condition holds. The plot in the next page:

- The larger η_∞, the stronger the condition.
- steepest increase happens around 0.

Simulations

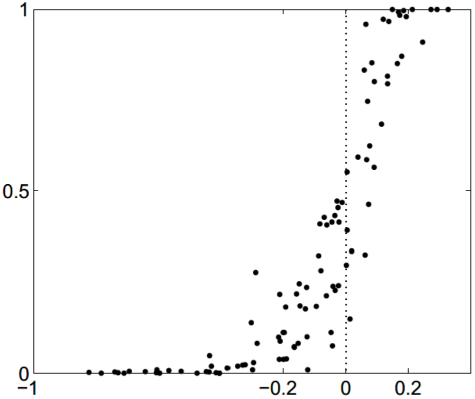
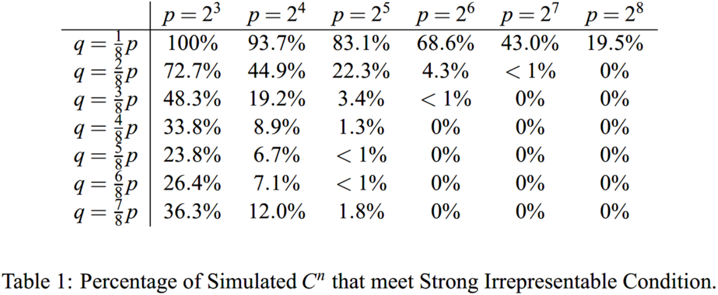


Figure 2: Comparison of Percentage of Lasso Selecting the Correct Model and η_∞. *X*-axis: η_∞. *Y*-axis: Percentage of Lasso Selecting the Correct Model.

Simulations

Simulation 3: How Strong is Irrepresentable Condition



	$p = 2^3$	$p = 2^4$	$p = 2^5$	$p = 2^6$	$p = 2^7$	$p = 2^8$
$q = \frac{1}{8}p$	100%	93.7%	83.1%	68.6%	43.0%	19.5%
$q = \frac{2}{8}p$	72.7%	44.9%	22.3%	4.3%	$< 1\%$	0%
$q = \frac{3}{8}p$	48.3%	19.2%	3.4%	$< 1\%$	0%	0%
$q = \frac{4}{8}p$	33.8%	8.9%	1.3%	0%	0%	0%
$q = \frac{5}{8}p$	23.8%	6.7%	$< 1\%$	0%	0%	0%
$q = \frac{6}{8}p$	26.4%	7.1%	$< 1\%$	0%	0%	0%
$q = \frac{7}{8}p$	36.3%	12.0%	1.8%	0%	0%	0%

Table 1: Percentage of Simulated C^n that meet Strong Irrepresentable Condition.

- When the true model is very sparse (q small), Strong Irrepresentable Condition has some probability to hold (Corollary 2)
- For the extreme case: $q = 1$, it holds (Corollary 4)
- For large p and q, the Condition rarely holds.