Journal of Machine Learning Research 7 (2006) 2541-2563 Submitted 4/06; Revised 9/06; Published 11/06

On Model Selection Consistency of Lasso

Peng Zhao PENGZHAO@STAT.BERKELEY.EDU
Bin Yu BINYU@STAT.BERKELEY.EDU
Department of Statistics
University of California, Berkeley
367 Evans Hall Berkeley, CA 94720-3860, USA

Editor: David Madigan

Abstract

Sparsity or parsimony of statistical models is crucial for their proper interpretations, as in sciences and social sciences. Model selection is a commonly used method to find such models, but usually involves a computationally heavy combinatorial search. Lasso (Tibshirani, 1996) is now being used as a computationally feasible alternative to model selection. Therefore it is important to study Lasso for model selection purposes.

In this paper, we prove that a single condition, which we 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. Based on these results, sufficient conditions that are verifiable in practice are given to relate to previous works and help applications of Lasso for feature selection and sparse representation.

This Irrepresentable Condition, which depends mainly on the covariance of the predictor variables, states that Lasso selects the true model consistently if and (almost) only if the predictors that are not in the true model are "irrepresentable" (in a sense to be clarified) by predictors that are in the true model. Furthermore, simulations are carried out to provide insights and understanding of this result.

Keywords: Lasso, regularization, sparsity, model selection, consistency

1. Introduction

A vastly popular and successful approach in statistical modeling is to use regularization penalties in model fitting (Hoerl and Kennard, 1970). By jointly minimizing the empirical error and penalty, one seeks a model that not only fits well and is also "simple" to avoid large variation which occurs in estimating complex models. Lasso (Tibshirani, 1996) is a successful idea that falls into this category. Its popularity is largely because the regularization resulting from Lasso's L_1 penalty leads to sparse solutions, that is, there are few nonzero estimates (among all possible choices). Sparse models are more interpretable and often preferred in the sciences and social sciences. However, obtaining such models through classical model selection methods usually involves heavy combinatorial search. Lasso, of which the entire regularization path can be computed in the complexity of one linear regression (Efron et al., 2004; Osborne et al., 2000b), provides a computationally feasible way for model selection (also see, for example, Zhao and Yu, 2004; Rosset, 2004). However, in order to use Lasso for model selection, it is necessary to assess how well the sparse model given by Lasso relates to the true model. We make this assessment by investigating Lasso's model selection consistency under linear models, that is, when given a large amount of data under what conditions Lasso does and does not choose the true model.

Assume our data is generated by a linear regression model

$$Y_n = \mathbf{X_n}\beta^n + \varepsilon_\mathbf{n}.$$

where $\varepsilon_\mathbf{n} = (\varepsilon_1, ..., \varepsilon_n)^T$ is a vector of i.i.d. random variables with mean 0 and variance σ^2. Y_n is an $n \times 1$ response and $\mathbf{X_n} = (X_1^n, ..., X_p^n) = ((x_1^n)^T, ..., (x_n^n)^T)^T$ is the $n \times p$ design matrix where X_i^n is its ith column (ith predictor) and x_j^n is its jth row (jth sample). β^n is the vector of model coefficients. The model is assumed to be "sparse", that is, some of the regression coefficients β^n are exactly zero corresponding to predictors that are irrelevant to the response. Unlike classical fixed p settings, the data and model parameters β are indexed by n to allow them to change as n grows.

The Lasso estimates $\hat{\beta}^n = (\hat{\beta}_1^n, ..., \hat{\beta}_j^n, ...)^T$ are defined by

$$\hat{\beta}^n(\lambda) = \arg\min_\beta \|Y_n - \mathbf{X_n}\beta\|_2^2 + \lambda\|\beta\|_1, \tag{1}$$

where $\|\cdot\|_1$ stands for the L_1 norm of a vector which equals the sum of absolute values of the vector's entries.

The parameter $\lambda \geq 0$ controls the amount of regularization applied to the estimate. Setting $\lambda = 0$ reverses the Lasso problem to Ordinary Least Squares which minimizes the unregularized empirical loss. On the other hand, a very large λ will completely shrink $\hat{\beta}^n$ to 0 thus leading to the empty or null model. In general, moderate values of λ will cause shrinkage of the solutions towards 0, and some coefficients may end up being exactly 0.

Under some regularity conditions on the design, Knight and Fu (2000) have shown estimation consistency for Lasso for fixed p and fixed β^n (i.e., p and β^n are independent of n) as $n \to \infty$. In particular, they have shown that $\hat{\beta}^n(\lambda_n) \to_p \beta$ and asymptotic normality of the estimates provided that $\lambda_n = o(n)$. In addition, it is shown in the work that for $\lambda_n \propto n^{\frac{1}{2}}$ (on the same order of $n^{\frac{1}{2}}$), as $n \to \infty$ there is a non-vanishing positive probability for lasso to select the true model.

On the model selection consistency front, Meinshausen and Buhlmann (2006) have shown that under a set of conditions, Lasso is consistent in estimating the dependency between Gaussian variables even when the number of variables p grows faster than n. Addressing a slightly different but closely related problem, Leng et al. (2004) have shown that for a fixed p and orthogonal designs, the Lasso estimate that is optimal in terms of parameter estimation does not give consistent model selection. Furthermore, Osborne et al. (1998), in their work of using Lasso for knot selection for regression splines, noted that Lasso tend to pick up knots in close proximity to one another. In general, as we will show, 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.

Since using the Lasso estimate involves choosing the appropriate amount of regularization, to study the model selection consistency of the Lasso, we consider two problems: whether there exists a deterministic amount of regularization that gives consistent selection; or, for each random realization whether there exists a correct amount of regularization that selects the true model. Our main result shows there exists an **Irrepresentable Condition** that, except for a minor technicality, is almost necessary and sufficient for both types of consistency. Based on this condition, we give sufficient conditions that are verifiable in practice. In particular, in one example our condition coincides with the "Coherence" condition in Donoho et al. (2004) where the L_2 distance between the Lasso estimate and true model is studied in a non-asymptotic setting.

After we had obtained our almost necessary and sufficient condition result, it was brought to our attention of an independent result in Meinshausen and Buhlmann (2006) where a similar condition to the Irrepresentable Condition was obtained to prove a model selection consistency result for Gaussian graphical model selection using the Lasso. Our result is for linear models (with fixed p and p growing with n) and it could accommodate non-Gaussian errors and non-Gaussian designs. Our analytical approach is direct and we thoroughly explain through special cases and simulations the meaning of this condition in various cases. We also make connections to previous theoretical studies and simulations on Lasso (e.g., Donoho et al., 2004; Zou et al., 2004; Tibshirani, 1996).

The rest of the paper is organized as follows. In Section 2, we describe our main result—the Irrepresentable Condition for Lasso to achieve consistent selection and prove that it is almost necessary and sufficient. We then elaborate on the condition by extending to other sufficient conditions that are more intuitive and verifiable to relate to previous theoretical and simulation studies of Lasso. Sections 3 contains simulation results to illustrate our result and to build heuristic sense of how strong the condition is. To conclude, Section 4 compares Lasso with thresholding and discusses alternatives and possible modifications of Lasso to achieve selection consistency when Irrepresentable Condition fails.

2. Model Selection Consistency and Irrepresentable Conditions

An estimate which is consistent in term of parameter estimation does not necessarily consistently select the correct model (or even attempt to do so) where the reverse is also true. The former requires

$$\hat{\beta}^n - \beta^n \to_p 0, \text{ as } n \to \infty$$

while the latter requires

$$P(\{i: \hat{\beta}_i^n \neq 0\} = \{i: \beta_i^n \neq 0\}) \to 1, \text{ as } n \to \infty.$$

In general, we desire our estimate to have both consistencies. However, to separate the selection aspect of the consistency from the parameter estimation aspect, we make the following definitions about Sign Consistency that does not 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

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

where 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.

Sign consistency is stronger than the usual selection consistency which only requires the zeros to be matched, but not the signs. The reason for using sign consistency is technical. It is needed for proving the necessity of the Irrepresentable Condition (to be defined) to avoid dealing with situations where a model is estimated with matching zeros but reversed signs. We also argue that an estimated model with reversed signs can be misleading and hardly qualifies as a correctly selected model.

Now we define two kinds of 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

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

Strong Sign Consistency implies one can use a preselected λ to achieve consistent model selection via Lasso. General Sign Consistency means for a random realization there exists a correct amount of regularization that selects the true model. Obviously, strong sign consistency implies general sign consistency. Surprisingly, as implied by our results, the two kinds of sign consistencies are almost equivalent to one condition. To define this condition we need the following notations on the design.

Without loss of generality, assume $\beta^n = (\beta_1^n, ..., \beta_q^n, \beta_{q+1}^n, \cdots \beta_p^n)^T$ where $\beta_j^n \neq 0$ for $j = 1, .., q$ and $\beta_j^n = 0$ for $j = q+1, ..., p$. Let $\beta_{(1)}^n = (\beta_1^n, ..., \beta_q^n)^T$ and $\beta_{(2)}^n = (\beta_{q+1}^n, ..., \beta_p^n)$. 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 $C_{11}^n = \frac{1}{n}\mathbf{X_n}(1)'\mathbf{X_n}(1)$, $C_{22}^n = \frac{1}{n}\mathbf{X_n}(2)'\mathbf{X_n}(2)$, $C_{12}^n = \frac{1}{n}\mathbf{X_n}(\mathbf{1})'\mathbf{X_n}(\mathbf{2})$ and $C_{21}^n = \frac{1}{n}\mathbf{X_n}(2)'\mathbf{X_n}(1)$. C^n can then be expressed in a block-wise form as follows:

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

Assuming C_{11}^n is invertible, we define the following Irrepresentable Conditions
Strong Irrepresentable Condition. There exists a positive constant vector η

$$|C_{21}^n(C_{11}^n)^{-1}\text{sign}(\beta_{(1)}^n)| \leq \mathbf{1} - \eta,$$

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

$$|C_{21}^n(C_{11}^n)^{-1}\text{sign}(\beta_{(1)}^n)| < \mathbf{1},$$

where the inequality holds element-wise.

Weak Irrepresentable Condition is slightly weaker than Strong Irrepresentable Condition. C^n can converge in ways that entries of $|C_{21}^n(C_{11}^n)^{-1}\text{sign}(\beta_{(1)}^n)|$ approach 1 from the below so that Weak Condition holds but the strict inequality fails in the limit. For a fixed p and $\beta^n = \beta$, the distinction disappears for random designs when, for example, x_i^n's are i.i.d. realizations with covariance matrix C, since then the two conditions are equivalent to $|C_{21}(C_{11})^{-1}\text{sign}(\beta(1))| < \mathbf{1}$ almost surely.

The Irrepresentable Conditions closely resembles a regularization constraint on the regression coefficients of the irrelevant covariates ($\mathbf{X_n}(2)$)) on the relevant covariates ($\mathbf{X_n}(1)$). In particular, when signs of the true β are unknown, for the Irrepresentable Condition to hold for all possible signs, we need the L_1 norms of the regression coefficients to be smaller than 1. To see this, recall for (2) to hold for all possible sign($\beta(1)$), we need

$$|((\mathbf{X_n}(1)^T\mathbf{X_n}(1))^{-1}\mathbf{X_n}(1)^T\mathbf{X_n}(2)| = |(C_{11}^n)^{-1}C_{12}^n| < \mathbf{1} - \eta, \tag{2}$$

that is, the total amount of an irrelevant covariate represented by the covariates in the true model is not to reach 1 (therefore the name "irrepresentable").

As a preparatory result, the following proposition puts a lower bound on the probability of Lasso picking the true model which quantitatively relates the probability of Lasso selecting the correct model and 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

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

It can be argued (see the proof of Proposition 1 in the appendix) that A_n implies the signs of of those of $\beta_{(1)}^n$ are estimated correctly. And given A_n, B_n further imply $\hat{\beta}_{(2)}^n$ are shrunk to zero. The regularization parameter λ_n trades off the size of these two events. Smaller λ_n leads to larger A_n but smaller B_n which makes it likely to have Lasso pick more irrelevant variables. On the other hand, larger constant η always leads to larger B_n and have no impact on A_n. So when Strong Irrepresentable Condition holds with a larger constant η, it is easier for Lasso to pick up the true model. This is quantitatively illustrated in Simulation 3.2.

Our main results relate Strong and Weak Irrepresentable Conditions with strong and general sign consistency. We describe the results for small q and p case next followed by results for large q and p in Section 2.2. Then, analysis and sufficient conditions are given in Section 2.3 to achieve a better understanding of the Irrepresentable Conditions and relate to previous works.

2.1 Model Selection Consistency for Small q and p

In this section, we work under the classical setting where q, p and β^n are all fixed as $n \to \infty$. In this setting, it is natural to assume the following regularity conditions:

$$C^n \to C, \text{ as } n \to \infty. \tag{3}$$

where C is a positive definite matrix. And,

$$\frac{1}{n}\max_{1\leq i\leq n}((x_i^n)^T x_i^n) \to 0, \text{ as } n \to \infty. \tag{4}$$

In practice, the covariates are usually scaled so that the diagonal elements of C^n are all 1's. The convergence in (3) and (4) are deterministic. However, the results in this is section also holds quite generally for random designs. Specifically, in the case of a random design, X can be conditioned on and the asymptotic results still apply if the probability of the set where (3) and (4) hold is 1. In general, (3) and (4) are weak in the sense that if one assumes x_i are i.i.d. with finite second moments then $C = E((x_i^n)^T x_i^n)$, $\frac{1}{n}\mathbf{X_n}^T\mathbf{X_n} \to_{a.s.} C$ and $\max_{1\leq i\leq n} x_i^T x_i = o_p(n)$, thus (3) and (4) hold naturally.

Under these conditions we have the following result.

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 $0 \le c < 1$, we have

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

A proof of Theorem 1 can be found in the appendix.

Theorem 1 states that, if Strong Irrepresentable Condition holds, then the probability of Lasso selecting the true model approaches 1 at an exponential rate while only the finite second moment of the noise terms is assumed. In addition, from Knight and Fu (2000) we know that for $\lambda_n = o(n)$ Lasso also has consistent estimation and asymptotic normality. Therefore Strong Irrepresentable Condition allows for consistent model selection and parameter estimation simultaneously. On the other hand, Theorem 2 shows that Weak Irrepresentable Condition is also necessary even 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 $n > N$.

A proof of Theorem 2 can be found in the appendix.

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.

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.

2.2 Model Selection Consistency for Large p and q

In the large p and q case, we allow the dimension of the designs C^n and model parameters β_n grow as n grows, that is, $p = p_n$ and $q = q_n$ are allowed to grow with n. Consequently, the assumptions and regularity conditions in Section 2.1 becomes inappropriate as C^n do not converge and β^n may change as n grows. Thus we need to control the size of the smallest entry of $\beta^n_{(1)}$, bound the eigenvalues of C^n_{11} and have the design scale properly. Specifically, we assume:

There exists $0 \le c_1 < c_2 \le 1$ and $M_1, M_2, M_3, M_4 > 0$ so the following holds:

$$\frac{1}{n}(X_i^n)'X_i^n \le M_1 \text{ for } \forall i, \tag{5}$$

$$\alpha' C^n_{11} \alpha \ge M_2, \text{ for } \forall \|\alpha\|_2^2 = 1, \tag{6}$$

$$q_n = O(n^{c_1}), \tag{7}$$

$$n^{\frac{1-c_2}{2}} \min_{i=1,..,q} |\beta_i^n| \ge M_3. \tag{8}$$

Condition (5) is trivial since it can always be achieved by normalizing the covariates. (6) requires the design of the relevant covariates have eigenvalues bounded from below so that the inverse of C_{11}^n behaves well. For a random design, if the eigenvalues of the population covariance matrix are bounded from below and $q_n/n \to \rho < 1$ then (6) usually follows Bai (1999).

The main conditions are (7) and (8) which are similar to the ones in Meinshausen (2005) for Gaussian graphical models. (8) requires a gap of size n^{c_2} between the decay rate of $\beta_{(1)}^n$ and $n^{-\frac{1}{2}}$. Since the noise terms aggregate at a rate of $n^{-\frac{1}{2}}$, this prevents the estimation to be dominated by the noise terms. Condition (7) is a sparsity assumption which requires square root of the size of the true model $\sqrt{q_n}$ to grow at a rate slower than the rate gap which consequently prevents the estimation bias of the Lasso solutions from dominating the model parameters.

Under these conditions, we have 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

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

A proof of Theorem 3 can be found in the appendix.

Theorem 3 states that Lasso can select the true model consistently given that Strong Irrepresentable Condition holds and the noise terms have some finite moments. For example, if only the second moment is assumed, p is allowed to grow slower than $n^{c_2-c_1}$. If all moments of the noise exist then, by Theorem 3, p can grow at any polynomial rate and the probability of Lasso selecting the true model converges to 1 at a faster rate than any polynomial rate. In particular, for Gaussian noises, we have:

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 $c_3 < c_4 < c_2 - c_1$,

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

A proof of Theorem 4 can be found in the appendix. As discussed in the introduction, this result has also been obtained independently by Meinshausen and Buhlmann (2006) in their study of high dimensional multivariate Gaussian random variables. This result is obtained more directly for linear models and differs from theirs by the use of fixed designs to accommodate non-Gaussian designs. p_n is also allowed to grow slightly faster than the polynomial rates used in that work.

It is an encouraging result that using Lasso we can allow p to grow much faster than n (up to exponentially fast) while still allow for fast convergence of the probability of correct model selection to 1. However, we note that this fast rate is not achievable for all noise distributions. In general, the result of Theorem 3 is tight in the sense that if higher moments of the noise distribution do not exist then the tail probability of the noise terms does not vanish quick enough to allow p to grow at higher degree polynomial rates.

Through Theorem 3 and 4, we have shown, for cases with large p—(polynomial in n given that noise have finite moments, exponential in n for Gaussian noises), Strong Irrepresentable Condition still implies the probability of Lasso selecting the true model converges to 1 at a fast rate. We have found it difficult to show necessariness of Irrepresentable Condition for the large p setting in a meaningful way. This is mainly due to the technical difficulty that arises from dealing with high dimensional design matrices. However, by the results for the small p case, the necessariness of Irrepresentable Condition is implied to some extent.

2.3 Analysis and Sufficient Conditions for Strong Irrepresentable Condition

In general, the Irrepresentable Condition is non-trivial when the numbers of zeros and nonzeros are of moderate sizes, for example, 3. Particularly since we do not know $\text{sign}(\beta)$ before hand, we need the Irrepresentable Condition to hold for every possible combination of different signs and placement of zeros. A closer look discloses that (2) does not depend on C^n_{22}, that is, the covariance of the covariates that are not in the true model. It linearly depends on C^n_{21}, the correlations between the covariates that are in the model and the ones that are not. For the C^n_{11} part, except for special cases (Corollary 1) we also want the correlations between covariates that are in the model to be small otherwise C^n_{11} may contain small eigenvalues which leads to large eigenvalues for $(C^n_{11})^{-1}$ and results in the violation of (2).

To further elaborate and relate to previous works, we give some sufficient conditions in the following corollaries such that Strong Irrepresentable Condition is guaranteed. All diagonal elements of C^n are assumed to be 1 which is equivalent to normalizing all covariates in the model to the same scale since Strong Irrepresentable Condition is invariant under any common scaling of C^n. Proofs of the corollaries are included in the appendix.

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.

Corollary 1 has particularly strong implications for applications of Lasso where the covariates of the regression are designed with a symmetry so that the covariates share a constant correlation. Under such a design, this result implies that Strong Irrepresentable Condition holds even for p growing with n as long as q remains fixed and consequently ensures that Lasso selects the true model asymptotically. However, when the design is random or, for example, arises from an observational study we usually do not have the constant correlation. Correspondingly, we have the following result on bounded correlations.

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.

Corollary 2 verifies the common intuition that when the design matrix is slightly correlated Lasso works consistently. And the larger q is, the smaller the bound on correlation becomes. For a q of considerable size, the bound becomes too small to meet in practice. Unfortunately, this bound is also tight in the following sense: when the bound is violated, one can construct

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

with $r \leq -\frac{1}{2q-1}$ and make the nonzero β_i's all positive then $|C_{21}(C_{11})^{-1}\text{sign}(\beta(1))| \geq \mathbf{1}$ holds element-wise which fails Strong Irrepresentable Condition.

In comparison, Donoho et al. (2004) showed that, in a non-asymptotic setup, the L^2 distance between the sparsest estimate and the true model is bounded by a linear multiple of the noise level if

$$q < (1/r+1)/2,$$

where $r = \max_{i,j} |C^n_{ij}|$ (called Coherence). This is equivalent to

$$\max_{i,j} |C^n_{ij}| < \frac{1}{2q-1}$$

which coincides with the condition of Corollary 2. Interestingly, for the same result to apply to the Lasso estimates, Donoho et al. (2004) required tighter bound on the correlation, that is, $\max_{i,j} |C^n_{ij}| < \frac{1}{4q-1}$.

Another typical design used for Lasso simulations (e.g., Tibshirani, 1996; Zou et al., 2004) is setting the correlation between X^n_i and X^n_j to be $\rho^{|i-j|}$ with an constant $0 < \rho < 1$. Although this design introduces more sophisticated correlation structure between the predictors and does not seem restrictive, the following corollary states under this design Strong Irrepresentable Condition holds for any q.

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

In addition, as instances of Corollary 2, under some simplified designs which are often used for theoretical studies, Lasso is consistent for model selection.

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.

One additional informative scenario to consider is a block-wise design. As it is commonly assumed in practice, this assumed scenario is a hybrid between the most highly structured designs like the orthogonal design and a general design. For this design, it can be shown that

Corollary 5. For a block-wise design such that

$$C^n = \begin{pmatrix} B^n_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & B^n_k \end{pmatrix}$$

with β^n written as $\beta^n = (b^n_1, ..., b^n_k)$ 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^n_j and b^n_j, $j = 1, ..., k$.

Combinations of Corollary 5 and Corollary 1-4 cover some interesting cases such as models with 2 × 2 design blocks and models where 0, 1 or all parameters out of each block are nonzero.

Through Corollaries 1 - 5, we have shown that under specific designs, which are commonly used or assumed in previous works, Irrepresentable Condition holds which leads to Lasso's consistency in model selection. Next, we demonstrate Lasso's model selection consistency and the Irrepresentable Conditions using simulations.

3. Simulation Studies

In this section, we give simulation examples to illustrate the established results. The first simulation illustrates the simplest case ($p = 3$, $q = 2$, cf. Corollary 4) under which Lasso is inconsistent for model selection. We also analyze the Lasso algorithm to explain how Lasso is misled into inconsistency when Irrepresentable Conditions fail. The second simulation quantitatively relates the consistency (inconsistency) of Lasso to how well the Strong Irrepresentable Condition holds (fails) by counting the percentages of Lasso selecting the true model and comparing it to $\eta_\infty = 1 - \|C^n_{21}(C^n_{11})^{-1}\text{sign}(\beta^n_{(1)})\|_\infty$. In the last simulation, we establish a heuristic sense of how strong our Strong Irrepresentable Condition is for different values of p and q by observing how often the condition holds when C is sampled from Wishart(p,p) distribution.

3.1 Simulation Example 1: Consistency and Inconsistency with 3 Variables

In this simple example, we aim to give some practical sense of the Lasso algorithm's behaviors when Strong Irrepresentable Condition holds and fails. We first generate i.i.d. random variables x_{i1}, x_{i2}, e_i and ε_i with variance 1 and mean 0 for $i = 1,\ldots,n$ and $n = 1000$. A third predictor x_{i3} is generated to be correlated with x_{i1} and x_{i2} by

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

then by construction, x_{i3} is also i.i.d. with mean 0 and variance 1.

The response is generated by

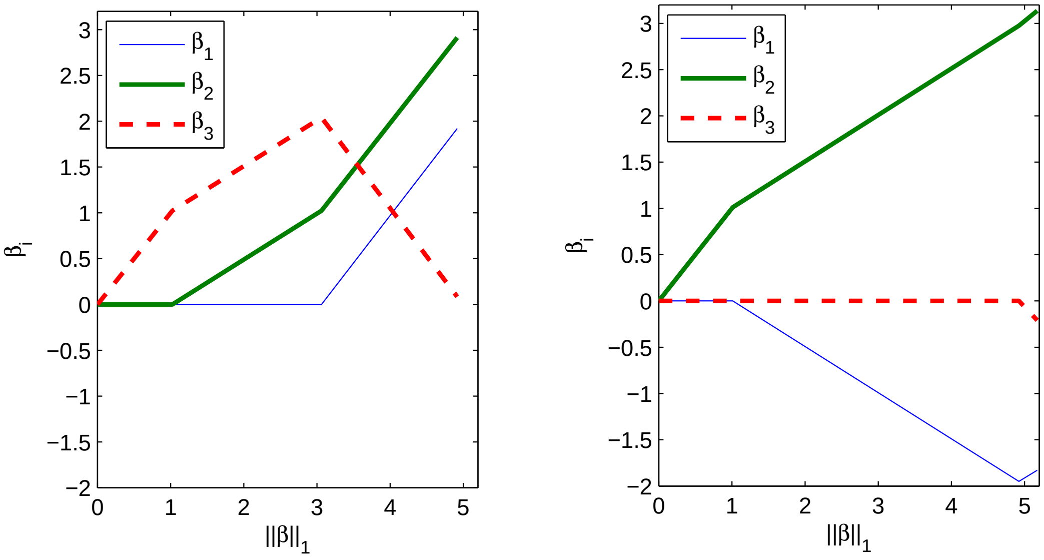
$$Y_i = x_{i1}\beta_1 + x_{i2}\beta_2 + \varepsilon_i.$$

Lasso is applied (through the LARS algorithm by Efron et al., 2004) on Y, X_1, X_2 and X_3 in two settings: (a) $\beta_1 = 2$, $\beta_2 = 3$; and (b) $\beta_1 = -2$, $\beta_2 = 3$. In both settings, $\mathbf{X}(1) = (X_1, X_2)$, $\mathbf{X}(2) = X_3$ and through (2), it is easy to get $C_{21}C_{11}^{-1} = (\frac{2}{3}, \frac{2}{3})$. Therefore Strong Irrepresentable Condition fails for setting (a) and holds for setting (b).

Now we investigate how these two set-ups lead to Lasso's sign consistency and inconsistency respectively. As we vary the amount of regularization (controlled by λ), we get different Lasso solutions which form the Lasso path (as illustrated by the left and right panels of Figure 1). This Lasso path follows the least angle direction (as described in for example, Efron et al. (2004) and Zhao and Yu (2004)), that is, $\hat{\beta}(\lambda)$ progresses in coordinates on which the absolute values of inner products between $Y^*(\lambda) := Y - X\hat{\beta}(\lambda)$ and the predictors are the largest while entries of $\hat{\beta}(\lambda)$ corresponding to smaller inner products are left at zero.

In this example,

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



(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.

For large n, $e' * Y^*$ is on a smaller order than the rest of the terms. If $\hat{\beta}_3$ is zero, the signs of X_1's and X_2's inner products with Y agree with the signs of $\hat{\beta}_1$ and $\hat{\beta}_2$. Therefore for Lasso to be sign consistent, the signs of β_1 and β_2 has to disagree which happens in setting (b) but not setting (a).

Consequently, in setting (a) Lasso does not shrink $\hat{\beta}_3$ to 0. Instead, the L_1 regularization prefers X_3 over X_1 and X_2 as Lasso picks up X_3 first and never shrinks it back to zero. For setting (b), Strong Irrepresentable Condition holds and with a proper amount of regularization, Lasso correctly shrinks $\hat{\beta}_3$ to 0.

3.2 Simulation Example 2: Quantitative Evaluation of Impact of Strong Irrepresentable Condition on Model Selection

In this example, we give some quantitative sense on the relationship between the probability of Lasso selecting the correct model and how well Strong Irrepresentable Condition holds (or fails). First, we take $n = 100$,$p = 32$, $q = 5$ and $\beta_1 = (7,4,2,1,1)^T$ and choose a small $\sigma^2 = 0.1$ to allow us to go into asymptotic quickly.

Then we would like to generate 100 designs of X as follows. We first sample a covariance matrix S from Wishart(p,p) (see section 3.3 for details), then take n samples of X_i from $N(0,S)$, and finally normalize them to have mean squares 1 as in common applications of Lasso. Such generated samples represent a variety of designs: some satisfy Strong Irrepresentable Condition with a large η, while others fail the condition badly. To evaluate how well the Irrepresentable condition holds we calculate $\eta_\infty = 1 - \|C^n_{21}(C^n_{11})^{-1}\text{sign}(\beta^n_{(1)})\|_\infty$. So if $\eta_\infty > 0$, Strong Irrepresentable holds otherwise it

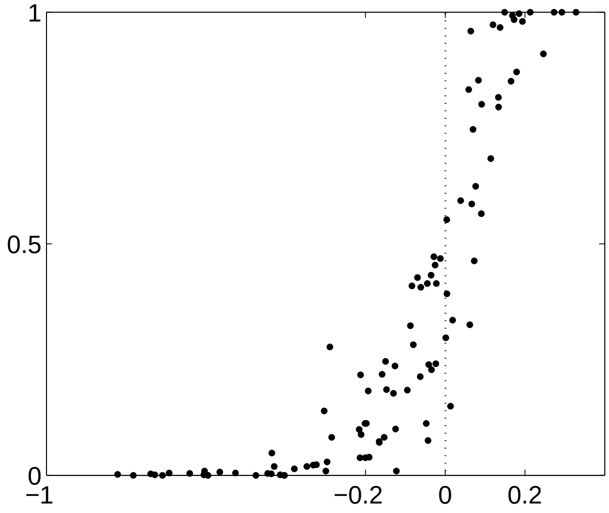


Figure 2: Comparison of Percentage of Lasso Selecting the Correct Model and η_∞. X-axis: η_∞. Y-axis: Percentage of Lasso Selecting the Correct Model.

fails. The η_∞'s of the 100 simulated designs are within $[-1.02, 0.33]$ with 67 of them being smaller than 0 and 33 of them bigger than 0.

For each design, we run the simulation 1000 times and examine general sign consistencies. Each time, n samples of ε are generated from $N(0,\sigma^2)$ and $Y = X\beta + \varepsilon$ are calculated. We then run Lasso (through the LARS algorithm by Efron et al., 2004) to calculate the Lasso path. The entire path is examined to see if there exists a model estimate that matches the signs of the true model. Then we compute the percentage of runs that generated matched models for each design and compare it to η_∞ as shown in Figure 2.

As can be seen from Figure 2, when η_∞ gets large, the percentage of Lasso selecting the correct model goes up with the steepest increase happening around 0. For η_∞ considerably larger than 0 (> 0.2) the percentage is close to 1. On the other hand, for η_∞ considerably smaller than 0 (< -0.3) there is little chance for Lasso to select the true model. In general, this is consistent with our result (Proposition 1 and Theorem 1 to 4), as for $\eta_\infty > 0$, if n is large enough, the probability of Lasso selects the true model gets close to 1 which does not happen if $\eta_\infty < 0$. This quantitatively illustrates the importance of Strong Irrepresentable Condition for Lasso's model selection performance.

3.3 Simulation Example 3: How Strong is Irrepresentable Condition?

As illustrated by Corollaries 1 to 4, Strong Irrepresentable Condition holds for some constrained special settings. While in Section 3.1 and 3.2, we have seen cases where Irrepresentable Condition fails. In this simulation, we establish some heuristic sense of how strong our Strong Irrepresentable Condition is for different values of p and q.

For a given p, the set of C^n is the set of nonnegative definite matrix of size p. To measure the size of the subset of C^n's on which Irrepresentable Condition holds, the Wishart measure family can be used. Since Strong Irrepresentable Condition holds for designs that are close to orthogonal

	$p=2^3$	$p=2^4$	$p=2^5$	$p=2^6$	$p=2^7$	$p=2^8$
$q=\frac{1}{8}p$	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.

(Corollary 2), we take the Wishart(p,p) measure which centers but does not concentrate around the identity matrix.

In this simulation study, we sample C^n's from white Wishart(p,p) and examine how often Irrepresentable Condition holds. For each $p=2^3,2^4,2^5,2^6,2^7,2^8$ and correspondingly $q=\frac{1}{8}p,\frac{2}{8}p,...,\frac{7}{8}p$ we generate 1000 C^n's from Wishart and re-normalize it to have 1's on the diagonal. Then we examine how often Irrepresentable Condition holds. The entries of $\beta(1)$ are assumed to be positive, otherwise a sign flip of the corresponding X_i's can make the corresponding β_i positive. The result is shown in Table 1.

Table 1 shows that, when the true model is very sparse (q small), Strong Irrepresentable Condition has some probability to hold which illustrates Corollary 2's conclusion. For the extreme case, $q=1$, it has been proved to hold (see Corollary 4). However, in general, for large p and q, Irrepresentable Condition rarely (measured by Wishart(p,p)) holds.

4. Discussions

In this paper, we 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. We have explored the meaning of the conditions through theoretical and empirical studies. Although much of Lasso's strength lies in its finite sample performance which is not the focus here, our 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). As a precaution, for data sets that can not be verified to satisfy the Irrepresentable Conditions, Lasso may not select the model correctly. In comparison, traditional all subset methods like BIC and MDL are always consistent but computationally intractable for p of moderate sizes. Thus, alternative computationally feasible methods that lead to selection consistency when the condition fails are of interest.

In particular, for small p cases, if consistency is the only concern then thresholding (either hard or soft) is an obvious choice that guarantees consistent selection. Since the OLS estimate $\hat{\beta}_{OLS}$ converges at a $1/\sqrt{n}$ rate, therefore a threshold that satisfies $t_n/\sqrt{n}\to\infty$ and $t_n\to 0$ leads to consistent selection. However, as emphasized earlier, consistency does not mean good performance in finite sample which is what matters in many applications where Lasso-type of technique is used. In particular, when the linear system is over determined $p>n$, the approach is no longer applicable since the OLS estimates are not well defined. On the other hand, Theorem 3 and Theorem 4 indicate that for cases where p may grow much faster then n, the Lasso still perform well.

To get some intuitive sense of how the thresholding performs comparing to the Lasso in finite sample, we ran the same simulations as in Section 3.2 and examined the sign matching rate of thresholding and compare it to the Lasso's performance. Our observation is, when the sample size is large, that is, in the asymptotic domain, even when Strong Irrepresentable Condition holds, Lasso does not perform better than simple thresholding in term of variable selection. In the small sample domain, however, Lasso seems to show an advantage which is consistent with the results reported in other publications (e.g., Tibshirani, 1996).

Another alternative that selects model consistently in our simulations is given by Osborne et al. (1998). They advise to use Lasso to do initial selection. Then a best subset selection (or a similar procedure, for example, forward selection) should be performed on the initial set selected by Lasso. This is loosely justified since, for instance, from Knight and Fu (2000) we know Lasso is consistent for $\lambda = o(n)$ and therefore can pick up all the true predictors if the amount of data is sufficient (although it may over-select).

Finally, we think it is possible to directly construct an alternative regularization to Lasso that selects model consistently under much weaker conditions and at the same time remains computationally feasible. This relies on understanding why Lasso is inconsistent when Strong Irrepresentable Condition fails: to induce sparsity, Lasso shrinks the estimates for the nonzero β's too heavily. When Strong Irrepresentable Condition fails, the irrelevant covariates are correlated with the relevant covariates enough to be picked up by Lasso to compensate the over-shrinkage of the nonzero β's. Therefore, to get universal consistency, we need to reduce the amount of shrinkage on the β estimates that are away from zero and regularize in a more similar fashion as l_0 penalty. However, as a consequence, this breaks the convexity of the Lasso penalty, therefore more sophisticated algorithms are needed for solving the minimization problems. A different set of analysis is also needed to deal with the local minima. This points towards our future work.

Acknowledgments

This research is partly supported by NSF Grant DMS-03036508 and ARO Grant W911NF-05-1-0104. Yu is also very grateful for the support of a 2006 Guggenheim Fellowship.

Appendix A. Proofs

To prove Proposition 1 and the rest of the theorems, we state Lemma 1 which is a direct consequence of KKT (Karush-Kuhn-Tucker) conditions:

Lemma 1. $\hat{\beta}^n(\lambda) = (\hat{\beta}_1^n, ..., \hat{\beta}_j^n, ...)$ are the Lasso estimates as defined by (1) if and only if

$$\begin{aligned} \frac{d\|Y_n - \mathbf{X_n}\beta\|_2^2}{d\beta_j}|_{\beta_j = \hat{\beta}_j^n} = & \quad \lambda \text{sign}(\hat{\beta}_j^n) \quad \text{for } j \text{ s.t. } \hat{\beta}_j^n \neq 0 \\ |\frac{d\|Y_n - \mathbf{X_n}\beta\|_2^2}{d\beta_j}|_{\beta_j = \hat{\beta}_j^n}| \leq & \quad \lambda \quad \text{for } j \text{ s.t. } \hat{\beta}_j^n = 0. \end{aligned}$$

With Lemma 1, we now prove Proposition 1.

Proof of Proposition 1. First, by definition

$$\hat{\beta}^n = \arg\min_{\beta}[\sum_{i=1}^{n}(Y_i - X_i\beta)^2 + \lambda_n\|\beta\|_1].$$

Let $\hat{u}^n = \hat{\beta}^n - \beta^n$, and define

$$V_n(u^n) = \sum_{i=1}^{n}[(\varepsilon_i - X_i u^n)^2 - \varepsilon_i^2] + \lambda_n\|u^n + \beta\|_1,$$

we have

$$\begin{aligned} \hat{u}^n &= \arg\min_{u^n}[\sum_{i=1}^{n}(\varepsilon_i - X_i u^n)^2 + \lambda_n\|u^n + \beta\|_1]. \\ &= \arg\min_{u^n} V_n(u^n). \end{aligned} \tag{9}$$

The first summation in $V_n(u^n)$ can be simplified as follows:

$$\begin{aligned} &\sum_{i=1}^{n}[(\varepsilon_i - X_i u^n)^2 - \varepsilon_i^2] \\ =\ &\sum_{i=1}^{n}[-2\varepsilon_i X_i u^n + (u^n)^T X_i^T X_i u^n], \\ =\ &-2W^n(\sqrt{n}u^n) + (\sqrt{n}u^n)^T C^n(\sqrt{n}u^n), \end{aligned} \tag{10}$$

where $W^n = (X^n)^T\varepsilon^n/\sqrt{n}$. Notice that (10) is always differentiable w.r.t. u^n and

$$\frac{d[-2W^n(\sqrt{n}u^n) + (\sqrt{n}u^n)^T C_n(\sqrt{n}u^n)]}{du^n} = 2\sqrt{n}(C^n(\sqrt{n}u^n) - W^n). \tag{11}$$

Let $\hat{u}^n(1)$, $W^n(1)$ and $\hat{u}^n(2)$, $W^n(2)$ denote the first q and last $p-q$ entries of $\hat{u}^n$ and W^n respectively. Then by definition we have:

$$\{\text{sign}(\hat{\beta}_j^n) = \text{sign}(\beta_j^n), \text{ for } j = 1, ..., q.\} \in \{\text{sign}(\beta_{(1)}^n)\hat{u}^n(1) > -|\beta_{(1)}^n|\}.$$

Then by Lemma 1, (9), (11) and uniqueness of Lasso solutions, if there exists $\hat{u}^n$, the following holds

$$C_{11}^n(\sqrt{n}\hat{u}^n(1)) - W^n(1) = -\frac{\lambda_n}{2\sqrt{n}}\text{sign}(\beta_{(1)}^n),$$

$$|\hat{u}^n(1)| < |\beta_{(1)}^n|,$$

$$-\frac{\lambda_n}{2\sqrt{n}}\mathbf{1} \leq C_{21}^n(\sqrt{n}\hat{u}^n(1)) - W^n(2) \leq \frac{\lambda_n}{2\sqrt{n}}\mathbf{1}.$$

then $\text{sign}(\hat{\beta}_{(1)}^n) = \text{sign}(\beta_{(1)}^n)$ and $\hat{\beta}_{(2)}^n = u^n(2) = 0$.

Substitute $\hat{u}^n(1)$, $\hat{u}^n(2)$ and bound the absolute values, the existence of such $\hat{\mu}^n$ is implied by

$$|(C_{11}^n)^{-1}W^n(1)| < \sqrt{n}(|\beta_{(1)}^n| - \frac{\lambda_n}{2n}|(C_{11}^n)^{-1}\text{sign}(\beta_{(1)}^n)|), \tag{12}$$

$$|C_{21}^n (C_{11}^n)^{-1} W^n(1) - W^n(2)| \leq \frac{\lambda_n}{2\sqrt{n}} (\mathbf{1} - |C_{21}^n (C_{11}^n)^{-1} \text{sign}(\beta_{(1)}^n)|) \tag{13}$$

(12) coincides with A_n and (13)$\in B_n$. This proves Proposition 1.

Using Proposition 1, we now prove Theorem 1.

Proof of Theorem 1. First, by Proposition 1 we have By Proposition 1, we have

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

Whereas

$$\begin{aligned} 1 - P(A_n \cap B_n) &\leq P(A_n^c) + P(B_n^c) \\ &\leq \sum_{i=1}^{q} P(|z_i^n| \geq \sqrt{n}(|\beta_i^n| - \frac{\lambda_n}{2n} b_i^n) + \sum_{i=1}^{p-q} P(|\zeta_i^n| \geq \frac{\lambda_n}{2\sqrt{n}} \eta_i). \end{aligned}$$

where $z^n = (z_1^n, ..., z_p^n)' = (C_{11}^n)^{-1} W^n(1)$, $\zeta^n = (\zeta_1^n, ..., \zeta_{p-q}^n)' = C_{21}^n (C_{11}^n)^{-1} W^n(1) - W^n(2)$ and $b = (b_1^n, ..., b^n) = (C_{11}^n)^{-1} \text{sign}(\beta_{(1)}^n)$.

It is standard result (see for example, Knight and Fu, 2000) that under regularity conditions (3) and (4),

$$(C_{11}^n)^{-1} W^n(1) \to_d N(0, C_{11}^{-1}),$$

and

$$C_{21}^n (C_{11}^n)^{-1} W^n(1) - W^n(2) \to_d N(0, C_{22} - C_{21} C_{11}^{-1} C_{12}).$$

Therefore all z_i^n's and ζ_i^n's converge in distribution to Gaussian random variables with mean 0 and finite variance $E(z_i^n)^2, E(\zeta_i^n)^2 \leq s^2$ for some constant $S > 0$.

For $t > 0$, the Gaussian distribution has its tail probability bounded by

$$1 - \Phi(t) < t^{-1} e^{-\frac{1}{2}t^2}. \tag{14}$$

Since $\frac{\lambda_n}{n} \to 0$, $\frac{\lambda_n}{n^{\frac{1+c}{2}}} \to \infty$ with $0 \leq c < 1$, p, q and β^n are all fixed, therefore

$$\begin{aligned} &\sum_{i=1}^{q} P(|z_i^n| \geq \sqrt{n}(|\beta_i^n| - \frac{\lambda_n}{2n} b_i^n) \\ \leq\ & (1 + o(1)) \sum_{i=1}^{q} (1 - \Phi((1 + o(1)) \frac{1}{s} n^{\frac{1}{2}} |\beta_i|)) \\ =\ & o(e^{-n^c}), \end{aligned}$$

and

$$\sum_{i=1}^{p-q} P(|\zeta_i^n| \geq \frac{\lambda_n}{2\sqrt{n}} \eta_i) = \sum_{i=1}^{p-q} (1 - \Phi(\frac{1}{s} \frac{\lambda_n}{2\sqrt{n}} \eta_i)) = o(e^{-n^c}).$$

Theorem 1 follows immediately.

Proof of Theorem 2. Consider the set F_1^n, on which there exists λ_n such that,

$$\begin{aligned} \text{sign}(\hat{\beta}_{(1)}^n) &= \text{sign}(\beta_{(1)}^n) \\ (\hat{\beta}_{(2)}^n) &= 0. \end{aligned}$$

General Sign Consistency implies that $P(F_1^n) \to 1$ as $n \to 1$.

Conditions of F_1^n imply that $\hat{\beta}_{(1)}^n \neq 0$ and $\hat{\beta}_{(2)}^n = 0$. Therefore by Lemma 1 and (11) from the proof of Proposition 1, we have

$$\begin{aligned} C_{11}^n(\sqrt{n}\hat{u}^n(1)) - W^n(1) &= -\frac{\lambda_n}{2\sqrt{n}}\text{sign}(\hat{\beta}_{(1)}^n) = -\frac{\lambda_n}{2\sqrt{n}}\text{sign}(\beta_{(1)}^n) && (15) \\ |C_{21}^n(\sqrt{n}\hat{u}^n(1)) - W^n(2)| &\leq \frac{\lambda_n}{2\sqrt{n}}\mathbf{1} && (16) \end{aligned}$$

which hold over F_1^n.

Re-write (16) by replacing $\hat{u}^n(1)$ using (15), we get

$$F_1^n \subset F_2^n := \{(\lambda_n/2\sqrt{n})L^n \leq C_{21}^n(C_{11}^n)^{-1}W^n(1) - W^n(2) \leq (\lambda_n/2\sqrt{n})U^n\}$$

where

$$\begin{aligned} L^n &= -1 + C_{21}^n(C_{11}^n)^{-1}\text{sign}(\beta_{(1)}^n), \\ U^n &= 1 + C_{21}^n(C_{11}^n)^{-1}\text{sign}(\beta_{(1)}^n). \end{aligned}$$

To prove by contradiction, if Weak Irrepresentable Condition fails, then for any N there always exists $n > N$ such that at least one element of $|C_{21}^n(C_{11}^n)^{-1}\text{sign}(\beta_{(1)}^n)| \geq 1$. Without loss of generality, assume the first element of $C_{21}^n(C_{11}^n)^{-1}\text{sign}(\beta_{(1)}^n) \geq 1$, then

$$[(\lambda_n/2\sqrt{n})L_1^n, (\lambda_n/2\sqrt{n})U_1^n] \subset [0, +\infty),$$

for any $\lambda_n \geq 0$. Since $C_{21}^n(C_{11}^n)^{-1}W^n(1) - W^n(2) \to_d N(0, C_{22} - C_{21}C_{11}^{-1}C_{12})$, there is a non-vanishing probability that the first element is negative, then the probability of F_2^n holds does not go to 1, therefore

$$\liminf P(F_1^n) \leq \liminf P(F_2^n) < 1.$$

This contradicts with the General Sign Consistency assumption. Therefore Weak Irrepresentable Condition is necessary for General Sign Consistency.

This completes the proof.

Proofs of Theorem 3 and 4 are similar to that of Theorem 1. The goal is to bound the tail probabilities in Proposition 1 using different conditions on the noise terms. We first derive the following inequalities for both Theorem 3 and 4.

Proof of Theorem 3 and Theorem 4. As in the proof of Theorem 1, we have

$$\begin{aligned} 1 - P(A_n \cap B_n) &\leq P(A_n^c) + P(B_n^c) \\ &\leq \sum_{i=1}^{q} P(|z_i^n| \geq \sqrt{n}(|\beta_i^n| - \frac{\lambda_n}{2n}b_i^n) + \sum_{i=1}^{p-q} P(|\zeta_i^n| \geq \frac{\lambda_n}{2\sqrt{n}}\eta_i). \end{aligned}$$

where $z^n = (z_1^n, ..., z_p^n)' = (C_{11}^n)^{-1}W^n(1)$, $\zeta^n = (\zeta_1^n, ..., \zeta_{p-q}^n)' = C_{21}^n(C_{11}^n)^{-1}W^n(1) - W^n(2)$ and $b = (b_1^n, ..., b^n) = (C_{11}^n)^{-1}\text{sign}(\beta_{(1)}^n)$.

Now if we write $z_i^n = H_A'\varepsilon^n$ where $H_A' = (h_1^a, ..., h_q^a)' = (C_{11}^n)^{-1}(n^{-\frac{1}{2}}\mathbf{X^n}(1))$, then

$$H_A'H_A = (C_{11}^n)^{-1}(n^{-\frac{1}{2}}\mathbf{X^n}(1)')((C_{11}^n)^{-1}(n^{-\frac{1}{2}}\mathbf{X^n}(1))')' = (C_{11}^n)^{-1}.$$

Therefore $z_i^n = (h_i^a)'\varepsilon$ with

$$\|h_i^a\|_2^2 \le \frac{1}{M_2} \text{ for } \forall i = 1,..,q. \tag{17}$$

Similarly if we write $\zeta^n = H_B'\varepsilon^n$ where $H_B' = (h_1^b,...,h_{p-q}^b)' = C_{21}^n(C_{11}^n)^{-1}(n^{-\frac{1}{2}}\mathbf{X^n}(1)')$ $-n^{-\frac{1}{2}}\mathbf{X^n}(2)'$, then

$$H_B'H_B = \frac{1}{n}(\mathbf{X^n}(2))'(I-\mathbf{X^n}(1)((\mathbf{X^n}(1)'(\mathbf{X^n}(1))^{-1}\mathbf{X^n}(1)')\mathbf{X^n}(2).$$

Since $I-\mathbf{X^n}(1)((\mathbf{X^n}(1)'(\mathbf{X^n}(1))^{-1}\mathbf{X^n}(1)'$ has eigenvalues between 0 and 1, therefore $\zeta_i^n = (h_i^b)'\varepsilon$ with

$$\|h_i^b\|_2^2 \le M_1 \text{ for } \forall i = 1,..,q. \tag{18}$$

Also notice that,

$$|\frac{\lambda_n}{n}b_n| = \frac{\lambda_n}{n}|(C_{11}^n)^{-1}\text{sign}(\beta_{(1)}^n)| \le \frac{\lambda_n}{nM_2}\|\text{sign}(\beta_{(1)}^n)\|_2 = \frac{\lambda_n}{nM_2}\sqrt{q} \tag{19}$$

Proof of Theorem 3. Now, given (17) and (18), it can be shown that $E(\varepsilon_i^n)^{2k} < \infty$ implies $E(z_i^n)^{2k} < \infty$ and $E(\zeta_i^n)^{2k} < \infty$. In fact, given constant n-dimensional vector α,

$$E(\alpha'\varepsilon^n)^{2k} \le (2k-1)!!\|\alpha\|_2^2 E(\varepsilon_i^n)^{2k}.$$

For radome variables with bounded $2k$'th moments, we have their tail probability bounded by

$$P(z_i^n > t) = O(t^{-2k}).$$

Therefore, for $\lambda/\sqrt{n} = o(n^{\frac{c_2-c_1}{2}})$, using (19), we get

$$\begin{aligned} & \sum_{i=1}^{q} P(|z_i^n| > \sqrt{n}(|\beta_i^n| - \frac{\lambda_n}{2n}b_i^n) \\ = \quad & qO(n^{-kc_2}) = o(\frac{pn^k}{\lambda_n^{2k}}). \end{aligned}$$

Whereas

$$\begin{aligned} & \sum_{i=1}^{p-q} P(|\zeta_i^n| > \frac{\lambda_n}{2\sqrt{n}}\eta_i) \\ = \quad & (p-q)O(\frac{n^k}{\lambda_n^{2k}}) = O(\frac{pn^k}{\lambda_n^{2k}}). \end{aligned}$$

Sum these two terms and notice for $p = o(n^{c_2-c_1})$, there exists a sequence of λ_n s.t. $\lambda/\sqrt{n} = o(n^{\frac{c_2-c_1}{2}})$ and $= o(\frac{pn^k}{\lambda_n^{2k}})$. This completes the proof for Theorem 3.

Proof of Theorem 4. Since ε_i^n's are i.i.d. Gaussian, therefore by (17) and (18), z_i's and ζ_i's are Gaussian with bounded second moments.

Using the tail probability bound (14) on Gaussian random variables , for $\lambda_n \propto n^{\frac{1+c_4}{2}}$, by (19) we immediately have

$$\begin{aligned} & \sum_{i=1}^{q} P(|z_i^n| > \sqrt{n}(|\beta_i^n| - \frac{\lambda_n}{2n} b_i^n) \\ = \quad & q \cdot O(1 - \Phi(\ (1+o(1))M_3 M_2 n^{c_2/2}\ ) = o(e^{-n^{c_3}}). \end{aligned}$$

(since $q < n = e^{\log n}$) and

$$\begin{aligned} & \sum_{i=1}^{p-q} P(|\zeta_i^n| > \frac{\lambda_n}{2\sqrt{n}} \eta_i) \\ = \quad & (p-q) \cdot O(1 - \Phi(\frac{1}{M_1} \frac{\lambda_n}{\sqrt{n}}) \eta) = o(e^{-n^{c_3}}). \end{aligned}$$

This completes the proof for Theorem 4.

Proof of Corollary 1. First we recall, for a positive definite matrix of the form

$$\begin{pmatrix} a & b & \cdots & b & b \\ b & a & \cdots & b & b \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ b & b & \cdots & a & b \\ b & b & \cdots & b & a \end{pmatrix}_{q \times q} .$$

The eigenvalues are $e_1 = a + (q-1)b$ and $e_i = a - b$ for $i \geq 2$. Therefore the inversion of

$$C_{11}^n = \begin{pmatrix} 1 & r_n & \cdots & r_n & r_n \\ r_n & 1 & \cdots & r_n & r_n \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ r_n & r_n & \cdots & 1 & r_n \\ r_n & r_n & \cdots & r_n & 1 \end{pmatrix}_{q \times q}$$

can be obtained by applying the formula and taking reciprocal of the eigenvalues which gets us

$$(C_{11}^n)^{-1} = \begin{pmatrix} c & d & \cdots & d & d \\ d & c & \cdots & d & d \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ d & d & \cdots & c & d \\ d & d & \cdots & d & c \end{pmatrix}_{q \times q}$$

for which $e_1' = c + (q-1)d = \frac{1}{e_1} = \frac{1}{1+(q-1)r_n}$.

Now since $C_{21}^n = r_n \times \mathbf{1}_{(p-q)\times q}$ so

$$C_{21}^n (C_{11}^n)^{-1} = r_n(c + (q-1)d)\mathbf{1}_{(p-q)\times q} = \frac{r_n}{1+(q-1)r_n} \mathbf{1}_{(p-q)\times q}.$$

By which, we get

$$\begin{aligned}|C_{21}^n(C_{11}^n)^{-1}\mathrm{sign}(\beta_{(1)}^n)| &= \frac{r_n}{1+(q-1)r_n}|\sum \mathrm{sign}(\beta_i)|\mathbf{1}_{q\times 1}\\ &\le \frac{qr_n}{1+(q-1)r_n}\mathbf{1}_{q\times 1} \le \frac{\frac{q}{1+cq}}{1+\frac{q-1}{1+cq}} = \frac{1}{1+c},\end{aligned}$$

that is, Strong Irrepresentable Condition holds.

Proof of Corollary 2. Without loss of generality, consider the first entry of $|C_{21}^n(C_{11}^n)^{-1}\mathrm{sign}(\beta_{(1)}^n)|$ which takes the form $|\alpha'(C_{11}^n)^{-1}\mathrm{sign}(\beta_{(1)}^n)|$ where α' is the first row of C_{21}^n. After proper scaling, this is bounded by the largest eigenvalue of $(C_{11}^n)^{-1}$ or equivalently the reciprocal of the smallest eigenvalue of C_{11}^n, that is,

$$|\alpha'(C_{11}^n)^{-1}\mathrm{sign}(\beta_{(1)}^n)| \le \|\alpha\|\|\mathrm{sign}(\beta_{(1)}^n)\|\frac{1}{e_1} < \frac{cq}{2q-1}\frac{1}{e_1}. \tag{20}$$

To bound e_1, we assume $C_{11}^n = c_{ij_{q\times q}}$. Then for a unit length $q\times 1$ vector $x=(x_1,...,x_q)'$, we consider

$$\begin{aligned}x'C_{11}^n x &= \sum_{i,j} x_i c_{ij} x_j = 1+\sum_{i\ne j} x_i c_{ij} x_j\\ &\ge 1-\sum_{i\ne j}|x_i||c_{ij}||x_j|\\ &\ge 1-\frac{1}{2q-1}\sum_{i\ne j}|x_i||x_j|\\ &= 1-\frac{1}{2q-1}(\sum_{i,j}|x_i||x_j|-1)\\ &\ge 1-\frac{q-1}{2q-1} = \frac{q}{2q-1}, \end{aligned}\tag{21}$$

where the last inequality is by Cauchy-Schwartz. Now put (21) through (20), we have $|C_{21}^n(C_{11}^n)^{-1}\mathrm{sign}(\beta_{(1)}^n)| < c\mathbf{1}$. This completes the proof for Corollary 2.

Proof of Corollary 3. Without loss of generality, let us assume x_j, $j=1,...,n$ are i.i.d. $N(0,C^n)$ random variables. Then the power decay design implies an AR(1) model where

$$\begin{aligned}x_{j1} &= \eta_{j1}\\ x_{j2} &= \rho x_{j1}+(1-\rho^2)^{\frac{1}{2}}\eta_{j2}\\ &\vdots\\ x_{jp} &= \rho x_{j(p-1)}+(1-\rho^2)^{\frac{1}{2}}\eta_{jp}\end{aligned}$$

where η_{ij} are i.i.d. $N(0,1)$ random variables. Thus, the predictors follow a Markov Chain:

$$x_{j1}\longrightarrow x_{j2}\longrightarrow\cdots\longrightarrow x_{jp}.$$

Now let

$$\begin{aligned}I_1 &= i:\beta_i\ne 0\\ I_2 &= i:\beta_i = 0.\end{aligned}$$

For $\forall k \in I_2$, assume

$$\begin{aligned} k_l &= \{i : i < k\} \cap I_1 \\ k_h &= \{i : i > k\} \cap I_1. \end{aligned}$$

Then by the Markov property, we have

$$x_{jk} \perp x_{jg} | (x_{jk_l}, x_{jk_h})$$

for $j = 1, .., n$ and $\forall g \in I_1 / \{k_l, k_h\}$. Therefore by the regression interpretation as in (2), to check Strong Irrepresentable Condition for x_{jk} we only need to consider x_{jk_l} and x_{jk_h} since the rest of the entries are zero by the conditional independence. To further simplify, we assume $\rho \geq 0$ (otherwise ρ can be modified to be positive by flipping the signs of predictors $1, 3, 5, ...$). Now regressing x_{jk} on (x_{jk_l}, x_{jk_h}) we get

$$\begin{aligned} & \mathrm{Cov}(\begin{pmatrix} x_{jk_l} \\ x_{jk_h} \end{pmatrix})^{-1} \mathrm{Cov}(x_{jk}, \begin{pmatrix} x_{jk_l} \\ x_{jk_h} \end{pmatrix}) \\ = & \begin{pmatrix} 1 & \rho^{k_h - k_l} \\ \rho^{k_h - k_l} & 1 \end{pmatrix}^{-1} \begin{pmatrix} \rho^{k_h - k} \\ \rho^{k - k_l} \end{pmatrix} \\ = & \begin{pmatrix} \frac{\rho^{k_l - k} - \rho^{k - k_l}}{\rho^{k_l - k_h} - \rho^{k_h - k_l}} \\ \frac{\rho^{k - k_h} - \rho^{k_h - k}}{\rho^{k_l - k_h} - \rho^{k_h - k_l}} \end{pmatrix}. \end{aligned}$$

Then sum of both entries follow

$$\frac{\rho^{k_l - k} - \rho^{k - k_l}}{\rho^{k_l - k_h} - \rho^{k_h - k_l}} + \frac{\rho^{k - k_h} - \rho^{k_h - k}}{\rho^{k_l - k_h} - \rho^{k_h - k_l}} = \frac{\rho^{k_l - k} + \rho^{k - k_h}}{1 + \rho^{k_l - k_h}} = 1 - \frac{(1 - \rho^{k_l - k})(1 - \rho^{k - k_h})}{1 + \rho^{k_l - k_h}} < 1 - \frac{(1 - c)^2}{2}.$$

Therefore Strong Irrepresentable Condition holds entry-wise. This completes the proof.

Proof of Corollary 4.

(a) Since the correlations are all zero so the condition of Corollary 2 holds for $\forall q$. Therefore Strong Irrepresentable Condition holds.

(b) Since $q = 1$, so $\frac{1}{2q-1} = 1$ therefore the condition of Corollary 2 holds. Therefore Strong Irrepresentable Condition holds.

(c) Since $p = 2$, therefore for $q = 0$ or 2, proof is trivial. When $q = 1$, result is implied by (b).

Proof of Corollary 5.

Let $\mathcal{M}$ be the set of indices of nonzero entries of β^n and $\mathcal{B}_j,\ j = 1,..,k$ be the set of indices of each block. Then the following holds

$$\begin{aligned}
& C^n_{21}(C^n_{11})^{-1}\mathrm{sign}(\beta^n_{(1)}) \\
= & \begin{pmatrix} C^n_{\mathcal{M}^c\cap\mathcal{B}_1,\mathcal{M}\cap\mathcal{B}_1} & \cdots & 0 \\ \cdots & \ddots & \cdots \\ 0 & \cdots & C^n_{\mathcal{M}^c\cap\mathcal{B}_k,\mathcal{M}\cap\mathcal{B}_k} \end{pmatrix} \\
& \times \begin{pmatrix} C^n_{\mathcal{M}\cap\mathcal{B}_1,\mathcal{M}\cap\mathcal{B}_1} & \cdots & 0 \\ \cdots & \ddots & \cdots \\ 0 & \cdots & C^n_{\mathcal{M}\cap\mathcal{B}_k,\mathcal{M}\cap\mathcal{B}_k} \end{pmatrix}^{-1} \mathrm{sign}(\begin{pmatrix} \beta^n_{\mathcal{M}\cap\mathcal{B}_1} \\ \vdots \\ \beta^n_{\mathcal{M}\cap\mathcal{B}_k} \end{pmatrix}) \\
= & \begin{pmatrix} C^n_{\mathcal{M}^c\cap\mathcal{B}_1,\mathcal{M}\cap\mathcal{B}_1}(C^n_{\mathcal{M}\cap\mathcal{B}_1,\mathcal{M}\cap\mathcal{B}_1})^{-1}\mathrm{sign}(\beta^n_{\mathcal{M}\cap\mathcal{B}_1}) \\ \vdots \\ C^n_{\mathcal{M}^c\cap\mathcal{B}_k,\mathcal{M}\cap\mathcal{B}_k}(C^n_{\mathcal{M}\cap\mathcal{B}_k,\mathcal{M}\cap\mathcal{B}_k})^{-1}\mathrm{sign}(\beta^n_{\mathcal{M}\cap\mathcal{B}_k}) \end{pmatrix}.
\end{aligned}$$

Corollary 5 is implied immediately from the shown equalities.

References

Z. D. Bai. Methodologies in spectral analysis of large dimensional random matrices: A review. *Statistica Sinica*, (9):611–677, 1999.

D. Donoho, M. Elad, and V. Temlyakov. Stable recovery of sparse overcomplete representations in the presence of noise. *Preprint*, 2004.

B. Efron, T. Hastie, and R. Tibshirani. Least angle regression. *Annals of Statistics*, 32:407–499, 2004.

A.E. Hoerl and R.W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12:55–67, 1970.

K. Knight and W. J. Fu. Asymptotics for Lasso-type estimators. *Annals of Statistics*, 28:1356–1378, 2000.

C. Leng, Y. Lin, and G. Wahba. A note on the lasso and related procedures in model selection. *Statistica Sinica*, (To appear), 2004.

N Meinshausen. Lasso with relaxation. *Technical Report*, 2005.

N. Meinshausen and P. Buhlmann. Consistent neighbourhood selection for high-dimensional graphs with the Lasso. *Annals of Statistics*, 34(3), 2006.

M.R. Osborne, B. Presnell, and B.A. Turlach. Knot selection for regression splines via the Lasso. *Computing Science and Statistics*, (30):44–49, 1998.

M.R. Osborne, B. Presnell, and B.A. Turlach. On the lasso and its dual. *Journal of Computational and Graphical Statistics*, (9(2)):319–337, 2000b.

S. Rosset. Tracking curved regularized optimization solution paths. *NIPS*, 2004.

R. Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B*, 58(1):267–288, 1996.

P. Zhao and B. Yu. Boosted Lasso. *Technical Report*, *Statistics Department, UC Berkeley*, 2004.

H. Zou, T. Hastie, and R. Tibshirani. On the ”degrees of freedom” of the Lasso. *submitted*, 2004.