ADAPTIVE LEARNING IN LASSO MODELS

A Thesis
Presented to
The Academic Faculty

by

Kaushik Patnaik

In Partial Fulfillment
of the Requirements for the Degree
Masters in Computer Science in the
School of Computational Science and Engineering

Georgia Institute of Technology
December 2015

ADAPTIVE LEARNING IN LASSO MODELS

Approved by:

Professor Le Song, Advisor
School of Computational Science and Engineering
Georgia Institute of Technology

Professor Bistra Dilkina
School of Computational Science and Engineering
Georgia Institute of Technology

Professor Polo Chau
School of Computational Science and Engineering
Georgia Institute of Technology

Professor Mark Davenport
School of Electrical and Computer Engineering
Georgia Institute of Technology

Date Approved: 15 Aug 2015

To my parents,

Saroj K Patnaik and Rita Das,

for inspiring and supporting me so far.

PREFACE

Regression with L1-regularization, Lasso, is a popular algorithm for recovering the sparsity pattern (also known as model selection) in linear models from observations contaminated by noise. We examine a scenario where a fraction of the zero co-variates are highly correlated with non-zero co-variates making sparsity recovery difficult. We propose two methods that adaptively increment the regularization parameter to prune the Lasso solution set. We prove that the algorithms achieve consistent model selection with high probability while using fewer samples than traditional Lasso. The algorithm can be extended to a broad set of L1-regularized M-estimators for linear statistical models.

ACKNOWLEDGEMENTS

I want to thank my adviser Prof Le Song, for guiding me the past year, and helping me understand the in and out of Machine Learning research. I would also like to thank Prof Nina Balcan for providing me the opportunity to start working on Machine Learning research.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER I

INTRODUCTION

We consider the linear regression problem, where we observe a set of input $X_1, ..., X_n \in R^p$ and output vectors $Y_1, ..., Y_n$. We assume our data is generated by a linear regression model, with target vector β^*.

$$Y = X\beta^* + \epsilon \tag{1}$$

where ϵ is the noise variable, a vector of n i.i.d random variables with mean zero and variance σ^2. Y is the nx1 outout variable and $X = (X_1, X_2, X_3, ..., X_p)$ is the n x p design matrix where X_i is the i^{th} predictor (column) and X^j is the j^{th} sample (row). The model is assumed to be "sparse", that is some of the regression coefficients β^* are exactly zero corresponding to predictors that are irrelevant to the response. The non-zero co-efficients, also called as the true set, are defined by the set $S : i : |\beta_i^*| \neq 0$ with s being the cardinality of the set S.

This paper focusses on the feature selection problem, where we are interested in estimating the set of non-zero coefficients of β^*. The standard method is subset selection, which computes the following estimator

$$\hat{\beta}_{L_0} = argmin_{\beta \in R^p} ||X\beta - Y||_2^2 \quad subject\ to\ ||\beta||_0 \leq k \tag{2}$$

where k is tuning parameter. However, the optimization problem in (2) is nonconvex, and the global solution to the problem cannot be efficiently computed.The most popular approximation to L_0 regularization is the L_1 regularization method, also

known as Lasso. The lasso estimates $\hat{\beta}$ are defined by

$$\hat{\beta} = \arg\min_{\beta} \frac{1}{2n}||Y - X\beta||_2^2 + \lambda||\beta||_1 \tag{3}$$

where > 0 is a regularization parameter and $||.||_1$ equals the sum of absolute values of the vector's entries.

The λ parameter controls the amount of regularization applied to the coefficients. A very large λ completely shrinks the coefficients β to zero thus leading to an empty model. In practice, the choice of λ is done by cross-validation.

The global optimum of (3) can be computed using standard convex optimization techniques. The performance of lasso, for both feature selection and parameter estimation, has been theoretically analyzed. For theoretical analysis of feature selection, the sign of estimated $\hat{\beta}$ is also compared with β^*, called model selection consistency.

$$P(Sign(\hat{\beta}) = Sign(\beta_S^*))$$

Among previous work, in particular Zhao and Yu[6] show, for fixed p and p growing with n, there exists an **Irrepresentable/Incoherence Condition** that is almost necessary and sufficient for model selection. The incoherence condition gives an upper bound on the correlation between non-zero co-variates (X_j for $j \in$ S) and zero co-variates (X_j for $j \in$ S^c).

1.1 Incoherence Condition

Let us assume, without loss of generality, for $\beta = (\beta_1, \beta_2, ..., \beta_p)^T$ where $\beta_j \neq 0$ for $j = 1, 2...s$ and $\beta_j = 0$ for $j = s+1, ...p$. Let $\beta_S = (\beta_1, \beta_2, ...\beta_s)^T$ and $\beta_{S^c} = (\beta_{s+1}, ...\beta_p)^T$. Therefore β_S contains all non-zero coefficients in β and β_{S^c} contains all zero coefficients. Similarly we can divide the design matrix X into smaller matrices

related to non-zero and zero coefficients. Let X_S contain the s columns that get multiplied with the non-zero co-efficients, and X_{S^c} contain the rest of the $(p-s)$ columns.

Based on these assumptions, we define the Incoherence condition as follows

Incoherence Condition There exists a positive constant vector η

$$|(\frac{1}{n}X_{S^c}^T X_S)(\frac{1}{n}X_S^T X_S)^{-1} sign(\beta_S)|_\infty \leq 1-\eta \tag{4}$$

where $|.|_\infty$ stands for the L_∞ vector norm (i.e the maximum value in the $(p-s)$ x 1 vector). The above equation can also be represented as an element wise inequality in the vector, with individual η_j for each $j \in S^c$. .

The incoherence condition closely resembles a constraint on the regression coefficients of the irrelevant/zero co-variates X_{S^c} with the relevant variates X_S. Assuming a bounded inverse of $(\frac{1}{n}X_S^T X_S)$, a lower η indicates higher correlation between co-variates in S^c and S. This condition impacts the Lasso model selection probability, as higher correlation leads to larger error in model selection.

1.2 Our Contribution

In this paper, we consider scenarios where incoherence is unevenly distributed among the zero co-variates. We assume two possible distributions of incoherence among the zero co-variates, and propose lasso procedures that exploit this structure to do model selection with fewer number of samples than traditional lasso methods. Our main results are

- a two-stage lasso procedure, where given the parameters $(\min_s |\beta_S^*|, \eta, \sigma)$ we can recover the true set with high probability

- a multi-stage procedure which adaptively updates the regularization parameter λ and recovers the true set with high probability in finite rounds

The rest of the paper is organized as follows. In chapter 2 we summarize work related to feature selection in Lasso and other feature screening methods. In chapter 3 we describe the problem setting and our proposed algorithm. In chapter 4, we summarize the theoretical proofs for feature selection of our methods. Finally in chapter 5 we demonstrate the advantages of our methods in simulation studies.

CHAPTER II

PREVIOUS WORK

The model selection consistency of lasso has been extensively studied by several authors, and some of the most notable papers are Meinhausen and Bhulmann[3], Wainwright [4], and Zhao and Yu[6]. Meinhausen and Bhulman first established a proof of lasso model selection using incoherence condition in Gaussian graphical models. Zhao and Yu later proved that incoherence condition is a necessary condition for model selection in lasso. Precise conditions on the problem dimension p, the number of nonzero elements s and number of samples n were proved by Wainwright. Wainwright also extended the proofs from deterministic models to random Gaussian ensembles.

In literature we have not yet found work exploiting the distribution of the incoherence parameter. Most work related to adaptive lasso focuses on improving the lasso bias and model selection probability rather than sample usage [7]. Such methods involve processing the entire data twice, and thus are data inefficient. Recently work by Ghaoui [1] have developed fast rules/methods that allow removal of zero co-variates that are guaranteed to be removed later by lasso (using properties of dual Lasso). However, these methods must be applied before the lasso algorithm, and thus are data inefficient. In the next section we summarize the important proofs from these papers related to model selection in Lasso.

2.1 Events needed for Lasso Model Selection

An estimate which is consistent in terms of parameter estimation may not be consistent in estimating the correct model. For model consistency we make the following definition about Sign Consistency that does not assume the estimates to be estimation consistent.

Sign Consistency: An estimate $\hat{\beta}$ is consistent with the true model β^* if an only if

$$sign(\hat{\beta}) = sign(\beta^*) \tag{5}$$

where the equality holds element wise. Sign(.) maps positive entries to 1, negative entries to -1, and zero to zero. Sign consistency is a stronger requirement than the usual selection consistency which only requires the zeros to be matched.

Using this definition we can define a lower bound on the probability of lasso selecting the correct model.

Lemma 1 *Assume strong irrepresentable condition holds with a constant* $\eta > 0$ *then*

$$P(Sign(\hat{\beta}^{\lambda}) = Sign(\beta^*)) \geq P(A_n \cap B_n) \tag{6}$$

for

$$A_n = |(\frac{1}{n}X_S^T X_S)^{-1}\frac{X_S^T \epsilon}{n}| < |\beta_S^*| - \lambda|(\frac{1}{n}X_S^T X_S)^{-1}Sign(\beta_S^*)| \tag{7}$$

$$B_n = |\frac{1}{n}X_{S^c}^T X_S(\frac{1}{n}X_S^T X_S)^{-1}\frac{X_S^T \epsilon}{n} - \frac{X_{S^c}^T \epsilon}{n}| \leq \lambda\eta \tag{8}$$

Proof: Since Lasso is a convex optimization problem, the solution to the Lasso should satisfy the sub-gradient equation

$$\frac{1}{n}X^TX(\hat{\beta}-\beta^*)-\frac{1}{n}X^T\epsilon+\lambda\hat{\gamma}=0 \tag{9}$$

Where γ is the sub-gradient associated with $|\beta|_1$. Let us assume, without loss of generality, for $\beta=(\beta_1,\beta_2,...,\beta_p)^T$ where $\beta_j\neq 0$ for $j=1,2...s$ and $\beta_j=0$ for $j=s+1,...p$. Let $\beta_S=(\beta_1,\beta_2,...\beta_s)^T$ and $\beta_{S^c}=(\beta_{s+1},...\beta_p)^T$. Similarly we can divide the design matrix X into smaller matrices related to non-zero and zero coefficients. Let X_S contain the s columns that get multiplied with the non-zero co-efficients, and X_{S^c} contain the rest of the p-s columns.

Then the X^TX term can be written down as $\begin{vmatrix} X_S^TX_S & X_S^TX_{S^c} \\ X_{S^c}^TX_S & X_{S^c}^TX_{S^c} \end{vmatrix}$. $(\hat{\beta}-\beta^*)$ can be written as $\begin{vmatrix} \hat{\beta}_S-\beta_S^* \\ 0 \end{vmatrix}$. The sub-gradient γ and $X_S^T\epsilon$ can also be written in the same format. This leads to two equations, one for all variables in S with the 1st row of the sub-gradient equation, and another for all variables in S^c. The 1st row of the sub-gradient equation can be written down as

$$(\hat{\beta}_S-\beta_S^*)=\left(\frac{1}{n}X_S^TX_S\right)^{-1}\left(\frac{1}{n}X_S^T\epsilon-\lambda\gamma_S\right) \tag{10}$$

Now for a solution to exist we have $|\hat{\beta}_S-\beta_S^*|<|\beta_S^*|$. Thus we get the equation

$$\left|\left(\frac{1}{n}X_S^TX_S\right)^{-1}\left(\frac{1}{n}X_S^T\epsilon-\lambda\gamma_S\right)\right|<|\beta_S^*| \tag{11}$$

which is implied by the following equation

$$\left|\left(\frac{1}{n}X_S^TX_S\right)^{-1}\frac{1}{n}X_S^T\epsilon\right|<|\beta_S^*|-\lambda\left|\left(\frac{1}{n}X_S^TX_S\right)^{-1}Sign(\beta_S)\right| \tag{12}$$

This proves the equation for event A_n. Similarly looking at the 2nd row of the sub-gradient equation, $\forall i\in S^c$ we have

$$\left(\frac{1}{n}X_{S^c}^T X_S\right)(\hat{\beta}_S - \beta^*) - \frac{X_{S^c}^T \epsilon}{n} + \lambda\gamma_{S^c} = 0 \tag{13}$$

Since $\gamma_{S^c} \in (-1, 1)$, we have

$$\left|\left(\frac{1}{n}X_{S^c}^T X_S\right)(\hat{\beta}_S - \beta^*) - \frac{X_{S^c}^T \epsilon}{n}\right| < \left|\lambda\right|$$

Substituting the values of $\hat{\beta}_S - \beta^*$ from Eq.(34) we get the following inequality

$$\left|\left(\frac{1}{n}X_{S^c}^T X_S\right)\left(\frac{1}{n}X_S^T X_S\right)^{-1}\frac{X_S^T \epsilon}{n} - \lambda\left(\frac{1}{n}X_{S^c}^T X_S\right)\left(\frac{1}{n}X_S^T X_S\right)^{-1} Sign(\beta_S^*) - \frac{X_{S^c}^T \epsilon}{n}\right| < \left|\lambda\right|$$

which can be realized by the following inequality

$$\left|\left(\frac{1}{n}X_{S^c}^T X_S\right)\left(\frac{1}{n}X_S^T X_S\right)^{-1}\frac{X_S^T \epsilon}{n} - \frac{X_{S^c}^T \epsilon}{n}\right| < \lambda\left(1 - \left|\left(\frac{1}{n}X_{S^c}^T X_S\right)\left(\frac{1}{n}X_S^T X_S\right)^{-1} Sign(\beta_S^*)\right|\right)$$

where 1 indicates a vector of 1's. Applying the irrepresentable condition (3), we get

$$\left|\left(\frac{1}{n}X_{S^c}^T X_S\right)\left(\frac{1}{n}X_S^T X_S\right)^{-1}\frac{X_S^T \epsilon}{n} - \frac{X_{S^c}^T \epsilon}{n}\right| < \lambda\eta \tag{14}$$

Where η represents the vector of η_i's. This proves event B_n.

CHAPTER III

SETTING AND PROPOSED ALGORITHM

In this paper, we study two scenarios where zero co-variates X_j for all $j \in S^c$ have different incoherence parameters:

- $\alpha(p-s)$ co-variates have incoherence parameter η_1, and $(1-\alpha)(p-s)$ co-variates have incoherence parameter η_2 with $\eta_1 < \eta_2$
- the value of incoherence follows an exponential distribution (i.e fewer zero co-variates have low incoherence values, and most zero co-variates have large incoherence values)

3.1 Two Stage Lasso Procedure

In the scenario where $\alpha(p-s)$ co-variates have incoherence parameter η_1, and $(1-\alpha)(p-s)$ co-variates have incoherence parameter η_2 with $\eta_1 < \eta_2$, we propose a two-stage lasso procedure with two regularization parameters λ_1 and λ_2. Let Y^{kn}, X^{kn} represent kXn rows of Y and X, and X_B, β_B represent X and β restricted to columns in set B. The algorithm proceeds as follows

Two-Stage Lasso Method

Input: p,n,s,k,min β_S^*,σ,η_1,η_2,α

Output: $\hat{\beta}_2$

Initialize $\lambda_1 = \lambda_2 = \lambda_{init} = n^{\frac{log(log(p))}{2log(n)} - \frac{1}{2}}$

- Calculate $\lambda_{ub} = \left(\frac{M_2}{4\sigma\sqrt{M_2}+\sqrt{s}}\right) \min_S |\beta_S^*|$

- Obtain λ_1, λ_2 and k using optimization procedure described below
- Estimate $\hat{\beta}_1$ using kn samples

$$\hat{\beta}_1 = \arg\min_{\beta} \frac{1}{2kn} ||Y^{kn} - X^{kn}\beta||_2^2 + \lambda_1 ||\beta||_1$$

- Define set $B = \{j : j \in p,\ \hat{\beta}_{1j} \neq 0\}$
- Estimate $\hat{\beta}_2$ using (1-k)n samples

$$\hat{\beta}_2 = \arg\min_{\beta} \frac{1}{2(1-k)n} ||Y^{(1-k)n} - X_B^{(1-k)n}\beta_B||_2^2 + \lambda_2 ||\beta_B||_1$$

where $\hat{\beta}_2$ is the final estimate used for feature selection. M_2 is the minimum eigenvalue of $\frac{1}{n} X_S^T X_S$. Since we use only kn samples in the first stage, given an appropriate λ_1, we are able to learn $(1-\alpha)(p-s)$ zero co-variates with high probability and eliminate them after the first stage. Thus the two-stage lasso procedure uses $(1-k)n(1-\alpha)(p - E(Supp(B)))$ fewer sample data than a single stage lasso with similar success probability, where $E(Supp(B))$ is the expected cardinality of set B.

Since we require the two-stage procedure to have high probability of feature selection while using fewer samples, we can run the following optimization algorithm which maximizes the number of samples we avoid using.

Optimization Algorithm:

$$\arg\max_{\lambda_1, \lambda_2, k} (1-k)(1-\alpha)(p - E(Supp(B)))n$$

$$s.t$$

$$\lambda_1 < \lambda_{ub}$$

$$\lambda_2 < \lambda_{ub}$$

$$k < 1$$

$$P(Sign(\hat{\beta}_2) = Sign(\beta^*)) \geq P(Sign(\hat{\beta}_1) = Sign(\beta^*))$$

To establish the proof for the two-stage procedure, we need to prove the existence of a certain $\lambda_1 \in (\lambda_{init}, \lambda_{ub})$ which removes the low correlation variables and another regularization parameter λ_2 which achieves better feature selection than λ_1.

3.2 Multi-Stage Lasso Procedure

In the scenario where, the incoherence values follow an exponential distribution, we can approximate the exponential distribution by selecting a set of η's and α's that lower bounds the continuous distribution. The choice of such η's and α's determine how well the algorithm will work in practice.

Given a set of $\eta = (\eta_1, \eta_2, ...\eta_l)$ and $\alpha = (\alpha_1, \alpha_2, ..., \alpha_l)$, we can extend the two-stage procedure to multiple stages. Let Y^{kn}, X^{kn} represent kn rows of Y and X, and X_B, β_B represent X and β restricted to columns in set B.

Multi-Stage Lasso Method

Input: p,n,s,min β_S^*,σ,η,α

Output: $\hat{\beta}_l$

Initialize $\lambda = \lambda_{init} = n^{\frac{log(log(p))}{2log(n)} - \frac{1}{2}}$

- Calculate $\lambda_{ub} = \left(\frac{M_2}{4\sigma\sqrt{M_2} + \sqrt{s}}\right) \min_S |\beta_S^*|$
- For $i = 1, 2, ...l$
  - Estimate $\lambda_i \geq 4log(\alpha_i(p-s))\sigma^2 M_1 / n\eta_i^2$
  - Estimate $\hat{\beta}_i$ using $n_i = n/2^i$ samples

$$\hat{\beta}_i = \arg\min_{\beta} \frac{1}{n_i} ||Y^{n_i} - X^{n_i}_{B_{i-1}} \beta_{B_{i-1}}||_2^2 + \lambda_i ||\beta_{B_{i-1}}||_1$$

  - Define set $B_i = \{j : j \in p,\ \hat{\beta}_{ij} \neq 0\}$

In this procedure $\hat{\beta}_l$ is the final estimate used for feature selection. M_2 is the minimum eigenvalue of $\frac{1}{n}X_S^T X_S$, and M_1 is the maximum $||(.)||_2$ of any column of X. The multi-stage lasso procedure is similar to the two-stage procedure, except for the optimization around sample usage.

CHAPTER IV

THEORETICAL ANALYSIS

We derive the model selection proofs using techniques from Zhao and Yu[6], Zhang[5] and Wainwright [4]. We restrict our proofs to deterministic design matrix X.We also allow the model parameters β, p, s to grow as n grows.

Assumption 1 *Assume that* $\{\epsilon_i\}_{i=1,2,..,n}$ *in (1) are independent sub-Gaussians. There exists* $\sigma \geq 0$ *such that* $\forall i$ *and* $\forall t \in R$

$$E_{\epsilon_i} e^{t\epsilon_i} \leq e^{\sigma^2 t^2/2}$$

Assumption 1 allows us to bound the overall noise in the Lasso estimation, and enables model selection by bounding the rate of growth of noise to $n^{-1/2}$

Assumption 2 *There exists* $0 \leq c_1 + c_3 < c_2 \leq 1$ *and* $M_1, M_2, M_3 > 0$ *so that the following holds:*

$$\frac{1}{n} X_i^T X_i \leq M_1 \ for \ \forall i \tag{15}$$

$$\alpha^T \frac{1}{n} X_S^T X_S \alpha \geq M_2, \ for \ \forall ||\alpha||_2^2 = 1 \tag{16}$$

$$s = O(n^{c_1}) \tag{17}$$

$$p = O(e^{-n^{c_3}}) \tag{18}$$

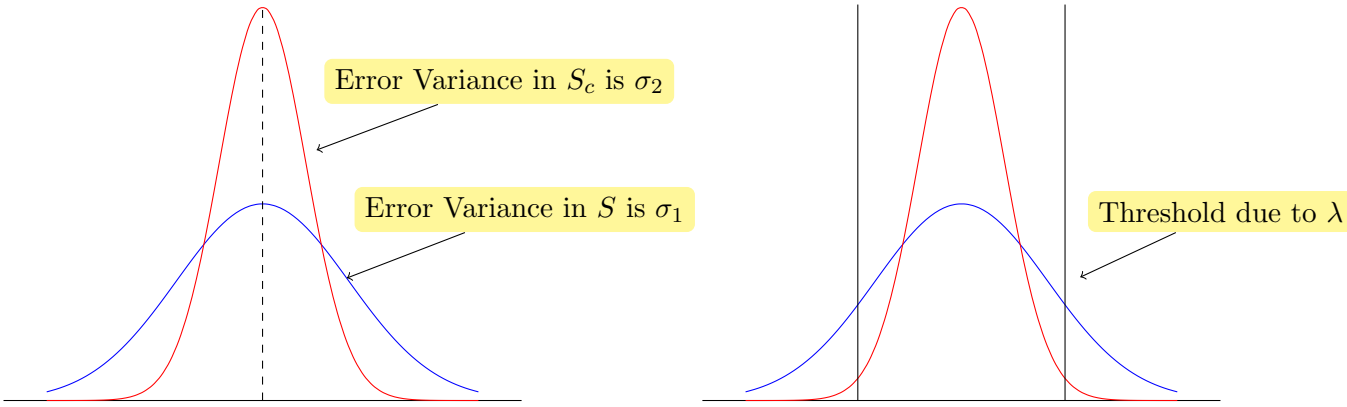

Figure 1: Left: The model selection error in Lasso can be divided into Type I error due to mistakes in set S, and Type II error due to mistakes in set S_c. Under the assumption of sub Gaussian noise, error probability is a zero mean sub Gaussian random variable with variance determined by $n, p, s, \min \beta$. Right: The upper bound on the error probability is determined by the value of the regularization parameter and the variance of the sub Gaussian random variable.

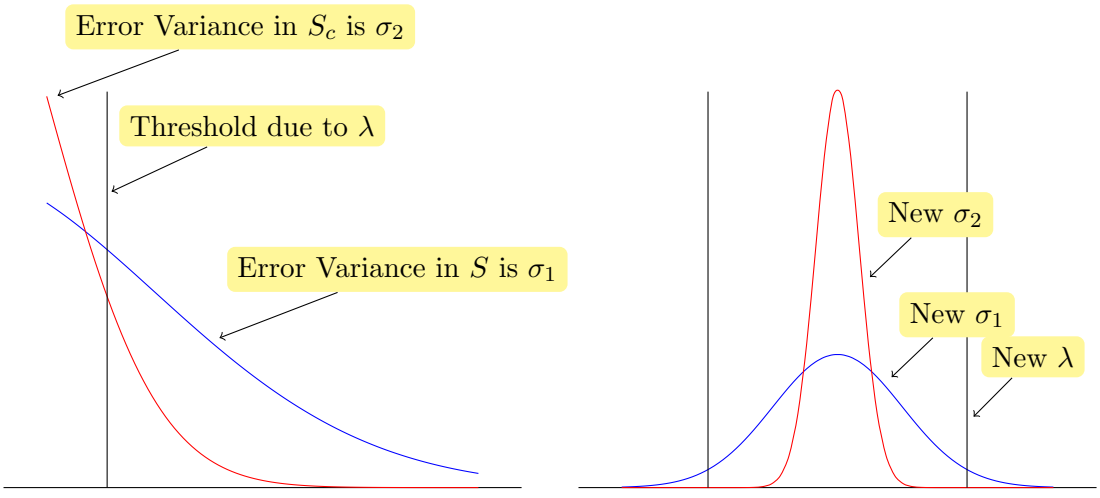

Figure 2: Left: Given the upper bound on error due to n, p, we divide the error into multiple stages by optimizing n, p, λ yielding advantages in overall data usage. Right: Modified error probability due to optimization of n, p in a single stage of the multi-stage lasso.

$$\min_{i=1,2,..s} |\beta_i| \geq M_3 n^{\frac{c_2-1}{2}} \tag{19}$$

Condition (15) requires the normalization of the covariates. (16) requires that the eigenvalues of $X_S^T X_S$ are bounded from below so that the inverse behaves well.The main conditions are (17) and (19). (19) requires a gap of n^{c_2} between β_S and noise, since noise terms aggregate at the rate of $n^{-1/2}$. (17) requires the $\sqrt{s}$ to grow at a rate slower than c_2, which prevents estimation bias from dominating the model. (18) allows exponential growth of p as a function of n.

Since we optimize around the number of sample used in different stages of the Lasso, we also extend the regularity conditions as

$$\frac{1}{kn} X^T X \leq M_1 \tag{20}$$

$$\alpha^T \frac{1}{kn} X_S^T X_S \alpha \geq M_2, \ for \ \forall ||\alpha||_2^2 = 1 \tag{21}$$

where the regularity conditions also holds for $k \in (logn, 1)$.

Theorem 1 ***Upper bound on*** λ***:*** *If the regularization parameter* $\lambda = O(n^{\frac{c_4-1}{2}})$*, where* $c_4 \in (c_3, c_2 - c_1)$*, and* λ *is upper bounded by*

$$\left(\frac{4\sigma\sqrt{M_2} + \sqrt{s}}{M_2} \right)^{-1} \min_S |\beta_S^*| \tag{22}$$

then

$$P(Sign(\hat{\beta}_S) = Sign(\beta^*)) \geq 1 - 2\exp(-cn\lambda^2)$$

for some constant $c > 0$*.*

Theorem 1 gives us an upper bound on the largest λ that can be used in lasso without shrinking the non-zero co-variates to zero. This upper bound prunes the solution set to our optimization algorithm for the two-stage and multi-stage procedure. Computation of this upper bound requires knowledge of several unknown constants, which are assumed to be known in the paper.

Theorem 2 ***Lower bound on*** λ ***based on*** η ***and*** n***:*** *If incoherence values of the zero co-variates are distributed discreetly* $\eta = \eta_1, \eta_2, ..\eta_l$*, with proportions* $\alpha = \alpha_1, \alpha_2, ...\alpha_l$*, then a choice of*

$$\lambda^2 \geq 4log(\alpha_i(p-s))\sigma^2 M_1/n\eta_i^2 \tag{23}$$

ensures that the zero co-variates with incoherences greater than η_i *are removed from the lasso solution set with probability*

$$\geq 1 - 2\exp(-c\lambda^2\eta_i^2 n)$$

for some constant $c > 0$*.*

Theorem 2 gives an lower bound on the regularization parameter that removes a set of zero co-variates with high probability. Also, if the incoherence parameters are known in advance then λ and n can be optimized to achieve a desired error. Theorem 2 also helps us simplify the co-variates being used in each stage of the multi-stage lasso procedure.

Theorem 3 ***Model Selection for Two-Stage and Multi-Stage Procedure*** *If Assumptions 1 and 2 hold, and if the incoherence values are distributed discreetly* $\eta = (\eta_1, \eta_2, ...\eta_l)$ *with* $\eta_1 > \eta_2 > \eta_3... > \eta_l$ *and with proportions* $\alpha = (\alpha_1, \alpha_2, ...\alpha_l)$*, such that* $\sum_{i=1}^{l} \alpha_i = 1$*.*

If λ_i, for stage i, is selected as

$$\lambda_i^2 \geq \max((log(l+1) - log\delta)2^{i-1}, log\alpha_i(p-s))\sigma^2 M_1/\eta_i^2 n \tag{24}$$

using $\frac{n}{2^i}$ samples and if for $i \in (1, 2, ..l)$

$$\lambda_i < \left(\frac{M_2}{4\sigma\sqrt{M_2} + \sqrt{s}}\right) \min_S |\beta_S^*|$$

then

$$P(Sign(\hat{\beta}_l) = Sign(\beta^*)) = P(Sign(\hat{\beta}) = Sign(\beta^*)) \geq 1 - \delta \tag{25}$$

where $\hat{\beta}$ represents lasso estimate using a single stage lasso and $\hat{\beta}_l$ is the lasso estimate using the multi-stage procedure.

Proof Sketch:

The proofs for the two-stage and multi-stage procedures proceed in the following stages: firstly using Theorem 1 we ensure that all λ_i's are upper bounded, so that the true co-variates do not shrink to zero in any stage. Next using Theorem 2 we identify conditions necessary to remove a specific subset of zero co-variates with incoherence η_i. Finally, we compare the failure probability of such a procedure with a single stage lasso procedure with given model selection error δ. The conditions required for removal of zero co-variates and model selection error gives us a lower bound for λ defined in Theorem 2. The detailed proof for Theorem 3 is given in Appendix.

Theorem 3 is the main result for the our proposed procedures. Given a upper bound on the error probability of a single stage procedure δ, we can determine the regularization parameter for each stage using equation (24) and (25). Theorem 3 provides a general proof for model selection of the multi-stage procedure without optimizing the number of samples used. However under specific conditions of η and α, we can

derive bounds on the sample usage. We prove two simple corollaries, for two-stage and three stage procedures, which provide conditions under which a quarter of the overall samples can be saved.

CHAPTER V

SIMULATION RESULTS

We have performed simulation studies to verify our theoretical analysis. Our comparison looks into feature selection accuracy using our predicted λ_2, and the number of samples saved due to the two-stage procedure and multi-stage procedure.

We first generate the true signal β^* as $6log(n)+randn$ which ensures condition (9) in our assumptions is satisfied. The sign of each non-zero element in β^* is decided by a binomial distribution $bin(1, 0.4)$. Following this, the covariance structure of X is setup as a random Gaussian matrix. This structure results is an exponential incoherence distribution as observed in Figures 3 and Figures 4.

To test the validity of our claims we approximated the incoherence distribution, with two and three distinct incoherence values, and tracked the model selection probability and sample usage for each setting. Theorems (1), (2) and (3) were used to estimate λ's for each stage. For the single stage lasso we used the following regularization parameter as suggested in Wainwright [4]

$$\lambda = \frac{2}{\eta}\sqrt{\frac{slogp}{n}}$$

For each figure, in the left subplot, we plot the original incoherence distribution and the approximation taken. In the right subplot we plot the model selection probability as a function of samples used.We can observe that the multi-stage lasso procedures are able to achieve higher model selection accuracy using lower number of samples than the passive algorithm.

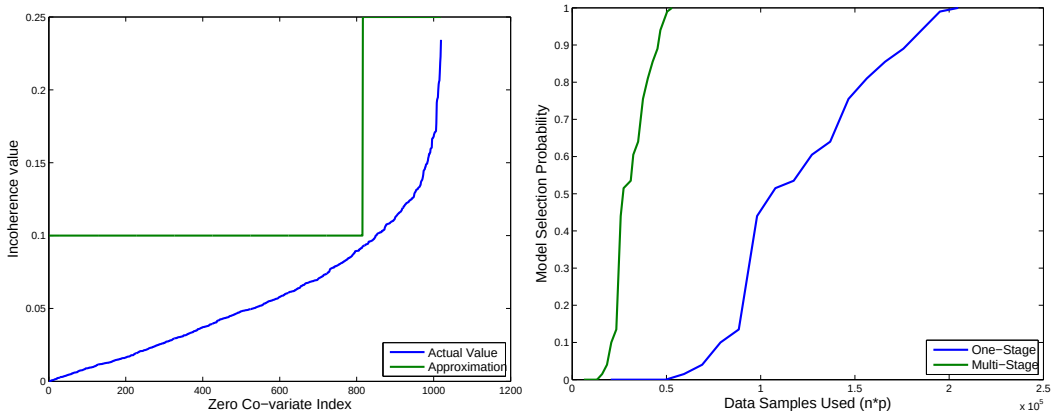

Figure 3: Two Stage approximation of an exponential incoherence distribution (Left) and Model Selection Performance (right) in the setting using single stage and multi-stage procedures for p=1024, s=5. Incoherence approximated as $\eta_1 = 0.1$, $\eta_2 = 0.25$

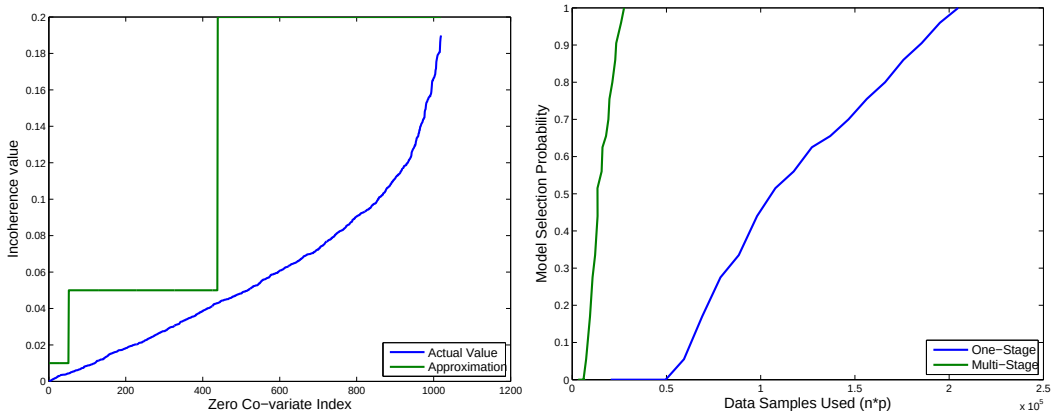

Figure 4: Three Stage approximation of an exponential incoherence distribution (Left) and Model Selection Performance (right) in the setting using single stage and multi-stage procedures for p=1024, s=5. Incoherence approximated as $\eta_1 = 0.01$, $\eta_2 = 0.05$, $\eta_3 = 0.2$

CHAPTER VI

CONCLUSION

In this paper we introduce two-stage and multi-stage lasso procedures which take advantage of variance in incoherence parameters to perform model selection with fewer number of samples than lasso methods. We prove the model selection performance of the proposed algorithms and also validate it though simulation studies.

Currently the algorithm requires knowledge of several parameters which are often unknown in advance in practical settings. This reduces the applicability of the algorithm to real datasets. In our future work, we would like to improve the algorithm to address these practical issues.

Also of interest, is extending the current proofs and theorems to a broad set of estimators. Recent work by Yen-Huan Li et al.[2] have shown generalized model selection theorems for L1-regularization. They show application of theorems towards linear regression, logistic regression, gamma regression and graphical model selection. Similar to Lasso, the model selection proofs for such procedures requires the existence of incoherence conditions. Thus under similar scenarios, multi-stage procedures for other L1-regularization methods can also be developed.

APPENDIX A

DETAILED PROOFS

A.1 Proof of Theorem 1

From the sub-gradient equation we have for $\forall i \in S$,

$$(\hat{\beta}_S - \beta_S^*) = (\frac{1}{n} X_S^T X_S)^{-1} \bigg(\frac{1}{n} X_S^T \epsilon - \lambda Sign(\beta_S^*) \bigg)$$

Using triangle inequality,

$$||\hat{\beta}_S - \beta_S^*||_\infty \leq ||(\frac{1}{n} X_S^T X_S)^{-1} \frac{1}{n} X_S^T \epsilon||_\infty + \lambda ||(\frac{1}{n} X_S^T X_S)^{-1} Sign(\beta_S^*)||_\infty$$

$$= ||(\frac{1}{n} X_S^T X_S)^{-1} \frac{1}{n} X_S^T \epsilon||_\infty + \lambda ||(\frac{1}{n} X_S^T X_S)^{-1}||_\infty$$

Now since ϵ is a zero mean sub-Gaussian vector with variance σ^2, it follows that $Z = (\frac{1}{n} X_S^T X_S)^{-1} \frac{1}{n} X_S^T \epsilon$ is zero mean sub-Gaussian with variance at most

$$\frac{\sigma^2}{n} |||(\frac{1}{n} X_S^T X_S)^{-1}|||_2 \leq \frac{\sigma^2}{n M_2}$$

Consequently by the sub-Gaussian tail bound and union bound we have,

$$P\bigg(\max_{i=1,2,..s} |Z_i| > t \bigg) \leq 2exp\bigg(- \frac{t^2 M_2 n}{2\sigma^2} + log(s) \bigg)$$

For $t = \frac{2\sigma\lambda}{\sqrt{M_2}}$, we have the exponential term simplifying as $-2n\lambda^2 + log(s)$. If we assume $2n\lambda^2 > log(s)$ which holds true base on Assumption 2, we have the exponential term simplifying as $-n\lambda^2$.

Now, we require $||\hat{\beta}_S - \beta^*_S||_\infty < \min \beta^*_S$ for sign consistency to hold. Thus we have,

$$\lambda\left(||(\frac{1}{n}X_S^T X_S)^{-1}||_\infty + \frac{4\sigma}{\sqrt{M_2}}\right) \leq \min \beta^*_S$$

Now for the first term

$$\lambda||(\frac{1}{n}X_S^T X_S)^{-1}||_\infty \leq \lambda\sqrt{s}||(\frac{1}{n}X_S^T X_S)^{-1}||_2 \leq \frac{\lambda\sqrt{s}}{M_2}$$

Thus simplifying we get,

$$\lambda\left(\frac{\sqrt{s}}{M_2} + \frac{4\sigma}{\sqrt{M_2}}\right) \leq \min \beta^*_S$$

with probability greater than $1 - 2exp(-cn\lambda^2)$.

A.2 Proof of Theorem 2

The error in the zero co-variates depend on the event

$$B_n = \left|\frac{1}{n}X_{S^c}^T X_S(\frac{1}{n}X_S^T X_S)^{-1}\frac{X_S^T\epsilon}{n} - \frac{X_{S^c}^T\epsilon}{n}\right| \leq \lambda\eta$$

Under assumptions (5)- (9), if we write the L.H.S term as $H^T\frac{\epsilon}{n}$, where $H^T = \frac{1}{n}X_{S^c}^T X_S(\frac{1}{n}X_S^T X_S)^{-1}X_S^T - X_{S^c}^T$ then

$$H^T H = X_{S^c}^T\left(I - X_S(X_S^T X_S)^{-1}X_S^T\right)X_{S^c}$$

Since $(I - X_S(X_S^T X_X)^{-1}X_S^T)$ has spectral norm one, it's eigenvalues lie between 0 and 1. Therefore the L.H.S term, will be a zero mean sub-Gaussian variable with maximum variance

$$\frac{\sigma^2}{n}||\frac{1}{n}X_{S^c}^T X_S(\frac{1}{n}X_S^T X_S)^{-1}X_S^T - X_{S^c}^T||_2^2 \leq \frac{\sigma^2 M_1}{n}$$

Thus the probability of error for event B_n can be upper bounded by

$$P(B_n^c) \leq 2(p-s)\exp\left(-\frac{1}{2}\frac{\lambda^2\eta^2 n}{\sigma^2 M_1}\right) \tag{26}$$

Where the error bound for each zero co-variate depends on its incoherence parameter. For a set of α_i zero co-variates with incoherence η_i, assuming $\lambda <$ upper bound proved in Theorem 1, if

$$\frac{\lambda^2\eta_i^2 n}{2\sigma^2 M_1} > 2log(\alpha_i(p-s))$$

then

$$P(B_{n\alpha_i}^c) \leq 2\exp(-\lambda^2\eta_i^2 n/2\sigma^2 M_1)$$

Which completes the proof.

A.3 Proof of Theorem 3

To ensure that the multi-stage or two-stage procedure achieves a similar rate of error as the single stage procedure, we can compare the model selection error probability of the two procedures and give conditions which ensure equal error selection probability. To simply the proof let us represent the error of a single stage procedure as δ.

For the multi-stage procedure with l rounds, let A_s be the event that the regularization parameter is upper bounded by the result given in Theorem 1. Also, let $B_{\alpha_i(p-s)}$ be the event that in stage i, $\alpha_i(p-s)$ zero co-variates with incoherence η_i are removed from the solution set using $\frac{n}{2^i}$ samples. Through union bound we can upper bound the overall failure probability as

$$\leq P(A_s^c) + \sum_{i=1}^{l} P(B_{\alpha_i(p-s)}^c)$$

Thus if

$$P(A_s^c) + \sum_{i=1}^{l} P(B^c_{\alpha_i(p-s)}) \leq \delta$$

the multi-stage procedure achieves similar model selection consistency results as a single stage lasso.This holds true if,

$$P(A_s^c) \leq \delta/(l+1)$$

and for all $i = (1, 2...l)$,

$$P(B^c_{\alpha_i(p-s)}) \leq \delta/(l+1)$$

Replacing expressions from Theorem 1 for $P(A_s^c)$, with $n/2^l$ samples we get the condition

$$2\exp(-n\lambda^2/2^l) \leq \delta/(l+1)$$

$$=> n\lambda^2 \geq 2^l(log(l+1) - log(\delta)) + 1$$

and for $P(B^c_{\alpha_i(p-s)})$

$$2\exp(-\lambda_i^2\eta_i^2 n/2^{i+1}\sigma^2 M_1) \leq \delta/(l+1)$$

$$=> \lambda_i^2\eta_i^2 n \geq (log(l+1) - log(\delta))2^{i+1}\sigma^2 M_1 + 1$$

Thus if these additional constraints on λ are satisfied along with conditions from Theorem 1 and Theorem 2, we will equivalent probability of model selection with a

one stage procedure.

Choosing λ_i's based on the following condition satisfies both the requirements

$$\lambda_i^2 \geq \max((log(l+1) - log\delta)2^{i-1}, log\alpha_i(p-s))\sigma^2 M_1/\eta_i^2 n$$

REFERENCES

[1] Ghaoui, L. E., Viallon, V., and Rabbani, T., "Safe feature elimination in sparse supervised learning," *Technical Report*, vol. 126, 2010.

[2] Li, Y.-H., Scarlett, J., Ravikumar, P., and Cevher, V., "Sparsistency of l1 regularized m-estimators," *Journal of Machine Learning Research*, vol. 38, 2015.

[3] Meinshausen, N. and Buhlmann, P., "High dimensional graphs and variable selection with the lasso," *Annals of Statistics*, vol. 55, 2009.

[4] Wainwright, M., "Sharp thresholds for high dimensional and noisy sparsity recovery using l1 - constrained quadratic programming," *IEEE Transactions on information theory*, vol. 55, 2006.

[5] Zhang, T., "Multi-stage convex relaxaton for feature selection," *Journal of the Machine Learning Research*, vol. 11, 2010.

[6] Zhao, P. and Yu, B., "On model selection consistency of lasso," *Journal of the Machine Learning Research*, vol. 7, 2006.

[7] Zou, H., "Adaptive lasso and its oracle properties," *Journal of American Statistical Organisation*, vol. 101, 2006.