

# When do Stepwise Algorithms Meet Subset Selection Criteria?

Xiaoming Huo <sup>a</sup> & Xuelei (Sherry) Ni  
*Georgia Institute of Technology*

## Abstract

Recent results in *homotopy* and *solution paths* demonstrate that certain well-designed greedy algorithms, with a range of values of the algorithmic parameter, can provide solution paths to a sequence of convex optimization problems. On the other hand, in regression, many existing criteria in subset selection (including  $C_p$ , AIC, BIC, MDL, RIC, etc.) involve optimizing an objective function that contains a counting measure. The two optimization problems are formulated as **(P1)** and **(P0)** in the present paper. The latter is generally combinatoric and proven to be NP-hard. We present the necessary and sufficient condition for a vector to be the optimal solution of **(P1)**. For **(P0)**, sufficient conditions are derived. We also study the conditions under which the two optimization problems have common solutions. Hence, in these situations, a greedy algorithm can be used to solve the seemingly unsolvable problem. We provide the results from three different angles: (1) a direct analysis on sufficiency and necessity, (2) results on covariates that are mostly correlated with the response, (3) results motivated by recent works in *sparse signal representation*. An extreme example with respect to the *least angle regressions* is constructed, which by itself is interesting. The applications, possible future research, and related works in statistics are discussed.

**AMS 2000 subject classification.** 62J07

**Key Words and Phrases.** subset selection, greedy algorithms, convex optimization, equivalence, concurrent optimal subset.

**Acknowledgements.** This work has been partially supported by National Science Foundation grants DMS 0140587.

## 1 Introduction

We consider two types of optimization problems:

- an optimization problem that is based on a counting measure,

$$\mathbf{(P0)} \quad \min_x \quad \|y - \Phi x\|_2^2 + \lambda_0 \cdot \|x\|_0,$$

where  $\Phi \in \mathbb{R}^{n \times m}$ ,  $x \in \mathbb{R}^m$ ,  $y \in \mathbb{R}^n$ , notation  $\|\cdot\|_2^2$  denotes the sum of squares of the entries of a vector, constant  $\lambda_0 \geq 0$  is an algorithmic parameter, and quantity  $\|x\|_0$  is the number of nonzero entries in vector  $x$ ;

- an optimization problem that depends on a sum of absolute values,

$$\mathbf{(P1)} \quad \min_x \quad \|y - \Phi x\|_2^2 + \lambda_1 \cdot \|x\|_1,$$

where  $\|x\|_1 = \sum_{i=1}^m |x_i|$  for vector  $x = (x_1, x_2, \dots, x_m)^T$ , and constant  $\lambda_1 \geq 0$  is another algorithmic parameter, whose role will be discussed later.

Note  $\|x\|_0$  (respectively,  $\|x\|_1$ ) is a quasi-norm (respectively, norm) in  $\mathbb{R}^m$ . In the literature of *sparse signal presentation*, they are called  $\ell_0$ -norm and  $\ell_1$ -norm, respectively. The numbers “0” and “1” in the notations **(P0)** and **(P1)** follow such a convention (Donoho and Huo, 2001; Donoho, Elad, and Temlyakov, 2004; Chen and Huo, 2005).

In subset selection under linear regression, many well known criteria – including  $C_p$  statistic, Akaike information criterion (AIC), Bayesian information criterion (BIC), minimum description length (MDL), risk inflation criterion (RIC), and so on – are special cases of **(P0)**, by assigning different values to  $\lambda_0$ . Details regarding the foregoing statement will be provided later. It is shown in this paper that problem **(P0)** in general is NP-hard (Theorem 2.1).

At the same time, **(P1)** is the mathematical problem that is called upon in Lasso (Tibshirani, 1996). Recent advances (whose details and references are provided in Section 2.2) demonstrate that some stepwise algorithms (e.g., least angle

regressions (LARS) presented in Efron et al. (2004)) reveal the solution paths of problem **(P1)**, while parameter  $\lambda_1$  takes a range of values. More importantly, most of these algorithms only take polynomial number of operations – i.e., they are polynomial-time algorithms. In fact, the complexity of finding a solution path for **(P1)** is the same as implementing an ordinary least square fit (Efron et al., 2004).

The main objective of this paper is to find when **(P0)** and **(P1)** give the same result in the subset selection under a regression model. A subset that corresponds to the nonzero subset of the minimizer of **(P0)** (respectively, **(P1)**) is called a *type-I* (respectively, *type-II*) *optimal subset with respect to*  $\lambda_0$  (respectively,  $\lambda_1$ ). A subset that is both type-I and type-II optimal is called a *concurrent optimal subset*. It will be shown that there is a necessary and sufficient condition for the type-II optimal subset (Theorem 4.2), and this condition can be verified in polynomial time. However, in general, there is no polynomial-time necessary and sufficient condition for the type-I optimal subset. We then search for easy-to-verify (i.e., polynomial-time) sufficient conditions for type-I optimal subsets. Two types of results are derived. The first is based on the assumption that the most correlated covariates form the optimal subset. The second result is motivated by a new advance in sparse signal representation, and is rather general.

The paper is organized as follows. Section 2 reviews the subset selection criteria that can be formulated as **(P0)**, as well as the solution paths property of Lasso and its solutions based on LARS. This material provides a starting point of the consequent work. Section 3 presents two case studies. In the first case, it is shown that a greedy algorithm (i.e., a version of LARS) can go totally wrong in an extreme situation. In the second case, it is shown that **(P0)** and **(P1)** give the same result in subset selection. These two opposing cases motivate us to analyze the conditions under which the two approaches choose the identical subset. Section 4 contains the main results. Our main results are organized in three groups. In Section 4.1, necessary and sufficient conditions are provided. For **(P0)**, such a condition is hard to verify in practice. In Section 4.2, a sufficient condition is derived. This condition started from a simple fact: the most correlated covariates

(with the response) form the concurrent optimal subset. This condition is easy to verify numerically. However, it is relatively restrictive. We use it as a preparation for more flexible sufficient conditions. In Section 4.3, a very general sufficient condition is derived. To our knowledge, this is the best known subset equivalence condition between **(P0)** and **(P1)**. Section 5 discusses related works and potential future research topics. A brief conclusion is provided in Section 6. To keep the flow of the paper, not-directly-required proofs are postponed into the appendices.

## 2 Formulation and Literature Review

We consider *subset selection* in regression. Recall in a regression setting,  $\Phi \in \mathbb{R}^{n \times m}$  ( $n > m$ ) denotes a model matrix. Vectors  $x \in \mathbb{R}^m$  and  $y \in \mathbb{R}^n$  are coefficient and response vectors. The columns of matrix  $\Phi$  are *covariates*. A regression model is  $y = \Phi x + \varepsilon$ , where  $\varepsilon$  is a random vector. Let  $\mathbf{I} = \{1, 2, \dots, m\}$  denote all the indices of the coefficients. A subset of coefficients (or, covariates) is denoted by  $\Omega \subseteq \mathbf{I}$ . Let  $|\Omega|$  denote the cardinality of the set  $\Omega$ . Let  $x_\Omega$  denote the coefficient vector that only takes nonzero values when the coefficient indices are in the subset  $\Omega$ . A subset selection problem has two competing objectives in choosing a subset  $\Omega$ : firstly, the residuals, which are in the vector  $y - \Phi x_\Omega$ , are close to zeros; secondly, the size of the set  $\Omega$  is small. Note that we differ from many statisticians, who emphasize the predictability of the selected models. We provide some discussions in Section 5.

### 2.1 Subset Selection Criteria and **(P0)**

There has been rich literature on the criteria regarding subset selection. Miller (1990) and George (2000) give an excellent overview. An interesting fact is that a majority of these criteria can be unified under **(P0)**, where  $\|y - \Phi x\|_2^2$  is the residual sum of squares (denoted by  $\text{RSS}(x)$ ) under the coefficient vector  $x$ , and constant  $\lambda_0$  depends on the criteria. The following summarizes some well-known results:

- Akaike (1973) defines his criterion by maximizing the expected log-likelihood  $E_{X,\hat{\theta}}(\log f(X|\hat{\theta}))$ , where  $\hat{\theta}$  is the estimate of parameter  $\theta$ ,  $f(X|\theta)$  is the density function. This is equivalent to maximizing the expected Kullback-Leibler's mean information for discrimination between  $f(X|\hat{\theta})$  and  $f(X|\theta)$ , i.e.,  $E_{X,\hat{\theta}}(\log \frac{f(X|\hat{\theta})}{f(X|\theta)})$ , for a known true  $\theta$ . Under a Gaussian assumption in the linear regression, the above leads to the Akaike information criterion (AIC) that minimizes

$$\text{AIC} = \frac{\text{RSS}(x)}{\sigma^2} + 2 \cdot \|x\|_0,$$

where  $\sigma^2$  is the noise variance, and other notations have been defined at the beginning of this section. It is a special case of **(P0)** by assigning  $\lambda_0 = 2\sigma^2$ .

- Mallows'  $C_p$  (Mallows, 1973; Gilmour, 1996), which is derived from the unbiased risk estimation, minimizes

$$C_p = \frac{1}{\hat{\sigma}^2} \text{RSS}(x) + 2 \cdot \|x\|_0 - n,$$

where  $\hat{\sigma}$  is an estimate of the parameter  $\sigma$ . When  $\hat{\sigma}^2 = \sigma^2$  is assumed, the  $C_p$  is equivalent with the AIC. Again  $C_p$  is a special case of **(P0)**.

- Motivated by the asymptotic behavior of Bayes estimators, Bayesian information criterion (BIC) (Schwarz, 1978) chooses to select the model that maximizes

$$\log f(X|\hat{\theta}) - \frac{1}{2} \cdot \log n \cdot \|x\|_0.$$

Again, under the squared error loss and the Gaussian model assumption with known variance  $\sigma^2$ , BIC is to minimize

$$\text{BIC} = \frac{\text{RSS}(x)}{\sigma^2} + \log n \cdot \|x\|_0.$$

The above is a special case of **(P0)** by assigning  $\lambda_0 = \sigma^2 \log n$ .

- According to Hastie, Tibshirani, and Friedman (2001, Section 7.8), the equivalence between BIC and the minimum description length (MDL) is well known. Hence MDL is a special case of **(P0)**.

- Risk inflation criterion (RIC) is suggested in George and Foster (1994) from a minimax estimation vantage point. RIC recommends the model that minimizes

$$\text{RIC} = \frac{RSS(x)}{\sigma^2} + 2 \log p \cdot \|x\|_0,$$

where  $p$  is the number of available predictors. This is derived from selecting the model with minimum risk inflation. Due to the different emphasis of the present paper, we do not include further details of RIC. However, readers can see that RIC is another special case of **(P0)**, by taking  $\lambda_0 = 2\sigma^2 \log p$ .

In this paper, the “subset selection criteria” that appears in the title encompasses all the aforementioned criteria, all adopting the formulation **(P0)**.

Solving **(P0)** generally requires exhaustive search of all the possible subsets. When  $\|x\|_0$  (i.e., the number of covariates) increases, the methods based on exhaustive search become rapidly impractical. In fact, solving **(P0)** in general is an NP-hard problem. The following theorem can be considered as an extension of a result that was originally presented in Natarajan (1995).

**Theorem 2.1** *Solving the problem **(P0)** with a fixed  $\lambda_0$  is an NP-hard problem.*

**Proof.** Let

$$f(m) = \min_{x: \|x\|_0 \leq m} \|y - \Phi x\|_2^2,$$

where all the symbols are defined in **(P0)**. It is evident that point array  $(m, f(m))$ ,  $m = 1, 2, \dots$ , forms a non-increasing curve in the positive quadrant.

We first establish the existence of an integer  $m_0$ , such that value  $f(m_0) + \lambda_0 m_0$  minimizes the objective in **(P0)**. Note that there are finite number of  $m$ 's such that  $\lambda_0 m \leq f(1) + \lambda_0 \cdot 1$ . This inequality gives an upper bound of  $m$ 's that satisfy  $f(m) + \lambda_0 m \leq f(1) + \lambda_0 \cdot 1$ . Among these finite number of  $m$ 's, there is at least one  $m_0$  that minimizes the value of function  $f(m) + \lambda_0 m$ .

Define  $\varepsilon = f(m_0)$ . In general, we can assume  $\varepsilon > 0$ , because if  $\varepsilon = 0$ , response  $y$  can be superposed by a small (more specifically, no more than  $m_0$ ) number of columns of matrix  $\Phi$ , which is a special case.

Using the idea of Lagrange multiplier, we can see that solving **(P0)** with  $\lambda_0$  is equivalent to solving the sparse approximate solution (SAS) problem in Natarajan

(1995, Section 2) with  $\varepsilon$ , which is proven in Natarajan (1995) to be NP-hard. Hence, in general, solving **(P0)** is NP-hard.  $\square$

## 2.2 Greedy Algorithms and **(P1)**

Due to the hardness of solving **(P0)**, a *relaxation* idea has been proposed. The relaxation replaces the  $\ell_0$  norm with the  $\ell_1$  norm in the objective, which leads to **(P1)**. The idea of relaxation started in *sparse signal representation* (Chen, Donoho, and Saunders, 1998). Theoretical properties are derived later in (Donoho and Huo, 2001; Donoho, Elad, and Temlyakov, 2004). A partial list of new representative results include Tropp (2004a), Tropp (2004b), Gribonval and Nielsen (2003), and Chen and Huo (2005). Being compared with this paper, the problem of sparse signal representation has a different emphasis. In sparse signal representations, researchers consider a redundant *dictionary* (Mallat, 1998; Gilbert, Muthukrishnan, and Strauss, 2003) and the conditions under which the sparsest representation can be solved via a linear programming. Their formulations of **(P0)** and **(P1)** are slightly different from ours. However, a group of results in this paper are certainly motivated by some recent results in sparse representation. More connections will be discussed when we present our findings in Section 4.3.

At the same time, **(P1)** has been proposed in statistics as a way of subset selection. The method is coined as Lasso (Tibshirani, 1996). An interesting recent development – the least angle regressions (LARS) (Efron, Hastie, Johnstone, and Tibshirani, 2004) – demonstrates that certain greedy algorithms can reveal the solutions to **(P1)** with varying values of  $\lambda_1$ , based on the idea of *homotopy* (Osborne, Presnell, and Turlach, 2000a). More recent analysis demonstrates further that greedy algorithms can literally render the entire solution path in a large class of problems, referring to Hastie, Rosset, Tibshirani, and Zhu (2004) and the references therein. A recent conference presentation (Malioutov, Cetin, and Willsky, 2005) gives the most succinct solution in generating solution paths, utilizing a homotopy continuation method (Osborne, Presnell, and Turlach, 2000b) and an analysis of *subdifferential*. Rockafellar (1970) is a standard reference for the

background of this material.

### 3 Motivations: Case Studies

#### 3.1 An Extremal Example for the Least Angle Regressions

Least Angle Regression (Efron et al., 2004) is a forward variable selection method. An extensive manual regarding forward selection can be found in Atkinson, Riani, and Cerioli (2004). As been indicated previously, LARS can give the solution path of **(P1)**. However, this homotopy does not guarantee that LARS always reveal the optimal solutions of **(P0)**. In this subsection, we present one particular case, in which LARS choose wrongly in the first iteration and end up correcting it inefficiently. As a result, LARS do not include the correct covariates until the last step. Initially, such an example motivated us to consider the conditions that will be presented later.

Details of LARS algorithm can be found in (Efron et al., 2004), Section 2. In simplicity, LARS start with zero coefficients, select the most correlated covariates with the signal  $s$ , then move along the direction that is equiangular among the selected covariates until some other covariates have as much correlation with the current residual, add these new covariates under consideration and move along the new equiangular direction. When the covariates and the response are standardized to have mean 0 and unit norm, correlation between vectors is proportional to the inner product. In this section, for clarity, we first give an example with nonstandardized vectors, and choose the covariates according to the inner products. The corresponding example with standardized covariates and signal is presented later in Section 3.1.1. Section 3.1.2 shows how to use the result in this section to come up with a dramatic example in presentation.

The first example is generated as follows. Let  $\phi_i \in \mathbb{R}^n, i = 1, 2, \dots, m$ , denote the  $i$ th column of the model matrix  $\Phi$ . Hence,  $\Phi = [\phi_1, \phi_2, \dots, \phi_m]$ . Let  $\delta_i \in \mathbb{R}^n, i = 1, 2, \dots, m$ , denote the dirac vector taking 1 at the  $i$ th position and zero elsewhere. For  $i = m - A + 1, m - A + 2, \dots, m$ , let  $\phi_i = \delta_i$ , where  $A$  is a positive integer. Consider a special signal  $s = \frac{1}{\sqrt{A}} \sum_{i=m-A+1}^m \phi_i$ . Obviously, in this case,

the optimal subset is  $\{m - A + 1, \dots, m\}$ . For the first  $m - A$  columns of  $\Phi$ , make  $\phi_j = a_j \cdot s + b_j \cdot \delta_j$ , where  $1 \leq j \leq m - A$  and  $a_j^2 + b_j^2 = 1$ . Note  $\phi_i$ 's and  $s$  are all unit-norm vectors. From now on, for simplicity, we always assume  $1 \leq j \leq m - A$  and  $m - A + 1 \leq i \leq m$ . It is easy to verify that

$$\langle s, \phi_j \rangle = a_j \quad \text{and} \quad \langle s, \phi_i \rangle = 1/\sqrt{A}.$$

In this example, we choose  $1 > a_1 > a_2 > \dots > a_{m-A} > 1/\sqrt{A} > 0$ .

Now consider the procedure of LARS. In the first step, since  $\phi_1$  has the largest inner product with  $s$ , evidently column  $\phi_1$  will be chosen. The next residual will be  $r_1 = s - c_1 \phi_1$ , where  $c_1$  is the coefficient to be determined. The following result about the consequent step in LARS will be proved in Appendix A.

**Lemma 3.1** *In the consequent step of LARS, covariate  $\phi_2$  is chosen, with  $c_1 = \frac{a_1 - a_2}{1 - a_1 a_2}$ .*

Hence, the residual of the first step becomes

$$\begin{aligned} r_1 &= s - c_1 \phi_1 \\ &= s - \frac{a_1 - a_2}{1 - a_1 a_2} (a_1 s + b_1 \delta_1) \\ &= \frac{b_1^2}{1 - a_1 a_2} s - \frac{(a_1 - a_2) b_1}{1 - a_1 a_2} \delta_1 \\ &= \frac{b_1^2}{1 - a_1 a_2} \left[ s - \frac{a_1 - a_2}{b_1} \delta_1 \right]. \end{aligned}$$

Note that in LARS, only the direction of a residual vector determines the selection of the next covariates. The amplitude of a residual vector does not change the variable selection. Hence, we introduce a surrogate residual with a simpler form:

$$\tilde{r}_1 = s - \frac{a_1 - a_2}{b_1} \delta_1.$$

Residuals  $\tilde{r}_1$  and  $r_1$  have the same direction. This is an important step to simplify our analysis. In the proof of the next theorem, the surrogate residuals with simpler forms are repeatedly called upon.

As a sanity check, the following calculations are performed:

1. For  $i$ ,  $\langle \phi_i, \tilde{r}_1 \rangle = 1/\sqrt{A}$ .
2. For  $j$ ,

$$\begin{aligned} \langle \phi_j, \tilde{r}_1 \rangle &= \langle a_j s + b_j \delta_j, s - \frac{a_1 - a_2}{b_1} \delta_1 \rangle \\ &= a_j - \frac{b_j(a_1 - a_2)}{b_1} \langle \delta_j, \delta_1 \rangle. \end{aligned}$$

As special cases:  $\langle \phi_1, \tilde{r}_1 \rangle = a_2$ ,  $\langle \phi_2, \tilde{r}_1 \rangle = a_2$ , and for  $j \geq 3$ ,  $\langle \phi_j, \tilde{r}_1 \rangle = a_j$ .

The above analysis demonstrates some basic techniques that will be used in the consequent LARS steps. Now we use induction to show the following.

**Theorem 3.2 (Case Study of LARS)** *In the example described in the beginning of this section, LARS choose covariates  $\phi_1, \phi_2, \dots, \phi_{m-A}$  one by one sequentially in the first  $m - A$  steps.*

It takes some energy to verify the above theorem. We postpone it to Appendix B. This example shows that LARS can choose all the covariates outside an intuitively optimal subset before it reaches any covariate inside the optimal subset.

### 3.1.1 Standardized Covariates

Readers may notice that LARS should proceed along the direction that depends on the correlations between  $\phi_i$ 's and the residual. Meanwhile, in our case study, the proceeding direction is determined due to the inner product. The inner product is not proportional to the correlation since the response  $s$  and the covariate vectors  $\phi_i$ 's are not standardized to have mean 0. However, this discrepancy can be easily remedied as follows. The key observation is that LARS only depend on geometric information. More specifically, the result depends only on  $\langle \phi_i, s \rangle$ ,  $i = 1, 2, \dots, m$ , and  $\langle \phi_i, \phi_j \rangle$ ,  $1 \leq i, j \leq m$ . For example, an orthogonal transform of both  $s$  and  $\phi_i$ 's will retain the results in LARS. We state this without a proof.

**Lemma 3.3** *After a simultaneously orthogonal transform on both response and covariates, the results of LARS from the transformed data is the same orthogonal transform of the LARS results from the original data.*

Hence, if we can find another set of standardized vectors, which retain the inner products and are the orthogonal transforms of  $\phi_i$ 's and  $s$  in the previous example, the same results can be predicted for LARS.

The standardization can be incorporated according to the following. The main idea is that an  $n$ -dimensional linear space can be treated as a subspace of  $\mathbb{R}^{n+1}$ , which is orthogonal to vector  $(1, 1, \dots, 1)$ . Let  $\{b_0, b_1, \dots, b_n\}$  denote an orthonormal basis of  $\mathbb{R}^{n+1}$ , with  $b_0 = \frac{1}{\sqrt{n+1}}(1, 1, \dots, 1)^T$ . Denote the unit-norm vectors  $s = (s_1, s_2, \dots, s_n)^T$  and  $\phi_i = (\phi_{i1}, \phi_{i2}, \dots, \phi_{in})^T$ ,  $i = 1, 2, \dots, m$ . Define  $s' = \sum_{j=1}^n s_j b_j$ ,  $\phi'_i = \sum_{j=1}^n \phi_{ij} b_j$ ,  $i = 1, 2, \dots, m$ . One can easily verify that  $\langle s', \phi'_i \rangle = \langle s, \phi_i \rangle$  for  $1 \leq i \leq m$ , and  $\langle \phi'_i, \phi'_j \rangle = \langle \phi_i, \phi_j \rangle$  for  $1 \leq i, j \leq m$ . Hence, applying LARS to  $s'$  and  $\phi'_i$ 's will produce the same result as in the first case study. It is not hard to verify that  $s'$  and  $\phi'_i$ 's are standardized. Hence, the conclusions in our case study can be extended to the case with standardized response and covariates.

**Theorem 3.4 (An Example with Standardized Covariates)** *There exists an orthogonal transform that can be applied to the previous example to create a case in which all the covariates and the response are standardized, and LARS select all the covariates outside the optimal subset before it chooses any covariate inside the optimal subset.*

### 3.1.2 A Dramatic Presentation

The foregoing example is developed in a fairly general form, with controlling parameters  $A$  and  $m$ . To illustrate how dramatic this example can be, let us consider the case where  $A = 10$  and  $m = 1,000,000$ . Based on the previous description, the LARS will select the first 999,990 covariates before it selects any of the last ten covariates. At the same time, the optimal subset is formed by the last ten covariates.

## 3.2 Variable Selection with Orthogonal Model Matrix

In order to provide some insights, a simple case in which  $\Phi$  is orthogonal is considered. Although this example has been studied in the original LARS paper (Efron

et al., 2004), the purpose of restating it here is to illustrate that there is a case in which LARS find the type-I optimal subset.

**Theorem 3.5 (Orthogonal Design)** *Let  $\tilde{x}_0$  and  $\tilde{x}_1$  denote the solutions to (P0) and (P1), respectively. When  $\Phi$  is orthogonal, we have*

$$\tilde{x}_{0,i} = \begin{cases} 0, & \text{if } |z_i| \leq \sqrt{\lambda_0}, \\ z_i, & \text{if } |z_i| > \sqrt{\lambda_0}, \end{cases}$$

and

$$\tilde{x}_{1,i} = \begin{cases} 0, & \text{if } |z_i| \leq \lambda_1/2, \\ \text{sign}(z_i)(|z_i| - \frac{\lambda_1}{2}), & \text{if } |z_i| > \lambda_1/2. \end{cases}$$

Here,  $\tilde{x}_{0,i}$  and  $\tilde{x}_{1,i}$  denote the  $i$ th entry of  $\tilde{x}_0$  and  $\tilde{x}_1$ , respectively, and  $z_i$  is the  $i$ th entry of  $z = \Phi^T y$ .

For readers who are familiar with soft-thresholding and hard-thresholding (Donoho and Johnstone, 1995), the above is not a surprise. A proof follows.

**Proof.** Both (P0) and (P1) can be decomposed into the univariate problems

$$\min_{x_i} (z_i - x_i)^2 + \lambda_0 \cdot 1(x_i \neq 0),$$

and

$$\min_{x_i} (z_i - x_i)^2 + \lambda_1 \cdot |x_i|.$$

From here, it is not hard to derive the formulae in the theorem.  $\square$

From the above, verifying the following becomes an easy task. Let  $\text{supp}(x)$  denote the set of indices of the nonzero entries in vector  $x$ .

**Corollary 3.6** *When  $\sqrt{\lambda_0} = \lambda_1/2$ , one has  $\text{supp}(\tilde{x}_0) = \text{supp}(\tilde{x}_1)$ , i.e., there is a concurrent optimal subset. Moreover,*

$$\tilde{x}_{0,i} - \tilde{x}_{1,i} = \begin{cases} 0, & \text{if } i \notin \text{supp}(\tilde{x}_0), \\ \frac{\lambda_1}{2} \cdot \text{sign}(z_i), & \text{if } i \in \text{supp}(\tilde{x}_0). \end{cases}$$

The proof is obvious and is omitted.

Now there are two opposing examples. On one hand, if  $\Phi$  is orthogonal, LARS and Lasso recover the optimal subset in **(P0)**. On the other hand, we found an example in which a version of LARS would choose all the covariates outside the optimal subset before choosing anything inside. These inconsistencies encourage us to analyze the solutions of **(P0)** and **(P1)**, and the conditions for a subset to be the concurrent optimal subset. We present the details and the results in the next section.

## 4 Main Results: Conditions of Equivalence

We present our findings in three subsections. In Section 4.1, we give a sufficient and necessary condition for a subset to be the concurrent optimal subset. Recall that **(P0)** in general is NP-hard. Checking the aforementioned condition can not be done via a polynomial-time algorithm. In Section 4.2, we ask when the  $k$  most correlated covariates form the concurrent optimal subset. A sufficient condition is derived. This result is easy to check but too restrictive. However, it inspires us to consider more general sufficient conditions. A more general sufficient condition for **(P0)** is derived in the next section – Section 4.3 – which is also motivated by a recent approach appeared in applied mathematics (Gribonval, Figueras i Ventura, and Vandergheynst, 2005). We modified their approach to solve a different mathematical problem.

### 4.1 Sufficient and Necessary Conditions

Before moving into the specific discussion, we introduce a sufficient and necessary condition for a concurrent optimal subset. Let  $I_1$  denote a subset of indices. Let  $\Phi_1$  and  $x_1$  denote columns of  $\Phi$  and entries of  $x$  with indices from  $I_1$ . Let  $\Phi = [\Phi_1 \ \Phi_2]$ . Here, a permutation that does not change the problem is implied.

**Theorem 4.1 (Sufficient and Necessary for **(P0)**)**  $I_1$  is the optimal subset of **(P0)** if and only if value

$$y^T y - y^T \Phi_1 (\Phi_1^T \Phi_1)^{-1} \Phi_1^T y + \lambda_0 \cdot \|x_1\|_0 \quad (4.1)$$

is the minimum of the objective in **(P0)**.

**Theorem 4.2 (Sufficient and Necessary (P1))**  $I_1$  is the optimal subset of **(P1)** if and only if there exists a vector  $\omega$ , such that

$$\Phi^T y = \begin{pmatrix} \Phi_1^T \Phi_1 \\ \Phi_2^T \Phi_1 \end{pmatrix} x_1 + \begin{pmatrix} \frac{\lambda_1}{2} \cdot \text{sign}(x_1) \\ \omega \end{pmatrix} \quad (4.2)$$

holds and  $\|\omega\|_\infty \leq \lambda_1/2$ .

**Theorem 4.3 (Sufficient and Necessary (Concurrent))**  $I_1$  is the concurrent optimal subset of **(P0)** and **(P1)** if and only if (4.1) and (4.2) are true. Moreover, recall  $\tilde{x}_0$  and  $\tilde{x}_1$  are the solutions of **(P0)** and **(P1)**, respectively. We have

$$(\tilde{x}_0 - \tilde{x}_1)_{I_1} = (\Phi_1^T \Phi_1)^{-1} \cdot \frac{\lambda_1}{2} \cdot \text{sign}((\tilde{x}_1)_{I_1}). \quad (4.3)$$

**Proof.** For the above theorems, Theorem 4.1 is from a direct derivation; and Theorem 4.2 is based on the argument of *subdifferential* (Tropp, 2004b; Malioutov, Cetin, and Willsky, 2005).

For Theorem 4.3, consider

$$(\tilde{x}_0)_{I_1} = (\Phi_1^T \Phi_1)^{-1} \Phi_1^T y,$$

and

$$\Phi_1^T y = (\Phi_1^T \Phi_1)(\tilde{x}_1)_{I_1} + \frac{\lambda_1}{2} \cdot \text{sign}((\tilde{x}_1)_{I_1}).$$

By combining the above two, (4.3) follows.  $\square$

The above theorem gives a necessary and sufficient condition for a concurrent optimal subset. The following provides some further comments.

**Remark 4.4** Equation (4.3) provides a methods of computing  $\tilde{x}_1$ , given that  $\tilde{x}_0$  is available and represents the optimal solution. Evidently,

$$(\tilde{x}_1)_{I_1} = (\tilde{x}_0)_{I_1} - \frac{\lambda_1}{2} (\Phi_1^T \Phi_1)^{-1} \cdot \text{sign}((\tilde{x}_1)_{I_1}).$$

**Remark 4.5** *Note*

$$\begin{aligned}\Phi(\tilde{x}_0 - \tilde{x}_1) &= \Phi_1(\tilde{x}_0 - \tilde{x}_1)_{I_1} \\ &= \frac{\lambda_1}{2} \cdot \Phi_1(\Phi_1^T \Phi_1)^{-1} \cdot \text{sign}((\tilde{x}_1)_{I_1}),\end{aligned}$$

which is an equiangular vector among the columns of  $\Phi_1$ . Hence, when optimality is achieved in both (4.2) and (4.3), the difference between the two predicted vectors is an equiangular vector.

Readers can compare the above results with those in Malioutov, Cetin, and Willsky (2005), who independently achieved the same results.

Verification of the sufficient and necessary conditions in (4.1) is difficult, requiring solving a combinatorial search problem. Because in general, solving **(P0)** is NP-hard (Theorem 2.1), it will be easy to verify that there should be no sufficient and necessary condition that can be verified by a polynomial time algorithm.

## 4.2 A Sufficient Condition for Covariates that are Mostly Correlated with the Response

Because it is generally impossible to have a necessary and sufficient condition that can be verified in polynomial time, we will focus on finding some easy-to-verify sufficient conditions.

We first introduce a set of sufficient conditions, which only depend on the correlations between the response  $y$  and the covariates  $\phi_i$ , as well as the maximum correlation between the covariates. For simplicity, we now assume that response  $y$  and covariates  $\phi_i$ 's are all standardized. It is not hard to see  $|\langle y, \phi_i \rangle| \leq 1$ ,  $i = 1, 2, \dots, m$ , and  $|\langle \phi_i, \phi_j \rangle| \leq 1$ ,  $1 \leq i, j \leq m$ . Denote  $z = \Phi^T y = (z_1, z_2, \dots, z_m)^T$ . Without loss of generality, we assume  $|z_1| > |z_2| > \dots > |z_m|$ . We want to find sufficient conditions such that subset  $A_1 = \{\phi_1, \phi_2, \dots, \phi_k\}$  is the solution to both **(P0)** and **(P1)**. In other words, the  $k$  most correlated covariates with the response form the optimal subset. Clearly, an optimal subset does not need to be the most correlated covariates with the response. Due to this additional condition, this set of conditions are *restrictive*. The restrictiveness is illustrated in an example in Section 4.2.1.

Denote

$$\mu = \max_{\substack{1 \leq i, j \leq m \\ i \neq j}} |\langle \phi_i, \phi_j \rangle|.$$

The following is a well-known result from linear algebra.

**Lemma 4.6** *Let  $\lambda(\Phi_1^T \Phi_1)$  denote an eigenvalue of matrix  $\Phi_1^T \Phi_1$ , where  $\Phi_1 = [\phi_1, \phi_2, \dots, \phi_k]$ . We have*

$$1 - (k-1)\mu \leq \lambda(\Phi_1^T \Phi_1) \leq 1 + (k-1)\mu, \quad (4.4)$$

and

$$\frac{1}{1 + (k-1)\mu} \leq \lambda((\Phi_1^T \Phi_1)^{-1}) \leq \frac{1}{1 - (k-1)\mu} \quad (4.5)$$

The above lemma will be used in proving the following two theorems.

**Theorem 4.7** *For a given  $\lambda_0$ , and correlations  $z_1, z_2, \dots, z_k$ , if the following three conditions are satisfied:*

$$[1 - (k-1)\mu]z_k^2 \geq 2(k-1)^2\mu + z_{k+1}^2[1 + (k-1)\mu], \quad (4.6)$$

$$z_{k+1}^2 \leq \lambda_0(1 - \Delta) - \frac{(2k-1)\mu}{1 + (k-1)\mu} \sum_{i=1}^k z_i^2, \quad (4.7)$$

$$z_k^2 \geq \lambda_0 + \frac{(2k-3)\mu}{1 + (k-1)\mu} \sum_{i=1}^k z_i^2, \quad (4.8)$$

where  $\Delta = n \cdot \mu$  in (4.7), then subset  $A_1$  is the type-I optimal subset.

To prove the above theorem, we will show that for subsets having sizes equal to  $k$ , or sizes greater than  $k$ , or sizes less than  $k$ , the above three conditions will guarantee that subset  $A_1$  is the type-I optimal subset.

**Proof.** For any subset of indices  $A$ , let  $z_A$  denote the subvector of  $z$  with indices in  $A$ , and let  $\Phi_A$  denote the submatrix of  $\Phi$  with the column indices in  $A$ . A simple derivation will show that the objective in **(P0)** can be rewritten as  $y^T y - z_A^T (\Phi_A^T \Phi_A)^{-1} z_A + \lambda_0 |A|$ . We only need to show that given (4.6), (4.7), and (4.8), function  $z_A^T (\Phi_A^T \Phi_A)^{-1} z_A - \lambda_0 |A|$  is maximized at  $A_1 = \{1, 2, \dots, k\}$ . In order to prove this, three situations are considered.

*Case 1.* If  $|A| = k$ , but  $A$  is not  $\{1, 2, \dots, k\}$ . Recall  $\Phi_1 = [\phi_1, \phi_2, \dots, \phi_k]$ . Let  $v_1 = (z_1, z_2, \dots, z_k)^T$ . From Lemma 4.6, we have

$$v_1^T (\Phi_1^T \Phi_1)^{-1} v_1 \geq \frac{\sum_{i=1}^k z_i^2}{1 + (k-1)\mu}.$$

On the other hand,

$$z_A^T (\Phi_A^T \Phi_A)^{-1} z_A \leq \frac{\sum_{i=1}^{k-1} z_i^2 + z_{k+1}^2}{1 - (k-1)\mu}.$$

If (4.6) is true, recall  $z_i \leq 1$ , we have

$$[1 - (k-1)\mu] z_k^2 \geq 2(k-1)\mu \sum_{i=1}^{k-1} z_i^2 + z_{k+1}^2 [1 + (k-1)\mu].$$

The above is equivalent to

$$[1 - (k-1)\mu] \sum_{i=1}^k z_i^2 \geq [1 + (k-1)\mu] \sum_{i=1}^{k-1} z_i^2 + z_{k+1}^2 [1 + (k-1)\mu],$$

which is equivalent to

$$\frac{\sum_{i=1}^k z_i^2}{1 + (k-1)\mu} \geq \frac{\sum_{i=1}^{k-1} z_i^2 + z_{k+1}^2}{1 - (k-1)\mu}.$$

Hence, we have proved that

$$z_A^T (\Phi_A^T \Phi_A)^{-1} z_A \leq v_1^T (\Phi_1^T \Phi_1)^{-1} v_1.$$

This proves that  $A_1$  minimizes the objective in **(P0)** among all  $k$ -subsets.

*Case 2.* If  $|A| > k$ , assume  $\ell = |A| - k$ . Using a similar argument as in the previous case, one only needs to prove

$$\frac{\sum_{i=1}^k z_i^2}{1 + (k-1)\mu} \geq \frac{\sum_{i=1}^{k+\ell} z_i^2}{1 - (k+\ell-1)\mu} - \ell \cdot \lambda_0,$$

which will guarantee that no subset with more than  $k$  covariates produces a larger value of

$$z_A^T (\Phi_A^T \Phi_A)^{-1} z_A - \lambda_0 \cdot |A|.$$

If (4.7) holds, we have

$$\begin{aligned} z_{k+1}^2 &\leq \sum_{i=1}^k z_i^2 \left[ \frac{1 - k\mu}{1 + (k-1)\mu} - 1 \right] = \lambda_0 \cdot (1 - \Delta) \\ &\leq \frac{1}{\ell} \sum_{i=1}^k z_i^2 \left[ \frac{1 - (k + \ell - 1)\mu}{1 + (k-1)\mu} - 1 \right] + \lambda_0 \cdot [1 - (k + \ell - 1)\mu]. \end{aligned}$$

Hence,

$$\sum_{i=k+1}^{k+\ell} z_i^2 \leq \sum_{i=1}^k z_i^2 \left[ \frac{1 - (k + \ell - 1)\mu}{1 + (k-1)\mu} - 1 \right] + \ell \cdot \lambda_0 \cdot [1 - (k + \ell - 1)\mu].$$

Hence,

$$\frac{\sum_{i=k+1}^{k+\ell} z_i^2}{1 - (k + \ell - 1)\mu} \leq \sum_{i=1}^k z_i^2 \left[ \frac{1}{1 + (k-1)\mu} - \frac{1}{1 - (k + \ell - 1)\mu} \right] + \ell \cdot \lambda_0.$$

Hence,

$$\frac{\sum_{i=1}^k z_i^2}{1 + (k-1)\mu} \geq \frac{\sum_{i=1}^{k+\ell} z_i^2}{1 - (k + \ell - 1)\mu} - \ell \cdot \lambda_0.$$

Again, subset  $A_1$  minimizes the objective in **(P0)** in this situation.

*Case 3.* If  $A$  has less than  $k$  covariates, defining  $\ell = k - |A|$ , one only needs

$$\frac{\sum_{i=1}^k z_i^2}{1 + (k-1)\mu} \geq \frac{\sum_{i=1}^{k-1} z_i^2}{1 - (k - \ell - 1)\mu} + \ell \cdot \lambda_0. \quad (4.9)$$

If (4.8) is true, we have

$$\ell \cdot z_k^2 [1 + (k-1)\mu] \geq \mu(2k - \ell - 2) \sum_{i=1}^k z_i^2 + \ell \cdot \lambda_0 [1 + (k-1)\mu] [1 - (k - \ell - 1)\mu].$$

Hence,

$$\ell \cdot z_k^2 \frac{1 + (k-1)\mu}{1 - (k - \ell - 1)\mu} \geq \frac{(2k - \ell - 2)\mu}{1 - (k - \ell - 1)\mu} \sum_{i=1}^k z_i^2 + \ell \cdot \lambda_0 \cdot [1 + (k-1)\mu].$$

Hence,

$$\sum_{i=k-\ell+1}^k z_i^2 \frac{1 + (k-1)\mu}{1 - (k - \ell - 1)\mu} \geq \sum_{i=1}^k z_i^2 \left[ \frac{1 + (k-1)\mu}{1 - (k - \ell - 1)\mu} - 1 \right] + \ell \cdot \lambda_0 [1 + (k-1)\mu].$$

Hence,

$$\sum_{i=1}^k z_i^2 \geq \frac{1 + (k-1)\mu}{1 - (k-\ell-1)\mu} \sum_{i=1}^{k-\ell} z_i^2 + \ell \cdot \lambda_0 [1 + (k-1)\mu],$$

which leads to (4.9).

Combining the above three cases, we have proved that  $A_1$  is the type-I optimal subset.  $\square$

**Remark 4.8** *Conditions (4.6), (4.7) and (4.8) are independent, i.e., none of them can be derived from the other two.*

The following theorem states the condition for set  $A_1 = \{\phi_1, \phi_2, \dots, \phi_k\}$  to be the type-II optimal subset.

**Theorem 4.9** *Given  $\lambda$  and  $k$ , if*

$$\frac{\lambda}{2} - |z_{k+1}| \geq \frac{\sqrt{k}\mu}{1 - (k-1)\mu} \sqrt{\sum_{i=1}^k \left(|z_i| + \frac{\lambda}{2}\right)^2}, \quad (4.10)$$

*then subset  $A_1$  is the type-II optimal subset.*

**Proof.** Recall  $v_1 = (z_1, z_2, \dots, z_k)^T$ . Define  $v_2 = (z_{k+1}, z_{k+2}, \dots, z_n)^T$ . From (4.2), we have

$$\omega = v_2 - (\Phi_2^T \Phi_1)(\Phi_1^T \Phi_1)^{-1} \left[ v_1 - \frac{\lambda_1}{2} \text{sign}(x_1) \right].$$

We want to show that when (4.10) holds,  $\|\omega\|_\infty \leq \frac{\lambda_1}{2}$ . This will imply that  $A_1$  satisfies (4.2). Hence,  $A_1$  is the type-II optimal subset.

One can show that for  $k < j \leq n$ ,

$$\begin{aligned} \|(\phi_j^T \Phi_1)(\Phi_1^T \Phi_1)^{-1}\|_2 &\leq \frac{1}{1 - (k-1)\mu} \|\phi_j^T \Phi_1\|_2 \\ &\leq \frac{\sqrt{k}\mu}{1 - (k-1)\mu}. \end{aligned}$$

Hence,

$$\begin{aligned}
\|\omega\|_\infty &\leq |z_{k+1}| + \max_j \|(\phi_j^T \Phi_1)(\Phi_1^T \Phi_1)^{-1}\|_2 \left\| \left( v_1 - \text{sign}(x_1) \frac{\lambda_1}{2} \right) \right\|_2 \\
&\leq |z_{k+1}| + \frac{\sqrt{k}\mu}{1 - (k-1)\mu} \sqrt{\sum_{i=1}^k \left( |z_i| + \frac{\lambda_1}{2} \right)^2} \\
&\leq \frac{\lambda_1}{2}.
\end{aligned}$$

A solution based on  $A_1$  satisfies (4.2). Hence, it minimizes **(P1)**.  $\square$

The following corollary gives a sufficient condition for  $A_1$  to be the concurrent optimal subset.

**Corollary 4.10** *Given (4.6), (4.7), (4.8), and (4.10), subset  $A_1$  is the concurrent optimal subset.*

#### 4.2.1 Restrictiveness of the Aforementioned Sufficient Conditions

Readers may notice that the four conditions in the previous section are restrictive. One can easily find an example that does not satisfy these conditions, however still has the concurrent optimal subset  $A_1$ .

A counter example can be established as follows. Suppose  $n, m$ , and  $k$  are three positive integers satisfying  $n > m > k$  and  $n \geq m + k$ . Let  $a_i$  denote the  $i$ th entry of vector  $\mathbf{a} \in \mathbb{R}^k$  with  $|a_1| \geq |a_2| \geq \dots \geq |a_k|$ . Let  $I_{m \times m} \in \mathbb{R}^{m \times m}$  be an identity matrix and  $\Phi_a \in \mathbb{R}^{k \times k}$  be the diagonal matrix with the  $i$ th diagonal entry being equal to  $a_i$ . Consider

$$\Phi = \text{standardized} \left\{ \left( \begin{array}{c} \Phi_a \mathbf{0}_{k \times (m-k)} \\ I_{m \times m} \\ \mathbf{0}_{(n-k-m) \times m} \end{array} \right) \right\}, \quad y = \sum_{i=1}^k \phi_i,$$

where  $\text{standardized}\{M\}$  refers to the standardization of all the columns of matrix  $M$ , matrices  $\mathbf{0}_{k \times (m-k)}$  and  $\mathbf{0}_{(n-k-m) \times m}$  are made by zeros, and  $\phi_i$  is the  $i$ th column of  $\Phi$ . The optimal solution is the first  $k$  covariates, and these covariates have larger correlations with  $y$ . However, there are many choices of  $m, n, k$  and vector  $\mathbf{a}$ , with

which condition (4.6) is not satisfied. As a special case, consider the following simple example:  $n = 10, m = 7, k = 3$ , and  $\mathbf{a} = (-1 \ 1 \ 0)^T$ . It's not hard to verify that  $\mu(\Phi) = 0.1667, z_3 = 0.7379, z_4 = -0.3162, [1 - (k - 1)\mu]z_k^2 = 0.3630$ , and  $2(k - 1)^2\mu + z_{k+1}^2[1 + (k - 1)\mu] = 0.9117$ . Hence, (4.6) does not hold for this case.

### 4.3 Sufficient Conditions based on the Model Matrix and the Correlations with Residuals

It is evident that the conditions in the previous subsection is restrictive. However, the derivation of the results (e.g., Theorem 4.7) demonstrates some key quantities that are required in the analysis: e.g., the correlations among the covariates, the correlations between the covariates and the response.

In order to come up with a practical subset selection scheme, it is helpful to have a sufficient condition for the type-I optimal subset. For example, when a solution path of **(P1)** is computed by an efficient stepwise algorithm, this sufficient condition can be used to test whether any of the solutions on this solution path is also type-I optimal. If yes, then a concurrent optimal subset is obtained.

We develop some sufficient conditions to identify whether a subset is a type-I optimal subset. Recall that  $x \in \mathbb{R}^m$  denote a coefficient vector. Denote the corresponding residual vector by  $\varepsilon = y - \Phi x$ . Recall that  $y \in \mathbb{R}^n$  and  $\Phi \in \mathbb{R}^{n \times m}$  are the response vector and the model matrix, respectively. Let  $\Omega$  denote the support of the vector  $x$ :  $\Omega = \text{supp}(x)$ . For an integer  $k \geq 1$ , let

$$\sigma_{\min, k}^2 = \inf \frac{\|\Phi\delta\|_2^2}{\|\delta\|_2^2}, \quad \text{subject to } \|\delta\|_0 \leq k.$$

The above quantity reflects certain property of the model matrix. Furthermore, for a vector  $v \in \mathbb{R}^n$  and an integer  $k \geq 1$ , we define

$$c(v, k) = \sqrt{\sum_{i=1}^k v_{(i)}^2},$$

where  $|v_{(1)}| \geq |v_{(2)}| \geq \dots \geq |v_{(n)}|$  are the non-increasing-ordered magnitudes of the entries of vector  $v$ . For finite  $k$ , we assume that quantities  $c^2(\Phi^T \varepsilon, k)$  and  $\sigma_{\min, k}^2$  are available.

The following theorem provides a sufficient condition for a subset being included in a type-I optimal subset with respect to  $\lambda_0$ .

**Theorem 4.11** *Given a subset of coefficient  $\Omega$ . Suppose that coefficient vector  $x$  is the minimizer of function  $\|y - \Phi x\|_2^2$  subject to  $\text{supp}(x) \subset \Omega$ . Let  $\varepsilon = y - \Phi x$ .*

(1) *If  $\min_{i \in \Omega} |x_i| > q_1(|\Omega|)$ , then with respect to  $\lambda_0$ , there is no type-I optimal subset whose size of the support is less than  $|\Omega|$ .*

(2) *Furthermore, if  $\min_{i \in \Omega} |x_i| > q(|\Omega|)$ , then with respect to  $\lambda_0$ , we have  $\Omega \subset \Omega'$ , where  $\Omega'$  is the type-I optimal subset with respect to  $\lambda_0$ .*

The quantities  $q_1(\cdot)$  and  $q(\cdot)$  are defined as follows. For an integer  $k \geq 1$ ,

$$q_1(k) = \sup_{m < k} \frac{c(\Phi^T \varepsilon, 1) + \sqrt{c^2(\Phi^T \varepsilon, k+m) + (k-m)\lambda_0\sigma_{\min, k+m}^2}}{\sigma_{\min, k+m}^2},$$

$$q_2(k) = \sup_{m \geq k} \frac{c(\Phi^T \varepsilon, 1) + \sqrt{c^2(\Phi^T \varepsilon, k+m) + (k-m)\lambda_0\sigma_{\min, k+m}^2}}{\sigma_{\min, k+m}^2},$$

and

$$q(k) = \max\{q_1(k), q_2(k)\}.$$

Note that quantities  $q_1(\cdot)$  and  $q_2(\cdot)$  have the same objective function. However, the ranges of variable  $m$  are different. Because  $q_1(k)$  only requires a finite choice of variable  $m$ , it is computable. It is not straightforward that for any  $k \geq 1$ , the quantity  $q_2(k)$  exists. In this paper, we assume the existence of this quantity.

**Proof.** Suppose  $\Omega'$  is the type-I optimal subset, with corresponding coefficient vector  $x'$ . We must have

$$\|y - \Phi x'\|_2^2 + \lambda_0 \|x'\|_0 \leq \|y - \Phi x\|_2^2 + \lambda_0 \|x\|_0. \quad (4.11)$$

Denote  $\delta = x' - x$ , we have  $\|\delta\|_0 \leq |\Omega| + |\Omega'|$ . We will prove that

$$\text{“if } |\Omega'| < |\Omega|, \text{ then } \|\delta\|_\infty \leq q_1(\Omega), \text{”} \quad (4.12)$$

and

$$\text{“for any } \Omega', \|\delta\|_\infty \leq q(\Omega). \text{”} \quad (4.13)$$

To see the above, a reformulation of (4.11) gives

$$\|\varepsilon - \Phi\delta\|_2^2 \leq \|\varepsilon\|_2^2 + \lambda_0(|\Omega| - |\Omega'|),$$

which is equivalent to

$$\|\Phi\delta\|_2^2 \leq 2\langle \Phi^T \varepsilon, \delta \rangle + \lambda_0(|\Omega| - |\Omega'|), \quad (4.14)$$

where  $\langle \cdot, \cdot \rangle$  denotes the inner product between two sequences. Define  $\delta' = \sigma_{\min, |\Omega|+|\Omega'|}^2 \delta$ . Because  $\|\Phi\delta\|_2^2 \geq \sigma_{\min, |\Omega|+|\Omega'|}^2 \|\delta\|_2^2$ , and (4.14), we have

$$\|\delta'\|_2^2 \leq 2\langle \Phi^T \varepsilon, \delta' \rangle + \lambda_0(|\Omega| - |\Omega'|) \cdot \sigma_{\min, |\Omega|+|\Omega'|}^2.$$

The above is equivalent to

$$\|\Phi^T \varepsilon - \delta'\|_2^2 \leq \|\Phi^T \varepsilon\|_2^2 + \lambda_0(|\Omega| - |\Omega'|) \cdot \sigma_{\min, |\Omega|+|\Omega'|}^2.$$

Define  $\varepsilon^* = \Phi^T \varepsilon$ . The above inequality leads to

$$\sum_{i \in \Omega \cup \Omega'} (\varepsilon_i^* - \delta'_i)^2 \leq \sum_{i \in \Omega \cup \Omega'} (\varepsilon_i^*)^2 + \lambda_0(|\Omega| - |\Omega'|) \cdot \sigma_{\min, |\Omega|+|\Omega'|}^2.$$

The above immediately leads to

$$\sup_{i \in \Omega \cup \Omega'} |\delta'_i| \leq \sup_{i \in \Omega \cup \Omega'} |\varepsilon_i^*| + \sqrt{\sum_{i \in \Omega \cup \Omega'} (\varepsilon_i^*)^2 + \lambda_0(|\Omega| - |\Omega'|) \cdot \sigma_{\min, |\Omega|+|\Omega'|}^2}.$$

Dividing both sides by  $\sigma_{\min, |\Omega|+|\Omega'|}^2$ , we have

$$\sup_{i \in \Omega \cup \Omega'} |\delta_i| \leq \frac{c(\Phi^T \varepsilon, 1) + \sqrt{c^2(\Phi^T \varepsilon, |\Omega| + |\Omega'|) + \lambda_0(|\Omega| - |\Omega'|) \cdot \sigma_{\min, |\Omega|+|\Omega'|}^2}}{\sigma_{\min, |\Omega|+|\Omega'|}^2}. \quad (4.15)$$

Recall the definitions of  $q_1(\cdot)$  and  $q(\cdot)$ , (4.12) and (4.13) can be derived directly from (4.15).

Now we are able to verify item (1) in the theorem. Suppose there is a type-I optimal subset  $\Omega'$  satisfying  $|\Omega'| < |\Omega|$ . We have

$$|x'_i| \geq |x_i| - |x_i - x'_i| \geq |x_i| - q_1(\Omega) > 0.$$

The second inequality is based on (4.12); and the last inequality is from the condition in item (1). The above implies  $\Omega \subset \Omega'$ , which contradicts  $|\Omega'| < |\Omega|$ . We have proved item (1).

The proof of item (2) is strongly similar to the proof of (1). We skip the obvious details.  $\square$

The above theorem is motivated by a recent related work in applied mathematics. Readers may compare it with the test proposed in Gribonval, Figueras i Ventura, and Vandergheynst (2005). Their test is related to the optimality in sparse signal representations.

In Theorem 4.11, quantities  $q_1(\cdot)$  and  $q(\cdot)$  require multiple values of  $\sigma_{\min,k}^2$ , for a range of values of  $k$ . Comparing to the quantities  $c(\cdot, k)$ , it is harder to compute  $\sigma_{\min,k}^2$ 's. Inspired by the derivation in Theorem 2 in Gribonval et al. (2005), we derive a sufficient condition, which only depends on  $\sigma_{\min,|\Omega|}^2$ , where  $\Omega$  is the subset that is tested. To state our result, the following quantity needs to be defined: for an integer  $m \geq 1$  and a given integral constant  $M$ , let

$$\lambda(m; M) = 1 - \frac{M}{\sqrt{m}} \sup_{|\mathcal{I}| \leq m} \sup_{k \notin \mathcal{I}} \|\Phi_{\mathcal{I}}^+ \phi_k\|_2,$$

where  $\mathcal{I}$  is a subset of indices,  $|\mathcal{I}|$  denotes the size of this subset, matrix  $\Phi_{\mathcal{I}}$  is a submatrix of  $\Phi$  whose column indices form the set  $\mathcal{I}$ ,  $\Phi_{\mathcal{I}}^+ = (\Phi_{\mathcal{I}}^* \Phi_{\mathcal{I}})^{-1} \Phi_{\mathcal{I}}^*$  is the Moore-Penrose pseudo-inverse (Golub and Loan, 1996) with  $(\cdot)^*$  denoting the adjoint, and  $\phi_k$  is the  $k$ th column (i.e., covariate) in  $\Phi$ . Given  $m$ , quantity  $\lambda(m)$  can be computed by enumerating all  $m$ -subset of the covariates.

Now we present another sufficient condition.

**Theorem 4.12** *Given a subset of coefficient  $\Omega$ . Suppose that coefficient vector  $x$  is the minimizer of function  $\|y - \Phi x\|_2^2$  subject to  $\text{supp}(x) \subset \Omega$ . Suppose it is known a priori that the size of the type-I optimal subset is no larger than  $M$ . If  $\min_i |x_i| > q'(|\Omega|, M)$ , then set  $\Omega$  is at least a subset of the type-I optimal subset. Here quantity  $q'(\cdot)$  is defined as, for integer  $k \geq 1$  and constant  $M$ ,*

$$q'(k, M) = \sup_{1 \leq m \leq M} \frac{c(\Phi^T \varepsilon, 1) + \sqrt{c^2(\Phi^T \varepsilon, k) + \lambda_0 \cdot \frac{k^2(k-m)}{(k+m)^2} \cdot \sigma_{\min,k}^2 \cdot \lambda^2(k; m)}}{\frac{k}{k+m} \sigma_{\min,k}^2 \cdot \lambda^2(k; m)}$$

**Proof.** The beginning of the proof is the same as the proof of the previous theorem. It starts to deviate at stage (4.14). For readers' convenience, we restate the inequality (4.14):

$$\|\Phi\delta\|_2^2 \leq 2\langle\Phi^T\varepsilon, \delta\rangle + \lambda_0(|\Omega| - |\Omega'|). \quad (4.16)$$

Readers are referred to the previous proof for the meanings of the notations.

First, we have

$$\langle\Phi^T\varepsilon, \delta\rangle \leq \sum_{i=1}^n |b_{(i)}| \cdot |\delta_{(i)}|, \quad (4.17)$$

where  $|\delta_{(1)}| \geq |\delta_{(2)}| \geq \dots \geq |\delta_{(n)}|$  is the ordered list of the magnitudes of the entries in vector  $\delta$ . Similarly,  $|b_{(1)}| \geq |b_{(2)}| \geq \dots \geq |b_{(n)}|$  is the ordered list of the magnitudes of the entries in vector  $\Phi^T\varepsilon$ . We denote  $\Phi^T\varepsilon$  by  $b$ . The following manipulations are needed:

$$\begin{aligned} \text{R.H.S. of (4.17)} &= \sum_{i=1}^{|\Omega|} |b_{(i)}| \cdot |\delta_{(i)}| + \sum_{i=|\Omega|+1}^n |b_{(i)}| \cdot |\delta_{(i)}| \\ &\leq \sum_{i=1}^{|\Omega|} |b_{(i)}| \cdot |\delta_{(i)}| + |b_{(|\Omega|+1)}| \cdot \sum_{i=|\Omega|+1}^n |\delta_{(i)}| \\ &\leq \sum_{i=1}^{|\Omega|} |b_{(i)}| \cdot |\delta_{(i)}| + |b_{(|\Omega|+1)}| \cdot \frac{|\Omega'|}{|\Omega|} \cdot \sum_{i=1}^{|\Omega|} |\delta_{(i)}| \\ &\leq \left(1 + \frac{|\Omega'|}{|\Omega|}\right) \sum_{i=1}^{|\Omega|} |b_{(i)}| \cdot |\delta_{(i)}| \\ &= \left(1 + \frac{|\Omega'|}{|\Omega|}\right) \langle b^*, \delta_{|\Omega|}^* \rangle, \end{aligned} \quad (4.18)$$

where vector  $\delta_{|\Omega|}^*$  takes the absolute values of  $\delta$  only at the positions where vector  $\delta$  has the  $|\Omega|$  largest magnitudes and zeros elsewhere. I.e.,

$$\delta_{|\Omega|,i}^* = \begin{cases} |\delta_i|, & \text{if } |\delta_i| \geq |\delta_{(|\Omega|)}|; \\ 0, & \text{elsewise.} \end{cases}$$

For vector  $b^*$ ,

$$b_i^* = |b_{(j)}|, \quad \text{where } \delta_i = \delta_{(j)}.$$

Putting (4.17) and (4.18) together, we have

$$\langle \Phi^T \varepsilon, \delta \rangle \leq \left(1 + \frac{|\Omega'|}{|\Omega|}\right) \langle b^*, \delta_{|\Omega|}^* \rangle. \quad (4.19)$$

Meanwhile, for any  $\Omega$ , we have

$$\begin{aligned} \|\Phi \delta\|_2^2 &\geq \|\Phi_\Omega \Phi_\Omega^+ \Phi \delta\|_2^2 \\ &\geq \sigma_{\min, |\Omega|}^2 \cdot \|\Phi_\Omega^+ \Phi \delta\|_2^2 \\ &= \sigma_{\min, |\Omega|}^2 \cdot \|\delta_\Omega + \Phi_\Omega^+ \Phi_{\Omega^c} \delta_{\Omega^c}\|_2^2, \end{aligned} \quad (4.20)$$

where set  $\Omega^c$  is the complement of set  $\Omega$ , matrices  $\Phi_\Omega$  and  $\Phi_{\Omega^c}$  are submatrices of matrix  $\Phi$  by taking columns whose indices are in  $\Omega$  and  $\Omega^c$ , respectively. As mentioned earlier, matrix  $\Phi_\Omega^+$  is a pseudo-inverse of  $\Phi_\Omega$ . Vector  $\delta_\Omega$  (respectively,  $\delta_{\Omega^c}$ ) only takes nonzero values when the index is in the set  $\Omega$  (respectively,  $\Omega^c$ ). In the above steps, the first inequality is true because the matrix  $\Phi_\Omega \Phi_\Omega^+$  is a projection matrix. The second inequality is based on the definition of  $\sigma_{\min, |\Omega|}^2$ . The last step is just a reorganization.

In the following, without loss of generality, we assume that set  $\Omega$  corresponds to the largest  $|\Omega|$  magnitudes in the vector  $\delta$ , i.e.,  $\delta_\Omega = \delta_{|\Omega|}^*$ . Note here  $\Omega$  can be any subset of the indices, which is different from the  $\Omega$  in the assumption at the beginning of the proof – we have an abuse of the notation. We can have

$$\begin{aligned} \|\delta_\Omega + \Phi_\Omega^+ \Phi_{\Omega^c} \delta_{\Omega^c}\|_2 &\geq \|\delta_\Omega\|_2 - \sum_{k \in \Omega^c} |\delta_k| \cdot \sup_{k \notin \Omega} \|\Phi_\Omega^+ \phi_k\|_2 \\ &\geq \|\delta_\Omega\|_2 - \sum_{k=|\Omega|+1}^n |\delta_{(k)}| \cdot \sup_{k \notin \Omega} \|\Phi_\Omega^+ \phi_k\|_2 \\ &\geq \|\delta_\Omega\|_2 - \frac{|\Omega'|}{|\Omega|} \cdot \|\delta_{|\Omega|}^*\|_1 \cdot \sup_{k \notin \Omega} \|\Phi_\Omega^+ \phi_k\|_2 \\ &\geq \|\delta_\Omega\|_2 - \frac{|\Omega'|}{\sqrt{|\Omega|}} \cdot \|\delta_{|\Omega|}^*\|_2 \cdot \sup_{k \notin \Omega} \|\Phi_\Omega^+ \phi_k\|_2 \\ &\geq \lambda(|\Omega|; |\Omega'|) \cdot \|\delta_{|\Omega|}^*\|_2. \end{aligned} \quad (4.21)$$

In the above, the first and the second steps are common maneuvers. The third inequality is based on  $\|\delta_{|\Omega|}^*\|_1/|\Omega| \geq \sum_{k=|\Omega|+1}^n |\delta_{(k)}|/|\Omega'|$ . The fourth inequality is

based on  $\|\delta_{|\Omega|}^*\|_1 \leq \sqrt{|\Omega|} \cdot \|\delta_{|\Omega|}^*\|_2$ . The last step recalls the definition of  $\lambda(\cdot, \cdot)$  and takes  $\Omega$  as the indices subset where  $\delta_{|\Omega|}^*$  having nonzero entries. Combining (4.20) and (4.21), we have

$$\|\Phi\delta\|_2^2 \geq \sigma_{\min,|\Omega|}^2 \cdot \lambda^2(|\Omega|; |\Omega'|) \cdot \|\delta_{|\Omega|}^*\|_2^2. \quad (4.22)$$

Now we put the above results together, and then maneuver back to the argument as in the proof of Theorem 4.11. Combining (4.16), (4.19), and (4.22), we have

$$\sigma_{\min,|\Omega|}^2 \cdot \lambda^2(|\Omega|; |\Omega'|) \cdot \|\delta_{|\Omega|}^*\|_2^2 \leq 2 \left(1 + \frac{|\Omega'|}{|\Omega|}\right) \langle b^*, \delta_{|\Omega|}^* \rangle + \lambda_0(|\Omega| - |\Omega'|).$$

Let

$$\delta' = \frac{|\Omega|}{|\Omega| + |\Omega'|} \sigma_{\min,|\Omega|}^2 \cdot \lambda^2(|\Omega|; |\Omega'|) \cdot \delta_{|\Omega|}^*.$$

We have

$$\|\delta'\|_2^2 \leq 2\langle b^*, \delta' \rangle + \lambda_0 \cdot \frac{|\Omega|^2(|\Omega| - |\Omega'|)}{(|\Omega| + |\Omega'|)^2} \cdot \sigma_{\min,|\Omega|}^2 \cdot \lambda^2(|\Omega|; |\Omega'|).$$

The above is equivalent to

$$\|\delta' - b^*\|_2^2 \leq \|b^*\|_2^2 + \lambda_0 \cdot \frac{|\Omega|^2(|\Omega| - |\Omega'|)}{(|\Omega| + |\Omega'|)^2} \cdot \sigma_{\min,|\Omega|}^2 \cdot \lambda^2(|\Omega|; |\Omega'|).$$

The above leads to the following

$$\|\delta'\|_\infty \leq \|b^*\|_\infty + \sqrt{c^2(b^*, |\Omega|) + \lambda_0 \cdot \frac{|\Omega|^2(|\Omega| - |\Omega'|)}{(|\Omega| + |\Omega'|)^2} \cdot \sigma_{\min,|\Omega|}^2 \cdot \lambda^2(|\Omega|; |\Omega'|)}.$$

Recall the definition of  $\delta'$  and  $b^*$ , we have

$$\begin{aligned} \|\delta\|_\infty &\leq \|b\|_\infty + \sqrt{c^2(b, |\Omega|) + \lambda_0 \cdot \frac{|\Omega|^2(|\Omega| - |\Omega'|)}{(|\Omega| + |\Omega'|)^2} \cdot \sigma_{\min,|\Omega|}^2 \cdot \lambda^2(|\Omega|; |\Omega'|)} \\ &\leq \frac{c(\Phi^T \varepsilon, 1) + \sqrt{c^2(\Phi^T \varepsilon, |\Omega|) + \lambda_0 \cdot \frac{|\Omega|^2(|\Omega| - |\Omega'|)}{(|\Omega| + |\Omega'|)^2} \cdot \sigma_{\min,|\Omega|}^2 \cdot \lambda^2(|\Omega|; |\Omega'|)}}{\frac{|\Omega|}{|\Omega| + |\Omega'|} \sigma_{\min,|\Omega|}^2 \cdot \lambda^2(|\Omega|; |\Omega'|)} \\ &\leq q'(|\Omega|; M). \end{aligned} \quad (4.23)$$

The above is equivalent to  $\|x - x'\|_\infty < q'(|\Omega|; M)$ . Using the same argument as in the last proof, we can argue that  $\Omega \subset \Omega'$ . Suppose  $x_i \neq 0$ , we have

$$|x'_i| \geq |x_i| - |x_i - x'_i| \geq |x_i| - q'(|\Omega|, M) > 0,$$

which implies that  $\Omega \subset \Omega'$ . □

### 4.3.1 Application in the Case with Orthonormal Covariates

If the model matrix  $\Phi$  is orthonormal, readers can verify that  $\sigma_{\min, k}^2 = 1$  and  $\lambda(m; M) = 1$ . It brings significantly simplified criteria in Theorem 4.11 and Theorem 4.12. Comparing with the result in Theorem 3.5, the new criteria are less attractive. We consider this a price of the generality.

It will be interesting to apply the above conditions to some applications with real data sets. However, due to the length of this paper, considering we are more focused on the formulation and theoretical developments in the present paper, we leave applications for future publications.

## 5 Discussion

### 5.1 Computing Versus Statistical Properties

The question that we addressed in this paper is quite different from some statistical works. In the present paper, we identify easy to verify (polynomial time) conditions for the type-I optimal subset. Our direct motivation is that certain greedy algorithm can find a path of type-II optimal subsets. If one of these type-II optimal subset is confirmed to be type-I optimal, then a concurrent optimal subset is obtained. In the above sense, our question is more statistical computing than prediction.

In traditional approaches of subset selection, researchers try to answer the questions regarding the consistency of variable selection, as well as the optimal accuracy rate in submodel prediction. There is a large scope of existing efforts. It is impossible and unnecessary for us to give a comprehensive survey here. We will

just list some publications that have been informative and inspiring to us. Efron et al. (2004), Weisberg (2004), Efron (2004), Shen, Huang, and Ye (2004), Zou, Hastie, and Tibshirani (2004), and the references therein give some interesting results in model estimation integrating the prediction accuracy. Consistency of variable selection has been studied in Zheng and Loh (1995).

Nowadays, due to the rapid rising of data sizes, it becomes increasingly important to develop computationally efficient statistical principle. Our idea of finding efficient sufficient conditions for otherwise unsolvable (i.e., NP-hard) subset selection principle is an incarnation of the aforementioned ideology.

## 5.2 Other Works in Variable Selection

Despite their generality, the formulations of **(P0)** and **(P1)** do not cover all the existing works in statistical model selection. We review some recent works that have attracted our attention.

Fan and Li (2001) proposes a family of new variable selection methods based on a nonconcave penalized likelihood approach. The criterion is to minimize

$$\text{Fan\&Li} = RSS(x) + 2n \cdot \sum_{j=1}^{\|x\|_0} p_\lambda(|\theta_j|),$$

where  $p_\lambda(\cdot)$  is a penalty function which is symmetric, nonconcave on  $(0, \infty)$  and has singularities at origin. With proper choice of  $\lambda$ , Fan and Li show that the estimators would have good statistical properties, such as sparsity and asymptotic normality.

Shen and Ye (2002) suggest an adaptive model selection procedure to estimate the algorithmic parameter  $\lambda$  from the data. In detail, the optimal value of  $\lambda$  is obtained by minimizing

$$\text{Shen\&Ye} = RSS(x) + \hat{g}_0(\lambda_0) \cdot \sigma^2,$$

which is derived from the optimal estimator of the loss  $l(\theta, \hat{\theta})$ . Quantity  $\hat{g}_0(\lambda_0)$  is the estimator of  $g_0(\lambda_0)$ , which is independent of the unknown parameter  $\theta$ . Value  $g_0(\lambda_0)/2$  is called the generalized degrees of freedom in Ye (1998).

At this moment, we do not know whether there are analogous conditions (to those in Section 4.3) that can be established in the above two settings. Examining possible connections will be an interesting topic for future research.

### 5.3 Back Elimination

Subset selections include at least three basic approaches: forward selection, backward elimination, and all subset selection. Problem **(P0)** is an all subset selection method. The greedy algorithms that have been discussed in this paper are assumed to be forward selection algorithms. Readers are referred to Section 2.2.

In Couvreur and Bresler (2000), a very interesting result is proved for backward elimination. It is shown that under certain conditions, back elimination finds the solution of **(P0)**. Such a result reveals the properties of problem **(P0)** from another angle.

It will be interesting to examine whether the approaches that are adopted in Section 4.3 can lead to stronger conditions in back elimination approaches. Again, this is left as a topic of future research.

### 5.4 Other Greedy Algorithms and Absolutely Optimal Subset in Variable Selection

We have treated LARS as a forward stepwise algorithm. Other greedy algorithms have made significant impact in other fields, such as signal processing. Two representative ones are matching pursuit (MP) (Davis, Mallat, and Zhang, 1994; Mallat and Zhang, 1993) and an improved version – orthogonal matching pursuit (OMP) (Pati, Rezaifar, and Krishnaprasad, 1993). MP and OMP do not generate the regularized solution path, while a version of LARS does. However, the intensive research effort following MP and OMP will provide researchers powerful tools.

Researchers have studied on the subsets that are unconditionally concurrent optimal, i.e., its concurrent optimality depends on neither the coefficients nor the corresponding residuals. The representative works include Donoho, Elad, and Temlyakov (2004), Tropp (2004a), and Tropp (2004b). The concept of exact recovery coefficient (ERC) (Tropp, 2004b) has inspired many recent works. Readers

can compare ERC with our quantity  $\lambda(m; M)$  that is defined right before Theorem 4.12 in Section 4.3.

Note that in our sufficient conditions, both coefficient and residuals are taken into account. This is due to the different emphasis of the problem. Comparing with our works, the results mentioned in the last paragraph can be considered as analysis of the worst cases.

## 5.5 Other Related Topics

An interesting model selection approach that adopts Bayesian computing is presented in Clyde and George (2004). This provides another interesting aspect of strategies. It will be interesting to analyze the connection with the contents of this paper.

Variable selection is a critical problem in supersaturated design. A citation search of Wu (1993) will provide most of existing literature. A numerically efficient condition on the optimality of subsets has the potential to identify a good design. Further study of this problem is left as a topic of future research.

## 6 Conclusion

Stepwise algorithms can be numerically efficient, i.e., polynomial time. Specially designed stepwise algorithms can find type-II optimal subset in subset selection. We derived sufficient conditions to test whether these type-II optimal subsets are also type-I optimal. Such an approach renders polynomial time algorithms to locate concurrent optimal subsets, which otherwise requires solving an NP-hard optimization problem in general.

## A Proof of Lemma 3.1

The choice of  $c_1$  depends on the following three correlations:

1. For  $m - A + 1 \leq i \leq m$ ,

$$\begin{aligned}\langle \phi_i, s - c_1 \phi_1 \rangle &= \langle \delta_i, s - c_1(a_1 s + b_1 \delta_1) \rangle \\ &= (1 - c_1 a_1) / \sqrt{A}.\end{aligned}\tag{A.24}$$

2. For  $1 \leq j \leq m - A$ ,

$$\begin{aligned}\langle \phi_j, s - c_1 \phi_1 \rangle &= \langle a_j s + b_j \delta_j, s - c_1(a_1 s + b_1 \delta_1) \rangle \\ &= a_j(1 - c_1 a_1) - c_1 b_j b_1 \langle \delta_j, \delta_1 \rangle.\end{aligned}$$

As the special cases, one has the following:

a. For  $j = 1$ ,

$$\langle \phi_1, s - c_1 \phi_1 \rangle = a_1 - c_1.\tag{A.25}$$

b. For  $j \geq 2$ ,

$$\langle \phi_j, s - c_1 \phi_1 \rangle = a_j(1 - c_1 a_1).\tag{A.26}$$

The choice of  $c_1$  is the maximum value that satisfies  $(A.25) \geq (A.24)$  and  $(A.25) \geq (A.26)$ . From  $(A.25) \geq (A.24)$ , we have  $a_1 - c_1 \geq (1 - c_1 a_1) / \sqrt{A}$ , which is equivalent to

$$a_1 - 1/\sqrt{A} \geq c_1(1 - a_1/\sqrt{A}).\tag{A.27}$$

From  $(A.25) \geq (A.26)$ , we have  $a_1 - c_1 \geq a_j(1 - c_1 a_1)$ , which is equivalent to

$$a_1 - a_j \geq c_1(1 - a_1 a_j).\tag{A.28}$$

Combining (A.27) and (A.28), we have

$$\begin{aligned}c_1 &= \min \left\{ \frac{a_1 - 1/\sqrt{A}}{1 - a_1/\sqrt{A}}, \frac{a_1 - a_j}{1 - a_1 a_j} \right\} \\ &= \frac{a_1 - a_2}{1 - a_1 a_2}.\end{aligned}$$

The last equality is based on the observation that function  $\frac{a_1 - x}{1 - x a_1}$  is a decreasing function of  $x$ .  $\square$

## B Proof of Theorem 3.2

In order to prove the theorem, we will need the following lemma.

**Lemma B.1** *The equiangular vector among  $\phi_1, \phi_2, \dots, \phi_k$  is*

$$u_k = [\phi_1 \ \phi_2 \ \cdots \ \phi_k] \left( D_k^{-1} \mathbf{1} - D_k^{-1} v_k \frac{v_k^T D_k^{-1} \mathbf{1}}{1 + v_k^T D_k^{-1} v_k} \right),$$

where

$$D_k = \begin{pmatrix} b_1^2 & 0 & \cdots & 0 \\ 0 & b_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & b_k^2 \end{pmatrix}, \quad v_k = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_k \end{pmatrix},$$

and  $\mathbf{1}$  is a  $k$ -dimensional all-one vector.

**Proof:** It is easy to verify that

$$[\phi_1 \ \phi_2 \ \cdots \ \phi_k]^T [\phi_1 \ \phi_2 \ \cdots \ \phi_k] = D_k + v_k v_k^T.$$

Using a known result in linear algebra, we have

$$(D_k + v_k v_k^T)^{-1} = D_k^{-1} - D_k^{-1} v_k v_k^T D_k^{-1} \frac{1}{1 + v_k^T D_k^{-1} v_k}.$$

Denoting  $\Phi_k = [\phi_1 \ \phi_2 \ \cdots \ \phi_k]$ , we have

$$\begin{aligned} u_k &= \Phi_k (\Phi_k^T \Phi_k)^{-1} \mathbf{1} \\ &= \Phi_k \left( D_k^{-1} \mathbf{1} - D_k^{-1} v_k v_k^T D_k^{-1} \mathbf{1} \frac{1}{1 + v_k^T D_k^{-1} v_k} \right). \end{aligned}$$

□

Note that in order to keep the formula simple, we do not normalize the vector  $u_k$ . In LARS, this does not change the selection of variables.

**Proof of Theorem 3.2** We apply induction to prove the theorem. Assume after step  $k-1$ ,  $k \leq m-A$ , the covariates  $\phi_1, \phi_2, \dots, \phi_k$  have been selected, and a surrogate residual is

$$\tilde{r}_{k-1} = s - \sum_{j=1}^{k-1} \frac{a_j - a_k}{b_j} \delta_j.$$

We will argue that in the next step, covariate  $\phi_{k+1}$  will be chosen, and the next surrogate residual has the form

$$\tilde{r}_k = s - \sum_{j=1}^k \frac{a_j - a_{k+1}}{b_j} \delta_j. \quad (\text{B.29})$$

Combining the above two, the theorem is proven. We first perform a sanity check:

1. For  $m - A + 1 \leq i \leq m$ ,  $\langle \phi_i, \tilde{r}_{k-1} \rangle = 1/\sqrt{A}$ .

2. For  $1 \leq j \leq k$ ,

$$\begin{aligned} \langle \phi_j, \tilde{r}_{k-1} \rangle &= \langle a_j s + b_j \delta_j, s - \sum_{t=1}^{k-1} \frac{a_t - a_k}{b_t} \delta_t \rangle \\ &= a_j - (a_j - a_k) \\ &= a_k. \end{aligned}$$

3. For  $k + 1 \leq j \leq m - A$ ,  $\langle \phi_j, \tilde{r}_{k-1} \rangle = a_j$ .

The next residual should be

$$r_k = \tilde{r}_{k-1} - c_k u_k,$$

where  $c_k$  is determined by considering the following three inner products:

1. For  $i \geq m - A + 1$ ,

$$\begin{aligned} \langle \phi_i, r_k \rangle &= \langle \phi_i, \tilde{r}_{k-1} - c_k u_k \rangle \\ &= \langle \phi_i, s - \sum_{t=1}^{k-1} \frac{a_t - a_k}{b_t} \delta_t - c_k u_k \rangle \\ &= \frac{1}{\sqrt{A}} - c_k \langle \phi, u_k \rangle \\ &= \frac{1}{\sqrt{A}} - c_k \frac{1}{\sqrt{A}} v_k^T \left( D_k^{-1} \mathbf{1} - D_k^{-1} v_k v_k^T D_k^{-1} \mathbf{1} \frac{1}{1 + v_k^T D_k^{-1} v_k} \right) \\ &= \frac{1}{\sqrt{A}} \left[ 1 - c_k \frac{v_k^T D_k^{-1} \mathbf{1}}{1 + v_k^T D_k^{-1} v_k} \right] \\ &= \frac{1}{\sqrt{A}} [1 - c_k g(k)], \end{aligned} \quad (\text{B.30})$$

where  $g(k) = \frac{v_k^T D_k^{-1} \mathbf{1}}{1 + v_k^T D_k^{-1} v_k}$ , and  $v_k, D_k$ , and  $\mathbf{1}$  are defined in Lemma B.1. These quantities will appear frequently in the following.

2. For  $j \leq k$ ,

$$\begin{aligned} \langle \phi_j, r_k \rangle &= \langle a_j s + b_j \delta_j, s - \sum_{t=1}^{k-1} \frac{a_t - a_k}{b_t} \delta_t - c_k u_k \rangle \\ &= a_k - c_k \langle a_j s + b_j \delta_j, u_k \rangle. \end{aligned}$$

From the definition of  $u_k$ , one has

$$\langle a_j s + b_j \delta_j, u_k \rangle = \langle \phi_j, u_k \rangle = 1.$$

Hence,

$$\langle \phi_j, r_k \rangle = a_k - c_k. \quad (\text{B.31})$$

3. For  $k+1 \leq j \leq m-A$ ,

$$\begin{aligned} \langle \phi_j, r_k \rangle &= \langle a_j s + b_j \delta_j, s - \sum_{t=1}^{k-1} \frac{a_t - a_k}{b_t} \delta_t - c_k u_k \rangle \\ &= a_j - c_k \langle a_j s + b_j \delta_j, u_k \rangle \\ &= a_j - c_k a_j \langle s, u_k \rangle \\ &= a_j - c_k a_j v_k^T \left( D_k^{-1} \mathbf{1} - D_k^{-1} v_k v_k^T D_k^{-1} \mathbf{1} \frac{1}{1 + v_k^T D_k^{-1} v_k} \right) \\ &= a_j - c_k a_j \frac{v_k^T D_k^{-1} \mathbf{1}}{1 + v_k^T D_k^{-1} v_k} \\ &= a_j [1 - c_k g(k)]. \end{aligned} \quad (\text{B.32})$$

In order to determine  $c_k$ , we consider two conditions:  $(B.31) \geq (B.30)$  and  $(B.31) \geq (B.32)$ . From  $(B.31) \geq (B.30)$ , we have  $a_k - c_k \geq \frac{1}{\sqrt{A}} [1 - c_k g(k)]$ , which is equivalent to

$$a_k - \frac{1}{\sqrt{A}} \geq c_k \left[ 1 - \frac{1}{\sqrt{A}} g(k) \right]. \quad (\text{B.33})$$

From  $(B.31) \geq (B.32)$ , we have  $a_k - c_k \geq a_j [1 - c_k g(k)]$ , which is equivalent to

$$a_k - a_j \geq c_k [1 - a_j g(k)]. \quad (\text{B.34})$$

Combining (B.33) and (B.34), we have

$$c_k = \min \left\{ \frac{a_k - \frac{1}{\sqrt{A}}}{1 - \frac{1}{\sqrt{A}}g(k)}, \frac{a_k - a_j}{1 - a_j g(k)}, j \geq k+1 \right\}.$$

It is not hard to verify that  $a_k < \frac{1}{g(k)}$ . One can also verify that function

$$f(x) = \frac{a_k - x}{1 - xg(k)} = \frac{1}{g(k)} + \frac{a_k - \frac{1}{g(k)}}{1 - xg(k)}$$

is a decreasing function of  $x$ . Hence,

$$c_k = \frac{a_k - a_{k+1}}{1 - a_{k+1}g(k)}.$$

It also indicates that  $\phi_{k+1}$  is selected in the next LARS step. This is the first result stated at the beginning of this proof. To verify (B.29), we need to compute the new residual:

$$\begin{aligned} r_k &= \tilde{r}_{k-1} - c_k u_k \\ &= s - \sum_{j=1}^{k-1} \frac{a_j - a_k}{b_j} \delta_j - c_k u_k \\ &= s - \sum_{j=1}^{k-1} \frac{a_j - a_k}{b_j} \delta_j - \frac{a_k - a_{k+1}}{1 - a_{k+1}g(k)} u_k. \end{aligned}$$

The coefficient of  $s$  in  $r_k$  is

$$\begin{aligned} &1 - \frac{a_k - a_{k+1}}{1 - a_{k+1}g(k)} v_k^T \left( D_k^{-1} \mathbf{1} - D_k^{-1} v_k v_k^T D_k^{-1} \mathbf{1} \frac{1}{1 + v_k^T D_k^{-1} v_k} \right) \\ &= 1 - \frac{a_k - a_{k+1}}{1 - a_{k+1}g(k)} g(k) \\ &= \frac{1 - a_k g(k)}{1 - a_{k+1}g(k)}. \end{aligned} \tag{B.35}$$

The coefficient of  $\delta_k$  in  $r_k$  is

$$\begin{aligned} &-\frac{a_k - a_{k+1}}{1 - a_{k+1}g(k)} \cdot \frac{1}{b_k} [1 - a_k g(k)] \\ &= -\frac{a_k - a_{k+1}}{b_k} \cdot \frac{1 - a_k g(k)}{1 - a_{k+1}g(k)}. \end{aligned} \tag{B.36}$$

The coefficient of  $\delta_j$ ,  $1 \leq j \leq k-1$ , in  $r_k$  is

$$\begin{aligned}
& -\frac{a_j - a_k}{b_j} - \frac{a_k - a_{k+1}}{1 - a_{k+1}g(k)} \cdot \frac{1}{b_j} [1 - a_j g(k)] \\
= & -\frac{1}{b_j} \left\{ a_j - a_k + \frac{a_k - a_{k+1}}{1 - a_{k+1}g(k)} [1 - a_j g(k)] \right\} \\
= & -\frac{a_j - a_{k+1}}{b_j} \cdot \frac{1 - a_k g(k)}{1 - a_{k+1}g(k)}. \tag{B.37}
\end{aligned}$$

By comparing (B.35), (B.36) and (B.37), one can conclude that the next surrogate residual, after getting rid of factor  $\frac{1 - a_k g(k)}{1 - a_{k+1}g(k)}$ , is

$$\tilde{r}_k = s - \sum_{j=1}^k \frac{a_j - a_{k+1}}{b_j} \delta_j.$$

This proves the second result stated at the beginning of this proof. From here, it is not hard to see that the theorem is proven.  $\square$

## References

- Akaike, H. (1973). Information theory and the maximum likelihood principle. In V. Petrov and F. Csáki (Eds.), *International Symposium on Information Theory*, Budapest. Akademiai Kiádo.
- Atkinson, A. C., M. Riani, and A. Cerioli (2004). *Exploring multivariate data with the forward search*. New York: Springer-Verlag. Springer series in statistics.
- Chen, J. and X. Huo (2005). Sparse representations for multiple measurement vectors (MMV) in an over-complete dictionary. In *International Conference on Acoustic, Speech, and Signal Processing (ICASSP)*.
- Chen, S. S., D. L. Donoho, and M. A. Saunders (1998). Atomic decomposition by basis pursuit. *SIAM J. Sci. Comput.* 20(1), 33–61. Reprinted at SIAM Rev. 43 (1): 129-159, (2001).
- Clyde, M. and E. I. George (2004, February). Model uncertainty. *Statist. Sci.* 19(1), 81–94.

- Couvreur, C. and Y. Bresler (2000). On the optimality of the backward greedy algorithm for the subset selection problem. *SIAM J. Matrix Anal. Appl.* 21(3), 797–808.
- Davis, G., S. Mallat, and Z. Zhang (1994). Adaptive time-frequency decompositions. *Optical Engrg.* 33, 2183–2191.
- Donoho, D. L., M. Elad, and V. Temlyakov (2004). Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Trans. Information Theory*. To appear. Available at <http://www-stat.stanford.edu/~donoho/Reports/2004/StableSparse-Donoho-et-al.pdf>.
- Donoho, D. L. and X. Huo (2001, November). Uncertainty principles and ideal atomic decomposition. *IEEE Transactions on Information Theory* 47(7), 2845–2862.
- Donoho, D. L. and I. M. Johnstone (1995). Adapting to unknown smoothness via wavelet shrinkage. *J. Amer. Statist. Assoc.* 90(432), 1200–1224.
- Efron, B. (2004, Sep.). The estimation of prediction error: Covariance penalties and cross-validation. *Journal of the American Statistical Association* 99(467), 619–632.
- Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani (2004). Least angle regression. *Ann. Statist.* 32(2), 407–499.
- Fan, J. and R. Li (2001). Variable selection via nonconvave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* 96(456), 1348–1360.
- George, E. I. and D. P. Foster (1994). The risk inflation criterion for multiple regression. *The Annals of Statistics* 22, 1947–1975.
- George, E. L. (2000, Dec.). The variable selection problem. *Journal of the American Statistical Association* 95(452), 1304–1308.

- Gilbert, A. C., M. Muthukrishnan, and M. J. Strauss (2003, Jan.). Approximation of functions over redundant dictionaries using coherence. In *The 14th Annual ACM-SIAM Symposium on Discrete Algorithms*.
- Gilmour, S. G. (1996). The interpretation of Mallows's  $C_p$ -statistic. *Statistician* 45(1), 49–56.
- Golub, G. H. and C. F. V. Loan (1996). *Matrix computations* (3rd ed.). Baltimore: Johns Hopkins University Press.
- Gribonval, R., R. M. Figueras i Ventura, and P. Vandergheynst (2005). A simple test to check the optimality of sparse signal approximations. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. [http://lts1pc19.epfl.ch/repository/Gribonval2005\\_1167.pdf](http://lts1pc19.epfl.ch/repository/Gribonval2005_1167.pdf); A longer version is available at <ftp://ftp.irisa.fr/techreports/2004/PI-1661.pdf>.
- Gribonval, R. and M. Nielsen (2003). Sparse representations in unions of bases. *IEEE Trans. Inform. Theory* 49(12), 3320–3325.
- Hastie, T., S. Rosset, R. Tibshirani, and J. Zhu (2004, October). The entire regularization path for the support vector machine. *Journal of Machine Learning Research* 5, 1391–1415. Also show in Neural Information Processing Systems (NIPS 2004).
- Hastie, T., R. Tibshirani, and J. Friedman (2001). *The elements of statistical learning: data mining, inference, and prediction*. New York: Springer.
- Malioutov, D. M., M. Cetin, and A. S. Willsky (2005, March). Homotopy continuation for sparse signal representation. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Volume 5, Philadelphia, PA, pp. 733–736.
- Mallat, S. (1998). *A wavelet tour of signal processing*. San Diego, CA: Academic Press, Inc.
- Mallat, S. and Z. Zhang (1993). Matching pursuit in a time-frequency dictionary. *IEEE Trans. Signal Proc.* 41, 3397–3415.

- Mallows, C. L. (1973). Some comments on  $C_p$ . *Technometrics* 15, 661–675.
- Miller, A. J. (1990). *Subset selection in regression*. New York: Chapman and Hall.
- Natarajan, B. K. (1995). Sparse approximate solutions to linear systems. *SIAM Journal on Computing* 24(2), 227–234.
- Osborne, M. R., B. Presnell, and B. Turlach (2000a). On the Lasso and its dual. *Journal of Computational and Graphical Statistics* 9(2), 319–337.
- Osborne, M. R., B. Presnell, and B. A. Turlach (2000b). A new approach to variable selection in least squares problems. *IMA J. Numer. Anal.* 20(3), 389–403.
- Pati, Y. C., R. Rezaifar, and P. S. Krishnaprasad (1993). Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. In A. Singh (Ed.), *Proc. 27th Asilomar Conference on Signals, Systems and Computers*, Los Alamitos, CA. IEEE Comput. Soc. Press.
- Rockafellar, R. T. (1970). *Convex analysis*. Princeton, NJ: Princeton University Press.
- Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics* 6, 461–464.
- Shen, X. T., H. C. Huang, and J. Ye (2004, Sep.). Inference after model selection. *Journal of the American Statistical Association* 99(467), 751–762.
- Shen, X. T. and J. M. Ye (2002, Mar.). Adaptive model selection. *Journal of the American Statistical Association* 97(457), 210–221.
- Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *J. Roy. Statist. Soc. Ser. B* 58(1), 267–288.
- Tropp, J. A. (2004a, October). Greed is good: Algorithmic results for sparse approximation. *IEEE Trans. Inform. Theory* 50(10), 2231–2242.

- Tropp, J. A. (2004b). Just relax: Convex programming methods for subset selection and sparse approximation. Technical report, ICES Report 04-04, UT-Austin.
- Weisberg, S. (2004). Discussion of Efron et al. (2004). *The Annals of Statistics* 32(2), 490–494.
- Wu, C. F. J. (1993, Sep.). Construction of supersaturated designs through partially aliased interactions. *Biometrika* 80(3), 661–669.
- Ye, J. M. (1998, Mar.). On measuring and correcting the effects of data mining and model selection. *Journal of the American Statistical Association* 93(441), 120–131.
- Zheng, X. D. and W. Y. Loh (1995, Mar.). Consistent variable selection in linear models. *Journal of the American Statistical Association* 90(429), 151–156.
- Zou, H., T. Hastie, and R. Tibshirani (2004). On the “degrees of freedom” of the Lasso. Submitted manuscript. Available at <http://www-stat.stanford.edu/~hastie/Papers/>.