Variable Selection by RIVAL

Er-Wei Bai, Kang Li and Paul Kump

Abstract—The paper considers variable selection problem and proposes an algorithm called the RIVAL (Removing Irrelevant Variables Amidst Lasso Iterations). For a given and fixed length of data points, the algorithm recursively updates the weights so that the ability of the algorithm in detecting zero coefficients is substantially improved. Theoretical convergence is established supported by numerical simulation results.

I. INTRODUCTION

We study a variable selection problem in this paper. Consider a linear system represented by

\[ y_n = X_n \beta^* + v_n \]  

(1)

where \( v_n = (v(1), ..., v(n))^T \in \mathbb{R}^n \) is an iid random noise sequence with zero mean and finite variance, \( y_n = (y(1), ..., y(n))^T \in \mathbb{R}^n \) is the output sequence or the response sequence and \( X_n \in \mathbb{R}^{n \times p} \) is the regressor matrix. The system (1) is assumed to be sparse, i.e., some of the unknown coefficients \( \beta^* = (\beta_1^*, ..., \beta_p^*)^T \in \mathbb{R}^p \) are exactly zero corresponding to the regression vectors that are irrelevant to the output. Without loss of generality by re-arranging indices, we may assume

\[
\begin{pmatrix}
\beta_1^* \\
\vdots \\
\beta_d^* \\
0 \\
0
\end{pmatrix}, \quad \beta_1^* \neq 0, ..., \beta_d^* \neq 0, \beta_{d+1}^* = ... = \beta_p^* = 0
\]

(2)

for some unknown \( 0 < d < p \). Also, let the index set be

\[ A^* = \{ j : \beta_j^* \neq 0 \} \]

(3)

Variable selection is to identify the index set \( A^* \) and remove those irrelevant variables that will simplify the model and enhance the prediction performance significantly. The ordinary least squares fails in this regard and usually gives small but non-zero estimates for \( \beta_j^*, j = d + 1, ..., p \). One very popular approach is the penalized approach referred to as Lasso [9],

\[
J_1(\beta) = \min_{\beta} \| y_n - X_n \beta \|^2 + \lambda_n \sum_{j=1}^p w_j |\beta_j|
\]

(4)

where

\[
\begin{pmatrix}
w_1 \\
w_2 \\
\vdots \\
w_p
\end{pmatrix} \in \mathbb{R}^p
\]

is the positive weighting vector and \( \lambda_n \) is the positive regularization parameter.

Lasso approach to variable selection has been intensively investigated [4], [5], [6], [8], [9], [11] and identifies the index set \( A^* \) if the irrepresentable condition [11] is satisfied. The problem is that in many applications the irrepresentable condition does not hold. To this end, a number of variants have been proposed including adaptive Lasso [12], non-negative garrote [10] and others [13]. These algorithms are asymptotic in nature. Under some technical conditions, these algorithms identify the correctly index set \( A^* \) as \( n \to \infty \). The problem is that in many applications, the number \( n \) of the data points is large but finite. Therefore, it would be desirable to have some results that could apply to a large but fixed \( n \).

The idea of the current paper is for a large but fixed \( n \), one generates a sequence of weights \( w(k) > 0, k = 0, 1, 2, ..., \). Let \( \beta(k), k = 0, 1, 2, ... \) be the corresponding solutions of Lasso for the given weights \( w(k), k = 0, 1, 2, ... \) respectively, where \( n \) is fixed. The hope is, for a fixed \( n \), by a properly generated sequence of \( w(k), k = 0, 1, 2, ... \), there exists an \( k_0 > 0 \) and constants \( 0 < \eta_1 < \eta_2 < \infty \) so that for all \( k \geq k_0, 0 < \eta_1 \leq |\beta_i(k)| \leq \eta_2 < \infty \) if \( \beta_i^* \neq 0 \) and \( \beta_i(k) = 0 \) if \( \beta_i^* = 0 \). Simply put, the unknown index set \( A^* \) is correctly identified. Note the iteration of \( \beta(k), k = 0, 1, 2, ... \) is with respect to a fixed \( n \). This is completely different from adaptive Lasso or non-negative garrote where only when a new observation is made or \( n \to n+1 \), a new \( \beta \) based on the \( n+1 \) data points is recalculated. The idea of the paper was first conceived in an unpublished note [1] which was further developed for an application where all unknown \( \beta_i^* \geq 0, i = 1, 2, ..., p \) [2]. Under this restrictive positive assumption, the algorithm has demonstrated substantial improvement in terms of variable selection over Lasso and its variants [2]. The contribution of this paper is to show that the restrictive positive condition that all unknown but true \( \beta_i^* \geq 0, i = 1, 2, ..., p \) can be removed and the RIVAL can be in fact applied to a general variable selection problem with some modifications.
II. RIVAL ALGORITHM

In this section, we first present the algorithm and then show the convergence results.

RIVAL algorithm:

Consider the system (1) and Lasso (2) for given \( n, y_n, X_n \) and \( \lambda_n \). Assume \( \frac{1}{n}X_n^TX_n > 0 \). Let \( 0 \leq q(k) \leq 1 \) be a sequence satisfying \( \sum_{k=1}^{\infty} q(k) = \infty \).

Step 1: Let \( \bar{w}_j(1) = (\bar{w}_{j1}(1), ..., \bar{w}_{jp}(1))' = (1, ..., 1)' \), \( w_j(1) = \xi_j \bar{w}_j(1), \) \( j = 1, ..., p \) and set \( k = 1 \), where \( \xi_j \) is the 2-norm of the \( j \)th column of \( \frac{1}{n}X_n^TX_n \).

Step 2: For the given \( w(k) \) at \( k \)'s iteration, apply Lasso as in (2) and denote \( \beta_k = (\beta_{k1}, \beta_{k2}, ..., \beta_{kp})' \) the solution of Lasso with respect to the weight vector \( w(k) \).

Step 3: If \( \beta_j(k) = 0 \), set \( \beta_j(k+1) = 0 \), for all \( i \geq 0 \). Remove \( \beta_j \) and \( w_j \) from \( \beta \) and \( w \) respectively as

\[
\begin{pmatrix}
\beta_1 \\
\vdots \\
\beta_{j-1} \\
\beta_j \\
\beta_{j+1} \\
\vdots \\
\beta_p 
\end{pmatrix} \rightarrow 
\begin{pmatrix}
\beta_1 \\
\vdots \\
\beta_{j-1} \\
\beta_{j+1} \\
\vdots \\
\beta_p 
\end{pmatrix}, \\
\begin{pmatrix}
w_1 \\
\vdots \\
w_{j-1} \\
w_{j+1} \\
\vdots \\
w_p 
\end{pmatrix} \rightarrow 
\begin{pmatrix}
w_1 \\
\vdots \\
w_{j-1} \\
w_{j+1} \\
\vdots \\
w_p 
\end{pmatrix}.
\]

Also remove the corresponding \( j \)th column from \( X_n \) so the dimension of the optimization is reduced by one. If \( |\beta_j(k)| > 0 \), let

\[
\bar{w}_j(k+1) = q(k) \cdot \frac{1}{|\beta_j(k)|} + (1 - q(k))\bar{w}_j(k), \\
w_j(k+1) = \xi_j \bar{w}_j(k+1).
\]

Repeat the process for all \( j = 1, 2, ..., p \) (dimension could be smaller if some \( \beta_j \)'s have been removed).

Step 4: Set \( k = k+1 \) and go back to Step 2. The dimension could be reduced if some of \( \beta_j = 0 \) at Step 3. (This step may be modified to add the stopping criterion using the standard one in the numerical analysis, e.g., the iteration stops if \( ||\beta(k+1) - \beta(k)|| / ||\beta(k)|| \) is smaller than the prescribed threshold).

The following result justifies the above algorithm.

Theorem 2.1: Consider the RIVAL. Assume \( n/\lambda_n \), \( \lambda_n \rightarrow \infty \) as \( n \rightarrow \infty \) and that \( \lambda_n \) is orthogonal, i.e., \( \frac{1}{n}X_n^TX_n = \\
\begin{pmatrix}
\xi_1 & 0 & \cdots & 0 \\
0 & \xi_2 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & \xi_p 
\end{pmatrix} > 0. \) Then, there is an integer \( n_0 > 0 \).

For any \( n \geq n_0 \), there exists an integer \( k_0 > 0 \) such that the sequence \( \beta(k) \) generated by the RIVAL satisfies

\[
0 < \eta_1 \leq |\beta_1(k)|, ..., |\beta_d(k)| \leq \eta_2 < \infty, \forall k \geq k_0, \\
\beta_{d+1}(k) = \beta_{d+2}(k) = ... = \beta_p(k) = 0, \forall k \geq k_0 \\
or equivalently \quad A(k) = A^* \quad \forall k \geq k_0
\]

where \( A(k) \) is the estimated index set at stage \( k \).

We make a few comments here.

- The result shows the convergence of the index set \( A = A^* \) but not the parameter convergence. However with \( A = A^* \), the asymptotic parameter convergence can be easily achieved. First, trim off all zero coefficient \( \beta_j = 0 \)'s identified in \( A \) and the corresponding regression vectors. Then, apply the ordinary least squares \( \beta_{AlS} = (\beta_{A1}, ..., \beta_{A2})' \) to estimate \( (\beta_1^*, ..., \beta_p^*)' \). The subscript \( AlS \) indicates that the estimate \( \beta_{AlS} \) is \( A \) dependent that is a result of the RIVAL. Clearly, if \( n \) is allowed to be large, the convergence of \( \beta_{AlS} \) to \( (\beta_1^*, ..., \beta_p^*)' \) is guaranteed by the least squares estimate properties and the convergence of \( A \) to \( A^* \).

- Similar to the standard Lasso, the choice of \( \lambda_n \) is critical. Though the increasing rate \( \lambda = \lambda_n \) as \( n \) gets larger is specified in the theorem, in a particular application \( n \) is always fixed and how to choose an optimal \( \lambda \) becomes issue. In general, the choice of \( \lambda \) is either based on information criterion, e.g., by AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion) [4] or based on cross-validation approach [9], [12]. We will follow the cross validation approach in this paper.

- The initial estimate \( \beta(1) \) is generated by the uniform weights \( \bar{w}_i(1) = 1, i = 1, ..., p \). To speed up, the inverse of the least square could be used \( \bar{w}_i(1) = 1/\beta_{AlS} \) if \( n \) is large, where \( \beta_{AlS} \) is the ordinary least squares estimate.

- At least in theory, the RIVAL has the perfect ability to find the index set if \( n \) is large but fixed and \( \frac{1}{n}X_n^TX_n \) is diagonal and positive. Simulations seem to suggest that the results also hold for a much larger classes.

III. NUMERICAL SIMULATIONS

In this section, we present numerical simulations of three extensively studied benchmark models [9], [12], [13] including a few large effects, all small effects and one large effect. For comparison, the results of Lasso [9], adaptive Lasso [12], non-negative garrote [10] and elastic net [13] are also included.

Model 1 (a few large effects): In this example, \( \beta^* = (3, 1.5, 0, 0, 2, 0, 0, 0, 0)' \). The regressor matrix \( X_n \) consists of iid normal columns. The pairwise correlation between \( i \)th and \( j \)th columns is \( \text{cor}(i, j) = (0.5)^{|i-j|} \). In simulations, noise is iid normal random variable with \( \sigma^2 = 1 \) and \( \lambda_n \) respectively. We set \( n = 20, 60, 100 \) to see the performance for small \( n = 20 \) to large \( n = 100 \). The tuning parameter \( \lambda_n \) is determined by 5-fold cross validation.

Model 2 (all small effects): The same as in Model 1 but with \( \beta^* = 0.85 \) for all \( j \).

Model 3 (one large effect): The same as in Model 1 but with \( \beta^* = (5, 0, 0, 0, 0, 0, 0, 0, 0)' \).

Tables 1 and 2 show the rates of correctly identifying the unknown index set \( A^* \) for 200 Monte Carlo simulations for Models 1 and 2 respectively. From the tables, the proposed RIVAL algorithm outperforms all 4 other well known algorithms for almost every pair of \( (n, \sigma^2) \), especially the improvement is obvious when \( n \) is small.
In Model 3, $\beta_j^*$‘s is either large or zero that makes variable selection problem easy. All 5 algorithms perform well.

IV. CONCLUDING REMARKS

In the paper, the RIVAL is proposed for variable selection along with its convergence analysis. The algorithm is particularly useful for practical applications where the length of data points is always finite. The theoretical results are supported by numerical simulations.

REFERENCES


V. APPENDIX

Lemma 5.1: Assume $\frac{n}{\lambda_n} \to \infty$ and $\frac{n}{\sqrt{n}} \to \infty$ as $n \to \infty$.

Further, assume uniformly for $n \geq n_0$ for some $n_0 > 0$,

$$\alpha_2 I \geq \frac{1}{n} X_n^T X_n = C_n = \begin{pmatrix} c_n(1, 1) & c_n(1, 2) \\ c_n^T(1, 2) & c_n(2, 2) \end{pmatrix} \geq \alpha_1 I > 0.$$ 

Define

$$s = \begin{pmatrix} w_{11} \cdot \text{sgn}(\beta_1^n) \\ w_{12} \cdot \text{sgn}(\beta_2^n) \\ \vdots \\ w_{1d} \cdot \text{sgn}(\beta_d^n) \end{pmatrix} \in \mathbb{R}^d,$$

and

$$a_n = c_n^T(1, 2)c_n^{-1}(1, 1)s \in \mathbb{R}^{p-d}.$$ 

Let $\epsilon_n > 0$ be a positive sequence satisfying $\epsilon_n \to 0$, $c_n/(\sqrt{\lambda_n}) \to \infty$, $\epsilon_n/(\sqrt{\lambda_n}) \to \infty$ as $n \to \infty$. Then, in probability as $n \to \infty$, a sufficient condition for $A \to A^*$ is

$$|a_{nj}| \leq w_{nj} - \epsilon_n, \quad j = 1, 2, ..., p-d. \quad (5)$$

<table>
<thead>
<tr>
<th>n</th>
<th>Lasso</th>
<th>Adaptive Lasso</th>
<th>Elastic Net</th>
<th>Garrote</th>
<th>RIVAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>0.47</td>
<td>0.485</td>
<td>0.51</td>
<td>0.695</td>
<td>0.85</td>
</tr>
<tr>
<td>60</td>
<td>0.74</td>
<td>0.77</td>
<td>0.79</td>
<td>0.88</td>
<td>0.91</td>
</tr>
<tr>
<td>100</td>
<td>0.83</td>
<td>0.835</td>
<td>0.885</td>
<td>0.895</td>
<td>0.93</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>n</th>
<th>Lasso</th>
<th>Adaptive Lasso</th>
<th>Elastic Net</th>
<th>Garrote</th>
<th>RIVAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>0.4</td>
<td>0.445</td>
<td>0.315</td>
<td>0.46</td>
<td>0.93</td>
</tr>
<tr>
<td>60</td>
<td>1</td>
<td>0.995</td>
<td>0.975</td>
<td>0.995</td>
<td>1</td>
</tr>
<tr>
<td>100</td>
<td>1</td>
<td>0.975</td>
<td>0.975</td>
<td>0.975</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>n</th>
<th>Lasso</th>
<th>Adaptive Lasso</th>
<th>Elastic Net</th>
<th>Garrote</th>
<th>RIVAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>0.868</td>
<td>0.89</td>
<td>0.95</td>
<td>0.855</td>
<td>0.94</td>
</tr>
<tr>
<td>60</td>
<td>0.984</td>
<td>0.988</td>
<td>1</td>
<td>0.915</td>
<td>0.965</td>
</tr>
<tr>
<td>100</td>
<td>0.988</td>
<td>0.992</td>
<td>1</td>
<td>0.93</td>
<td>0.975</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>n</th>
<th>Lasso</th>
<th>Adaptive Lasso</th>
<th>Elastic Net</th>
<th>Garrote</th>
<th>RIVAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>0.47</td>
<td>0.51</td>
<td>0.695</td>
<td>0.85</td>
<td>0.93</td>
</tr>
<tr>
<td>60</td>
<td>0.74</td>
<td>0.77</td>
<td>0.79</td>
<td>0.88</td>
<td>0.91</td>
</tr>
<tr>
<td>100</td>
<td>0.83</td>
<td>0.835</td>
<td>0.885</td>
<td>0.895</td>
<td>0.93</td>
</tr>
</tbody>
</table>

TABLE I

RATE OF CORRECTLY IDENTIFYING THE INDEX SET $A^*$, MODEL 1.

<table>
<thead>
<tr>
<th>n</th>
<th>Lasso</th>
<th>Adaptive Lasso</th>
<th>Elastic Net</th>
<th>Garrote</th>
<th>RIVAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>0.4</td>
<td>0.445</td>
<td>0.315</td>
<td>0.46</td>
<td>0.93</td>
</tr>
<tr>
<td>60</td>
<td>1</td>
<td>0.995</td>
<td>0.975</td>
<td>0.995</td>
<td>1</td>
</tr>
<tr>
<td>100</td>
<td>1</td>
<td>0.975</td>
<td>0.975</td>
<td>0.975</td>
<td>1</td>
</tr>
</tbody>
</table>

TABLE II


<table>
<thead>
<tr>
<th>n</th>
<th>Lasso</th>
<th>Adaptive Lasso</th>
<th>Elastic Net</th>
<th>Garrote</th>
<th>RIVAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>0.868</td>
<td>0.89</td>
<td>0.95</td>
<td>0.855</td>
<td>0.94</td>
</tr>
<tr>
<td>60</td>
<td>0.984</td>
<td>0.988</td>
<td>1</td>
<td>0.915</td>
<td>0.965</td>
</tr>
<tr>
<td>100</td>
<td>0.988</td>
<td>0.992</td>
<td>1</td>
<td>0.93</td>
<td>0.975</td>
</tr>
</tbody>
</table>

TABLE III

RATE OF CORRECTLY IDENTIFYING THE INDEX SET $A^*$, MODEL 3.
Lemma 5.2: Consider a sequence of scalar minimization problems

\[ J = \min_{\beta(k)} \{ a \beta^2(k) - c \beta(k) + w(k) | \beta(k) | \}, \quad a > 0, \ w(1) = 1 \]

Let \( \beta(k) \) be the solution of \( J \) for the given \( w(k) \). Construct \( \beta(k+1) \) as follows. If \( \beta(k) = 0 \), set \( \beta(k+i) = 0 \) for \( i \geq 0 \) and stop the algorithm. If \( \beta(k) \neq 0 \), let \( \bar{w}(k) = 1/|\beta(k)| \) and

\[ w(k+1) = q(k) \bar{w}(k) + (1 - q(k)) w(k), \]

where \( 0 \leq q(k) \leq 1 \) is a sequence satisfying \( q(k) \to 0 \) and \( \sum_{k=1}^{\infty} q(k) = \infty \). Denote \( \beta(k+1) \) the solution of \( J \) for given \( w(k+1) \). Assume for some constant \( \delta > 0 \)

\[ a \geq \frac{(c + \delta)^2}{8} \]

Then, there exists a finite integer \( k_0 \geq 1 \) such that the sequence generated above satisfies

\[ \beta(k) = 0, \ \forall k \geq k_0 \]

Proof: If \( 1 = w(1) \geq c \), the minimum \( \beta(1) = 0 \) and this implies \( k_0 = 1 \) and \( \beta(k) = 0, \ k \geq k_0 = 1 \). If \( c > w(1) \), the minimum \( \beta(1) \) is achieved at some \( \beta(k) > 0 \). The first order necessary condition

\[ \frac{\partial J}{\partial \beta} = 2a\beta - (c - w) \]

implies

\[ \beta(1) = \frac{c - w(1)}{2a} > 0 \to \bar{w}(1) = \frac{2a}{c - w(1)} \]

From the hypothesis, we have

\[ 2a - \frac{(c - \delta)^2}{4} \geq \delta c \to (w(1) - \frac{(c - \delta)^2}{2})^2 - \frac{(c - \delta)^2}{4} \]

\[ + 2a \geq \delta c \]

Thus,

\[ \frac{2a - cw(1) + w^2(1)}{c - w(1)} \geq \delta \quad \text{or} \quad \bar{w}(1) \geq w(1) + \delta \]

and

\[ w(2) = q(1)\bar{w}(k) + (1 - q(1)) w(k) \]

\[ \geq q(1) w(1) + q(1) \delta + (1 - q(1)) w(k) \]

\[ = w(1) + q(1) \delta \]

By induction, if \( w(k) < c, \ w(k+1) \geq w(k) + q(k) \delta \geq w(1) + \sum_{i=0}^{k} p(i) \delta \) or equivalently, there is an integer \( k_0 > 0 \) such that \( w(k_0) > c \) and the corresponding solution of \( J \) is \( \beta(k_0) = 0 \) and \( \beta(k_0 + i) = 0, \ i \geq 0 \). This completes the proof.

Lemma 5.3: Consider a sequence of scalar minimization problems

\[ J = \min_{\beta(k)} \{ a \beta^2(k) - 2ab \beta(k) + w(k) | \beta(k) | \}, \quad a > 0, \ b \neq 0, \ w(1) = 1 \]

Let \( \beta(k) \) be the solution of \( J \) for the given \( w(k) \). Construct \( \beta(k+1) \) as follows. If \( \beta(k) = 0 \), set \( \beta(k+i) = 0 \) for \( i \geq 0 \) and stop the algorithm. If \( \beta(k) \neq 0 \), let \( \bar{w}(k) = 1/|\beta(k)| \) and

\[ w(k+1) = q \bar{w}(k) + (1 - q) w(k), \ 0 \leq q \leq 1 \]

Denote \( \beta(k+1) \) the solution of \( J \) for given \( w(k+1) \). Assume that there exists a constant \( \delta > 0 \)

\[ |b| > \delta = \frac{|b| - \sqrt{b^2 - 2/a}}{2} > 0 \]

and

\[ 0 < 1 = w(1) \leq 2a |b| - \frac{1}{|b| - \delta} \]

Then, the sequence \( \beta(k) \) is uniformly bounded above and also uniformly bounded away from zero

\[ 0 < \eta_1 \leq |\beta(k)| \leq \eta_2 < \infty, \ \forall k \]

Proof: By the assumption that \( b \neq 0 \), there are two scenarios \( b < 0 \) and \( b > 0 \). We only show the case that \( b > 0 \). The case that \( b < 0 \) can be shown similarly. The idea of the proof is to show that \( w(k) \)'s are bounded, \( b > 0 \) and \( w(k) < 2ab \) imply that the minimum is achieved at some \( \beta(1) > 0 \). The first order necessary condition

\[ \frac{\partial J}{\partial \beta} = 2a\beta - 2ab + w \]

implies

\[ \beta(1) = b - \frac{w(1)}{2a} > 0 \to \bar{w}(1) = \frac{1}{b - \frac{w(1)}{2a}} > 0 \]

Further, \( w(1) \leq 2ab - \frac{1}{b - \delta} \) leads to

\[ (b - \delta)w(1) \leq 2ab(b - \delta) - 1 \to \]

\[ \bar{w}(1) - b = \frac{1}{2ab - w(1)} - b \leq -\delta \]

or

\[ \bar{w}(1) \leq 2ab - 2a \delta \]

On the other hand, from the definition of \( \delta \), it is easily verified that

\[ \delta^2 - \delta b + 1/(2a) = 0 \to 2a \delta = 1/(b - \delta) \]

Hence,

\[ 0 < w(1) \leq 2ab - \frac{1}{b - \delta} \to 0 < \bar{w}(1) \leq 2ab - \frac{1}{b - \delta} \]

and this implies

\[ 0 < w(2) = q \bar{w}(1) + (1 - q) w(k) \leq 2ab - \frac{1}{b - \delta} \]

By the induction, we have for all \( k \geq 1 \),

\[ 0 < w(k) \leq 2ab - \frac{1}{b - \delta} \]

and

\[ \beta(k) = b - \frac{w(k)}{2a} \geq \delta > 0 \]
This shows that $\beta(k)$ is bounded away from zero. The upper bound can be derived easily,

$$|\beta(k)| \leq b + |w(k)/(2a)| \leq 2b - 1/(b - \delta) < \infty, \forall k$$

This finishes the proof.

Proof of Theorem (2.1): From the lemmas, whether the index set $A \to A^*$ depends on $|a_{j2}| \leq w_{j2} - \epsilon_n$. Observe that $a_n$ depends on the data $X_n$ and the unknown $\beta_j^*$'s. In general, the above condition is not satisfied for a finite $n$. Thus, the question is if some modifications can be made. It is interesting to observe that the conditions depend on the choice of the weights $w_{j1} > 0$ and $w_{j2} > 0$,

$$|c_n^T(1,2)c_n^{-1}(1,1)
\begin{pmatrix}
w_{11} & \cdots \\
\vdots & \\
w_{1d} & \\
w_{d1} & \cdots & \cdots & \cdots & \cdots \\
w_{21} & \cdots \end{pmatrix}
\begin{pmatrix}
sgn(\beta_j^*) \\
\vdots \\
sgn(\beta_j^*) \\
\end{pmatrix}
|$$

$$\leq \left|\begin{pmatrix}w_{21} \\
w_{2(p-d)} \end{pmatrix}\right| - \epsilon_n$$

that is automatically satisfied if the weights were chosen in such a way that $w_{j1}$'s are small and $w_{j2}$'s are large enough, giving rise to the right index set $A^*$. The problem is of course that we do not know which $\beta_j^* = 0$ and which $\beta_j^* \neq 0$. The RIVAL is designed to overcome this difficulty. Now minimizing

$$J_1 = (y_n - X_n\beta)^T (y_n - X_n\beta) + \lambda_n \sum_{j=1}^p w_j |\beta_j|$$

is equivalent to minimizing

$$J_2 = -2n\beta^T C_n\beta - 2v_n^T X_n\beta + n^2 \beta^T C_n\beta + \lambda_n \sum_{j=1}^p w_j |\beta_j|$$

which is equivalent to minimizing

$$J_3 = -2n\beta^T C_n\beta + 2v_n^T X_n\beta + \lambda_n \sum_{j=1}^p w_j |\beta_j|$$

where

$$a = \frac{n}{\lambda_n}, \ d = \frac{2v_n^T X_n}{\lambda_n} = \frac{2v_n^T X_n \sqrt{n}}{\lambda_n} = (d_1, d_2, \ldots, d_p)$$

$$\sum_{j=1}^d [\xi_j a(\beta_j^* - \beta_j)^2 + d_j (\beta_j^* - \beta_j) + w_j |\beta_j|]$$

Therefore, minimizing $J_3$ is achieved by minimizing $J_4$ and $J_5$ separately, where

$$J_4 = \sum_{j=1}^d (a\beta_j^2 - 2ab_j \beta_j + w_j/\xi_j |\beta_j|), \ b_j = \beta_j^* + \frac{1}{2a\xi_j} d_j$$

$$J_5 = \sum_{j=1}^d [a\beta_j^2 - c_j \beta_j + w_j/\xi_j |\beta_j|]$$

Since $a = \frac{n}{\lambda_n}, \ d = \frac{2v_n^T X_n \sqrt{n}}{\lambda_n}, \ c_j = d_j/\xi_j, b_j = \beta_j^* + \frac{1}{2a\xi_j} d_j$, there always exists a constant $\delta > 0$ for large enough $n$ so that the conditions of the above lemmas are simultaneously satisfied

$$a \geq \frac{(c + \delta)^2}{8}, \ b |\beta| > \delta$$

Then, the conclusions follow.