IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 2, 2020, accepted July 12, 2020, date of publication July 15, 2020, date of current version July 24, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3009377

An Improved Forward Regression Variable Selection Algorithm for High-Dimensional Linear Regression Models

YANXI XIE, YUEWEN LI, ZHIJIE XIA, AND RUIXIA YAN
School of Management Studies, Shanghai University of Engineering Science, Shanghai 201620, China
Corresponding author: Yuewen Li (sueslyw@gmail.com)

This work was supported by the Ministry of Education of Humanities and Social Science Project under Grant 17YJCZH199.

ABSTRACT Variable selection plays an important role in various fields, such as process modeling and process monitoring. It generally involves a large number of predictor variables, usually with the number of predictor variables d much larger than the sample size n. Therefore, how to filter useful variables and extract useful information in high-dimensional setup is a critical issue in the era of big data. This paper proposes an improved Forward Regression algorithm for variable selection under the high-dimensional setup. The proposed improved Forward Regression method demonstrates good performance in relevant-variable selection by introducing a predefined stopping rule. The stopping rule links the residual sum of squares to the noise ratio so that the relevant predictors can be distinguished from the random noises. Throughout theoretical analysis and simulations, it is confirmed that the improved Forward Regression algorithm can identify relevant predictors to ensure selection consistency in variable selection. Compared with the traditional Forward Regression method, the proposed Forward Regression algorithm can improve prediction accuracy and reduce computational cost by selecting only the relevant variables.

INDEX TERMS Improved forward regression, high-dimensional setup, variable selection, selection consistency, big data.

I. INTRODUCTION

With the rapid development in information technology, contemporary data from various fields such as finance and gene expressions tend to be extremely large in terms of the number of variables. Sometimes the number of variables or parameters d can be much larger than the sample size n. For this kind of high dimensional problems, it is challenging to identify important variables out of thousands of predictors, with a number of observations usually in tens or hundreds. In other words, it becomes critical to investigate the existence of complex relationships and dependencies in high-dimensional data, in the aim of building a relevant model for future prediction.

Statistically, a traditional method is to conduct variable selection, which is a technique of selecting a subset of relevant features for building robust learning models, under small n and large d situation. By removing most irrelevant and redundant variables from the data, variable selection helps improve the performance of learning models in terms of obtaining higher estimation accuracy.

The associate editor coordinating the review of this manuscript and approving it for publication was Qingchao Jiang.

A. BACKGROUND

In regression analysis, a linear model is commonly used to link a response variable to explanatory variables. The resulting the ordinary Least Squares Estimates (LSE) have a closed form, which is easy to compute. However, the LSE fails when the number of linear predictors d is greater than the sample size n. The Best Subset Selection is one of the standard techniques for improving the performance of the LSE. A Best Subset Selection algorithm usually uses criteria such as the Akaike's information criterion (AIC) and the Byesian information criterion (BIC), to perform either forward or backward stepwise selection procedures to select variables. Among all the subset selection procedures,the Orthogonal Matching Pursuit (OMP), of which the selection consistency property was investigated in Zhang (2009), is of great interest to us. In fact, the OMP is an iteratively greedy algorithm that selects at each step the column which is most correlated with

the current residuals. In addition, various shrinkage methods have gained a lot of popularity during the past decades and the Least Absolute Shrinkage and Selection Operator (Lasso) in Tibshirani (1996) has been the most popularly used method. The fundamental explanation of these shrinkage methods is to make the bias-variance tradeoff, to overcome the limitations of the LSE and the Best Subset Selection methods. In the context of variable selection, screening approaches have also gained a lot of attention besides the Lasso. The Sure Independence Screening (SIS) proposed in Fan and Lv (2008) and the Forward Regression (FR) in Wang (2009) are the popular ones. When the predictor dimension is much larger than the sample size, the story changes drastically in the sense that the conditions for most of the Lasso type algorithms can not be satisfied. Therefore, to conduct model selection in high dimensional setup, variable screening is a reasonable solution. Wang (2009) proposed the Forward Regression (FR) method for ultrahigh dimensional variable screening. As one type of important greedy algorithms, the FR's theoretical properties have been studied in the previous literature.

B. MOTIVATION

There are two fundamental goals in statistical learning: identifying relevant predictors and ensuring high prediction accuracy. The first goal, by means of variable selection, is of particular importance when the true underlying model has a sparse representation. Discovering relevant predictors can enhance the performance of the prediction from the fitted model. Usually an estimate $\hat{\beta}$ is considered desirable if it is consistent in terms of both the coefficient estimate and the variable selection. Hence, before we try to estimate the regression coefficients β_js, it is preferable that we have a set of useful predictors in hand. Our task in this paper is to propose novel methods, in the aim of identifying relevant predictors to ensure consistency in variable selection.

Motivated by the current studies on variable selection, we are interested in showing the consistency property of the FR under certain conditions. We would like to restrict the technical conditions stated in Wang (2009) and hence select one relevant predictor at each step until all the relevant predictors are selected. A key component here is the stopping rule which depends on the noise structure.

The rest of this paper is organized as follows. Section 2 provides the literature review on current variable selection methods. Section 3 explains a variable selection technique based on the FR. In Section 4, the asymptotic results of the estimators are studied. Section 5 demonstrates via simulation that our proposed technique exhibits desired sample properties and can be useful in practical applications. Finally, Section 5 concludes the paper and some future research direction.

II. LITERATURE REVIEW OF VARIABLE SELECTION

Let $(\mathbf{x_i}, Y_i)$ be the observation collected from the i^{th} subject $(1 \leq i \leq n)$, where $Y_i \in R^1$ is the response variable and $\mathbf{x_i} = (x_{i1}, \ldots, x_{id})^T \in R^d$ is a vector of high dimensional predictors with $d > n$ and $cov(Y) = \Sigma$. Moreover, $\beta = (\beta_1, \ldots, \beta_d)^T$ is the regression coefficient. Without loss of generality, we assume that the data are centered, that is, the columns of X are orthonormal and Y_i's are conditionally independent given the design matrix X.In matrix representation, the design matrix is $X \in R^{n\times d}$ and the response vector is $Y \in R^n$. Consider the linear regression model

$$y = X\beta + \varepsilon. \tag{1}$$

Moreover, the error terms ϵ are independently and identically distributed with mean zero and finite variance σ^2. A model fitting procedure produces the vector of estimated coefficients $\hat{\beta} = (\hat{\beta_1}, \ldots, \hat{\beta_d})^T$.

The ordinary LSE are obtained by minimizing the residual sum of squared errors

$$\hat{\beta}_{LSE} = \min_{\beta}\{(Y - X\beta)^T(Y - X\beta)\}. \tag{2}$$

Though the LSE are easy to compute, there are two main disadvantages pointed out in Tibshirani (1996). Firstly, all the LSE are non-zero but only a subset of predictors is relevant to exhibit the strongest effects on response variable Y. Secondly, since the LSE often have low bias and large variance, the prediction accuracy is low. In fact, we can sacrifice a little bias to reduce the variance of the predicted values, and hence the overall prediction accuracy can be improved substantially. On top of the disadvantages, the LSE completely fail when the number of linear predictors d is greater than the sample size n.

The Best Subset Selection is one of the standard techniques for improving the performance of the LSE. The Best Subset Selection, such as Akaike's information criterion (AIC) and Bayesian information criterion (BIC), following either forward or backward stepwise selection procedures to select variables. Nevertheless, the stepwise best subset selection procedure has been identified as extremely variable since it usually results in very different models.

To overcome the limitations of the LSE and the Best Subset Selection, various penalization methods were proposed. They usually shrink estimates to make trade-off between bias and variance.The penalized estimates are obtained by minimizing the residual sum of squared errors plus a penalty term, i.e.

$$\hat{\beta}_{penalized} = \min_{\beta}\{(Y - X\beta)^T(Y - X\beta) + \lambda\sum_{j=1}^{d} p_\lambda(|\ \beta_j\ |)\}, \tag{3}$$

where $\lambda \geq 0$ is a tuning parameter and p_λ represents a penalty function.

Fan(1997) and Antoniadis(1997) both introduced the hard thresholding penalty function

$$p_\lambda(|\ \beta\ |) = \lambda^2 - (|\ \beta\ | - \lambda)^2\mathbf{I}(|\ \beta\ | < \lambda). \tag{4}$$

The resulting thresholding estimator is given by

$$\hat{\beta}_{HT} = \hat{\beta}_{LSE}\mathbf{I}(|\ \hat{\beta}_{LSE}\ | > \lambda). \tag{5}$$

Frank and Friedman (1993) mentioned the Bridge Regression with L_q penalty function $\lambda \mid \beta \mid^q$, where q is a positive constant. There are two cases in this Bridge Regression. When $q \leq 1$, the L_q penalty functions lead to sparse solutions with relatively large biases. On the other hand, when $q > 1$, the resulting penalized estimates shrink the solution to reduce variability without sparsity. Ridge regression, which is a special case of bridge regression, uses the penalty function $\lambda \mid \beta \mid^2$. It gives the estimates

$$\hat{\beta}_{ridge} = \frac{\hat{\beta}_{LSE}}{1+\gamma}, \tag{6}$$

under the condition that the design matrix is orthonormal, and γ being a positive number. One point to note is that ridge regression does not set any coefficients to 0 and therefore does not give an easily interpretable model.

The most frequently employed method is the Lasso Algorithm, which was proposed in Tibshirani (1996). Under the linear regression model (1), for a given λ, the Lasso estimator of β is

$$\hat{\beta} = \min_{\beta}\{(Y - X\beta)^T(Y - X\beta) + \lambda\|\beta\|_1\}, \tag{7}$$

where $\lambda = 0$ corresponds to the LSE $\hat{\beta}_{LSE}$ and$\|\beta\|_1 = \sum_j |\beta_j|$ is the L_1 norm of β. This L_1 penalty leads to a solution

$$\hat{\beta}_{Lasso} = sgn(\hat{\beta}_{LSE})(\mid \hat{\beta}_{LSE} \mid -\frac{\lambda}{2})^+, \tag{8}$$

for $X^TX = I$where $(\pi)^+ = \pi, \pi > 0; 0, \pi \leq 0$, and π is an arbitrage number.The Lasso does both continuous shrinkage and automatic variable selection simultaneously based on the nature of the L_1 penalty. Osbornel et al (2000) detected the conditions for the existence, uniqueness and the number of non-zero coefficients of the Lasso estimator and developed efficient algorithms for calculating the Lasso estimates and its covariance matrix. Consider the optimization problem mentioned: the objective function $f = (Y - X\beta)^T(Y - X\beta)$ is continuous and convex and the feasible region $\{\beta : \|\beta\|_1 \leq \lambda\}$ is compact, which ensures the existence of $\hat{\beta}$;the assumption $\lambda < \lambda_0$ implies any solution must lie on the boundary of the feasible region; the strict convexity leads to the uniqueness of $\hat{\beta}$.

There are some good properties of the Lasso. First, as an estimator of β, the Lasso's consistency was investigated in Knight and Fu (2000), stating that the Lasso is consistent for estimating β under appropriate conditions. In addition, as variable selection becomes increasingly important in modern data analysis, the Lasso is much more appealing because of its sparse representation. Last but not least, the entire Lasso solution paths can be computed by LARS algorithm, which was proposed by Efron et al (2004), when the design matrix X is given.

However, when the Lasso enjoys great computational advantages and excellent performances, it has three main disadvantages at the same time. First of all, the Lasso can not handle collinearity problem. When the pairwise correlations among a group of variables are very high, the Lasso tends to select only one variable from the group and ignore the rest of the variables in that group. In addition, the Lasso is not suitable for general factor selection since it can only select individual input variables. Thirdly, the Lasso lacks the oracle property stated in Fan and Li (2001).

In fact, Fan and Li (2001) defined that a good penalty function should return an estimator with three properties. The first property is unbiasedness, which means the resulting estimator has no over penalization for large parameters to avoid unnecessary modeling bias. Furthermore, sparsity is another property that an estimator enjoys. In other words, the resulting estimator automatically set insignificant parameters to 0. Last, continuity is the third property, meaning that the resulting estimator is continuous in data in order to avoid instability in model prediction.

Together with the idea of oracle property, Fan and Li (2001)proposed the Smoothly Clipped Absolute Deviation Penalty (SCAD)

$$p'_\lambda(\beta) = \lambda\{I(\beta \leq \lambda) + \frac{a\lambda - \beta}{(a-1)\lambda}I(\beta > \lambda)\}, \tag{9}$$

for some $a > 2$ and $\beta > 0$. The penalty function above is continuous and symmetric, leaving large values of the parameter λ not excessively penalized. Under the condition that the design matrix X is orthonormal, the resulting estimator is given by

$$\hat{\beta}_{SCAD} = \begin{cases} sgn(\hat{\beta}_{LSE})(\mid \hat{\beta}_{LSE} \mid -\lambda)^+, \text{when} \quad \mid \hat{\beta}_{LSE} \mid \leq 2\lambda \\ \frac{\{(a-1)\hat{\beta}_{LSE} - sgn(\hat{\beta}_{LSE})a\lambda\}}{a-2}, \text{when} \quad 2\lambda < \mid \hat{\beta}_{LSE} \mid < a\lambda \\ \hat{\beta}_{LSE}, \text{when} \quad \mid \hat{\beta}_{LSE} \mid \geq a\lambda \end{cases} \tag{10}$$

This solution actually reduces the least significant variables to zero and hence produces less complex and easier to implement models. Moreover, Fan and Li (2001) showed that the SCAD penalty can result in estimates with the oracle property. In other word, the non-zero coefficients are estimated as well as they would have been if the correct model were known in advance. In addition, when a true parameter is 0, it is estimated as 0 with probability tending to one. In terms of the two tuning parameters (λ, a), they can be searched by some criteria, such as the cross validation, the generalized cross validation, and the BIC. Fan and Li (2001) suggested that choosing $a = 3.7$ works reasonably well. Furthermore, using the language of Fan and Li (2001), we call δ an oracle procedure if $\hat{\beta}(\delta)$ has the following oracle properties:

- It can identify the right subset model, $\{j : \hat{\beta}_j \neq 0\} = \mathbf{A}$;
- It has the optimal estimation rate, $\sqrt{n}(\hat{\beta}(\delta)_{\mathbf{A}} - \beta_{\mathbf{A}}) \rightarrow_d N(0, \Sigma^*)$, where Σ^* is the covariance matrix knowing the true subset model.

It has been shown that hard thresholding and the L_q penalty functions do not satisfy the three properties defined in

Fan and Li (2001). Therefore, the SCAD actually improves these non-concave penalties in terms of the oracle properties. Fan and Li (2001) established oracle properties of the SCAD for only finite parameter cases. Fan and Peng (2004) generalized the situations to diverge number of parameters, where oracle properties can still be incorporated.

Zou (2006) proposed an improved version of the Lasso for simultaneous estimation and variable selection, called the Adaptive Lasso, where adaptive weights are used for penalizing different coefficients in the L_1 penalty. The adaptive Lasso estimators of β_js are

$$\hat{\beta}_{AdapLasso} = \min_{\beta}(Y - X\beta)^T(Y - X\beta) + \lambda \sum_{j=1}^{d} w_j|\beta_j|, \quad (11)$$

where $\mathbf{w} = \{w_1, w_2, \ldots, w_d\}$ is a known weights vector. It has been shown if the weights are data-dependent and cleverly chosen, the weighted Lasso can achieve the oracle properties, or in other words, it performs as well as if the true underlying model were known in advance. Furthermore, the adaptive Lasso solution is continuous from its definition, which makes the oracle procedure to be optimal. Finally, the Adaptive Lasso shrinkage results in a near-minimax-optimal estimator.

Zou and Hastie (2005) introduced the elastic net method, which is a regularization technique. The naive elastic net estimator can be obtained by minimizing $(Y - X\beta)^T(Y - X\beta)$ subject to

$$(1 - \alpha)\sum_{j=1}^{d} \mid \beta_j \mid + \alpha \sum_{j=1}^{d} \beta_j^2 \leq t, \quad (12)$$

where $\alpha = \frac{\lambda_2}{\lambda_1 + \lambda_2}$, $\lambda_1 and \lambda_2$ were defined in Zou and Hastie (2005). From its definition, it is obvious that the Elastic Net is a convex combination of the Lasso and the ridge regression. In fact, we have three scenarios to consider. The first case is when $\alpha = 0$. Then the naive elastic net becomes the Lasso. The second case is when $\alpha \in (0, 1)$. We need to consider a two-stage procedure for this case: for each fixed λ_2, we find the ridge regression coefficients in the first step, and then perform the Lasso in the following step. In consequence, a double amount of shrinkage happens, and it brings unnecessary additional bias compared with the pure Lasso or the ridge regression. To compromise the extra shrinkage, the naive elastic net coefficients are rescaled by a constant $(1 + \lambda_2)$. The third case is when $\alpha = 1$, and then the naive elastic net is equivalent to the ridge regression. In all, the elastic net estimator for β is given

$$\hat{\beta}_{Enet} = sgn(\hat{\beta}_{LSE})(\mid \hat{\beta}_{LSE} \mid - \frac{\lambda_1}{2})^+. \quad (13)$$

In a similar way to the Lasso, the Elastic Net does automatic variable selection and continuous shrinkage at the same time. Moreover, the Elastic Net tends to potentially select all d variables and groups of correlated predictors. This solves the collinearity problem for the Lasso. However, the Elastic Net lacks one oracle property in terms of variable selection consistency even though it has high prediction accuracy. This was pointed out and discussed in various papers(Meinshausen and Buhlmann (2006); Leng et al (2006); Zou (2006)).

Zou and Zhang (2009) pointed out that the adaptive Lasso outperforms the Lasso in terms of achieving the oracle property even though the collinearity problem for the Lasso remains. Although, as discussed in the previous paragraphs, the Elastic Net can handle the collinearity problem for the Lasso but does not have the oracle property. These two penalties advance the Lasso in two different ways. Hence, Zou and Zhang (2009) combined the adaptive lasso and elastic net and introduced a better estimator that can handle the collinearity problem while enjoying the oracle property at the same time. This improved estimator is called the adaptive elastic-net, and has the following representation:

$$\hat{\beta}_{AdapEnet} = (1 + \frac{\lambda_2}{n})\{\min_{\beta}(Y - X\beta)^T(Y - X\beta) + \lambda_1 \sum_{j=1}^{d} w_j|\beta_j| + \lambda_2\|\beta\|^2\}. \quad (14)$$

To handle high-dimensional and highly correlated data, Wold (1966) introduced partial least squares (PLS), which was among the most popular data-driven soft-sensor development methods. For batch processes, Nomikos and Macgrego (1995) proposed the multiway PLS (MPLS) that unfolds the three-way data. However, there are several drawbacks of classical MPLS method, which may degrade the prediction performance. For example, not all predictor variables are beneficial for predicting the final quality. The existence of irrelevant variables may damage useful information and degrade prediction performance. Therefore, it is important to select the relevant variables and eliminate the irrelevant variables in PLS-based modeling. To overcome the above-mentioned drawbacks in MPLS, Jiang et al (2020) proposed an optimized sparse PLS (OSPLS) modeling approach for efficient batch-end quality prediction and relevant-variable selection. The OSPLS achieved simultaneous quality prediction and relevant-variable selection by optimizing the variable resolution before SPLS modeling through a stochastic optimization approach. To make process monitoring more purposeful and more accurate, Song (2019) proposed a novel performance-indicator-oriented concurrent subspace (PIOCS) process monitoring method containing three subspaces with different degrees of importance. Moreover, Song et al (2019) proposed a novel multimode quality related process monitoring method called multi-subspace elastic network (MSEN), which is a novel clustering algorithm based on the neighborhood information and subtractive clustering algorithm.

III. SELECTION CONSISTENCY OF THE IMPROVED FR

Donoho and Stodden (2006) and Barron and Cohen (2008) both investigated the theoretical properties of Forward Regression, which is a very popular yet classical variable screening method in the literature. As one type of important

greedy algorithms, the screening consistency of Forward Regression, under an ultra-high dimensional setup, was not established by those pioneer researches. Wang (2009) investigated Forward Regression's screening consistency property under some technical conditions. Motivated by the idea of the Forward Regression in ultrahigh dimensional setup proposed by Wang (2009), we are interested in investigating the selection consistency of the improved FR, we try to impose some technical conditions on the linear regression model to derive the theoretical selection consistency property of the FR. A key point here is the stopping rule which links the residual sum of squares to the noise ratio so that the relevant predictors can be distinguished from the random noises.

A. MODEL SETUP AND TECHNICAL CONDITIONS

We now consider model (1). Without loss of generality, we assume that the data are centered, that is, the columns of X are orthonormal and Y_i's are conditionally independent given the design matrix X. Moreover, we assume $E(\epsilon_{ij}) = 0$ and $Var(\epsilon_{ij}) = \sigma^2$, i.e. the error terms ϵ are independently and identically distributed with mean zero and finite variance σ^2.

To show the selection consistency of the improved FR, four technical conditions are imposed:

Technical Conditions

(C1) Normality assumption. Assume that ε follow the normal distribution.

(C2) Eigenvalues of $\frac{1}{n}X_T^T X_T$ need to be properly bounded with λ_{min} and λ_{max}. Moreover, λ_{min} needs to be bounded away from zero. Here, X_T is the sub-matrix of X.

(C3) $\max_{j\notin T} \parallel X_T^T(X_T^T X_T)^{-1}X_T \; x_j \parallel^2 < c$ for some proper $c \in (0, 1)$.

(C4) Divergence speed of d and d_0. We assume $log(d) = O(n^{\xi})$ and $d_0 = O(n^{\xi_0})$. In other words, there exists constants ξ, ξ_{min}, and ν, such that $log(d) \leq \nu n^{\xi}$, $d_0 \leq \nu n^{\xi_0}$, and $\xi + 6\xi_0 + 12\xi_{min} < 1$.

Therefore, C1 is the normality assumption. C2 provides lower and upper bounds for the eigenvalues of $\frac{1}{n}X_T^T X_T$. C3 gives a bound for the projection of irrelevant predictors on the space spanned by the true set T. To be more specific, norm of the projection of irrelevant predictors on the space spanned by the true set T needs to be within a pre-set value. C4 allows the predictor dimension d to diverge to infinity at an exponentially fast speed, which implies that the predictor dimension can be substantially larger than the sample size n mentioned in Fan and Lv (2008).

B. IMPROVED FR ALGORITHM

In fact, under the assumption that the true model T exists, our Proposed FR algorithm, which is an improved version of Wang (2009), aims of discovering all relevant predictors consistently in a stepwise manner.

The first two steps are identical to those in Wang (2009). In fact, the only difference between our improved FR algorithm and the counterpart in Wang (2009) is that we set a stopping rule instead of screening the variables by repeating

Algorithm 1 The Improved FR Algorithm

Step 1(Initialization). Set $S^{(0)} = \emptyset$.

Step 2(Forward Regression).

- (2.1) Evaluation. In the kth step ($k \geq 1$), we are given $S^{(k-1)}$. Then, for every $j \in F \setminus S^{(k-1)}$, we construct a candidate model $M_j^{(k-1)} = S^{(k-1)} \cup \{j\}$. We then compute $RSS_j^{(k-1)} = Y^T\{I_n - \tilde{H}_j^{(k-1)}\}Y$, where $\tilde{H}_j^{(k-1)} = X_{(M_j^{(k-1)})}\{X^T_{(M_j^{(k-1)})}X_{(M_j^{(k-1)})}\}^{-1}X^T_{(M_j^{(k-1)})}$ is a projection matrix and $I_n \in R^{n\times n}$ is the identity matrix.
- (2.2) Selection. We then find

$$a_k = arg \min_{j\in F\setminus S^{(k-1)}} RSS_j^{(k-1)}$$

and update $S^{(k)} = S^{(k-1)} \cup \{a_k\}$ accordingly. In addition, we update residual sum of squares $RSS_j^{(k)} = Y^T\{I_n - H_{S^{(k)}}\}Y$, where $H_{S^{(k)}} = X_{(S^{(k)})}\{X^T_{(S^{(k)})}X_{(S^{(k)})}\}^{-1}X^T_{(S^{(k)})}$.

Step 3 (Solution Path). Iterating Step 2 until we come to the stopping rule with $RSS_j^{(k)} \leq \sigma^2(n + 2\sqrt{nlog(n)})$.

Step 2 n times. By following this stopping rule, computational time can be shortened since the algorithm stops as soon as all the relevant predictors are selected from the full model.

In fact, with probability tending to one, the improved FR algorithm can detect all relevant predictors within O $(n^{\xi_0+4\xi_{min}})$ steps. This number of steps is much smaller than the sample size n under condition (C4). In particular, if the dimension of the true model is finite with $\xi_0 = \xi_{min} = 0$, only a finite number of steps are needed to discover the entire relevant variable set.

C. MAIN RESULT

To prove Theorem 1, the following lemma is needed.

Lemma 1: Consider $S^{(k)} \subset T \subset \{1, 2, \ldots, d\}$. Let $\beta^{(S^{(k)})}$ be the parameter estimate of the coefficient β such that

$$\beta^{(S^{(k)})} = \min_{\beta\in R^d} \frac{1}{n} \parallel X\beta - y \parallel_2^2$$

subject to that we have k relevant variables. Then

$$\inf_{\alpha\in R, j\in T-S^{(k)}} \parallel X\beta + \alpha x_j - y \parallel_2^2$$
$$\leq \parallel X\beta - y \parallel_2^2 - \frac{\lambda_{min}}{\mid T - S^{(k)} \mid} \parallel X\beta^{(S^{(k)})} - X\beta \parallel_2^2 .$$

Proof of Lemma 1 For all $j \in S^{(k)}$, we have $\parallel X\beta + \alpha x_j - y \parallel_2^2$ achieves the minimum at $\alpha = 0$. This implies that$(X\beta^{(S^{(k)})} - y)x_j = 0$ for $j \in S^{(k)}$. Therefore, we have

$$(X\beta^{(S^{(k)})} - y)^T \sum_{j\in T-S^{(k)}} (\beta_j - \beta_j^{(S^{(k)})})\, x_j$$
$$= (X\beta^{(S^{(k)})} - y)^T \sum_{j\in T\cup S^{(k)}} (\beta_j - \beta_j^{(S^{(k)})})\, x_j$$

$$
\begin{aligned}
&= (X\beta^{(S^{(k)})} - y)^T (X\beta - X\beta^{(S^{(k)})}) \\
&= - \parallel X\beta^{(S^{(k)})} - X\beta \parallel_2^2 \\
&\quad + (X\beta - y)^T (X\beta - X\beta^{(S^{(k)})}) \\
&= - \parallel X\beta^{(S^{(k)})} - X\beta \parallel_2^2 .
\end{aligned}
$$

The last quality follows from the definition of β and $S^{(k)} \subset T$, which implies $(X\beta - y)^T (X\beta - X\beta^{(S^{(k)})}) = 0$. Now let $s = \mid T - S^{(k)} \mid$, then the above equality leads to the following derivation for all $\eta > 0$:

$$
\begin{aligned}
&\underset{j \in T - S^{(k)}}{s \ inf} \parallel X\beta^{(S^{(k)})} + \eta(\beta_j - \beta_j^{(S^{(k)})}) x_j - y \parallel_2^2 \\
&\leq \sum_{j \in T - S^{(k)}} \parallel X\beta^{(S^{(k)})} \\
&\quad + \eta(\beta_j - \beta_j^{(S^{(k)})}) x_j - y \parallel_2^2 \\
&= s \parallel X\beta^{(S^{(k)})} - y \parallel_2^2 \\
&\quad + \eta^2 \sum_{j \in T - S^{(k)}} (\beta_j - \beta_j^{(S^{(k)})})^2 \parallel x_j \parallel_2^2 \\
&\quad + 2\eta \, (X\beta^{(S^{(k)})} - y)^T \sum_{j \in T - S^{(k)}} (\beta_j - \beta_j^{(S^{(k)})}) x_j \\
&= s \parallel X\beta^{(S^{(k)})} - y \parallel_2^2 \\
&\quad + \eta^2 \sum_{j \in T - S^{(k)}} (\beta_j - \beta_j^{(S^{(k)})})^2 \\
&\quad - 2\eta \parallel X\beta^{(S^{(k)})} - X\beta \parallel_2^2 .
\end{aligned}
$$

Note that in the last equation, we have used $\parallel x_j \parallel_2^2 = 1$. By optimizing over η, we obtain

$$
\begin{aligned}
&\underset{j \in T - S^{(k)}}{s \ inf} \parallel X\beta^{(S^{(k)})} + \eta(\beta_j - \beta_j^{(S^{(k)})}) x_j - y \parallel_2^2 \\
&\leq s \parallel X\beta^{(S^{(k)})} - y \parallel_2^2 - \frac{\parallel X\beta^{(S^{(k)})} - X\beta \parallel_2^4}{\sum_{j \in T} (\beta_j - \beta_j^{(S^{(k)})})^2} \\
&\leq s \parallel X\beta^{(S^{(k)})} - y \parallel_2^2 \\
&\quad - \lambda_{min} \parallel X\beta^{(S^{(k)})} - X\beta \parallel_2^2 .
\end{aligned}
$$

This completes the proof for the Lemma 1.

Besides Lemma 1, the following result shown in Cai et al (2009) is useful in deriving the proof for Theorem 1. We define a bound set

$$
B_\infty(\eta) = \{\varepsilon : \parallel X^T \varepsilon \parallel_\infty \leq \sigma\sqrt{2(1+\eta) log \, d}\},
$$

where ε is the noise vector, which follows a Gaussian distribution $\varepsilon \sim N(0, \sigma^2 I_n)$ and $\eta \geq 0$. The following result, which follows from standard probability calculations, shows that the Gaussian noise is essentially bounded with

$$
P(\varepsilon \in B_\infty(\eta)) \geq 1 - \frac{1}{2d^\eta \sqrt{\pi log \, d}}.
$$

Theorem 1: Let $\mu = \max_{j \notin T} \mid X_T^T (X_T^T X_T)^{-1} X_T x_j \mid$. Suppose all the nonzero coefficients β_j satisfy

$$
\mid \beta_j \mid \geq \frac{2\sigma\sqrt{n^{-1}(1 + 2\sqrt{logn/n})}}{(1-\mu)\lambda_{min}},
$$

then the Forward Regression algorithm with the stopping rule $\parallel r_i \parallel \leq \sigma\sqrt{n + 2\sqrt{n \, logn}}$ selects a correct variable at each step until all the variables in T are selected.

In fact, since $\sigma\sqrt{n^{-1}(1 + 2\sqrt{logn/n})}$ is the noise level, if there exists a target coefficient β_j that is smaller than $O(\sigma n^{-1/2})$ in absolute value, then we can not distinguish such a small coefficient from zero or noise with large probability.

In other words, under the assumption of Theorem 1, it is possible to identify all features correctly using the improved FR algorithm as long as the target coefficients β_j are larger than $O(\sigma n^{-1/2})$.

The emphasis of Theorem 1 is to let the improved FR identify all the relevant variables before it stops, i.e. to recover exactly.

Proof of Theorem 1:

Step 1 Let $\mu(T) = \max_{j \notin T} \parallel X_T^T (X_T^T X_T)^{-1} X_T x_j \parallel_1 < 1$. This condition ensures that the algorithm chooses a relevant variable at the first step, i.e. $S^{(1)} \subset T$. By definition of $\mu(T)$, there exists $\nu = X_T^T X_T \, u \in R^{|T|}$ such that

$$
\begin{aligned}
\mu(T) &= \max_{j \notin T} \parallel X_T^T (X_T^T X_T)^{-1} X_T \, x_j \parallel_1 \\
&= \max_{j \notin T} \frac{\mid \nu^T (X_T^T X_T)^{-1} X_T \, x_j \mid}{\parallel \nu \parallel_\infty} \\
&= \max_{j \notin T} \frac{\mid u^T X_T \, x_j \mid}{\parallel (X_T^T X_T) \, u \parallel_\infty} \\
&= \frac{\max_{j \notin T} \mid x_j^T X_T \, u \mid}{\max_{i \in T} \mid x_i^T X_T \, u \mid}.
\end{aligned}
$$

Therefore, if $\mu(T) < 1$, we can find $u \in R^{|T|}$ such that

$$
\max_{j \notin T} \mid x_j^T X_T \, u \mid < \max_{i \in T} \mid x_i^T X_T \, u \mid .
$$

Consider an arbitrary $\delta_n > 0$, and β_T such that

$$
\max_{j \notin T} \mid x_j^T X_T \, \beta_T \mid < \max_{i \in T} \mid x_i^T X_T \, \beta_T \mid - 2\delta_n. \tag{15}
$$

Moreover, with probability larger than $1 - \eta$,

$$
\max_j \mid x_j^T (y - X_T \beta_T) \mid \leq \delta_n = \sigma\sqrt{2n \, ln(2d/\eta)}.
$$

Therefore, equation 15 implies

$$
\begin{aligned}
\max_{j \notin T} \mid x_j^T y \mid &\leq \max_{j \notin T} \mid x_j^T X_T \beta_T \mid \ + \max_{j \notin T} \mid x_j^T (y - X_T \beta_T) \mid \\
&< \max_{i \in T} \mid x_i^T X_T \beta_T \mid \ - \max_{i \in T} \mid x_i^T (y - X_T \beta_T) \mid \\
&\leq \max_{i \in T} \mid x_i^T y \mid .
\end{aligned}
$$

Therefore, we have proven

$$
\max_{j \notin T} \mid x_j^T y \mid < \max_{i \in T} \mid x_i^T y \mid .
$$

It guarantees that the algorithm chooses a relevant variable at the first step, i.e. $S^{(1)} \subset T$.

Step 2 We now proceed by induction on k to show that $S^{(k+1)} \subset T$ before the process stops. Assume the claim is true after k steps for $k \geq 2$. By induction hypothesis, we have $S^{(k)} \subset T$ at the end of step k. Define

$$\begin{aligned}\Omega(k) &= RSS(S^{(k)}) - RSS(S^{(k+1)}) \\ &= \frac{|\ x_j^T\{I_n - H_{(S^{(k)})}\}Y\ |^2}{\|\ x_j^{(k)}\ \|^2},\end{aligned}$$

where $H_{(S^{(k)})} = X_{(S^{(k)})}\{X^T_{(S^{(k)})}X_{(S^{(k)})}\}^{-1}X^T_{(S^{(k)})}$ is a projection matrix, $X_j^{(k)} = \{I_n - H_{(S^{(k)})}\}X_j$ and $RSS(S^{(k)}) = Y^T\{I_n - H_{(S^{(k)})}\}Y$.

Aim: $\max\limits_{j\in T}\ \Omega(k) > \max\limits_{j\notin T}\ \Omega(k)$

$$\begin{aligned}\max_{j\in T}\ \Omega(k) &= \max_{j\in T}\ \frac{|\ x_j^T\{I_n - H_{(S^{(k)})}\}Y\ |^2}{\|\ x_j^{(k)}\ \|^2} \\ &\geq \frac{\max\limits_{j\in T}\ |\ x_j^T\{I_n - H_{(S^{(k)})}\}Y\ |^2}{\max\limits_{j\in T}\ \|\ x_j^{(k)}\ \|^2} \\ &\geq \frac{\max\limits_{j\in T}\ |\ x_j^T\{I_n - H_{(S^{(k)})}\}Y\ |^2}{\max\limits_{j\in T}\ \|\ x_j\ \|^2} \\ &= \max_{j\in T}\ |\ x_j^T\{I_n - H_{(S^{(k)})}\}Y\ |^2\ .\end{aligned}$$

On the other hand,

$$\begin{aligned}\max_{j\notin T}\ \Omega(k) &= \max_{j\notin T}\ \frac{|\ x_j^T\{I_n - H_{(S^{(k)})}\}Y\ |^2}{\|\ x_j^{(k)}\ \|^2} \\ &\leq \frac{\max\limits_{j\notin T}\ |\ x_j^T\{I_n - H_{(S^{(k)})}\}Y\ |^2}{\min\limits_{j\notin T}\ \|\ x_j^{(k)}\ \|^2} \\ &\leq \frac{\max\limits_{j\notin T}\ |\ x_j^T\{I_n - H_{(S^{(k)})}\}Y\ |^2}{1-c},\end{aligned}$$

because $\max\limits_{j\notin T}\ \|\ X_T^T(X_T^TX_T)^{-1}X_T\ x_j\ \|^2 < c < 1$ implies $\min\limits_{j\notin T}\ \|\ x_j^{(k)}\ \|^2 > 1-c$. From Lemma 1, it implies

$$\begin{aligned}&\min_{\alpha,i\in T}\ \|\ X\beta^{(S^{(k)})} + \alpha x_i - y\ \|_2^2 \\ &\quad\leq \|\ X\beta^{(S^{(k)})} - y\ \|_2^2\ - \frac{\lambda_{min}}{|\ T - S^{(k)}\ |}\ \|\ X\beta^{(S^{(k)})} - X\beta\ \|_2^2\ .\end{aligned}$$

Therefore,

$$\begin{aligned}&\max_{j\in T}\ \ |\ x_j^T\{I_n - H_{(S^{(k)})}\}Y\ |^2 \\ &\quad\geq (\max_{j\in T}\ |\ (X\beta^{(S^{(k)})} - y)^Tx_j\ |)^2 \\ &\quad= \|\ X\beta^{(S^{(k)})} - y\ \|_2^2 - \min_{\alpha,j\in T}\ \|\ X\beta^{(S^{(k)})} + \alpha x_j - y\ \|_2^2 \\ &\quad\geq \frac{\lambda_{min}}{|\ T - S^{(k)}\ |}\ \|\ X\beta^{(S^{(k)})} - X\beta\ \|_2^2 \\ &\quad\geq \frac{\lambda_{min}^2}{|\ T - S^{(k)}\ |}\ \|\ \beta^{(S^{(k)})} - \beta\ \|_2^2 \\ &\quad\geq \frac{\lambda_{min}^2}{|\ T - S^{(k)}\ |}\ \|\ \beta_{T\setminus S^{(k)}}\ \|_2^2 \\ &\quad> \lambda_{min}^2\ |\ \beta_{min}\ |^2 \\ &\quad> \lambda_{min}^2\ \frac{4\sigma^2\ (n + 2\sqrt{n\ log\ n})}{(1-\mu)^2\ \lambda_{min}^2} \\ &\quad= \frac{4\sigma^2\ (n + 2\sqrt{n\ log\ n})}{(1-\mu)^2}.\end{aligned} \tag{16}$$

On the other hand,

$$\begin{aligned}&\max_{j\notin T}\ |\ (X\beta^{(S^{(k)})} - y)^Tx_j\ | \\ &\quad= \max_{j\notin T}\ |\ (X\beta^{(S^{(k)})} - X\beta + X\beta - y)^Tx_j\ | \\ &\quad\leq \max_{j\notin T}\ |\ (X\beta^{(S^{(k)})} - X\beta)^Tx_j\ |\ + \max_{j\notin T}\ |\ (X\beta - y)^Tx_j\ |\end{aligned} \tag{17}$$

Part 1 of right hand side of equation 17:

$$\begin{aligned}&\max_{j\notin T}\ |\ (X\beta^{(S^{(k)})} - X\beta)^Tx_j\ | \\ &\quad\leq \mu\ \max_{j\in T}\ |\ (X\beta^{(S^{(k)})} - X\beta)^Tx_j\ | \\ &\quad= \mu\ \max_{j\in T}\ |\ (X\beta^{(S^{(k)})} - y)^Tx_j\ | \\ &\quad\leq \mu\ \max_{j\in T}\ \ |\ x_j^T\{I_n - H_{(S^{(k)})}\}Y\ | \\ &\quad\leq \mu\ \max_{j\in T}\ \ \|\ x_j\ \|_2\|\ X\ \|_2\|\ \beta^{(S^{(k)})} - \beta\ \|_2 \\ &\quad\leq \mu\ \lambda_{max}\ \|\ \beta^{(S^{(k)})} - \beta\ \|_\infty \\ &\quad\leq \mu\ \lambda_{max}\sigma\sqrt{2log\ (2d_0/\eta_0)/\lambda_{min}},\end{aligned}$$

with probability larger than $1 - \eta_0$ for $\eta_0 \geq 0$.

Part 2 of right hand side of equation 17: $\max\limits_{j\notin T}\ |\ (X\beta - y)^Tx_j\ | \leq\ \sigma\sqrt{2(1+\eta)log\ d}$ with probability larger than $1 - \frac{1}{2d^\eta\sqrt{\pi log\ d}}$ for $\eta \geq 0$. (Please refer to Cai, Xu and Zhang (2009), see details in Appendix of the above mentioned paper.)

Hence, with probability larger than $min(1 - \frac{1}{2d^\eta\sqrt{\pi log\ d}}, 1 - \eta_0)$, we have

$$\begin{aligned}&\max_{j\notin T}\ \ |\ x_j^T\{I_n - H_{(S^{(k)})}\}Y\ |^2 \\ &\leq (\mu\ \lambda_{max}\sigma\sqrt{2log\ (2d_0/\eta_0)/\lambda_{min}}\ +\ \sigma\sqrt{2(1+\eta)log\ d})^2 \\ &< (1-c)(2\sigma\sqrt{2(1+\eta)log\ d})^2 \\ &= 4\sigma^2 2(1+\eta)\ log\ d\ (1-c).\end{aligned}$$

Now we have,

$$\begin{aligned}\max_{j\notin T}\ \Omega(k) &\leq 4\sigma^2 2(1+\eta)\ log\ d \\ &< \frac{4\sigma^2\ (n + 2\sqrt{n\ log\ n})}{(1-\mu)^2} \\ &< \max_{j\in T}\ \Omega(k),\end{aligned}$$

with probability tending to 1 as $n \to \infty$. This completes the proof for the induction part. Therefore, the algorithm selects a relevant variable at each step until the algorithm stops.

Step 3 Stopping Rule: $\sigma^2(n + 2\sqrt{n\ log\ n})$. Consider the Gaussian error $\varepsilon \sim N(0, \sigma^2 I_n)$, it satisfies

$$P(\| \varepsilon \|_2 \leq \sigma\sqrt{n + 2\sqrt{n\ log\ n}}) \geq 1 - \frac{1}{n}.$$

Suppose $X = \frac{\|\varepsilon\|_2^2}{\sigma^2}$ is a χ_n^2 random variable. Then for any $\lambda > 0$,

$$P(X > (1+\lambda)n) \leq \frac{1}{\lambda\sqrt{\pi n}}\ exp\ \{-\frac{n}{2}(\lambda - log\ (1+\lambda))\}.$$

Please refer to Cai (2002), lemma 4 for a detailed proof. Hence,

$$\begin{aligned} P(\| \varepsilon \|_2 &\leq \sigma\sqrt{n + 2\sqrt{n\ log\ n}}) \\ &= 1 - P(X > (1+\lambda)n) \\ &\geq 1 - \frac{1}{\lambda\sqrt{\pi n}}\ exp\ \{-\frac{n}{2}(\lambda - log\ (1+\lambda))\}, \end{aligned}$$

where $\lambda = 2\sqrt{n^{-1} log\ n}$. It follows from the fact that $log(1+\lambda) \leq \lambda - \frac{1}{2}\lambda^2 + \frac{1}{3}\lambda^3$. Therefore,

$$\begin{aligned} P(\| \varepsilon \|_2 &\leq \sigma\sqrt{n + 2\sqrt{n\ log\ n}}) \\ &\geq 1 - \frac{1}{n}\frac{1}{2\sqrt{\pi log\ n}}\ exp\ \{\frac{4(log\ n)^{\frac{3}{2}}}{3\sqrt{n}}\} \\ &\geq 1 - \frac{1}{n}, \end{aligned}$$

since $\frac{1}{2\sqrt{\pi log\ n}}\ exp\ \{\frac{4(log\ n)^{\frac{3}{2}}}{3\sqrt{n}}\} \leq 1$ for all $n \geq 2$.

Let $b_2 = \sigma\sqrt{n + 2\sqrt{n\ log\ n}}$. We have $| \beta_i | > \frac{2\ b_2}{(1-\mu)\lambda_{min}}$. Suppose the algorithm has run k steps for some $k < d_0 =| T |$. We will verify that $\| r_k \|_2 > b_2$ where $\| r_k \|_2$ is the square root of the residual sum of squares RSS, and so Forward Regression does not stop at the current step. Again let $X_{T\setminus S^{(k)}}$ denote the set of unselected but correct variables and $\beta_{T\setminus S^{(k)}}$ be the corresponding coefficients. Note that

$$\begin{aligned} \| r_k \|_2 &= \| (I - H_{(S^{(k)})})X\beta + (I - H_{(S^{(k)})})\varepsilon \|_2 \\ &\geq \| (I - H_{(S^{(k)})})X\beta \|_2 - \| (I - H_{(S^{(k)})})\varepsilon \|_2 \\ &\geq \| (I - H_{(S^{(k)})})X_{T\setminus S^{(k)}}\beta_{T\setminus S^{(k)}} \|_2 - \| \varepsilon \|_2 \\ &\geq \lambda_{min} \| \beta_{T\setminus S^{(k)}} \|_2 - \| \varepsilon \|_2 \\ &> \frac{2\ b_2}{1-\mu} - b_2. \\ &> b_2. \end{aligned}$$

Therefore, by all the three steps, the theorem is proved.

In all, we have also provided the theoretical proof that the improved FR is variable selection consistent under some proper conditions. In other words, by the time stopping rule is satisfied, all the relevant predictors are included in the selected model with probability tending to one. Then estimation accuracy can be improved a lot based on the reduced and correctly selected model.

IV. NUMERICAL ANALYSIS

A. SIMULATION SETUP

For reliable numerical comparison, we present the following three simulation examples on both the improved FR proposed in this paper and the FR algorithm in Wang(2009), to examine the performances of the selection consistency property of the improved FR. For each parameter setup, a total of $N = 100$ simulation replications are conducted.

Let $\hat{S}^{(k)} = \{j : \hat{\beta}_{j(k)} \neq 0\}$ be the model selected in the *kth* simulation replications and the corresponding Average Model size $= 100^{-1} \sum_k | \hat{S}^{(k)} |$. Recall T represents the true model, we evaluate the Coverage Probability $= 100^{-1} \sum_k I(\hat{S}^{(k)} \supset T)$, which measures how likely all relevant variables will be discovered by one particular method. This defined Coverage probability characterizes the screening property of a particular method.

To characterize the capability of a method in producing sparse solutions, we define

$$\begin{aligned} &Percentage\ of\ Correct\ Zeros(\%) \\ &= \frac{1}{d - d_0}\{\frac{1}{100}\sum_{k=1}^{100}\sum_{j=1}^{d} I(\hat{\beta}_{j(k)} = 0) \times I(\beta_j = 0)\} \end{aligned} \quad (18)$$

To characterize the method's underfitting effect, we further define

$$\begin{aligned} &Percentage\ of\ Incorrect\ Zeros(\%) \\ &= \frac{1}{d_0}\{\frac{1}{100}\sum_{k=1}^{100}\sum_{j=1}^{d} I(\hat{\beta}_{j(k)} = 0) \times I(\beta_j \neq 0)\} \end{aligned} \quad (19)$$

If all sparse solutions are correctly identified for all irrelevant predictors and no sparse solution is mistakenly produced for all relevant variables, the true model is perfectly identified, that is $\hat{S}^{(k)} = T$. To measure the performance, we define the Percentage of Correctly Fitted (%) $= 100^{-1} \sum_k I(\hat{S}^{(k)} = T)$, which characterizes the selection consistency property of a particular method.

As we need to know which variables are truly relevant or irrelevant, we create sparse regression vectors by setting $\beta_i = 0$ for all $i = 1, \ldots, d$, except for a chosen set T of coefficients, where β_i are defined in advance for every $1 \leq i \leq d_0$. Moreover, the noise vector $(\epsilon_1, \ldots, \epsilon_n)$ is chosen i.i.d. $N(0, 1)$. Note that all the simulation runs are conducted in Matlab.

Example 1 (Independent Predictors): This is an example borrowed from Fan and Lv (2008). X_i is generated independently according to a standard multivariate normal distribution. Thus, different predictors are mutually independent. $(n, d, d_0) = (100, 5000, 8)$ with $\beta_j = (-1)^{U_j}(4\ logn\sqrt{n} + | Z_j |)$, where U_j is a binary random variable with $P(U_j) = 0.4$ and Z_j is a standard normal random variable.

Example 2 (Autoregressive Correlation): X_i is generated from a multivariate normal distribution with mean 0 and $Cov(X_{ij_1}, X_{ij_2}) = 0.5^{|j_1 - j_2|}$. This is called an autoregressive type correlation structure. Such a correlation structure might be useful if a natural order exists among the predictors. As a

consequence, the predictors with large distances in order are expected to be approximately mutually independent. This is an example from Tibshrani (1996) with $(n, d, d_0) = (100, 5000, 3)$. In addition, the first, fourth, and seventh components of β are set to be 3, 1.5 and 2, respectively.

Example 3 (Grouped Variables): X_i is generated by the following rule. $X_{ij} = \sqrt{3/20}Z_1 + \sqrt{17/20}\epsilon_{x,j}$ for $j \in \{1, \ldots, d_0\}$, $X_{ij} = \sqrt{19/20}Z_2 + \sqrt{1/20}\epsilon_{x,j}$ for $j \in \{d_0 + 1, \ldots, d_0 + 5\}$, and $X_{ij} = \epsilon_{x,j}$ otherwise, where $Z_1 \sim N(0, 1)$, $Z_2 \sim N(0, 1)$, and $\epsilon_{x,j} \sim N(0, 1)$ are independent. This creates within-group correlations of $\rho_{ij} = 0.15$ for $i, j \in \{1, \ldots, d_0\}$ and $\rho_{ij} = 0.95$ for $i, j \in \{d_0+1, \ldots, d_0+5\}$. This example presents an interesting scenario where a group of significant variables are mildly correlated and simultaneously a group of insignificant variables are strongly correlated. The settings are similar to those in Example 2. $(n, d, d_0) = (100, 5000, 3)$. In addition, the three nonzero components of β are set to be 3, 1.5 and 2, respectively.

B. SIMULATION RESULTS OF THE IMPROVED FR SELECTION CONSISTENT PROPERTY

Sample performances of the improved FR and the traditional FR selection consistency property are investigated based on the three examples described above. Simulation results are presented in Table 1 and Table 2. In addition, the stopping rule is set to be $\| r_i \| \leq \sigma\sqrt{n + 2\sqrt{n\, logn}}$ for the improved FR.

TABLE 1. Simulation results of the improved FR with (n,d)=(100,5000).

	d_0	Coverage probability (%)	Percentage of correct zeros(%)	Percentage of incorrect zeros(%)	Percentage of correctly fitted(%)	Average model size
IP	8	100	99.9	0	97	8.04
AC	3	100	100	0	100	3
GV	3	94	97.4	1.8	61	3.25

TABLE 2. Simulation results of the FR algorithm with (n,d)=(100,5000).

	d_0	Coverage probability (%)	Percentage of correct zeros(%)	Percentage of incorrect zeros(%)	Percentage of correctly fitted(%)	Average model size
IP	8	99.9	100	0.1	97.5	8.0
AC	3	94.2	100	5.8	82.5	2.8
GV	3	94.3	100	5.7	76	3.3

First of all, the simulation results shown in Table 1 for the Independent Predictor Example demonstrate a good performance in terms of variable selection consistency for improved FR. Specifically, we have 100% Coverage Probability, which means all relevant variables can be discovered by improved FR method with the prescribed stopping rule. In addition, 97% of correctly fitted means that with the prescribed stopping rule, FR recovers all the relevant variables exactly and correctly 97 times out of 100 simulation replications. Further more, the percentages of correct and incorrect zeros are 99.9% and 0%, respectively, meaning that a few irrelevant predictors are selected into the final model. Last but not least, the average model size is 8.04, which is slightly above $d_0 = 8$.

The simulation results shown in Table 2 for the Independent Predictor Example demonstrate a good performance in terms of variable selection consistency for FR as well. Specifically, we have 99.9% Coverage Probability, which means 99.9% of the relevant variables can be discovered by the FR method. In addition, 97.5% of correctly fitted means that FR recovers all the relevant variables exactly and correctly 97.5 times out of 100 simulation replications. Further more, the percentages of correct and incorrect zeros are 100% and 0.1%, respectively, meaning that a few irrelevant predictors are selected into the final model. Last but not least, the average model size is 8.0, which is the same as $d_0 = 8$.

Second, the simulation results shown in Table 1 for Autoregressive Correlation Example demonstrate an excellent performance in terms of variable selection consistency for improved FR. Both of the Coverage Probability and the Percentage of Correctly Fitted are 100 %. Especially, 100% of correctly fitted means improved FR selects the true set of variables exactly and correctly 100 times out of 100 simulation replications. This is good news since the number of nonzero β_j d_0 is 3, which is a very sparse representation given $d = 5000$. On top of that, the percentages of correct and incorrect zeros are 100% and 0%, respectively. Last but not least, the average model size is 3. Therefore, it can be concluded that our improved FR algorithm works well under this Autoregressive Correlation setup with a sparse representation of β.

The simulation results shown in Table 2 for Autoregressive Correlation Example demonstrate good performance in terms of variable selection consistency for the FR algorithm. The Coverage Probability is 94.2% and the Percentage of Correctly Fitted is 82.5 %. Especially, 82.5% of correctly fitted means the FR selects the true set of variables exactly and correctly 82.5 times out of 100 simulation replications. On top of that, the percentages of correct and incorrect zeros are 100% and 5.8%, respectively. Last but not least, the average model size is 2.8, which is slightly smaller than $d_0 = 3$.

Third, the simulation results shown in Table 1 for Grouped Variables Example show the worst performance among all the three examples in terms of variable selection consistency for improved FR. However, the performance itself is still acceptable. Coverage Probability is 94%, meaning that not all the relevant predictors can be discovered by the improved FR algorithm with the prescribed stopping rule, in some of the simulation replications. In addition, 61% of correctly fitted means that FR selects the true set of variables correctly 61 times out of 100 simulation replications. The percentages of correct and incorrect zeros are 97.4% and 1.8%, respectively. Moreover, the average model size is 3.25.

The simulation results shown in Table 2 for Grouped Variables Example show the performance in terms of variable selection consistency for the FR. The performance is not as great as that in Example 3 but is still acceptable. Coverage Probability is 94.3%, meaning that not all the relevant predictors can be discovered by the FR algorithm, in some of the simulation replications. In addition, 76% of correctly fitted means that FR selects the true set of variables correctly 76 times out of 100 simulation replications.

The percentages of correct and incorrect zeros are 100% and 5.7%, respectively. Moreover, the average model size is 3.3, which is slightly larger than $d_0 = 3$.

In conclusion, simulation performances in terms of variable selection consistency for the improved FR are good under all the three examples. This means that our theories proposed earlier are supported. Moreover, the overall performances of the FR algorithm are good as well. Similarly, both the FR and the improved FR demonstrate good performances in Example 1 and 3. However, the improved FR performs better than the algorithm in Example 2.

As a cautionary note, we should not claim the improved FR as the only good method for variable selection. However, our extensive simulation studies do confirm that the improved FR is a very promising method, as compared with the traditional FR.

V. CONCLUDING REMARKS

In this paper, we show the theoretical result that the improved FR algorithm is variable selection consistent under some proper conditions. In particular, if the time stopping rule is satisfied, all the relevant predictors are included in the selected model with probability close to one. Then estimation accuracy can be greatly improved based on the reduced and correctly selected model. In addition, our simulation results confirm the theoretical result. The computational cost in variable selection has been reduced due to the simplified steps compared to the FR in Wang (2009).

On the other hand, we have only discussed the improved FR algorithm in the linear model setup. It is possible to investigate the improved FR in multi-linear models, such as partial linear models. Moreover, there is a normality assumption (technical condition C1) discussed in the improved FR algorithm. This normality assumption may not be proper sometimes since not all real-life dataset are normally distributed. It might be interesting to investigate the property of the improved FR under other distributional setup.

As for future research, it would be interesting to investigate the improved FR algorithm under multi-linear regression setup or in partial linear models. For the partial linear models, it is possible to introduce the penalized h-likelihood approach, which can be extended for more complicated circumstances. The model is assumed to be a simple one-component structure for the random effects, such that only a random intercept is considered. For possible future research, we may consider a Partial Linear Model for modeling the conditional mean with more than one random effects.

REFERENCES

[1] Y. Xie, "Variable selection procedures in linear regression models," Ph.D. dissertation, Stats Dept., NUS, Singapore, 2013.

[2] D. L. Donoho, "High-dimensional data analysis: The curses and blessings of dimensionality," in *Proc. AMS Conf. Math Challenges 21st Century*, 2000, p. 32.

[3] X. Sun, L. Liu, C. Geng, and S. Yang, "Fast data reduction with granulation-based instances importance labeling," *IEEE Access*, vol. 7, pp. 33587–33597, 2019.

[4] D. L. Donoho, M. Elad, and V. N. Temlyakov, "Stable recovery of sparse overcomplete representations in the presence of noise," *IEEE Trans. Inf. Theory*, vol. 52, no. 1, pp. 6–18, Jan. 2006.

[5] T. Zhang, "On the consistency of feature selection using greedy least squares regression," *J. Mach. Learn. Res.*, vol. 10, no. 3, pp. 555–568, 2009.

[6] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Stat. Soc., Ser. B (Methodol.)*, vol. 58, no. 1, pp. 267–288, Jan. 1996.

[7] J. Fan and J. Lv, "Sure independence screening for ultrahigh dimensional feature space," *J. Roy. Stat. Soc., Ser. B (Stat. Methodol.)*, vol. 70, no. 5, pp. 849–911, Nov. 2008.

[8] H. Wang, "Forward regression for ultra-high dimensional variable screening," *J. Amer. Stat. Assoc.*, vol. 104, no. 488, pp. 1512–1524, Dec. 2009.

[9] L. E. Frank and J. H. Friedman, "A statistical view of some chemometrics regression tools," *Technometrics*, vol. 35, no. 2, pp. 109–135, May 1993.

[10] M. R. Osborne, B. Presnell, and B. A. Turlach, "On the LASSO and its dual," *J. Comput. Graph. Statist.*, vol. 9, no. 2, p. 319, Jun. 2000.

[11] W. Fu and K. Knight, "Asymptotics for lasso-type estimators," *Ann. Statist.*, vol. 28, no. 5, pp. 1356–1378, Oct. 2000.

[12] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *Ann. Statist.*, vol. 32, no. 2, pp. 407–499, 2004.

[13] J. Fan and R. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties," *J. Amer. Stat. Assoc.*, vol. 96, no. 456, pp. 1348–1360, Dec. 2001.

[14] J. Fan and H. Peng, "Nonconcave penalized likelihood with a diverging number of parameters," *Ann. Statist.*, vol. 32, no. 3, pp. 928–961, Jun. 2004.

[15] H. Zou, "The adaptive lasso and its oracle properties," *J. Amer. Stat. Assoc.*, vol. 101, no. 476, pp. 1418–1429, Dec. 2006.

[16] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *J. Roy. Stat. Soc., Ser. B (Stat. Methodol.)*, vol. 67, no. 2, pp. 301–320, Apr. 2005.

[17] N. Meinshausen and P. Bühlmann, "High-dimensional graphs and variable selection with the lasso," *Ann. Statist.*, vol. 34, no. 3, pp. 1436–1462, Jun. 2006.

[18] C. Leng, Y. Lin, and G. Wahba, "A note on lasso and related procedures on model selection," *Statistica Sinica*, vol. 16, no. 4, pp. 1273–1284, 2004.

[19] H. Zou and H. H. Zhang, "On the adaptive elastic-net with a diverging number of parameters," *Ann. Statist.*, vol. 37, no. 4, pp. 1733–1751, Aug. 2009.

[20] T. Cai, G. Xu, and J. Zhang, "On recovery of sparse signals via l1 minimization," *IEEE Trans. Inf. Theory*, vol. 55, no. 7, pp. 3388–3397, Jul. 2009.

[21] Y. Zhou, H. Hang, and L. Tian, "Distributed dimensionality reconstruction algorithm for high dimensional data in Internet of brain things," *IEEE Access*, vol. 6, pp. 75789–75797, 2018.

[22] W. Zhang, Z. Xu, Y. Chen, and J. Yang, "A new algorithm for learning large Bayesian network structure from discrete data," *IEEE Access*, vol. 7, pp. 121665–121674, 2019.

[23] B. Song, H. Yan, H. Shi, and S. Tan, "Multisubspace elastic network for multimode quality-related process monitoring," *IEEE Trans. Ind. Informat.*, vol. 16, no. 9, pp. 5874–5883, Sep. 2020.

[24] B. Song, X. Zhou, H. Shi, and Y. Tao, "Performance-indicator-oriented concurrent subspace process monitoring method," *IEEE Trans. Ind. Electron.*, vol. 66, no. 7, pp. 5535–5545, Jul. 2019.

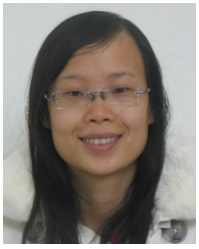
[25] Q. Jiang, X. Yan, H. Yi, and F. Gao, "Data-driven batch-end quality modeling and monitoring based on optimized sparse partial least squares," *IEEE Trans. Ind. Electron.*, vol. 67, no. 5, pp. 4098–4107, May 2020.

[26] P. Nomikos and J. F. MacGregor, "Multi-way partial least squares in monitoring batch processes," *Chemometric Intell. Lab. Syst.*, vol. 30, no. 1, pp. 97–108, Nov. 1995.

[27] A. R. Barron and A. Cohen, "Approximation and learning by greedy algorithms," *Ann. Statist.*, vol. 36, no. 1, pp. 64–94, 2008.

[28] H. Wold, "Estimation of principal components and related models by iterative least squares," in *PMultivariate Analysis*. New York, NY, USA: Academic, 1996.

[29] D. Donoho and V. Stodden, "Breakdown point of model selection when the number of variables exceeds the number of observations," in *Proc. IEEE Int. Joint Conf. Neural Netw.* Los Alamitos, CA, USA: IEEE, Jul. 2006, pp. 1916–1921.



YANXI XIE received the B.S. and Ph.D. degrees in statistics from the National University of Singapore, Singapore, in 2009 and 2013, respectively. She has been working as an Assistant Professor with the School of Management Studies, Shanghai University of Engineering Science, Shanghai, since 2014. Her research interests include high dimensional data management and dimension reduction for social media information, the credibility of social media information in emergencies.

YUEWEN LI received the B.S. and M.S. degrees in applied mathematics and information from the National University of Belarus, Russia, in 2001. He is currently working as an Associate Professor with the School of Management Studies, Shanghai University of Engineering Science. His research interests include enterprise information management, E-commerce, and computer applications.

ZHIJIE XIA received the B.S. degree in mathematics and the M.S. degree in mathematics and computer science from Anhui University, China, in 2002 and 2005, respectively, and the Ph.D. degree in economics and management from Tongji University, China, in 2008. He is currently working as a Professor with the School of Management Studies, Shanghai University of Engineering Science. His research interests include IT investment decision-making, IT governance, information system value management, emergency information management, and Web 2.0 technology applications.

RUIXIA YAN received the B.S. and M.S. degrees from Liaocheng University, China, in 2005 and 2008, respectively, and the Ph.D. degree from Donghua University, China, in 2012. She has been working as an Associate Professor with the School of Management Studies, Shanghai University of Engineering Science, since 2012. Her research interests include rough sets theory and data mining.