Research Article

An Orthogonal Matching Pursuit Variable Screening Algorithm for High-Dimensional Linear Regression Models

Yanxi Xie, Yuewen Li, Victor Shi, and Quan Lu

1 School of Management, Shanghai University of Engineering Science, Shanghai, China
2 Lazaridis School of Business and Economics, Wilfrid Laurier University, Waterloo, Canada N2L 3C5
3 School of Finance and Trade, Wenzhou Business College, Wenzhou, China

Correspondence should be addressed to Quan Lu; luq@wzbc.edu.cn

Received 22 April 2022; Accepted 22 June 2022; Published 1 August 2022

Academic Editor: Hongguang Ma

Copyright © 2022 Yanxi Xie et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Variable selection plays an important role in data mining. It is crucial to filter useful variables and extract useful information in a high-dimensional setup when the number of predictor variables $d$ tends to be much larger than the sample size $n$. Statistical inferences can be more precise after irrelevant variables are moved out by the screening method. This article proposes an orthogonal matching pursuit algorithm for variable screening under the high-dimensional setup. The proposed orthogonal matching pursuit method demonstrates good performance in variable screening. In particular, if the dimension of the true model is finite, OMP might discover all relevant predictors within a finite number of steps. Throughout theoretical analysis and simulations, it is confirmed that the orthogonal matching pursuit algorithm can identify relevant predictors to ensure screening consistency in variable selection. Given the sure screening property, the BIC criterion can be used to practically select the best candidate from the models generated by the OMP algorithm. Compared with the traditional orthogonal matching pursuit method, the resulting model can improve prediction accuracy and reduce computational cost by screening out the relevant variables.

1. Introduction

Variable screening is an important technique in data mining. It captures informative variables by reducing the dimension in a high-dimensional setup when the number of predictor variables $d$ tends to be much larger than the sample size $n$. However, statistical inference is difficult to compute in ultra-high dimensional linear models before variable screening due to the computational complexity. It is necessary to remove the irrelevant variables from the model before statistical inference. The core idea is to screening out the informative variables with the aim of building a relevant model for future prediction. 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 [1]. Then the AIC [2] or BIC [3] can be applied to further guarantee the accuracy of the relevant model.

The focus of this article is on ultra-high dimensional linear models, in which the number of predictor variables $d$ tends to be much larger than the sample size $n$. In particular, the number of covariates may increase at an exponential rate. Such linear models have gained a lot of attention in practical areas, such as sentiment analysis and finance. Existing techniques in the past literature include forward selection [1], least absolute shrinkage and selection operator (Lasso) [4], smoothly clipped absolute deviation penalty (SCAD) [5], etc. These efforts have been devoted to the challenging ultra-high dimensionality problem, which is motivated by contemporary applications such as bioinformatics, genomics, finance, etc. In other words, it is becoming a major issue to investigate the existence of complex relationships and dependencies in data with the aim of building a relevant model for inference. A practically attractive approach is to first use a quick screening procedure to reduce the dimensionality of the covariates to a
reasonable scale, for example below the sample size, and then apply variable selection techniques such as LASSO and SCAD in the second stage.

Motivated by the current studies on variable screening approaches in ultra-high dimensional linear models, it is of our interests in showing the screening consistency property of the OMP under certain conditions, by restricting the technical conditions stated in Wang [6] and hence selecting a subset of predictors, which includes all relevant predictors, to ensure variable screening results.

The rest of this article is organized as follows: Section 2 provides the literature review on current variable screening methods. Section 3 demonstrates a variable screening algorithm based on the OMP. Furthermore, the asymptotic results of the estimators are studied. Section 4 examines via simulation that our proposed technique exhibits desired sample properties and can be useful in practical applications. Finally, Section 5 concludes the article and provides some future research direction. In particular, the proof of the asymptotic theories and lemmas can be found in Appendix.

2. Literature Review

In the content of variable selection, screening approaches have gained a lot of attention besides the penalty approaches such as Lasso [4] and SCAD [5]. 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 cannot be satisfied. Therefore, to conduct model selection in the high-dimensional setup, variable screening is a reasonable solution.

Sure independence screening (SIS), which is proposed by Fan and Lv [7], has gained popularity under the condition when the number of predictor variables $d$ tends to be much larger than the sample size $n$. Sure screening means a property that all the important variables are selected after applying a variable screening procedure with probability tending to 1. It is desired to have a dimensionality reduction method with the sure screening property. There are three facts why sure screening is of great importance and usage. First of all, if dimension $d$ is larger than sample size $n$, which is clearly stated in Fan and Lv [7]. First of all, the design matrix $X$ is rectangular, having more columns than rows. In this case, the matrix $X^TX$ is giant in dimension and singular. The maximum spurious correlation between a covariate and a response can be large due to the dimensionality and the fact that an unimportant predictor can be highly correlated with the response variable owing to the presence of important predictors associated with the predictor. In addition, the population covariance matrix $\Sigma$ may become ill-conditioned as $n$ grows, and it makes variable selection difficult. Third, the minimum nonzero absolute coefficient $|\beta_j|$ may decay with $n$ and fall close to the noise level, say, the order $\log(d/n)^{−1/2}$. Hence, in general, it becomes challenging to estimate the sparse parameter vector $\beta$ accurately when $d \gg n$.

To solve the abovementioned difficulties in variable selection, Fan and Lv [7] proposed a simple sure screening method using componentwise regression or equivalently correlation learning, to reduce dimensionality from high to moderate scale that is below sample size. Below is the description of the SIS method.

Let $\omega = (\omega_1, \ldots, \omega_d)^T$ be a $d$-vector that is obtained by componentwise regression, that is

$$\omega = X^T y,$$

where the $n \times d$ data matrix $X$ is first standardized columnwise. For any given $y \in (0,1)$, we sort the $d$ componentwise magnitudes of the vector $\omega$ in a descending order and define a submodel

$$M_y = \{1 \leq i \leq d: \ |\omega_i| \text{ is among the first } \left \lfloor yn \right \rfloor \text{ largest of all}\},$$

where $\lfloor yn \rfloor$ denotes the integer part of $yn$. It shrinks the full model $\{1, 2, \ldots, d\}$ down to a submodel $M_y$ with size $\lfloor yn \rfloor$ smaller than the sample size $n$. This correlation learning ranks the importance of features according to their marginal correlations with the response variable. Moreover, it is called the independence screening because each feature is used independently as a predictor to decide the usefulness for predicting the response variable. The computational cost of SIS is of order $O(nd)$.

With dimension reduced accurately from high to below sample size, variable selection can be improved on both speed and accuracy, and can then be accomplished by a well-developed method such as SCAD, Lasso, or adaptive Lasso [8, 9], denoted by SIS-SCAD, SIS-Lasso, or SIS-AdapLasso, respectively. Moreover, sure screening property has been proven in Fan and Lv [7]. Intuitively, the core idea of SIS is to select the variables by two stages. In the first stage, an easy-to-implement method is used to remove the least important variables. In the second stage, a more sophisticated and accurate method is applied to reduce the variables further.

Though SIS enjoys sure screening property and is easy to be applied, it has several potential problems. First of all, if there is an important predictor jointly correlated but marginally uncorrelated with the response variable, it is not selected by SIS and thus cannot be included in the estimated model. Second, similar to Lasso, SIS cannot handle the collinearity problem between predictors in terms of variable selection. Third, when there are some unimportant predictors which are highly correlated with the important predictors, these unimportant predictors can have higher chance of being selected by SIS than other important predictors that are relatively weakly related to the response variable. In all, these three potential issues can be carefully treated when some extensions of SIS are proposed. In particular, iterative SIS (ISIS) is designed to overcome the weakness of SIS.

ISIS works in two steps. In the first step, a subset of $k_1$ variables $A_1 = \{X_{i_1}, \ldots, X_{i_{k_1}}\}$ is selected by using an SIS-based model selection method such as SIS-SCAD or SIS-Lasso methods. There is an $n$-vector of residuals from regressing the response $y$ over $X_{i_1}, \ldots, X_{i_{k_1}}$. In the second step, the residuals are treated as the new response variable and the previous step is repeated to the remaining $d - k_1$
variables. It returns a subset of \( k_2 \) variables \( A_2 = \{ X_{j_1}, \ldots, X_{j_{k_2}} \} \). Fitting the residuals from the previous step on \( \{ X_{j_1}, \ldots, X_{j_{k_2}} \} \) can significantly weaken the prior selection of those unimportant variables that are highly correlated with the response through their relations with \( X_{j_1}, \ldots, X_{j_{k_2}} \). In addition, the second step also makes those important variables which are missed out in the first step possible to be selected. Iteratively, the second step is iterated until \( l \) disjoint subsets \( A_1, \ldots, A_l \) are obtained with the union \( A = \bigcup_{j=1}^{l} A_j \) has a size \([yn]\).

If SIS is used to select only one variable at each iteration, that is \( |A_j| = 1 \), ISIS is equivalent to orthogonal matching pursuit (OMP) [10], which is a greedy algorithm for variable selection. This is discussed in the study by Barron and Cohen [11].

Kim [12] proposed a filter ranking method using the elastic net penalty with sure independence screening (SIS) on resampling technique to overcome the overfitting and high-performance computational issues. It is demonstrated via extensive simulation studies that SIS-LASSO, SIS-MCP, and SIS-SCAD with the proposed filtering method achieve superior performance of not only accuracy but also true positive detection compared to those with the marginal maximum likelihood ranking (MMLR) method.

Another very popular yet classical variable screening method is the forward regression (FR). As one type of important greedy algorithms, FR’s theoretical properties have been investigated in the studies by Barron and Cohen [11], Donoho and Stodden [13], and Wang [6]. In particular, Wang [6] investigated FR’s screening consistency property, under an ultra-high dimensional setup, by introducing the four technical conditions.

There are a few comments on those four technical conditions introduced in the study by Wang [6]. First of all, the normality assumption has been popularly used in the past literature for theory development. Second, the smallest and largest eigenvalues of the covariance matrix \( \Sigma \) need to be properly bounded. This bounded condition together with the normality assumption ensures the sparse Riesz condition (SRC) defined in the study by Zhang and Huang [14]. Third, the standard \( L_2 \) norm of the regression coefficients \( \beta \) is bounded above by some proper constant. It guarantees that the signal-to-noise ratio is convergent. Moreover, the minimum value of the nonzero \( \beta_j \) needs to be bounded below. This constraint on the minimal size of the nonzero regression coefficient ensures that relevant predictors can be correctly selected. Otherwise, if some of the nonzero coefficients converge too fast, they cannot be selected consistently. Last but not least, \( \log(d) \) is bounded above in the order of \( n^\xi \) for some small constant \( \xi \). This condition allows the predictor dimension \( d \) to diverge to infinity at an exponential fast speed, which implies that the predictor dimension can be substantially larger than the sample size \( n \).

Under the assumption that the true model \( T \) exists, Wang [6] introduces the FR algorithm in the aim of discovering all relevant predictors consistently. The main step of FR algorithm is the iterative forward regression part. Consider the case where \( k - 1 \) relevant predictors have been selected accordingly. Then the next step is to construct a candidate model that include one more predictor that belongs to the full set but excluding the selected \( k - 1 \) predictors and calculate the residual sum of squares based on the constructed candidate model. This step is repeated for each predictor that belongs to the full set but excluding the selected \( k - 1 \) predictors, and all the residual sum of squares are recorded accordingly. The minimum value of all the recorded residual sum of squares are found, and the \( k \)th relevant predictor is updated based on the index of the corresponding minimum residual sum of squares. A detailed algorithm can be found in the study by Wang [6].

Wang [6] showed the theoretical proof that FR can identify all relevant predictors consistently, even if the predictor dimension is considerably larger than the sample size. In particular, if the dimension of the true model is finite, FR might discover all relevant predictors within a finite number of steps. In other words, sure screening property can be guaranteed under the four technical conditions. Given the sure screening property, the recently proposed BIC of Chen and Chen [3] can be used to practically select the best candidate from the models generated by the FR algorithm. The resulting model is good in the sense that many existing variable selection methods, such as Adaptive Lasso and SCAD, can be applied directly to increase the estimation accuracy.

The extended Bayes information criterion (EBIC) proposed by Chen and Chen [3] is suitable for large model spaces. It has the following form:

\[
\text{BIC}(M) = \log \hat{\sigma}^2_{(M)} + n^{-1}|M| (\log n + 2\log d),
\]

where \( M \) is an arbitrary candidate model with \(|M| \leq n\), \( \hat{\sigma}^2_{(M)} = n^{-1}\text{RSS}(M) = n^{-1}YY'\{I_n - H_{(M)}\}Y \), and \( H_{(M)} = \sum_{(M)}^{-1}X_{(M)}'X_{(M)}\). We then select the best model \( \hat{S} = S^{(\hat{m})} \), where \( \hat{m} = \arg \min_{m \leq |M|} \text{BIC}(S^{(m)}) \).

EBIC, which includes the original BIC as a special case, examines both the number of unknown parameters and the complexity of the model space. The model in Chen and Chen [3] is defined to be identifiable if no model of comparable size other than the true submodel can predict the response almost equally well. It has been shown that EBIC is selection consistent under some mild conditions. It also handles the heavy collinearity problem for the covariates. Furthermore, EBIC is easy to implement due to the fact the extended BIC family does not require a data adaptive tuning parameter procedure.

Other screening approaches include tournament screening (TS) [15], sequential Lasso [16], quantile-adaptive model-free variable screening [17], and conditional screening [18]. When \( P \gg n \), the tournament screening possesses the sure screening property to reduce spurious correlation. Furthermore, the asymptotic properties of sequential Lasso for feature selection in linear regression models with ultra-high dimensional feature spaces are investigated. The advantage of sequential Lasso is that it is not restricted by the dimensionality of the feature space.
Quantile-adaptive model-free variable screening has two distinct features, allowing the set of active variables to vary across quantiles and overcoming the difficulty in specifying the form of a statistical model in a high-dimensional space. Baranowski [19] proposed a workflow representation for scheduling, provenance, or visualization to resolve variable and method dependencies and evaluated the performance of screening properties. Samudrala [20] proposed a parallel algorithm by identifying key components for dimensionality reduction of large-scale data. It shows better performance for dimension reduction compared to the existing methods. Chen [21] proposed a model-free feature screening method when the censored response and error-prone covariates both exist. An iterative algorithm is developed in the presence of the censored response and error-prone covariates. In addition, we also develop the iteration method to improve the accuracy of selecting all important covariates. Choudalakis [22] proposed appropriate numerical methods for parameter estimation under the high-dimensional setup. A thorough comparison is considered among existing methods for both coefficient estimations and variable selection for supersaturated designs. Xu et al. [23–25] proposed several multi-objective robust optimization models for MDVRLS in refined oil distribution. Ren [26] proposed an asymmetric learning to hash with variable bit encoding algorithm (AVBH) to solve the high-dimensional data problem, and a real data application is applied for the finite performance of the proposed AVBH algorithm. We proposed a parallel framework for dimensionality reduction of large-scale data. We also identified key components underlying the model fitting procedure produces the vector of coefficients \( \hat{\beta} = (\hat{\beta}_1, \ldots, \hat{\beta}_d)^T \).

Before the main result for the screening property of OMP is presented, four technical conditions are needed as follows:

3.1.1. Assumption 1. Technical Conditions.

(C1) Normality assumption. Assume that \( \varepsilon \) follows the normal distribution.

(C2) Covariance matrix. \( \lambda_{\min}(A) \) and \( \lambda_{\max}(A) \) represent, respectively, the smallest and largest eigenvalues of an arbitrary positive definite matrix \( A \). We assume that there exist a positive constant \( 1 < c < \infty \), such that \( c^{-1} < \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) < c \).

(C3) Regression coefficients. We assume that \( \|\beta\| \leq C_\beta \) for some constant \( C_\beta > 0 \), where \( \|\cdot\| \) denotes the standard L_2 norm. Also assume \( \beta_{\min} \geq \gamma_0 \xi_{\min}^2 \) for some \( \xi_{\min} > 0 \), with \( \beta_{\min} = \min_{j \in [T]} |\beta_j| \).

(C4) Divergence speed of \( d \) and \( d_0 \). We assume \( \log(d) = O(n^{1/2}) \) and \( d_0 = O(n^{1/2}) \). In other words, there exists constants \( \xi, \xi_{\min}, \) and \( \nu \), such that \( \log(d) \leq n^{1/2} \), \( d_0 \leq n^{1/2} \), and \( \xi + 6\xi_0 + 12\xi_{\min} < 1 \).

3.2. OMP Algorithm. Under the assumption that the true set \( T \) exists, our main objective is to discover all relevant predictors consistently. To this end, we consider the following OMP algorithm (Algorithm1):

3.3. Theoretical Properties. To prove Theorem 1, the following lemmas are needed. For convenience, we define \( \bar{\Sigma} = n^{-1}X^TX \). Moreover, \( \bar{\Sigma}^{(M)} \) and \( \Sigma^{(M)} \) are the submatrices of \( \bar{\Sigma} \) and \( \Sigma \) (corresponding to \( M \)), respectively.

Lemma 1. If \( \beta \) and \( \lambda_{\min} \leq \lambda_{\min}(\Sigma) < \lambda_{\max}(\Sigma) < c \), then \( \lambda_{\min}(\Sigma) \) are also bounded. Suppose \( \Sigma_{11} \) contains \( m \) variables, that is \( \Sigma_{11} \) is a \( m \times m \) matrix. Moreover, with probability tending to 1, we have

\[
(1/c) < \lambda_{\min}(\Sigma_{11}) < \lambda_{\max}(\Sigma_{11}) < c, \quad (5)
\]

as long as \( m = O(n^{(1-\delta)/3}) \).

The Proof of Lemma 1 can be found in Appendix.

Lemma 2. Assume conditions (C1), (C2), (C4), and \( m = O(n^{(1-\delta)/3}) \). Then, with probability tending to 1, we have

\[
\tau_{\min} \leq \min_{|M| \leq m} \lambda_{\min}(\bar{\Sigma}^{(M)}) \leq \max_{|M| \leq m} \lambda_{\min}(\bar{\Sigma}^{(M)}) \leq \tau_{\max}. \quad (6)
\]

Note that proof of Lemma 2 can be found in the study by Wang (2009), with only slight changes.

Before the theorems are established, we follow Wang (2009)’s idea on screening consistency of a solution path and define the solution path \( S \) to be screening consistent, if
Step 1 (Initialization). Set $S^{(0)} = \emptyset$. Set the residual $r_0 = y$.
Step 2 (Forward selection).
(i) (2.1) Evaluation. In the $k$th step ($k \geq 1$), we are given $S^{(k-1)}$. Then, for every $j \in F \setminus S^{(k-1)}$, we compute
$$X_T^{(k-1)} r_{k-1},$$
where $r_{k-1} = (I_n - H_{S^{(k-1)}}) y$, $H_{S^{(k-1)}} = X_{S^{(k-1)}} (X_T^{(k-1)} X_{S^{(k-1)}})^{-1} X_T^{(k-1)}$ is the projection onto the linear space spanned by the elements of $X_{S^{(k-1)}}$ and $I_n \in \mathbb{R}^{n \times n}$ is the identity matrix.
(ii) (2.2) Screening. We then find $a_k = \arg \max_{j \in F \setminus S^{(k-1)}} |X_T^{(k-1)} r_{k-1}|$, and update $S^{(k)} = S^{(k-1)} \cup \{a_k\}$ accordingly.
Step 3 (Solution path). Iterating Step 2 for $n$ times leads a total of $n$ nested candidate models. We then collect those models by a solution path $S = \{S^{(k)}: 1 \leq k \leq n\}$ with $S^{(k)} = \{a_1, \ldots, a_k\}$.

**Algorithm 1: The OMP algorithm.**

$$P(T \subset S^{(k)} \in S\text{ for some } 1 \leq k \leq n) \longrightarrow 1. \quad (7)$$

Then OMP’s screening consistency can be formally established by the following theorem.

**Theorem 1.** Under model (4) and conditions (C1)–(C4), we have as $n \longrightarrow \infty$

$$P\left(T \subset S\left(\left[Kn^2 + 4n_{\text{min}}\right]\right)\right) \longrightarrow 1,$$ \quad (8)

where the constant $K = 2 \tau_{\text{max}} \tau_{\text{min}} C_0^2 \gamma^4 \phi^4 y$ is independent of $n$, the constants $\tau_{\text{max}}$, $\tau_{\text{min}}$, $C_0$, $\gamma$, $\phi$, and $y$ are defined in conditions (C2)–(C4), and $\lfloor t \rfloor$ is the smallest integer no less than $t$.

Theorem 1 proves that within $Kn^2 + 4n_{\text{min}}$ steps, all relevant predictors will be identified by the OMP algorithm. 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 $\xi_0 = \xi_{\text{min}} = 0$, only a finite number of steps are needed to discover the entire relevant variable set.

Furthermore, Theorem 1 provides a theoretical basis for OMP, which enables us to empirically select the best model from $S$. On the other hand, the solution path $S$ contains a total of $n$ nested models. To further select relevant variables from the solution path $S$, the following BIC (Chen and Chen, 2008) is considered.

$$\text{BIC}(M) = \log \hat{\sigma}_0^2(M) + n^{-1}|M|(\log n + 2\log d),$$ \quad (9)

where $M$ is an arbitrary candidate model with $|M| \leq n$, $\hat{\sigma}_0^2(M) = n^{-1} \text{RSS}(M) = n^{-1} Y^T \left[n - H(M)\right] Y$, and $H(M) = X(M) (X_T(M) X(M))^{-1} X_T(M)$. We then select the best model $\hat{S} = S^{(\hat{m})}$, where $\hat{m} = \arg \min_{1 \leq m \leq n} \text{BIC}(S^{(m)})$. We typically do not expect $\hat{S}$ to be selection consistent (i.e., $P(\hat{S} = T) \longrightarrow 1$). However, we are able to show that $\hat{S}$ is indeed screening consistent.

**Theorem 2.** Under model (4) and conditions (C1)–(C4), then as $n \longrightarrow \infty$

$$P(T \subset \hat{S}) \longrightarrow 1.$$ \quad (10)

Define $k_{\text{min}} = \min \{k: T \subset S^{(k)}\}$. By Theorem 2.1, we know that $k_{\text{min}}$ satisfies $k_{\text{min}} \leq Kn^2 + 4n_{\text{min}}$ with probability tending to 1. Therefore, our aim is to prove that $P(\hat{m} \leq k_{\text{min}}) \longrightarrow 0$ as $n \longrightarrow \infty$. Then the theorem conclusion follows. Equivalently, it suffices to show that

$$P\left(\min_{1 \leq k \leq k_{\text{min}}} \left\{\text{BIC}(S^{(k)}) - \text{BIC}(S^{(k+1)})\right\} > 0\right) \longrightarrow 1.$$ \quad (11)

A detailed proof can be found in Appendix.

**4. Numerical Analysis**

4.1. **Simulation Setup.** For a reliable numerical comparison, the following three simulation examples on OMP algorithm are presented, to examine the finite performances of the screening consistent property of OMP. For each parameter setup, a total of $N = 100$ simulation replications are conducted.

Let $S^{(k)} = \{j: \hat{\beta}_j^{(k)} \neq 0\}$ be the model selected in the $k$th simulation replications and the corresponding average model size $= 100^{-1}\sum_{k=1}^{100}|S^{(k)}|$. Recall $T$ represents the true model, and the coverage probability $= 100^{-1}\sum_{k=1}^{100}I(S^{(k)} T)$, which measures how likely all relevant variables is discovered by the method, is evaluated. 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
Example 3. (grouped variables). This example is borrowed from Fan and Lv [7]. $X_i$ is generated independently according to a standard multivariate normal distribution. Thus, different predictors are mutually independent. $(n, d, d_0) = (100, 5000, 8)$ with $\beta_{ij} = (-1)^{i+j}/(4d_0\sqrt{n} + |Z_j|)$, where $U_j$ is a binary random variable with $P(U_j) = 0.4$ and $Z_j$ is a standard normal random variable.

Example 4. (autoregressive correlation). $X_i$ is generated from a multivariate normal distribution with mean 0 and Cov$(X_{ij}, X_{ij}') = 0.5^{1-j-i}$. This is called an autoregressive type correlation structure. Such type of 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 mutually independent approximately. This is an example from Tibshirani [4] with $(n, d, d_0) = (100, 5000, 3)$. In addition, the first, fourth, and seventh components of $\beta$ are set to be 3, 1.5, and 2, respectively.

Example 5. (grouped variables). $X_i$ is generated by the following rule. $X_{ij} = \sqrt{3/20} Z_j + \sqrt{17/20} \epsilon_{x_{ij}}$ for $j \in \{1, \ldots, d_0\}$, $X_{ij} = \sqrt{19/20} Z_j + \sqrt{17/20} \epsilon_{x_{ij}}$ for $j \in \{d_0 + 1, \ldots, d_0 + 5\}$, and $X_{ij} = \epsilon_{x_{ij}}$ otherwise, where $Z_j \sim N(0, 1)$, $Z_j \sim N(0, 1)$, and $\epsilon_{x_{ij}} \sim N(0, 1)$ 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 $\beta$ are set to be 3, 1.5, and 2, respectively.
These graphs are not representable as a whole; however, they do provide trends of BIC casewise. Take BIC of Example 1 as an example. Please refer to Figure 1(a). BIC decreases as the number of variables included in the model increases. BIC reaches a local minimum when the number of variables included hits \( d_0 = 8 \). Thereafter, BIC increases as the number of variables included in the model increases until the number of variables gets near to the sample size. This is not surprising since BIC decreases as the model complexity increases. Similar trends can be observed for Example 2 and Example 3. Please refer to Figures 1(b) and 1(c).

One possible suggestion for OMP algorithm is that after the screening process with \( n \) candidate models, only \( (n/2) \) BIC candidate models for minimum BIC values are compared. By doing so, computational time can be saved without loss of correctness of screening consistent property of OMP.

In conclusion, finite simulation performances in terms of screening selection consistency for OMP are good under all the three examples. Those performances support our theories proposed in Section 3.3.

5. Conclusion and Future Research

To conclude, this article shows the theoretical proof that OMP can identify all relevant predictors consistently, even if the predictor dimension is considerably larger than the sample size. In particular, if the dimension of the true model is finite, OMP might discover all relevant predictors within a finite number of steps. In other words, sure screening property can be guaranteed under the four technical conditions. Given the sure screening property, the recently proposed BIC of Chen and Chen (2008) can be used to practically select the best candidate from the models generated by the OMP algorithm. The resulting model is good in the sense that many existing variable selection methods, such as adaptive Lasso and SCAD, can be applied directly to increase the estimation accuracy. Compared with the traditional orthogonal matching pursuit method, the resulting model can improve prediction accuracy and reduce computational cost by screening out the relevant variables.

The abovementioned variable selection procedure only considers the fixed effect estimates in the linear models. However, in real life, a lot of existing data have both the fixed effects and random effects involved. For example, in the clinic trials, several observations are taken for a period of time for one particular patient. After collecting the data needed for all the patients, it is natural to consider random effects for each individual patient in the model setting since a common error term for all the observations is not sufficient to capture the individual randomness. Future research may include random effects in the model by imposing penalized hierarchical likelihood algorithm for accurate variable selection.

Appendix

Proof of lemmas and theorems

Proof of Lemma 1. Let \( r = (r_1, \ldots, r_d)^T \in \mathbb{R}^d \) be an arbitrary \( d \)-dimensional vector and \( r_i \) be the subvector corresponding to \( \Sigma_{i1} \). By the condition, we know immediately,
\[ \frac{1}{c} \leq \min \inf \frac{r^T}{||r||_1} \Sigma_{11} r_1 \leq \max \sup \frac{r^T}{||r||_1} \Sigma_{11} r_1 \leq c. \] (A.1)

The argument is presented in the following. Suppose that the matrix \( \Sigma \) is positive definite and has the partition as given by \( (\Sigma_{11} \: \Sigma_{12} \: \Sigma_{21} \: \Sigma_{22}) \). Then the inverse of \( \Sigma \) has the following form:

\[ \Sigma^{-1} = \begin{pmatrix} \Sigma_{11}^{-1} + \Sigma_{12}^{-1} \Sigma_{22}^{-1} & \Sigma_{12}^{-1} - \Sigma_{11}^{-1} \Sigma_{12} \Sigma_{22}^{-1} \\ -\Sigma_{21} \Sigma_{11}^{-1} & \Sigma_{22}^{-1} \end{pmatrix} \] (A.2)

where \( \Sigma_{22,1} = \Sigma_{22} - \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12} \). In fact, the above formula can be derived from the following identity:

\[ (I \: 0) \Sigma^{-1} (I \: 0)^{-1} = (I \: 0) \Sigma_{22} (I \: 0)^{-1}. \] (A.4)

Moreover, the largest singular value is referred to the operator norm of the linear operator (matrix) in a Hilbert space. If \( A \) is a \( p \times n \) matrix of complex entries, then its singular values \( s_1 \geq \ldots \geq s_q \geq 0, q = \min(p, n) \) are defined as the square roots of the \( q \) largest eigenvalues of the nonnegative definite Hermitian matrix \( AA^* \). If \( A \) (\( n \times n \)) is Hermitean, then let \( \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n \) denote its eigenvalues.

**Stage 1.** To prove \( S_{\max}(\Sigma_{11}) < c \).

Let \( T = (I \: 0) \), so \( \Sigma_{11} = (I \: 0) \Sigma_{11} (I \: 0)^{-1} = T \Sigma T^t \). Then we have

\[ S_{\max}(\Sigma_{11}) \leq S_{\max}(T) S_{\max}(\Sigma) S_{\max}(T^t) \]

\[ = S_{\max}(\Sigma) \]

\[ < c, \]

where \( S_{\max}(T) = \sqrt{\lambda_{\max}(TT^t)} \). This part has been proven in page 334 of the book Spectral Analysis of Large Dimensional Random Matrices.

**Stage 2.** To prove \( S_{\min}(\Sigma_{11}) > 1/c \).

\( \Sigma^{-1} \) is positive definite. So is the submatrix of \( \Sigma^{-1} \), that is diagonal entries. Let \( A = \Sigma_{11}, B = \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21}, C = A + B \). We have that \( C \) is positive definite. We want to prove \( A = \Sigma_{11} \) is positive definite. We have \( B \) is also positive definite because for any vector \( x \),

\[ x^T B x \]

\[ = a^T \Sigma_{22}^{-1} a. \] (A.6)

Since \( \Sigma_{22}^{-1} \) is positive definite. Therefore, for any vector \( \beta \)

\[ \beta^T A \beta \]

\[ = S_{\max}(C) \leq S_{\max}(\Sigma^{-1}). \] (A.7)

So

\[ S_{\max}(\Sigma_{11}) \leq S_{\max}(\Sigma^{-1}) \Rightarrow S_{\min}(\Sigma_{11}) > S_{\min}(\Sigma) > \frac{1}{c}. \] (A.8)

Now, the desired conclusion of Lemma 1 is implied by

\[ P \left( \max_{||r||_1=1} \left| r^T \right| \left( \Sigma_{11} - \Sigma_{12} \right) r_1 \right| > \epsilon \right) \to 0, \] (A.9)

where \( \epsilon > 0 \) is an arbitrary positive number. The left-hand side of (A.9) is bounded by

\[ \left| r^T \left( \Sigma_{11} - \Sigma_{12} \right) r_1 \right| \]

\[ \leq \max_{1 \leq j_1, j_2 \leq d} \left| \sigma_{j_1 j_2} - \sigma_{j_1 j_2} \right| \sum_{j_1, j_2 \leq M} |r_{j_1}| \times |r_{j_2}| \]

\[ = \max_{1 \leq j_1, j_2 \leq d} \left| \sigma_{j_1 j_2} - \sigma_{j_1 j_2} \right| \left( \sum_{j \in M} r_j \right)^2 \]

\[ \leq |M| \max_{1 \leq j_1, j_2 \leq d} \left| \sigma_{j_1 j_2} - \sigma_{j_1 j_2} \right| \]

\[ = m \max_{1 \leq j_1, j_2 \leq d} \left| \sigma_{j_1 j_2} - \sigma_{j_1 j_2} \right|, \]

where \( M \) is the set that contains \( m \) variables. Hence,

\[ \leq P \left( \max_{1 \leq j_1, j_2 \leq d} \left| \sigma_{j_1 j_2} - \sigma_{j_1 j_2} \right| > \frac{\epsilon}{m} \right) \]

\[ \leq \sum_{1 \leq j_1, j_2 \leq d} P \left( \left| \sigma_{j_1 j_2} - \sigma_{j_1 j_2} \right| > \frac{\epsilon}{m} \right) \]

\[ \leq d^2 C_1 \exp (-C_2 n^2 m^{-2}). \]

By lemma A3 in Bickel and Elizaveta [28], there exists constants \( C_1 > 0 \) and \( C_2 > 0 \), such that \( P (|\sigma_{j_1 j_2} - \sigma_{j_1 j_2}| > \epsilon) \leq C_1 \exp (-C_2 n^2 m^{-2}). \)

\[ = C_1 ((m + 3) \log d - C_2 n^2 m^{-2}). \] (A.12)

By C4, we have \( d \leq n^\ell \). Therefore, \( \leq C_1 \exp ((m + 3) n^\ell - C_2 n^2 m^{-2}) \)

Since \( m = O(n^{(1 - \ell)/3}) \), \( m + 3 \leq \nu_m n^{(1 - \ell)/3} \) for some constant \( \nu_m \geq 0 \).

\[ = C_1 \exp \nu_m n^{((1 - \ell)/3) + \epsilon} \left( 1 - C_2 n^2 n^{-1} n^{((1 - \ell)/3) - \ell} \right). \] (A.13)
Proof of Theorem 1

Proof. For every \( k \leq [Kn^{1/n+\varepsilon}] \), we have

\[
\Omega(k) = \frac{\left\| H_{a_{k+1}} \left[ I_n - H \left( S^{(i)} \right) \right] Y \right\|^2}{Y^T \left[ I_n - H \left( S^{(i)} \right) \right] X_{a_{k+1}}^T \left[ I_n - H \left( S^{(i)} \right) \right] Y} = \frac{Y^T \left[ I_n - H \left( S^{(i)} \right) \right] X_{a_{k+1}}^T \left[ I_n - H \left( S^{(i)} \right) \right] Y}{\left\| X_{a_{k+1}} \right\|^2}.
\]

where \( H_{k} = X_j^T X_j^T \| X_j \|^2 \) and \( X_j^T \left[ I_n - H \left( S^{(i)} \right) \right] X_j \). Note that \( a_{k+1} \) is selected to solve the maximization problem:

\[
\max_{j \in S_k^c} \left| X_j^T \left[ I_n - H \left( S^{(i)} \right) \right] X_j \right|^2 \geq \max_{j \in S_k^c} \left| X_j^T \left[ I_n - H \left( S^{(i)} \right) \right] X_j \right|^2 \geq \frac{\left\| X_{a_{k+1}} \right\|^2}{\left\| X_{a_{k+1}} \right\|^2}
\]

where \( S_k^c = T \backslash S^{(i)} \neq \emptyset \), \( \varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^T \in \mathbb{R}^n \). In what follows the two terms involved in (A.18) will be carefully evaluated separately. □

Step 1. (The first term in (A.18)). For convenience, we define \( Q \left( S^{(i)} \right) = I_n - H \left( S^{(i)} \right) \). Then,

\[
\max_{j \in S_k^c} \left| X_j^T Q \left( S^{(i)} \right) X_j \right|^2 = \max_{j \in S_k^c} \left| X_j^T Q \left( S^{(i)} \right) X_j \right|^2 \leq C_\beta \times \| X_j^T Q \left( S^{(i)} \right) X_j \| ^2 \leq C_\beta \times \| X_j^T Q \left( S^{(i)} \right) X_j \| ^2.
\]

Applying (A.20) back to (A.19), we have

\[
\max_{j \in S_k^c} \left| X_j^T \left[ I_n - H \left( S^{(i)} \right) \right] X_j \right|^2 \geq \frac{\left\| X_j^T \left[ I_n - H \left( S^{(i)} \right) \right] X_j \right\|^2}{d_0 C_\beta}.
\]

Recall that \( S_k^c = T \backslash S^{(i)} \neq \emptyset \). Then by conditions (C1)–(C3), and Lemma 2.1, we have

\[
\max_{j \in P \left( S^{(i)} \right)} X_j^T RSS \left( S^{(i)} \right) X_j = \max_{j \in P \left( S^{(i)} \right)} \left| X_j^T \left[ I_n - H \left( S^{(i)} \right) \right] Y \right|^2.
\]
\[
\left\| Q^{(S_{(i)})} X \right\| \leq \sum_{j \in S_{(i)}} \beta_j \chi_j^I Q^{(S_{(i)})} X_j \\
\geq \beta_{\min} \left\| Q^{(S_{(i)})} X \right\| \\
= \beta_{\min} \left\| X^T_{(S_{(i)})} \right\| \\
= \beta_{\min} \left\| X_{(S_{(i)})} \right\| \\
\geq n \tau_{\min} \beta_{\min}.
\]

with probability tending to 1. Applying this result back to (A.21) and also noting the technical conditions (C3) and (C4), we have

\[
\max_{j \in S_{(i)}} \left\| X^T_{(k)} \left[ I_n - H^{(S_{(i)j})} \right] X \right\| \leq \frac{\gamma^4 H^2 \mu^2 \tau_{\min}}{\nu^4 C_{\beta}} \\
= \gamma^4 \tau_{\min} \nu^{-1} C_{\beta}^{-2} n^{-2} \xi_0 - 4\xi_{\max}.
\]

(23)

Step 2. (The second term in (A.18)).

\[
\left\| X^T_{(j)} \left[ I_n - H^{(S_{(i)j})} \right] X \right\| ^2 \\
= \left\| X^T_{(j)} \left[ I_n - H^{(S_{(i)j})} \right] X \right\| ^2 \\
\leq \max_{j \in T} \max_{|M| \geq m} \chi^I_{(M)} \\
\leq \sum_{j \in T} \sum_{|M| \geq m} P \left( \chi^I_{(M)} \geq 2(m^* + 2) \log(d) \right) \\
= \frac{d^{m^* + 2}}{2(m^* + 2) \log(d)} \int_0^\infty X^{(−1/2)} e^{−x/2} dx \\
\leq \frac{\gamma^2 H^2 \mu^2 \tau_{\min}}{\nu^4 C_{\beta}} \\
= \frac{1}{\pi (m^* + 2) \log(d)} \\
\longrightarrow 0,
\]

which implies that \( \max \chi^I_{(M)} \leq 2(m^* + 2) \log(d) \) with probability tending to 1 as \( d \rightarrow \infty \). Then in conjunction with (C4), we have

\[
\max_{j \in T} \left\| X \right\| ^2 \max_{j \in T} \max_{|M| \geq m} \chi^I_{(M)} \\
\leq n \tau_{\max} \times 3K \nu^{1/4} \xi_0 + 4\xi_{\max} \\
= n \tau_{\max} \times 3K \nu^{1/4} \xi_0 + 4\xi_{\max},
\]

(27)

with probability tending to 1. Combining (A.23) and (A.27) and putting them back to (A.18), we can show

\[
n^{-1} \Omega(k) \\
= \tau_{\max} \tau_{\min} C_{\beta}^{-1} \nu^{-1} n^{-\xi_0 - 4\xi_{\max}} \\
\times \left\{ 1 - 3K \nu^2 \tau_{\max} \tau_{\min} C_{\beta}^{-2} n^{\xi_0 + 2\xi_0 + 8\xi_{\max}} \right\} \\
= \frac{2}{Kn^{\xi_0 + 4\xi_{\max}}},
\]

(28)

uniformly for every \( k \leq Kn^{\xi_0 + 4\xi_{\max}} \). Recall that \( K = 2\tau_{\max} \tau_{\min} C_{\beta}^{-1} \nu^{1/4} \). Under condition (C4), we have

\[
n^{-1} \left\| Y \right\| ^2 \\
\geq 2 \times \left\{ 1 - 3K \nu^2 \tau_{\max} \tau_{\min} C_{\beta}^{-2} n^{\xi_0 + 2\xi_0 + 8\xi_{\max}} \right\} \\
\longrightarrow 2.
\]

(29)
In contrast, under the assumption $\text{Var}(Y_i) = 1$, we have $n^{-1}\|Y\|^2 \rightarrow P_1$. This contradicts with the result of (A.29). Hence, it implies that it is impossible to have $S^{(k)} \cap T = \emptyset$ for every $1 \leq k \leq Kn^{\xi+4\xi_{\text{min}}}$. Consequently, with probability tending to 1, all relevant predictors should be recovered within a total of $Kn^{\xi+4\xi_{\text{min}}}$ steps. This completes the proof.

**Proof of Theorem 2**

Proof. It suffices to show that

$$ P(\min_{1 \leq k \leq k_{\text{num}}} \{\text{BIC}(S^{(k)}) - \text{BIC}(S^{(k+1)})\} > 0) \rightarrow 1. \quad (A.30) $$

Note that

$$ \text{BIC}(S^{(k)}) - \text{BIC}(S^{(k+1)}) $$

$$ = -n^{-1}(\log n + 2\log d) $$

$$ \geq \log \left( \frac{\hat{\sigma}^2(S^{(k)})}{\hat{\sigma}^2(S^{(k+1)})} - 3n^{-1}\log d \right) $$

$$ = \log \left( 1 + \frac{\hat{\sigma}^2(S^{(k)}) - \hat{\sigma}^2(S^{(k+1)})}{\hat{\sigma}^2(S^{(k+1)})} \right) $$

$$ - 3n^{-1}\log d, $$

under the assumption $d > n$. By definition, we have $\hat{\sigma}^2(S^{(k+1)}) \leq n^{-1}\|Y\|^2 \rightarrow P_1$. Then, with probability tending to 1, the right-hand side of (A.14) is

$$ = \log \left[ 1 + \frac{\text{RSS}(S^{(k)}) - \text{RSS}(S^{(k+1)})}{2n} \right] - 3n^{-1}\log d \quad (A.32) $$

$$ = \log \left[ 1 + \frac{\Omega(k)}{2n} \right] - 3n^{-1}\log d, $$

by the definition of $\Omega(k) = \text{RSS}(S^{(k)}) - \text{RSS}(S^{(k+1)})$. In addition, we use the fact that $\log(1 + x) \geq \min\{\log 2, 0.5x\}$ for any $x > 0$. Then the right-hand side of (A.32) is

$$ \geq \min\{\log 2, 4^{-1}K^{-1}n^{-1} - 4\xi_{\text{min}} \} - 3n^{-1}\log d, \quad (A.33) $$

according to (A.29). Moreover, the right-hand side of (A.33) is independent of $k$, hence, it is a uniform lower bound for $\text{BIC}(S^{(k)}) - \text{BIC}(S^{(k+1)})$ with $1 \leq k \leq k_{\text{num}}$. Thus, it suffices to show that the right-hand side of (A.33) is positive with probability tending to 1. To this end, we first note that

$$ \log 2 - 3n^{-1}\log d \rightarrow 0, \quad (A.34) $$

under condition (C4). Therefore, we can look at

$$ 4^{-1}K^{-1}n^{-1} - 4\xi_{\text{min}} - 3n^{-1}\log d $$

$$ = 4^{-1}K^{-1}n^{-1} - 4\xi_{\text{min}} $$

$$ \cdot \left( 1 - 12\nu Kn^{\xi+4\xi_{\text{min}}-1} \right) \quad (A.35) $$

$$ \geq 0, $$

with probability tending to 1 under condition (C4). This completes the proof. \hfill \Box

**Data Availability**

The data used to support the findings of the study are available from the corresponding author upon request.

**Conflicts of Interest**

The authors declare that they have no conflicts of interest.

**Acknowledgments**

A preprint has previously been published as a part of the Ph.D. thesis of the first author [27]: https://scholarbank.nus.edu.sg/bitstream/10635/43427/1/thesis.pdf. This work was funded by the Shanghai Planning Project of Philosophy and Social Science (Grant no. 2021EGL004). The authors gratefully acknowledge the Shanghai Planning Office of Philosophy and Social Science for the technical and financial support.

**References**

[1] Y. Xie, Y. Li, Z. Xia, and R. Yan, “An improved forward regression variable selection algorithm for high-dimensional linear regression models,” IEEE Access, vol. 8, Article ID 129032, 2020.

[2] H. Akaike, “Information theory and an extension of the maximum likelihood principle,” in Proceedings of the Second International Symposium of Information Theory, pp. 199–213, Springer, Budapest, Hungary, 1973.

[3] J. Chen and Z. Chen, “Extended Bayesian information criteria for model selection with large model spaces,” Biometrika, vol. 95, no. 3, pp. 759–771, 2008.

[4] R. Tibshirani, “Regression shrinkage and selection via the lasso,” Journal of the Royal Statistical Society: Series B, vol. 58, no. 1, pp. 267–288, 1996.

[5] J. Fan and R. Li, “Variable selection via nonconcave penalized likelihood and its oracle properties,” Journal of the American Statistical Association, vol. 96, no. 456, pp. 1348–1360, 2001.

[6] H. Wang, “Forward regression for ultra-high dimensional variable screening,” Journal of the American Statistical Association, vol. 104, no. 488, pp. 1512–1524, 2009.

[7] J. Fan and J. Lv, “Sure independence screening for ultra-high dimensional feature space,” Journal of the Royal Statistical Society: Series B, vol. 70, no. 5, pp. 849–911, 2008.
[8] H. Zou, "The adaptive lasso and its oracle properties," *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1418–1429, 2006.

[9] H. Zou and R. Li, "One-step sparse estimates in nonconcave penalized likelihood models," *Annals of Statistics*, vol. 36, no. 4, pp. 1509–1533, 2008.

[10] T. T. Cai and L. Wang, "Orthogonal matching pursuit for sparse signal recovery with noise," *IEEE Transactions on Information Theory*, vol. 57, no. 7, pp. 4680–4688, 2011.

[11] A. R. Barron, A. Cohen, W. Dahmen, and R. DeVore, "Approximation and learning by greedy algorithms," *Annals of Statistics*, vol. 36, no. 1, pp. 64–94, 2008.

[12] S. Kim and J. M. Kim, "Two-stage classification with sis using a new filter ranking method in high throughput data," *Mathematics*, vol. 7, no. 6, p. 493, 2019.

[13] D. Donoho and V. Stodden, "Breakdown point of model selection when the number of variables exceeds the number of observations, the 2006," in *Proceedings of the IEEE International Joint Conference on Neural Network Proceedings IEEE*, Vancouver, BC, Canada, July 2006.

[14] C. H. Zhang and J. Huang, "The sparsity and bias of the lasso selection in high-dimensional linear regression," *Annals of Statistics*, vol. 36, no. 4, pp. 1567–1594, 2008.

[15] Z. Chen and J. Chen, "Tournament screening cum ebic for feature selection with high-dimensional feature spaces," *Science in China(Series A:Mathematics)*, vol. 52, no. 6, pp. 1327–1341, 2009.

[16] S. Luo and Z. Chen, "Sequential lasso cum EBIC for feature selection with ultra-high dimensional feature space," *Journal of the American Statistical Association*, vol. 109, no. 507, pp. 1229–1240, 2014.

[17] F. He, L. Wang, and H. G. Hong, "Quantile-adaptive model-free variable screening for high-dimensional heterogeneous data," *Annals of Statistics*, vol. 41, no. 1, pp. 342–369, 2013.

[18] H. G. Hong, L. Wang, and X. He, "A data-driven approach to conditional screening of high-dimensional variables," *Stat*, vol. 5, no. 1, pp. 200–212, 2016.

[19] M. Baranowski, A. Belloum, M. Bubak, and M. Malawski, "Constructing workflows from script applications," *Scientific Programming*, vol. 20, no. 4, pp. 359–377, 2012.

[20] S. K. Samudrala, J. Zola, S. Aluru, and B. Ganapathysubramanian, "Parallel framework for dimensionality reduction of large-scale datasets," *Scientific Programming*, vol. 2015, Article ID 180214, 12 pages, 2015.

[21] L. P. Chen, "Feature screening based on distance correlation for ultrahigh-dimensional censored data with covariate measurement error," *Computational Statistics*, vol. 36, no. 2, pp. 857–884, 2021.

[22] S. Choudalakis, M. Mitrouli, A. Polychronou, and P. Roupa, "Solving high-dimensional problems in statistical modelling: a comparative study," *Mathematics*, vol. 9, no. 15, pp. 1806–1816, 2021.

[23] X. Xu, Z. Lin, X. Li, C. Shang, and Q. Shen, "Multi-objective robust optimisation model for mdvrpls in refined oil distribution," *International Journal of Production Research*, pp. 1–21, 2021.

[24] X. Xu, J. Hao, and Y. Zheng, "Multi-objective artificial bee colony algorithm for multi-stage resource leveling problem in sharing logistics network," *Computers & Industrial Engineering*, vol. 142, no. 2, Article ID 106338, 2020.

[25] X. Xu, Z. Lin, and J. Zhu, "DVRP with limited supply and variable neighborhood region in refined oil distribution," *Annals of Operations Research*, vol. 309, no. 2, pp. 663–687, 2022.

[26] Y. Ren, J. Qian, Y. Dong, Y. Xin, and H. Chen, "Avbh: asymmetric learning to hash with variable bit encoding," *Scientific Programming*, vol. 2020, Article ID 2424381, 11 pages, 2020.

[27] Y. Xie, *Variable Selection Procedures in Linear Regression Models*, Ph.D. dissertation, Stats Dept., National University of Singapore, Singapore, 2013.

[28] P. J. Bickel and E. Levina, "Covariance regularization by thresholding," *Annals of Statistics*, vol. 36, no. 6, pp. 2577–2604, 2008.