Correlation-based Ultrahigh-dimensional Variable Screening

Talal Ahmed and Waheed U. Bajwa
Department of Electrical and Computer Engineering
Rutgers University–New Brunswick, NJ 08854, USA
{talal.ahmed, waheed.bajwa}@rutgers.edu

***Abstract*—Statistical inference can be computationally prohibitive in ultrahigh-dimensional linear models. Correlation-based variable screening, in which one leverages marginal correlations for removal of irrelevant variables from the model prior to statistical inference, can be used to overcome this challenge. Prior works on correlation-based variable screening either impose strong statistical priors on the linear model or assume specific post-screening inference methods. This paper extends the analysis of correlation-based variable screening to arbitrary linear models and post-screening inference techniques. In particular, (i) it shows that a condition—termed the screening condition—is sufficient for successful correlation-based screening of linear models, and (ii) it provides insights into the dependence of marginal correlation-based screening on different problem parameters. Finally, numerical experiments confirm that the insights of this paper are not mere artifacts of analysis; rather, they are reflective of the challenges associated with marginal correlation-based variable screening.**

I. Introduction

Consider the ordinary linear model $y = X\beta + \text{noise}$ where the dimension, p, of β (henceforth, referred to as the number of features/predictors/variables) greatly exceeds the dimension, n, of y (henceforth, referred to as the sample size). While this high-dimensional setting should ordinarily lead to ill-posed problems, the *principle of parsimony*—which states that only a small number of variables typically affect the response y—helps obtain unique solutions to inference problems based on high-dimensional linear models. Our focus in this paper is on *ultrahigh-dimensional* linear models, in which the number of variables can scale exponentially with the sample size: $\log p = \mathcal{O}(n^\alpha)$ for $\alpha \in (0, 1)$. Such linear models are increasingly becoming common in application areas ranging from genomics [1], [2] and proteomics [3] to sentiment analysis [4], [5] and hyperspectral imaging [6], [7]. While there exist a number of techniques in the literature—such as forward selection/matching pursuit and backward elimination [8], least absolute shrinkage and selection operator (LASSO) [9], elastic net [10], smoothly clipped absolute deviation (SCAD) [11], bridge regression [12], [13], adaptive LASSO [14], group LASSO [15], and Dantzig selector [16]—that can be employed for inference from high-dimensional linear models, all these techniques have super-linear (in the number of variables p) computational complexity, and thus these methods can quickly become computationally prohibitive in the ultrahigh-dimensional setting.

This work is supported in part by the National Science Foundation under grants CCF-1453073 and CCF-1525276, and by the Army Research Office under grant W911NF-17-1-0546.

Variable selection-based dimensionality reduction, commonly referred to as *variable screening*, has been put forth as a practical means of overcoming this *curse of dimensionality* [17]: since only a small number of (independent) variables actually contribute to the response (dependent variable) in the ultrahigh-dimensional setting, one can first—in principle—discard most of the variables (the screening step) and then carry out inference on a relatively low-dimensional linear model using any one of the sparsity-promoting techniques. There are two main challenges that arise in the context of variable screening in ultrahigh-dimensional linear models. First, the screening algorithm should have low computational complexity (ideally, $\mathcal{O}(np)$). Second, the screening algorithm should be accompanied with mathematical guarantees that ensure the reduced linear model contains *all* relevant variables that affect the response. Our goal in this paper is to revisit one of the simplest screening algorithms, which uses marginal correlations between the variables $\{X_i\}_{i=1}^p$ and the response y for screening purposes [18], [19], and provide a theoretical understanding of its screening performance for arbitrary ultrahigh-dimensional linear models.

A. Relationship to Prior Work

Researchers have long intuited that the (absolute) marginal correlation $|X_i^\top y|$ is a strong indicator of whether the i-th variable contributes to the response variable. One of the earliest screening works in this regard that is agnostic to the choice of the subsequent inference techniques is termed *sure independence screening* (SIS) [20]. SIS is based on simple thresholding of marginal correlations and satisfies the so-called *sure screening* property—which guarantees that all important variables survive the screening stage with high probability—for the case of normally distributed variables. An iterative variant of SIS, termed ISIS, is also discussed in [20], while [21] presents variants of SIS and ISIS that can lead to reduced false selection rates of the screening stage. Extensions of SIS to generalized linear models are discussed in [21], [22], while its generalizations for semi-parametric (Cox) models and non-parametric models are presented in [23], [24] and [25], [26], respectively.

The defining characteristics of the works referenced above is that they are agnostic to the inference technique that follows

the screening stage. In recent years, screening methods have also been proposed for specific optimization-based inference techniques. To this end, [27] formulates a marginal correlations-based screening method, termed SAFE, for the LASSO problem and shows that SAFE results in zero false selection rate. In [28], the so-called *strong rules* for variable screening in LASSO-type problems are proposed that are still based on marginal correlations and that result in discarding of far more variables than the SAFE method. The screening tests of [27], [28] for the LASSO problem are further improved in [29]–[31] by analyzing the dual of the LASSO problem.

Notwithstanding these prior works, we have holes in our understanding of variable screening in ultrahigh-dimensional linear models. Works such as [27]–[31] necessitate the use of LASSO-type inference techniques after the screening stage. In addition, these works do not help us understand the relationship between the problem parameters and the dimensions of the reduced model. Similar to [20], [21], [32], [33], and in contrast to [27]–[31], our focus in this paper is on screening that is agnostic to the post-screening inference technique. To this end, [32] lacks a rigorous theoretical understanding of variable screening using the generalized correlation. While [20], [21], [33] overcome this shortcoming of [32], these works have two major limitations. First, their results are derived under the assumption of restrictive statistical priors on the linear model (e.g., normally distributed X_i's). In many applications, however, it can be a challenge to ascertain the distribution of the independent variables. Second, the analyses in [20], [21], [33] assume the variance of the response variable to be bounded by a constant; this assumption, in turn, imposes the condition $\|\beta\|_2 = \mathcal{O}(1)$. In contrast, defining $\beta_{\min} := \min_i |\beta_i|$, we establish in the sequel that the ratio $\frac{\beta_{\min}}{\|\beta\|_2}$ (and not $\|\beta\|_2$) directly influences the performance of marginal correlation-based screening procedures.

B. Our Contributions

Our focus in this paper is on marginal correlation-based screening of high-dimensional linear models that is agnostic to the post-screening inference technique. To this end, we provide an extended analysis of the thresholding-based SIS procedure of [20]. The resulting screening procedure, which we term *extended sure independence screening* (ExSIS), provides new insights into marginal correlation-based screening of arbitrary high-dimensional linear models. Specifically, we first provide a simple, distribution-agnostic sufficient condition—termed the *screening condition*—for (marginal correlation-based) screening of linear models. This sufficient condition, which succinctly captures joint interactions among both the active and the inactive variables, is then leveraged to explicitly characterize the performance of ExSIS as a function of various problem parameters, including noise variance, the ratio $\frac{\beta_{\min}}{\|\beta\|_2}$, and model sparsity. The numerical experiments reported at the end of this paper confirm that the dependencies highlighted in this screening result are reflective of the actual challenges associated with marginal correlation-based screening and are not mere artifacts of our analysis.

C. Notation and Organization

The following notation is used throughout this paper. Lowercase letters are used to denote scalars and vectors, while uppercase letters are used to denote matrices. Given $a \in \mathbb{R}$, $\lceil a \rceil$ denotes the smallest integer greater than or equal to a. Given $q \in \mathbb{Z}_+$, we use $[[q]]$ as a shorthand for $\{1, \ldots, q\}$. Given a vector v, $\|v\|_p$ denotes its ℓ_p norm. Given a matrix A, A_j denotes its j-th column. Further, given a set $\mathcal{I} \subset \mathbb{Z}_+$, $A_{\mathcal{I}}$ (resp., $v_{\mathcal{I}}$) denotes a submatrix (resp., subvector) obtained by retaining columns of A (resp., entries of v) corresponding to the indices in $\mathcal{I}$. Finally, the superscript $(\cdot)^\top$ denotes the transpose operation.

The rest of this paper is organized as follows. We formulate the problem of marginal correlation-based screening in Sec. II. Next, in Sec. III, we define the screening condition and present our main result that establishes the screening condition as a sufficient condition for successful variable screening. Finally, results of numerical experiments are reported in Sec. IV, while concluding remarks are presented in Sec. V.

II. Problem Formulation

Our focus in this paper is on the ultrahigh-dimensional ordinary linear model $y = X\beta + \eta$, where $y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n \times p}$ and $p \ggg n$. In the statistics literature, X is referred to as data/design/observation matrix with the rows of X corresponding to individual observations and the columns of X corresponding to individual features/predictors/variables, y is referred to as observation/response vector with individual responses given by $\{y_i\}_{i=1}^n$, β is referred to as the parameter vector, and η is referred to as modeling error or observation noise. Throughout this paper, we assume X has unit ℓ_2-norm columns, $\beta \in \mathbb{R}^p$ is k-sparse with $k < n$ (i.e., $\big|\{i \in [[p]] : \beta_i \neq 0\}\big| = k < n$), and $\eta \in \mathbb{R}^p$ is a zero-mean Gaussian vector with (entry-wise) variance σ^2 and covariance $C_\eta = \sigma^2 I$. Here, η is taken to be Gaussian with covariance $\sigma^2 I$ for the sake of this exposition, but our analysis is trivially generalizable to other noise distributions and/or covariance matrices. Further, we make no a priori assumption on the distribution of X. Finally, we define $\mathcal{S} := \{i \in [[p]] : \beta_i \neq 0\}$ to be the set that indexes the non-zero components of β. Using this notation, the linear model can equivalently be expressed as

$$y = X\beta + \eta = X_{\mathcal{S}}\beta_{\mathcal{S}} + \eta. \tag{1}$$

Given (1), the goal of variable screening is to reduce the number of variables in the linear model from p (since $p \ggg n$) to a moderate scale d ($\lll p$) using a fast and efficient method. Our focus here is on screening methods that satisfy the so-called *sure screening* property [20]; specifically, a method is said to carry out sure screening if the d-dimensional model returned by it is guaranteed with high probability to retain all the columns of X that are indexed by $\mathcal{S}$. In this paper, we study sure screening using marginal correlations between the response vector and the columns of X. The resulting screening procedure is outlined in Algorithm 1.

The term *sure independence screening* (SIS) was coined in [20] to refer to screening of ultrahigh-dimensional *Gaussian*

Algorithm 1: Marginal Correlation-based Screening

Input: $X \in \mathbb{R}^{n\times p}$, $y \in \mathbb{R}^n$, and $d \in \mathbb{Z}_+$

1: $w \leftarrow X^\top y$

2: $\widehat{\mathcal{S}}_d \leftarrow \{i \in [[p]] :$ $|w_i|$ is among the d largest of all correlations$\}$

Output: $\widehat{\mathcal{S}}_d \subset [[p]]$ such that $|\widehat{\mathcal{S}}_d| = d$

linear models using Algorithm 1. Our goal in this paper is to provide an understanding of the screening performance of Algorithm 1 for *arbitrary* (and, thus, not just Gaussian) design matrices. We use the term *extended sure independence screening* (ExSIS) to refer to screening of arbitrary linear models using Algorithm 1.

Before proceeding further, it is worth highlighting the computational savings associated with the use of Algorithm 1. To this end, we make use of the IMDb movie reviews dataset [34], with the response being either a *positive* or a *negative* review and the features extracted using the *term frequency-inverse document frequency method* [35]. The original dataset of 25K reviews is first randomly divided into five bins for five independent trials, with each bin further divided into 3K train and 2K test reviews. We then use LASSO and elastic net, both with and without screening, for training a linear data model for movie reviews. In each of the four cases, Table I summarizes both the average predictive power of the trained model in terms of the percentage of correctly classified reviews for train and test data and the average computational time needed for training the model. It can be seen from this table that Algorithm 1 reduces the training time by a factor of more than two, while there is negligible change in predictive power of the trained model. This is despite the fact that the design matrix in this problem is non-Gaussian, as verified by the Q–Q (quantile–quantile) plot (not shown here due to space limitations).

TABLE I
AVERAGE TRUE POSITIVE (TP) RATES AND COMPUTATIONAL TIMES FOR EXPERIMENTS ON THE IMDB DATASET.

Training method	Train TP rate	Test TP rate	Training time
LASSO	91.35 %	83.01 %	388.35 s
ExSIS-LASSO	98.39 %	82.23 %	177.43 s
Elastic net	96.69 %	84.35 %	272.46 s
ExSIS-Elastic net	99.71 %	82.06 %	111.20 s

III. MAIN RESULT

In this section, we derive the most general sufficient conditions for ExSIS of ultrahigh-dimensional linear models. The result reported in this section provides important insights into the workings of ExSIS *without* imposing any statistical priors on X and β. We begin with a definition of the *screening condition* for the design matrix X.

Definition 1 ((k,b)-Screening Condition)**.** *Fix an arbitrary $\beta \in \mathbb{R}^p$ that is k-sparse. The (normalized) matrix X satisfies the (k,b)-screening condition if there exists $0 < b(n,p) < \frac{1}{\sqrt{k}}$ such that the following hold:*

$$\max_{i\in\mathcal{S}} |\sum_{\substack{j\in\mathcal{S}\\ j\neq i}} X_i^\top X_j\beta_j| \leq b(n,p)\|\beta\|_2, \tag{SC-1}$$

$$\max_{i\in\mathcal{S}^c} |\sum_{j\in\mathcal{S}} X_i^\top X_j\beta_j| \leq b(n,p)\|\beta\|_2. \tag{SC-2}$$

The screening condition is a statement about the collinearity of the independent variables in the design matrix. The parameter $b(n,p)$ in the screening condition captures the similarity between (i) the columns of $X_{\mathcal{S}}$, and (ii) the columns of $X_{\mathcal{S}}$ and $X_{\mathcal{S}^c}$; the smaller the parameter $b(n,p)$ is, the less similar the columns are. Furthermore, since $k < (b(n,p))^{-2}$ in the screening condition, the parameter $b(n,p)$ also reflects constraints on the sparsity parameter k.

We now present our main screening result for arbitrary design matrices, which highlights the significance of the screening condition and the role of the parameter $b(n,p)$ within ExSIS. The proof of the following theorem is omitted in this paper for the sake of brevity, but it can be found in the journal version [36] of this work.

Theorem 1 (Sufficient Conditions for ExSIS)**.** *Let $y = X\beta + \eta$ with β a k-sparse vector and the entries of η independently distributed as $\mathcal{N}(0,\sigma^2)$. Define $\beta_{\min} := \min_{i\in\mathcal{S}} |\beta_i|$. Suppose X satisfies the screening condition and assume $\frac{\beta_{\min}}{\|\beta\|_2} > 2b(n,p) + 4\frac{\sqrt{\sigma^2\log p}}{\|\beta\|_2}$. Then, Algorithm 1 satisfies $\Pr(\mathcal{S}\subset\widehat{\mathcal{S}}_d) \geq 1-2p^{-1}$ as long as $d \geq \left\lceil \frac{\sqrt{k}}{\frac{\beta_{\min}}{\|\beta\|_2} - 2b(n,p) - \frac{4\sqrt{\sigma^2\log p}}{\|\beta\|_2}} \right\rceil$.*

A. Discussion

Theorem 1 highlights the dependence of ExSIS on the observation noise, the ratio $\frac{\beta_{\min}}{\|\beta\|_2}$, the parameter $b(n,p)$ and model sparsity. We further note from the statement of Theorem 1 that the higher the *signal-to-noise ratio* (SNR), defined here as SNR $:= \frac{\|\beta\|_2}{\sigma}$, the more Algorithm 1 can screen irrelevant/inactive variables. It is also worth noting here trivial generalizations of Theorem 1 for other noise distributions. In the case of η distributed as $\mathcal{N}(0,C_\eta)$, Theorem 1 has σ^2 replaced by the largest eigenvalue of the covariance matrix C_η. In the case of η following a non-Gaussian distribution, Theorem 1 has $2\sqrt{\sigma^2\log p}$ replaced by distribution-specific upper bound on $\|X^\top\eta\|_\infty$ that holds with high probability.

In addition to the noise distribution, the performance of ExSIS also seems to be impacted by the *minimum-to-signal ratio* (MSR), defined here as MSR $:= \frac{\beta_{\min}}{\|\beta\|_2} \in \left(0, \frac{1}{\sqrt{k}}\right]$. Specifically, the higher the MSR, the more Algorithm 1 can screen inactive variables. Stated differently, the independent variable with the weakest contribution to the response determines the size of the screened model. Finally, the parameter $b(n,p)$ in the screening condition also plays a central role in characterization of the performance of ExSIS. First, the smaller the parameter $b(n,p)$,

the more Algorithm 1 can screen inactive variables. Second, the smaller the parameter $b(n,p)$, the more independent variables can be active in the original model; indeed, we have from the screening condition that $k < (b(n,p))^{-2}$. Third, the smaller the parameter $b(n,p)$, the lower the smallest allowable value of MSR; indeed, we have from the theorem statement that MSR $> 2b(n,p) + 4\frac{\sqrt{\sigma^2 \log p}}{\|\beta\|_2}$.

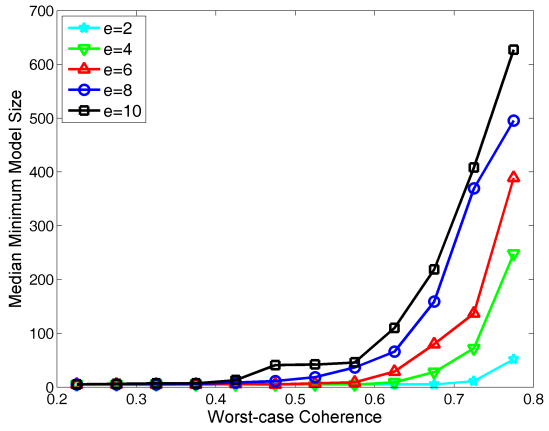
IV. Numerical Experiments

In order to ensure the insights offered by Theorem 1 are not mere artifacts of our analysis, we carry out numerical experiments to study the impact of relevant parameters on the screening performance of an *oracle* that has perfect knowledge of the minimum value of d required in Algorithm 1 to ensure $\mathcal{S} \subset \widehat{\mathcal{S}}_d$. In particular, we use these oracle-based experiments to verify the role of $b(n,p)$ and MSR in screening using Algorithm 1, as specified by Theorem 1. Before we describe our experiments, let us define the notion of worst-case coherence, μ, of X as defined in [37]: $\mu := \max_{i,j:i\neq j} \left|X_i^\top X_j\right|$. Since worst-case coherence is an indirect measure of pairwise similarity among the columns of X, we use μ as a surrogate for the value of $b(n,p)$ in our experiments.

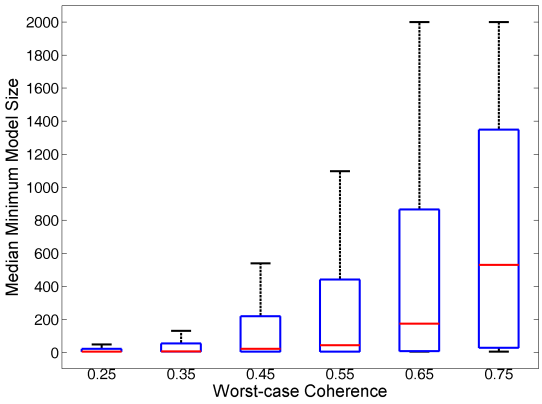
The design matrix $X \in \mathbb{R}^{n\times p}$ in our experiments is generated such that it consists of independent and identically distributed Gaussian entries, followed by normalization of the columns of X. Among other parameters, $n = 500$, $p = 2000$, $k = 5$, and $\sigma = 0$ in the experiments. The entries of $\mathcal{S}$ are chosen uniformly at random from $[[p]]$. Furthermore, the non-zero entries in the parameter vector β are sampled from a uniform distribution $U[a,e]$; the value of a is set at 1 whereas $e \in [2,10]$. Finally, the experiments comprise the use of an oracle to find the minimum possible value of d that can be used in Algorithm 1 while ensuring $\mathcal{S} \subset \widehat{\mathcal{S}}_d$. We refer to this minimum value of d as the *minimum model size* (MMS), and we use median of MMS over 400 runs of the experiment as a metric of difficulty of screening.

To analyze the impact of increasing μ (equivalently, $b(n,p)$) and MSR on screening using Algorithm 1, the numerical experiments are repeated for various values of μ and MSR. In particular, the worst-case coherence of X is varied by scaling its largest singular value, followed by normalization of the columns of X, while the MSR is increased by decreasing the value of e. In Fig. 1(a), we plot the median MMS against μ for different MSR values. The experimental results of the oracle performance offer two interesting insights. First, the median MMS increases with μ; this shows that any analysis for screening using Algorithm 1 needs to account for the similarity between the columns of X. This relationship is captured by the parameter $b(n,p)$ in Theorem 1. Second, the difficulty of screening for an oracle increases with decreasing MSR values. This relationship is also reflected in Theorem 1: as $\|\beta\|_2$ increases for a fixed e, MSR decreases and the median MMS increases.

More interestingly, if we focus on the plot in Fig. 1(a) for $b = 10$, and we plot the relationship between μ and median MMS along with the interquartile range of MMS for each value



(a)



(b)

Fig. 1. Understanding the limitations of correlation-based screening through the use of an oracle. (a) Relationship between the worst-case coherence and the MMS for various values of MSR. (b) Boxplot of the MMS versus the worst-case coherence for $e = 10$.

of μ, it can be seen that there are instances when the oracle has to select all 2000 predictors to ensure $\mathcal{S} \subset \widehat{\mathcal{S}}_d$ (see boxplot for $\mu = 0.65$ and 0.75). In other words, no screening can be performed at all in these cases. This phenomenon is also reflected in Theorem 1: when $b(n,p)$ becomes too large, the condition imposed on MSR is no longer true and our analysis cannot be used for screening using Algorithm 1.

V. Conclusion

In this paper, we provided mathematical guarantees for variable screening of arbitrary linear models using a marginal correlation-based approach, without imposing any statistical prior on the linear model. Moreover, our experiments demonstrated that the insights from the main result are reflective of the actual challenges involved with screening of arbitrary linear models using marginal correlations.

REFERENCES

[1] X. Huang and W. Pan, "Linear regression and two-class classification with gene expression data," *Bioinformatics*, vol. 19, no. 16, pp. 2072–2078, 2003.

[2] F. C. Stingo, Y. A. Chen, M. G. Tadesse, and M. Vannucci, "Incorporating biological information into linear models: A Bayesian approach to the selection of pathways and genes," *Ann. Appl. Stat.*, vol. 5, no. 3, 2011.

[3] A. I. Nesvizhskii, "A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics," *J. Proteomics*, vol. 73, no. 11, pp. 2092–2123, 2010.

[4] B. Liu, "Sentiment analysis and opinion mining," *Synthesis Lectures on Human Language Technologies*, vol. 5, no. 1, pp. 1–167, 2012.

[5] R. Feldman, "Techniques and applications for sentiment analysis," *Commun. ACM.*, vol. 56, no. 4, pp. 82–89, 2013.

[6] A. Plaza et al., "Recent advances in techniques for hyperspectral image processing," *Remote Sens. Environ.*, vol. 113, pp. S110–S122, 2009.

[7] J. Bioucas-Dias et al., "Hyperspectral unmixing overview: Geometrical, statistical, and sparse regression-based approaches," *IEEE J. Select. Topics Appl. Earth Observ. Remote Sensing*, vol. 5, no. 2, pp. 354–379, 2012.

[8] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*. Springer Texts in Statistics, 2013.

[9] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Statist. Soc. B.*, vol. 58, no. 1, pp. 267–288, 1996.

[10] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *J. Roy. Statist. Soc.*, vol. 67, no. 2, pp. 301–320, 2005.

[11] J. Fan and R. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties," *J. Am. Statist. Ass.*, vol. 96, no. 456, pp. 1348–1360, 2001.

[12] W. J. Fu, "Penalized regressions: The bridge versus the lasso," *J. Computnl Graph. Statist.*, vol. 7, no. 3, pp. 397–416, 1998.

[13] J. Huang, J. L. Horowitz, and S. Ma, "Asymptotic properties of bridge estimators in sparse high-dimensional regression models," *Ann. Stat.*, vol. 36, no. 2, pp. 587–613, 2008.

[14] H. Zou, "The adaptive lasso and its oracle properties," *J. Am. Statist. Ass.*, vol. 101, no. 476, pp. 1418–1429, 2006.

[15] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *J. Royal Statistical Soc. B*, vol. 68, no. 1, pp. 49–67, 2006.

[16] E. Candes and T. Tao, "The Dantzig selector: Statistical estimation when p is much larger than n," *Ann. Stat.*, vol. 35, no. 6, pp. 2313–2351, 2007.

[17] D. L. Donoho, "High-dimensional data analysis: The curses and blessings of dimensionality," *AMS Math Challenges Lecture*, vol. 1, p. 32, 2000.

[18] ——, "For most large underdetermined systems of linear equations the minimal ℓ_1-norm solution is also the sparsest solution," *Commun. Pure Appl. Math.*, vol. 59, no. 6, pp. 797–829, 2006.

[19] C. R. Genovese, J. Jin, L. Wasserman, and Z. Yao, "A comparison of the lasso and marginal regression," *J. Machine Learning Res.*, vol. 13, pp. 2107–2143, 2012.

[20] J. Fan and J. Lv, "Sure independence screening for ultrahigh dimensional feature space," *J. Roy. Statist. Soc. B.*, vol. 70, no. 5, pp. 849–911, 2008.

[21] J. Fan, R. Samworth, and Y. Wu, "Ultrahigh dimensional feature selection: Beyond the linear model," *J. Machine Learning Res.*, vol. 10, pp. 2013–2038, 2009.

[22] J. Fan and R. Song, "Sure independence screening in generalized linear models with NP-dimensionality," *Ann. Stat.*, vol. 38, no. 6, pp. 3567–3604, 2010.

[23] J. Fan, Y. Feng, and R. Song, "Nonparametric independence screening in sparse ultra-high dimensional additive models," *J. Am. Statist. Ass.*, vol. 106, no. 494, pp. 544–557, 2011.

[24] J. Fan, Y. Ma, and W. Dai, "Nonparametric independence screening in sparse ultra-high dimensional varying coefficient models," *J. Am. Statist. Ass.*, vol. 109, no. 507, pp. 1270–1284, 2014.

[25] J. Fan, Y. Feng, and Y. Wu, "High-dimensional variable selection for Coxs proportional hazards model," in *Borrowing Strength: Theory Powering Applications–A Festschrift for Lawrence D. Brown*. Institute of Mathematical Statistics, 2010, pp. 70–86.

[26] S. D. Zhao and Y. Li, "Principled sure independence screening for Cox models with ultra-high-dimensional covariates," *J. Multivariate Anal.*, vol. 105, no. 1, pp. 397–411, 2012.

[27] L. E. Ghaoui, V. Viallon, and T. Rabbani, "Safe feature elimination for the lasso and sparse supervised learning problems," *arXiv preprint arXiv:1009.4219*, 2010.

[28] R. Tibshirani et al., "Strong rules for discarding predictors in lasso-type problems," *J. Roy. Statist. Soc. B.*, vol. 74, no. 2, pp. 245–266, 2012.

[29] Z. J. Xiang and P. J. Ramadge, "Fast lasso screening tests based on correlations," in *Proc. IEEE Int. Conf. on Acoustics Speech and Sig. Proc. (ICASSP)*, 2012, pp. 2137–2140.

[30] L. Dai and K. Pelckmans, "An ellipsoid based, two-stage screening test for BPDN," in *Proc. 20th Eur. Sig. Proc. Conf.*, 2012, pp. 654–658.

[31] J. Wang, P. Wonka, and J. Ye, "Lasso screening rules via dual polytope projection," *J. Mach. Learn. Res.*, vol. 16, pp. 1063–1101, 2015.

[32] P. Hall and H. Miller, "Using generalized correlation to effect variable selection in very high dimensional problems," *J. Computnl Graph. Statist.*, vol. 18, no. 3, pp. 533–550, 2009.

[33] G. Li, H. Peng, J. Zhang, and L. Zhu, "Robust rank correlation based screening," *Ann. Stat.*, vol. 40, no. 3, pp. 1846–1877, 2012.

[34] A. L. Maas et al., "Learning word vectors for sentiment analysis," in *Proc. 49th Ann. Meeting of the Assoc. for Computational Linguistics: Human Language Technologies*, June 2011, pp. 142–150.

[35] C. D. Manning, P. Raghavan, and H. Schtze, *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[36] T. Ahmed and W. U. Bajwa, "ExSIS: Extended sure independence screening for ultrahigh-dimensional linear models," *arXiv preprint arXiv:1708.06077*, August 2017.

[37] G. Davis, S. Mallat, and M. Avellaneda, "Adaptive greedy approximations," *J. Construct. Approx.*, vol. 13, no. 1, pp. 57–98, 1997.