Electronic Journal of Statistics
Vol. 10 (2016) 2039–2065
ISSN: 1935-7524
DOI: 10.1214/16-EJS1157

Screening-based Bregman divergence estimation with NP-dimensionality

Chunming Zhang

Department of Statistics, University of Wisconsin, Madison, WI 53706
USA
e-mail: cmzhang@stat.wisc.edu

Xiao Guo

Department of Statistics and Finance, School of Management,
University of Science and Technology of China, Hefei, Anhui 230026
China
e-mail: xiaoguo@ustc.edu.cn

and

Yi Chai

Biogen, Cambridge MA 02142
USA
e-mail: ychai.stat@gmail.com

Abstract: Feature screening via the marginal screening ([5]; [7]) has gained special attention for high dimensional regression problems. However, their results are confined to the generalized linear model (GLM) with the exponential family of distributions. This inspires us to explore the suitability of applying screening procedures to more general models, for example without assuming either the explicit form of distributions or parametric forms between response and covariates. In this paper, we extend the marginal screening procedure, by means of Bregman divergence (BD) as the loss function, to include not only the GLM but also the quasi-likelihood model. A sure screening property for the resulting screening procedure is established under this very general framework, assuming only certain moment conditions and tail properties, where the dimensionality p_n is allowed to grow with the sample size n as fast as $\log(p_n) = O(n^a)$ for some $a \in (0, 1)$. Simulation and real data studies illustrate that a two-step procedure, which combines the feature screening in the first step and a penalized-BD estimation in the second step, is practically applicable to identifying the set of relevant variables and achieving good estimation of model parameters, with the computational cost much less than those without using the screening step.

MSC 2010 subject classifications: Primary 62F35; secondary 62F30, 62F12.

Keywords and phrases: Bregman divergence, exponential family, NP-dimensionality, sure screening, variable selection.

Received August 2014.

Contents

1. Introduction

In the recent literature, there has been a tremendous amount of work on the high dimensional regression estimation and classification. These types of studies arise frequently from many different areas of scientific research, such as fMRI brain images, microarrays, genomics, financial data, and internet traffic data. With the development of new technologies, we are now able to collect data sets which are much larger and more complex than we could have imagined a few years ago. In certain applications, we can even see that the dimensionality $p = p_n$ can grow much faster than the sample size n. Particularly, if p_n can grow at $\log(p_n) = O(n^a)$ for some $a > 0$, we call p_n the non-polynomial dimensionality or "NP-dimensionality".

For problems with NP-dimensionality, the classical regression model with p_n parameters is not identifiable. On the other hand, in many applications only a small number of variables among all p_n covariates would really have actual impact on the response variable. Thus, a sparse structure is usually assumed in such cases. As a result, those techniques that can generate sparse solutions are preferred and extensively studied. Regularization is one of the most commonly used techniques aiming at obtaining well behaved solutions to overparameterized

estimation problems. Numerous variable selection methods, based on regularization/penalization, have been developed, including the bridge regression ([8]), the Lasso ([17]), the SCAD ([6]), the MCP ([21]), and the Dantzig selector ([3]), among many others.

[5] proposed another approach, which is a screening procedure to select relevant variables based on their marginal correlations. The "sure independence screening" property was established under certain conditions in their work. [7] extended the sure independence screening procedure to the generalized linear model (GLM) ([15]). Their result works well for the GLM, but is somewhat restrictive, since their arguments largely depend on the nice properties associated with the exponential family and the canonical link function. [24] proposed a model-free feature screening approach (SIRS) using a special marginal utility measure based on the conditional distribution of the response given covariates. They showed that their approach can rank the relevant variables above irrelevant ones asymptotically in multi-index models. [14] studied another model-free feature screening method by considering the distance correlation and demonstrated the sure screening property for their method. [13] developed the robust rank correlation based screening procedure for the linear model and discussed the possibility of extending their method to the generalized linear model. In [4], an independence feature screening procedure based on the marginal empirical likelihood is studied for the generalized linear models under the exponential family distribution assumption.

Those works inspire us to explore the suitability of applying the screening procedures to more general models, for example without either the explicit form of distributions or any parametric forms between response and covariates. In this paper, we extend the marginal screening procedure, by means of the notation of Bregman divergence (BD) as the loss function ([22]), to a wider class of screening procedure, including not only ranking by marginal maximum likelihood estimate in the GLM, which has been studied by [7], but also ranking by the quasi-likelihood ([16]), which has been less developed, and a lot more. An interesting example is given in Section 5.1.1 for overdispersed Poisson responses, to which the conventional GLM is not applicable. Hence, compared with the methods in [13] and [4], our proposed method is applicable to a more general setting than the generalized linear model. Furthermore, there are a few BD loss functions widely used in machine learning systems, for example hinge loss for the support vector machine ([19]) and exponential loss for boosting ([12]), but not generally fulfilling GLM model assumptions.

The proposed method utilizes the marginal regression minimum-BD estimator of each covariate and ranks their importance according to the absolute values of the marginal estimates. The sure screening property can be established based on a non-asymptotic probability bound for the occurrences of selection inconsistency. This means that all the truly relevant variables will be selected with overwhelming probabilities and the results are applicable under NP dimensionality which allows the dimension p_n to grow as fast as $\log(p_n) = O(n^a)$ for some $a \in (0, 1)$. Our results do not require that the covariate either follows an elliptically contoured distribution as in [7] or satisfies linearity condition as in [24].

While the above screening procedure is able to identify relevant variables, it does not directly provide good estimates of the non-zero parameters in a given parametric model. It is thus natural to devise a two-step procedure, which combines the "feature screening" in the first step with the "parameter estimation" in the second step, to obtain the final model. To our knowledge, the theoretical properties of such two-step procedure, where the second step uses the penalized-BD estimation of parameters ([23]), have not been studied in the existing works on screening procedures. We carry out numerical assessment and comparison of the proposed screening method and the resulting estimation performance of several popular methods for the final model, via both simulation studies and real data analysis. From the simulation studies, the performance of our proposed screening method is better than the other model-free alternative methods like those in [24] and [14] in ultra-high dimensional settings. Our screening method based marginal regression minimum-BD estimator performs well, especially in the most stringent simulation setting in which both response and covariates are binary and very sparse. The results show that the two-step procedure is practically applicable. Another considerable advantage enjoyed by this two-step procedure is that by filtering out most of the irrelevant variables in the first step, we can greatly reduce the computational expense for parameter estimation in the second step which is usually more costly. Thus, the computation time of the two-step procedure in the simulation is just a fraction of those without using the screening step.

It is relevant to note that our main contribution in this paper is using Bregman divergence as a powerful tool to unify many commonly used loss functions and simultaneously study their asymptotic behavior under a very general framework in which the distribution of the response, conditional on the covariates, is allowed to be incompletely or not fully specified, and only certain moment conditions and tail properties are assumed. The main result of sure screening property does not require a particular parametric model, and it reveals that different choices of BD will only affect some constants in the probability bound.

The rest of the paper is organized as follows. Section 2 introduces the setup of the general regression model and briefly reviews the Bregman divergence (BD). Section 3 develops a screening procedure, based on componentwise regression minimum-BD estimation, and justifies its sure screening property. Section 4 proposes a two-step procedure combined with the penalized-BD estimation and demonstrates the oracle property of parameter estimation. Simulation results are presented in Section 5, and real data applications are given in Section 6. Details of technical assumptions and proofs are relegated to the Appendix.

2. Regression model and Bregman divergence (BD)

2.1. A general framework

Assume that the observed data $\{(\boldsymbol{X}_i, Y_i) : i = 1, \ldots, n\}$ are random samples from the population distribution of a p-dimensional covariate vector $\boldsymbol{X}$ and a

scalar response Y, where $\boldsymbol{X} = (X_1, \ldots, X_p)^T$ and $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ip})^T$. The number of variables p is allowed to grow with the sample size n, thus we denote it as p_n when needed. In this paper, we are interested in predicting the response Y by its conditional expectation given $\boldsymbol{X}$,

$$m(\boldsymbol{X}) = E(Y \mid \boldsymbol{X}). \tag{2.1}$$

While it is possible that p_n can grow much faster than n, we assume that the true underlying model is sparse, which means that $m(\boldsymbol{X})$ functionally only depends on a small fraction of the covariates, denoted by

$$\mathcal{M}_n = \{1 \le j \le p_n : m(\boldsymbol{X}) \text{ functionally depends on } X_j\}$$

with cardinality $s_n = |\mathcal{M}_n|$. Without loss of generality, we can re-arrange the covariates such that $\mathcal{M}_n = \{1, \ldots, s_n\}$. Write $\boldsymbol{X} = (\boldsymbol{X}^{(\mathrm{I})T}, \boldsymbol{X}^{(\mathrm{II})T})^T$, where $\boldsymbol{X}^{(\mathrm{I})}$ collects all truly relevant covariates and $\boldsymbol{X}^{(\mathrm{II})}$ is just noise. Under this framework, our goal is to investigate the performance of a wide class of feature screening methods, by means of Bregman divergence that will be introduced in the next section.

2.2. Bregman divergence

[2] introduced a device for constructing a bivariate function which can be used as a general loss function. For a given concave function q, define the Bregman divergence as

$$Q(\nu, \mu) = -q(\nu) + q(\mu) + (\nu - \mu)q'(\mu). \tag{2.2}$$

Conversely, for a given Q-loss, [22] provided necessary and sufficient conditions for Q being a BD, and further derived an explicit formula for solving the generating q-function. They also showed that the quadratic function, the Kullback-Leibler divergence (or the deviance loss) for the exponential family of probability functions, the (negative) quasi-likelihood function, and many margin-based loss functions, such as the misclassification loss, the hinge loss for the support vector machine ([19]), and the exponential loss used in AdaBoost ([12]), are all special cases of BD.

As an illustration, when we relax the distributional assumption on the response Y by only assuming $\mathrm{var}(Y \mid \boldsymbol{X} = \boldsymbol{x}) = \sigma^2 V\{m(\boldsymbol{x})\}$ for a known continuous function $V(\cdot) > 0$, the quasi-likelihood function Q, given by the partial differential equation

$$\partial Q(Y, \mu)/\partial \mu = (Y - \mu)/V(\mu),$$

for a nuisance parameter $\sigma^2 > 0$, is usually used as an alternative of complete log-likelihood function. [22] verified that the (negative) quasi-likelihood function belongs to the BD and derived the generating q-function, given by

$$q(\mu) = \int_a^\mu \frac{s - \mu}{V(s)} ds, \tag{2.3}$$

where a is a finite constant such that the integral is well-defined.

3. Screening via componentwise regression minimum-BD estimation

Our proposed screening procedure is based on the componentwise regression minimum-BD estimators, defined as

$$\widehat{m}_j(\cdot) = \arg\min_m \frac{1}{n}\sum_{i=1}^{n} Q\{Y_i, m(X_{ij})\}, \text{ for } j = 1, \ldots, p_n, \tag{3.1}$$

where the loss function Q is a BD as defined in (2.2) with a generating q-function. Furthermore, we restrict $m(x)$ in (3.1) to be of the form,

$$F^{-1}(\alpha_j + x\beta_j), \tag{3.2}$$

where α_j and β_j are two parameters to be estimated, and F is a known link function for appropriate data type. Usually, an identity link $F(\mu) = \mu$ corresponds to the linear regression model for continuous responses; a logit link $F(\mu) = \log(\frac{\mu}{1-\mu})$ is utilized in the logistic regression for binary responses; a log link $F(\mu) = \log(\mu)$ is used in the Poisson regression of count responses.

The functional form in (3.2) is a linear approximation to the problem in (3.1) which appears somewhat restrictive, however our later theoretical results show that such class of functions is actually rich enough to detect the marginal importance of covariates for the screening purpose.

Thus, the minimization problem in (3.1) is equivalent to estimating $(\widehat{\alpha}_j^{\mathrm{CR}}, \widehat{\beta}_j^{\mathrm{CR}})$, for $j = 1, \ldots, p_n$, which are defined as

$$(\widehat{\alpha}_j^{\mathrm{CR}}, \widehat{\beta}_j^{\mathrm{CR}}) = \arg\min_{(\alpha_j, \beta_j) \in \mathbb{R}^2} \frac{1}{n}\sum_{i=1}^{n} Q\{Y_i, F^{-1}(\alpha_j + X_{ij}\beta_j)\}. \tag{3.3}$$

We select the variables by choosing those whose componentwise coefficient estimators $|\widehat{\beta}_j^{\mathrm{CR}}|$ exceed a predefined threshold value $\gamma_n > 0$, i.e. variables X_j with indices j belonging to the set

$$\widehat{\mathcal{M}}_{\gamma_n} = \{1 \le j \le p_n : |\widehat{\beta}_j^{\mathrm{CR}}| \ge \gamma_n\}$$

will be selected; the remaining variables will be screened out.

The minimization problem (3.3) only involves a univariate covariate and an intercept. Thus fast and robust computation would be feasible even in NP-dimensional problems. When an appropriate γ_n is chosen, we can significantly reduce the dimension of the original parameter space to a much smaller one and thus make it more manageable. After the screening step, other variable selection methods, like those (mentioned in Section 1) based on penalization, would be more feasible on survived variables.

In our screening procedure, the magnitude of the componentwise regression coefficient estimator, $\widehat{\beta}_j^{\mathrm{CR}}$, serves as a proxy for the importance of the corresponding feature X_j. Two questions arise naturally:

(I) how well the set $\widehat{\mathcal{M}}_{\gamma_n}$ preserves all relevant covariates, given that the estimators $\widehat{\beta}_j^{\mathrm{CR}}$ from the componentwise minimization problem (3.3) only approximate the importance of covariates in the original model (2.1);
(II) how small the size of $\widehat{\mathcal{M}}_{\gamma_n}$ can be, given that $\widehat{\mathcal{M}}_{\gamma_n}$ should still include all truly relevant variables.

We will answer these two questions, by showing that the sure screening property holds under certain conditions, in the following sections.

*3.1. Population version of componentwise regression minimum-*BD *estimator*

Recall that the estimators in (3.3) are based on empirical minimization. To gain further insights, we define a population analogue of (3.3), denoted by

$$(\alpha_j^{\mathrm{CR}}, \beta_j^{\mathrm{CR}}) = \underset{(\alpha_j,\beta_j)\in\mathbb{R}^2}{\arg\min}\, E[Q\{Y, F^{-1}(\alpha_j + X_j\beta_j)\}], \tag{3.4}$$

where the expectation is taken with respect to the underlying joint distribution of $(\boldsymbol{X}, Y)$.

Note that the componentwise minimum-BD estimator $\widehat{\beta}_j^{\mathrm{CR}}$ will converge in probability to the population version β_j^{CR}. To guarantee the validity of the screening procedure, it is necessary that β_j^{CR} should at least preserve the significance of truly relevant covariates, i.e. those whom $m(\boldsymbol{X})$ functionally depends on. Theorem 1 confirms that the significance of β_j^{CR} (i.e. $\beta_j^{\mathrm{CR}} \neq 0$) only depends on the correlation between Y and X_j.

Theorem 1. *Assume Conditions* A1–A4 *in the Appendix. For any* $j = 1, \ldots, p_n$, *it follows that* $\beta_j^{\mathrm{CR}} = 0$ *if and only if* $\mathrm{cov}(Y, X_j) = \mathrm{cov}\{m(\boldsymbol{X}), X_j\} = 0$.

Theorem 1 implies that if the response variable Y is correlated with a relevant variable X_j, then the componentwise regression coefficient β_j^{CR} will be non-zero. In contrast, for those irrelevant variables X_j which are uncorrelated with Y, β_j^{CR} will be zero. Theorem 2 further indicates that the magnitude of β_j^{CR} is also closely related to the magnitude of correlation between X_j and Y.

Theorem 2. *Assume Conditions* A1–A5 *in the Appendix. For any positive sequences* $\mathcal{A}_n$ *and* $\mathcal{B}_n$,

(i) *if* $\min_{j=1,\ldots,s_n} |\mathrm{cov}(Y, X_j)| \geq \mathcal{A}_n$, *then there exists a positive constant* c_1 *such that*

$$\min_{j=1,\ldots,s_n} |\beta_j^{\mathrm{CR}}| \geq c_1 \mathcal{A}_n;$$

(ii) *if* $\max_{j=s_n+1,\ldots,p_n} |\mathrm{cov}(Y, X_j)| = O(\mathcal{B}_n)$, *then*

$$\max_{j=s_n+1,\ldots,p_n} |\beta_j^{\mathrm{CR}}| = O(\mathcal{B}_n).$$

The conditions used in Theorem 2 are typically regarded as mild and are often assumed in the literature ([22]; [7]). Assumptions A1 and A2 are related to the tail behavior of the population distribution. Assumptions A3 and A4 are about the convexity and smoothness of BD. The requirement of covariance between Y and X_j's for $j \in \mathcal{M}_n$ is to ensure that the minimal signal strength of relevant variables should not be too weak and still identifiable.

If those conditions hold and $\mathcal{A}_n \succeq \mathcal{B}_n$, naturally we could utilize the gap between two groups of $\{|\beta_j^{\mathrm{CR}}|\}_{j=1}^{p_n}$ to identify the relevant variables, where $a_n \succeq b_n$ denotes that there exists a constant $c > 0$ such that $a_n \geq cb_n$ for all $n \geq 1$.

3.2. Sure screening property of componentwise BD *regression*

We start by giving the uniform convergence of componentwise regression minimum-BD estimator (3.3). To facilitate the derivation, we assume $E(X_j) = 0$ and $E(X_j^2) = 1$, for $j = 1, \ldots, p_n$ in the following results.

Theorem 3. *Assume Conditions* A1–A5 *in the Appendix. Then for any positive sequence* $\mathcal{A}_n$ *satisfying* $\mathcal{A}_n\sqrt{n}/\log(n) \to \infty$, *there exists some positive constant* c_2 *such that*

$$\mathrm{P}\Big(\max_{1\leq j\leq p_n} |\widehat{\beta}_j^{\mathrm{CR}} - \beta_j^{\mathrm{CR}}| \geq \mathcal{A}_n\Big) \leq p_n\{\exp(-c_2\mathcal{A}_n^2 n) + nm_0\exp(-m_1\mathcal{A}_n^2 n)\},$$

where m_0 *and* m_1 *are the constants given in Condition* A2 *of the Appendix.*

Theorem 3 is an application of the exponential bound for the Quasi-MLE in [7] (Theorem 1). It guarantees that the empirical estimator $\widehat{\beta}_j^{\mathrm{CR}}$ will be close enough to the population version β_j^{CR} with large probability. With Theorem 3, we obtain Corollary 1 below which demonstrates the sure screening property of componentwise BD regression.

Corollary 1. *Assume conditions in Theorem 3. Set* $\gamma_n = c_1\mathcal{A}_n/2$, *where* c_1 *is the constant given in Theorem 2.*

(i) (*Sure screening property*) *If* $\min_{j=1,\ldots,s_n} |\mathrm{cov}(Y, X_j)| \geq \mathcal{A}_n$, *then*

$$\mathrm{P}(\mathcal{M}_n \subseteq \widehat{\mathcal{M}}_{\gamma_n}) \geq 1 - s_n\{\exp(-c_2c_1^2\mathcal{A}_n^2 n/4) + nm_0\exp(-m_1c_1^2\mathcal{A}_n^2 n/4)\}.$$

(ii) *If* $\max_{j=s_n+1,\ldots,p_n} |\mathrm{cov}(Y, X_j)| \leq \mathcal{B}_n = o(\mathcal{A}_n)$, *then*

$$\begin{aligned}&\mathrm{P}(\widehat{\mathcal{M}}_{\gamma_n} \subseteq \mathcal{M}_n)\\ &\quad\geq 1 - (p_n - s_n)\{\exp(-c_2c_1^2\mathcal{A}_n^2 n/4) + nm_0\exp(-m_1c_1^2\mathcal{A}_n^2 n/4)\}.\end{aligned}$$

Corollary 1 addressed the first question raised at the beginning of Section 3. It is easy to see that if we assume $\mathcal{A}_n = c_0 n^{-\alpha}$ with a constant $0 < \alpha < 1/2$ and $\log(p_n) = o(n^{1-2\alpha})$, then the probability bounds in Corollary 1 are approaching one with the order $1 - O\{p_n\exp(-c_3 n^{1-2\alpha})\}$ for a positive constant c_3, which is the same rate obtained in [5] and [7]. This implies that the correct model will be selected with probability tending to one even under NP-dimensionality, where p_n is permitted to be as large as $\log(p_n) = o(n^{1-2\alpha})$.

Remark 1. *In Corollary 1, the conclusion from part* (i) *and the conclusion from part* (ii) *hold separately. In some cases, the assumption about the covariance between the response and covariates in part* (ii) *of Corollary 1 needs to be relaxed, but the sure screening property given in part* (i) *of Corollary 1 will still hold. It means that even if we can not eliminate all irrelevant covariates due to certain correlation between covariates, it is still guaranteed that we will not miss any truly relevant variables.*

3.3. Comparison with sure independence screening in GLM

While our motivation is from [7]'s work on the generalized linear model (GLM), our results largely enhance the capability of marginal screening methods by extending it to a broader class of models with any BD as a loss function. In fact, proving the sure screening property under such general framework is by no means straightforward. The main challenge is that certain relationships in GLM are not applicable under the arbitrary choice of BD and link function in our setting. For example, the following equality holds under GLM combined with its canonical link F,

$$E\{F^{-1}(\alpha_j^{\mathrm{CR}} + \beta_j^{\mathrm{CR}} X_j)X_j\} = E\{F^{-1}(\beta_{0;0} + \boldsymbol{X}^T\boldsymbol{\beta}_0)X_j\}, \tag{3.5}$$

where $\beta_{0;0}$ and $\boldsymbol{\beta}_0 = (\beta_{1;0}, \ldots, \beta_{p;0})^T$ are unknown true parameters (see parameterization in (4.1)). Actually, (3.5) is the same as equation (14) in [7] which is a significant part of the proofs for Theorems 2, 3 and 5 therein. However, (3.5) no longer holds in the BD estimation when F could be an arbitrary link. To overcome such technical challenge, we need to introduce a different way to express the componentwise regression minimum BD estimate and also impose some assumptions on the uniform bound of covariates. For details, please see the Appendix.

Although the proposed screening procedure is based on certain linear form, the sure screening property of the proposed screening method actually does not require any particular parametric form of relationship between the response Y and the covariates $\boldsymbol{X}$. Instead, the sure screening property is mainly built on the assumption about minimal signal strength of relevant variables measured by marginal covariance. Our results also reveal that different choices of BD will only affect constants c_1 and c_2 in the probability bounds in Corollary 1.

4. Two-step procedure with penalized-BD estimation

The results in Section 3 show that the screening procedure based on componentwise regression minimum-BD estimation works well in selecting the truly relevant variables. However, it may not be a good way to build a predictive model and provide estimates. In the absence of screening, [23] investigated the penalized-BD estimation and its oracle property in a large-dimensional model with the following form,

$$m(\boldsymbol{X}) = E(Y \mid \boldsymbol{X}) = F^{-1}(\beta_{0;0} + \boldsymbol{X}^T\boldsymbol{\beta}_0), \tag{4.1}$$

where $\beta_{0;0}$ and $\boldsymbol{\beta}_0 = (\beta_{1;0}, \ldots, \beta_{p;0})^T$ are unknown true parameters, and F is a known link function. Particularly, their penalized-BD estimator using weighted L_1 penalties minimizes the criterion function,

$$\ell(\beta_0, \boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} Q\{Y_i, F^{-1}(\beta_0 + \boldsymbol{X}_i^T\boldsymbol{\beta})\} + \lambda_n \sum_{j=1}^{p_n} w_j|\beta_j|, \tag{4.2}$$

where λ_n is the tuning parameter and $\{w_j\}_{j=1}^{p_n}$ are given weights for parameters $\{\beta_j\}_{j=1}^{p_n}$. In this section, we adopt the same setting.

Now we could answer the second question given at the beginning of Section 3 by Theorem 4 which is to control the number of the selected variables in the set $\widehat{\mathcal{M}}_{\gamma_n}$.

Theorem 4. *Assume the conditions in Theorem 3 and Condition* C *in the Appendix. Let* $\gamma_n = c_1\mathcal{A}_n/2$, *where* c_1 *is the constant given in Theorem 2. It holds that*

$$\begin{aligned}&\mathrm{P}[|\widehat{\mathcal{M}}_{\gamma_n}| \le O\{\mathcal{A}_n^{-2}\lambda_{\max}(\Sigma)\}]\\ &\quad\ge 1 - p_n\{\exp(-c_2c_1^2\mathcal{A}_n^2 n/4) + nm_0\exp(-m_1c_1^2\mathcal{A}_n^2 n/4)\},\end{aligned}$$

where $\Sigma = \mathrm{var}(\boldsymbol{X})$ *and* $\lambda_{\max}(\Sigma)$ *denotes the maximum eigenvalue of* Σ.

When $\mathcal{A}_n = O(n^{-\alpha})$ and $\lambda_{\max}(\Sigma) = O(n^{\tau})$ with $2\alpha+\tau < 1$, Theorem 4 indicates that the number of selected covariates will not exceed the order $n^{2\alpha+\tau} = o\{n/\log(n)\}$. Therefore, we propose a two-step procedure from screening features to estimating coefficients of selected variables as follows. Since the cutoff value γ_n involves some unknown constant, in practice we propose another easier and more straightforward scheme that choose γ_n to be the p_n'th largest values of $|\widehat{\beta}_j^{\mathrm{CR}}|$, where $p_n' = \lfloor n/\log(n)\rfloor$ and $\lfloor\cdot\rfloor$ denotes the floor function. The choice of p_n' is large enough so that all truly relevant covariates will be selected, and also suitable for further estimation method in the second stage.

Step 1: Obtain the componentwise regression minimum-BD estimators $\widehat{\beta}_j^{\mathrm{CR}}$ in (3.3) and select sufficiently many covariates, corresponding to p_n' largest values of $|\widehat{\beta}_j^{\mathrm{CR}}|$. Denote by $\widehat{\mathcal{M}}$ the set of indices of selected variables, where $|\widehat{\mathcal{M}}| = p_n'$.

Step 2: Set the coefficients of $(p_n - p_n')$ variables not in $\widehat{\mathcal{M}}$ equal to zero. Use (4.2) to estimate the other parameters of those p_n' features selected in Step 1.

The idea of two-step procedures is widely used in the literature, for example the multi-stage method in [20]. After the first step, we can greatly reduce the dimensionality and at the same time, by the sure screening property, still preserve all truly relevant variables with high probability.

Proposition 1 also indicates that our two-step procedure enjoys the oracle property under certain conditions.

Proposition 1. *Assume the conditions in Theorem 4 and* $\mathcal{A}_n = O(1)$, $\mathcal{B}_n = o(\mathcal{A}_n)$. *Suppose* $s_n = o(n^{1/5})$, $s_n(p'_n - s_n) = o(n)$ *and* $\lambda_{\max}(\Sigma)/(p'_n\mathcal{A}_n^2) = o(1)$. *Let* $\lambda_n = \mathcal{A}_n/\sqrt{n}$ *and select weights in* (4.2) *by*

$$\widehat{w}_j = |\widehat{\beta}_j^{\mathrm{PCR}}|^{-1} \quad \text{for } j \in \widehat{\mathcal{M}},$$

where $\widehat{\beta}_j^{\mathrm{PCR}}$ *is based on the penalized componentwise regression* BD *estimation,*

$$(\widehat{\alpha}_j^{\mathrm{PCR}}, \widehat{\beta}_j^{\mathrm{PCR}}) = \underset{(\alpha_j,\beta_j)\in\mathbb{R}^2}{\arg\min} \Big[\frac{1}{n}\sum_{i=1}^{n} Q\{Y_i, F^{-1}(\alpha_j + X_{ij}\beta_j)\} + \kappa_n|\beta_j|\Big] \tag{4.3}$$

with $\kappa_n = \mathcal{A}_n$. *Let* $\widetilde{\boldsymbol{X}}^{(\mathrm{I})} = (1, \boldsymbol{X}^{(\mathrm{I})T})^T$, $\widetilde{\boldsymbol{\beta}}_0 = (\beta_{0;0}, \boldsymbol{\beta}_0^T)^T$, $\widetilde{\boldsymbol{\beta}}_0^{(\mathrm{I})} = (\beta_{0;0}, \boldsymbol{\beta}_0^{(\mathrm{I})T})^T$. *Then we have the following results for the two-step estimator.*

(i) *There exists a local minimizer* $\widehat{\widetilde{\boldsymbol{\beta}}}$ *such that* $\|\widehat{\widetilde{\boldsymbol{\beta}}} - \widetilde{\boldsymbol{\beta}}_0\|_2 = O_{\mathrm{P}}\{(s_n/n)^{1/2}\}$.

(ii) *Any* $(n/s_n)^{1/2}$*-consistent local minimizer* $\widehat{\widetilde{\boldsymbol{\beta}}} = (\widehat{\widetilde{\boldsymbol{\beta}}}^{(\mathrm{I})T}, \widehat{\boldsymbol{\beta}}^{(\mathrm{II})T})^T$ *satisfies* $\mathrm{P}(\widehat{\boldsymbol{\beta}}^{(\mathrm{II})} = \mathbf{0}) \to 1$.

(iii) *Assume Condition* D *in the Appendix. If* $\min_{j=1,\ldots,s_n}|\beta_j|/(s_n/n)^{1/2} \to \infty$, *then for any fixed integer* k *and any* $k \times (s_n + 1)$ *matrix* A_n *such that* $A_n A_n^T \to G$ *with* G *being a* $k \times k$ *nonnegative-definite symmetric matrix, we have that*

$$\sqrt{n}A_n\boldsymbol{\Omega}_n^{-1/2}\{\mathbf{H}_n(\widehat{\widetilde{\boldsymbol{\beta}}}^{(\mathrm{I})} - \widetilde{\boldsymbol{\beta}}_0^{(\mathrm{I})}) + \lambda_n\mathbf{W}_n\mathrm{sign}(\widetilde{\boldsymbol{\beta}}_0^{(\mathrm{I})})\} \xrightarrow{\mathcal{L}} N(\mathbf{0}, G),$$

where

$$\begin{aligned}
\boldsymbol{\Omega}_n &= E[\mathrm{var}(Y \mid \boldsymbol{X})\{q''(m(\boldsymbol{X}))/F'(m(\boldsymbol{X}))\}^2\widetilde{\boldsymbol{X}}^{(\mathrm{I})}\widetilde{\boldsymbol{X}}^{(\mathrm{I})T}],\\
\mathbf{H}_n &= -E[q''(m(\boldsymbol{X}))/\{F'(m(\boldsymbol{X}))\}^2\widetilde{\boldsymbol{X}}^{(\mathrm{I})}\widetilde{\boldsymbol{X}}^{(\mathrm{I})T}],\\
\mathbf{W}_n &= \mathrm{diag}(0, \widehat{w}_1, \ldots, \widehat{w}_{s_n}),
\end{aligned}$$

and $\mathrm{sign}(\widetilde{\boldsymbol{\beta}}_0^{(\mathrm{I})}) = (\mathrm{sign}(\beta_{0;0}), \mathrm{sign}(\beta_{1;0}), \ldots, \mathrm{sign}(\beta_{s_n;0}))^T$.

Note that the proposed componentwise BD regression weight selection method in [23] excludes an intercept term. In contrast, our current weight selection method (4.3) includes the intercept term. Nevertheless, the assumptions for the oracle property are still satisfied and thus our procedure would also enjoy the oracle property.

5. Simulation study

In this section, we assess the performance of both the screening step and the estimation step in the two-step procedure. Two different settings of (n, p_n) are used in our simulation,

$$(250, 250) \quad \text{and} \quad (350, 15000),$$

which represent the high dimensionality and ultra-high dimensionality of data, respectively.

5.1. Performance of feature screening

To evaluate the performance of screening methods, we will measure the accuracy of the importance ranking of the covariates by the minimum model size (MMS) needed to include all truly relevant covariates. We also provide a coverage measure as the percentage of runs that all truly non-zero coefficients are picked up when setting $p'_n = \lfloor n/\log(n) \rfloor$. The following methods are compared:

$$\begin{aligned} &\text{“SIS-BD”} : \text{our proposed screening method,} \\ &\text{“DC-SIS”} : \text{method in [14],} \\ &\text{“EL-SIS”} : \text{method in [4],} \\ &\text{“RRCS”} : \text{method in [13],} \\ &\text{“SIRS”} : \text{method in [24].} \end{aligned} \tag{5.1}$$

All the results are averaged over 400 simulation runs.

5.1.1. Overdispersed Poisson responses

Here we consider the overdispersed Poisson model with the response Y generated according to $\text{var}(Y \mid \boldsymbol{X} = \boldsymbol{x}) = 2\, m(\boldsymbol{x})$, where $m(\boldsymbol{x}) = E(Y \mid \boldsymbol{X} = \boldsymbol{x})$. The link function used for count data is the log link. Thus,

$$\log\{m(\boldsymbol{x})\} = \beta_{0;0} + \boldsymbol{x}^T \boldsymbol{\beta}_0,$$

where $\beta_{0;0} = 1$ and $\boldsymbol{\beta}_0 = (2.5, 2, 2, 1.5, 0, \ldots, 0)^T$. The covariates are generated by $X_{ij} = \Phi(Z_{ij}) - 0.5$, $i = 1, \ldots, n$ and $j = 1, \ldots, p_n$, where Φ is the standard normal distribution function, and

$$(Z_{i1}, \ldots, Z_{ip_n})^T \sim N(\mathbf{0}, \rho \mathbf{J}_{p_n} + (1-\rho)\mathbf{I}_{p_n}), \tag{5.2}$$

with $\mathbf{J}_d$ a $d \times d$ matrix in which all entries are ones and $\mathbf{I}_d$ a $d \times d$ identity matrix. Thus $(X_{i1}, \ldots, X_{ip_n})$ are marginally Uniform$(-0.5, 0.5)$ random variables and correlated if $\rho \neq 0$. The type of BD we used here is

$$Q(Y, \mu) = \mu - Y \log(\mu) - Y + Y \log(Y)$$

which is generated by the q-function in (2.3) when $V(\mu) = \mu$, explicitly, $q(\mu) = \mu - a - \mu\{\log(\mu) - \log(a)\}$.

Table 1 presents the mean, standard deviation along with a five number summary of the MMS as well as the coverage percentage for the screening methods in different settings. For the case of $(n, p_n) = (250, 250)$ and $\rho = 0.2$, all the procedures work well in this nonlinear model and rank the truly relevant covariates at the very top of the list, as the resulting MMS's are very close to the true model size. Also for $(n, p_n) = (250, 250)$, when the dependence parameter ρ increases from 0.2 to 0.5, the correlation between the covariates becomes larger and the irrelevant covariates can be easily confounded with the relevant covariates. In this case, the MMS becomes a little bigger and the coverage percentage

TABLE 1

(Simulation results: overdispersed Poisson count responses) *Mean, standard deviation* (std)*, and five number summary of the minimum model size, out of* 400 *replications with* Q_1*: first quartile;* Q_2*: median;* Q_3*: third quartile;* CP*: percentage of runs that all truly non-zero coefficients are covered by* $\widehat{\mathcal{M}}$*. Methods* SIS-BD, DC-SIS, EL-SIS, RRCS *and* SIRS *are described in* (5.1).

n / p_n	ρ	Method	Mean (std)	Min	Q_1	Q_2	Q_3	Max	CP
250 / 250	0.2	SIS-BD	4.62 (0.2)	4	4	4	4	71	.998
		DC-SIS	4.44 (0.1)	4	4	4	4	34	1.00
		EL-SIS	5.53 (0.2)	4	4	4	5	50	.998
		RRCS	4.43 (0.1)	4	4	4	4	17	1.00
		SIRS	4.52 (0.1)	4	4	4	4	26	1.00
	0.5	SIS-BD	7.11 (0.4)	4	4	4	6	89	.990
		DC-SIS	6.45 (0.4)	4	4	4	6	86	.988
		EL-SIS	8.43 (0.7)	4	4	5	7	174	.983
		RRCS	7.06 (0.5)	4	4	4	6	127	.988
		SIRS	7.39 (0.5)	4	4	4	6	121	.993
350 / 15000	0.2	SIS-BD	7.81 (0.9)	4	4	4	5	242	.988
		DC-SIS	8.40 (1.6)	4	4	4	5	590	.988
		EL-SIS	43.27 (10.9)	4	4	4	11	2937	.893
		RRCS	10.13 (1.6)	4	4	4	5	402	.980
		SIRS	11.37 (1.9)	4	4	4	5	595	.968
	0.5	SIS-BD	48.02 (6.5)	4	4	8	32	1176	.865
		DC-SIS	69.61 (12.0)	4	4	7	26	2936	.835
		EL-SIS	181.68 (32.8)	4	6	13	62	7113	.738
		RRCS	66.71 (11.3)	4	4	7	28	2615	.838
		SIRS	81.75 (13.5)	4	4	9	39.5	3043	.800

becomes slightly smaller, while all methods still perform comparably well. For $(n, p_n) = (350, 15000)$, our proposed method performs better than or as well as the other methods. Particularly, when $\rho = 0.5$, SIS-BD has a significantly smaller mean MMS and a larger coverage percentage than the other methods. It's also seen that, in the ultra-high dimensional case, as ρ increases, the MMS increases and the coverage percentage decreases as expected. By comparing the results for $(n, p_n) = (350, 15000)$ with those for $(n, p_n) = (250, 250)$, the higher dimensionality makes the feature selection problem harder, but the values of the MMS do not increase much, which supports our theoretical results in Section 3.

5.1.2. Bernoulli binary responses

We further investigate the logistic regression model with a binary response Y, which is generated as a Bernoulli random variable with

$$\mathrm{P}(Y = 1 \mid \boldsymbol{X} = \boldsymbol{x}) = \frac{1}{1 + \exp\{-(\beta_{0;0} + \boldsymbol{x}^T \boldsymbol{\beta}_0)\}},$$

where $\beta_{0;0} = 3$ and $\boldsymbol{\beta}_0 = (3, 2, -3, -2, 0, \ldots, 0)^T$. The covariates are also generated by independent Bernoulli random variables with

$$\mathrm{P}(X_{ij} = 1) = r, \qquad i = 1, \ldots, n,\ j = 1, \ldots, p_n. \tag{5.3}$$

The logit link function is used.

Results from two types of BD will be presented in this section, the Bernoulli deviance (DEV) loss,

$$Q(Y, \mu) = -2\{Y \log(\mu) + (1 - Y) \log(1 - \mu)\}$$

which corresponds to $q(\mu) = -2\{\mu \log(\mu) + (1 - \mu) \log(1 - \mu)\}$, and the exponential (EXP) loss,

$$Q(Y, \mu) = \exp[-(Y - .5) \log\{\mu/(1 - \mu)\}]$$

which corresponds to $q(\mu) = 2\{\mu(1 - \mu)\}^{1/2}$.

Table 2 reports the mean and the five number summary of the MMS from our simulation. Our proposed methods SIS-BD(DEV) and SIS-BD(EXP), combined with two choices of loss functions, DEV and EXP respectively, give similar results in all cases as expected and both perform better than or as well as the other methods. When $r = 0.2$, the signal from the data is more scarce and it is more difficult to identify the truly relevant variables. Under this stringent situation, the mean values of the MMS for our methods remain at a low level compared with the total number of covariates, and are smaller than those of the other methods. For $r = 0.5$, the performance of the screening methods is improved compared with that for $r = 0.2$. As the dimensionality grows, the MMS's for our methods increase, but at a much smaller rate so that the identification of relevant covariates and further parameter estimation would still be possible. However, because of the ultra-high dimensionality and the sparse signals from the covariates, there is a dramatic decrease in the coverage percentage as p_n increases from 250 to 15000 for $r = 0.2$. Compared with the corresponding results in Table 1 for the overdispersed Poisson model, the values of the MMS in Table 2 are much larger, because, in the logistic regression model, the response and covariates are all binary which provide very limited information from either part. It's worth noting that the RRCS method fails to identify the truly non-zero coefficients. This doesn't contradict the conclusions in [13], where the sure screening properties of RRCS are demonstrated for linear models but not for generalized linear models due to technical difficulties as discussed in their Section 4.2.

5.2. Performance of parameter estimation

Here we will compare the performance of the two-step procedure with those variable selection methods using penalization which are directly applied to all variables. Namely, they minimize a criterion function similar to (4.2), except that the choice of the penalty can be the following:

TABLE 2
(Simulation results: Bernoulli binary responses) *Mean, standard deviation, and five number summary of the minimum model size, out of* 400 *replications. Methods* SIS-BD, DC-SIS, EL-SIS, RRCS *and* SIRS *are described in* (5.1).

n p_n	r	Method	Mean (std)	Min	Q_1	Q_2	Q_3	Max	CP
250 250	0.2	SIS-BD (DEV)	24.92 (1.7)	4	8	12.5	25	240	.858
		SIS-BD (EXP)	25.03 (1.7)	4	8	13	25	239	.855
		DC-SIS	28.02 (1.7)	4	9	15	32	241	.843
		EL-SIS	25.73 (1.7)	4	7	12	28	239	.855
		RRCS	244.53 (0.6)	168	245	249	250	250	.000
		SIRS	28.14 (1.7)	4	9	15	32	240	.835
	0.5	SIS-BD (DEV)	9.21 (0.6)	4	4	5	9	96	.975
		SIS-BD (EXP)	9.20 (0.6)	4	4	5	9	96	.975
		DC-SIS	9.13 (0.6)	4	4	5	9	97	.978
		EL-SIS	9.13 (0.6)	4	4	5	9	95	.978
		RRCS	250 (0.0)	250	250	250	250	250	.000
		SIRS	9.15 (0.6)	4	4	5	9	97	.980
350 15000	0.2	SIS-BD (DEV)	597.11 (67.5)	4	81.5	185	434.5	10627	.193
		SIS-BD (EXP)	597.37 (67.5)	4	81.5	186	441.5	10612	.190
		DC-SIS	713.77 (71.5)	8	78.5	234	669.5	10803	.200
		EL-SIS	616.30 (70.7)	4	36	142.5	481.5	10631	.335
		RRCS	14814.73 (22.3)	9949	14821	14960	14995	15000	.000
		SIRS	716.39 (71.8)	8	78.5	234.5	670	10836	.200
	0.5	SIS-BD (DEV)	125.24 (24.2)	4	5	10.5	53.5	6099	.773
		SIS-BD (EXP)	125.39 (24.3)	4	5	10.5	54	6137	.775
		DC-SIS	125.15 (24.4)	4	4	10.5	49.5	6294	.765
		EL-SIS	124.86 (24.2)	4	4	11	51.5	6128	.765
		RRCS	14999.99 (0.0)	14999	15000	15000	15000	15000	.000
		SIRS	125.62 (24.5)	4	4	10.5	49.5	6306	.763

(I) (SCAD) the SCAD penalty, with an accompanying parameter $a = 3.7$ ([6]);
(II) (MCP) the MCP penalty, with an accompanying parameter $a = 3.7$ ([21]);
(III) (L_1) the L_1 penalty ([17]);
(IV) (WL1PCR) the weighted-L_1 penalty with weights selected by (4.3).

Let $p'_n = \lfloor n/\log(n) \rfloor$ in the first step. For brevity, the two-step procedures are referred to as S-SCAD, S-MCP, S-L_1, and S-WL1PCR, respectively. The tuning constants λ_n and κ_n are selected via a grid search separately to minimize the BIC. All the results are averaged over 100 simulation runs.

5.2.1. Overdispersed Poisson responses

The setting is similar to that in Section 5.1.1, except that the dependence parameter between covariates is fixed at $\rho = 0.2$. To compare the accuracy of the

TABLE 3

(Simulation results: overdispersed Poisson count responses) *Covariates are marginally* Uniform$(-0.5, 0.5)$ *with dependence parameter* $\rho = 0.2$ *in* (5.2). *Results are averaged over* 100 *replications. Here* TE *is the test error obtained from an independent test set; time is the average running time in seconds. Timing was carried out on an Intel* 3.60 *GHz processor.*

n p_n	Method	$\|\widehat{\widetilde{\boldsymbol{\beta}}} - \widetilde{\boldsymbol{\beta}}_0\|_2$	TE	#C-Z	#C-NZ	Time (sec)
250 250	SCAD	0.72	1.02	231.1 (0.6)	4.0 (0.0)	0.53
	S-SCAD	0.54	0.97	237.3 (0.3)	4.0 (0.0)	0.05
	MCP	0.60	0.99	234.9 (0.7)	4.0 (0.0)	0.48
	S-MCP	0.51	0.97	239.4 (0.3)	4.0 (0.0)	0.05
	L_1	0.80	1.04	231.7 (0.6)	4.0 (0.0)	0.18
	S-L_1	0.70	1.01	235.6 (0.4)	4.0 (0.0)	0.05
	WL1PCR	0.53	0.97	241.4 (0.3)	4.0 (0.0)	1.45
	S-WL1PCR	0.53	0.97	241.3 (0.3)	4.0 (0.0)	0.34
350 15000	SCAD	1.09	1.15	14959.6 (1.2)	4.0 (0.0)	29.54
	S-SCAD	0.65	1.01	14982.1 (0.4)	4.0 (0.0)	1.72
	MCP	0.83	1.04	14969.1 (1.0)	4.0 (0.0)	29.64
	S-MCP	0.63	1.01	14984.2 (0.4)	4.0 (0.0)	1.70
	L_1	1.13	1.16	14962.2 (1.2)	4.0 (0.0)	12.27
	S-L_1	0.90	1.10	14977.9 (0.4)	4.0 (0.0)	1.69
	WL1PCR	0.72	1.02	14984.4 (0.5)	4.0 (0.0)	102.52
	S-WL1PCR	0.75	1.04	14984.8 (0.5)	4.0 (0.0)	6.62

estimated parameters by different methods, the average of $\|\widehat{\widetilde{\boldsymbol{\beta}}} - \widetilde{\boldsymbol{\beta}}_0\|_2$ across those 100 training sets is calculated. The test error (TE) is obtained from an independently generated test set $\{(\boldsymbol{x}_\ell, y_\ell)\}_{\ell=1}^{L=10000}$ by $\sum_{\ell=1}^{L} Q\{y_\ell, \widehat{m}(\boldsymbol{x}_\ell)\}/L$. We also provide the model selection performance via C-Z which is the total number of coefficients which are correctly estimated to be zero when the true coefficients are zero, and C-NZ which is the total number of coefficients which are correctly estimated to be non-zero when the true coefficients are non-zero. Finally, we record the average running time of each method under different settings. The high dimensional problem usually imposes a big challenge in computations as well as model selection and estimation, so considerably faster speed can be viewed as an advantage. Table 3 summarizes the simulation results.

By comparing the average losses of the estimates for $\widetilde{\boldsymbol{\beta}}_0$ and the test errors, we can see that all methods have satisfactory performance, while most two-step procedures are slightly better than their counterparts which directly apply the estimation step. Besides, every method is able to select all of the truly non-zero parameters. While the gain of the accuracy from the screening step does not seem to be very dramatic, we notice that the speed of the two-step procedures is much faster, where the screening step can reduce the computation time by a factor of 3 to 20. This indicates that the screening step can indeed filter out most irrelevant covariates without sacrificing the accuracy, so that we can make better use of the computational resources. When p_n grows from 250 to 15000, for each method, the average loss of the estimates for $\widetilde{\boldsymbol{\beta}}_0$ and the test error increase slightly, indicating that the two-step procedures work conceivably

well for ultra-high dimensional case. Compared with all the other methods, the screening-based estimation method using the MCP penalty corresponds to the smallest values of $\|\widehat{\widetilde{\boldsymbol{\beta}}} - \widetilde{\boldsymbol{\beta}}_0\|_2$ and the test error, and enjoys the shortest computational time. But, the difference between the performance of different two-step procedures is indeed not significant.

5.2.2. Bernoulli binary responses

The setting is similar to that in Section 5.1.2, except that we only present the results with $r = 0.5$. For this type of classification problem, we calculate the misclassification rate (MR) for an independent test set instead of the test error. Other metrics are similar to those in Section 5.2.1. The results are summarized in Table 4.

From Table 4, for most of the methods, the two-step procedures perform better than or as well as the counterparts without applying the screening step in each setting. Among all the two-step procedures, the one using the weighted-L_1 penalized estimation with weights selected by the PCR corresponds to the smallest averaged loss of the estimates for $\widetilde{\boldsymbol{\beta}}_0$ and the lowest misclassification rate for each setting. As p_n increases, for each method and loss function, there is a reasonable increase in the loss of the estimate for $\widetilde{\boldsymbol{\beta}}_0$ and a slight increase in the misclassification rate. Therefore, the two-step procedures have satisfactory performance under the ultra-high dimensional situation. Again, in our comparison, the screening step can make the computation much faster than those methods that directly work with all possible covariates. Compared with Table 3, Table 4 has much larger values of $\|\widehat{\widetilde{\boldsymbol{\beta}}} - \widetilde{\boldsymbol{\beta}}_0\|_2$ and smaller values of C-NZ, which is due to the specific setting for the logistic regression model where the covariates and response are binary.

6. Real data application

In this section, we apply the methods considered in Section 5.2 to real data to illustrate the practical usefulness of the screening procedures. The tuning constants λ_n and κ_n in the second step is selected by the Akaike's information criterion (AIC).

6.1. Colon data

The classification of colon cancer is discussed in [1] and the data set can be downloaded from `http://genomics-pubs.princeton.edu/oncology/`. It consists of $p = 2000$ genes and $n = 62$ samples, in which 22 samples are from normal colon tissues and 40 samples are from tumor tissues. In our analysis, the data set is randomly split into two parts, with 45 samples as training samples and the rest 17 as test samples. We repeat the random split 100 times and calculate the

TABLE 4

(Simulation results: Bernoulli binary responses) *Covariates are independent Bernoulli random variables with* $r = 0.5$ *in* (5.3). MR *is the misclassification rate on an independent test set. Results are averaged over* 100 *replications. Timing was carried out on an Intel* 3.60 *GHz processor.*

n / p_n	Loss	Method	$\|\widehat{\widetilde{\boldsymbol{\beta}}} - \widetilde{\boldsymbol{\beta}}_0\|_2$	MR	#C-Z	#C-NZ	Time (sec)
250 / 250	DEV	SCAD	24.45	0.17	236.1 (0.4)	3.9 (0.0)	0.44
		S-SCAD	11.58	0.16	238.8 (0.4)	3.9 (0.0)	0.44
		MCP	22.63	0.17	236.1 (0.3)	3.9 (0.0)	0.45
		S-MCP	10.93	0.16	239.3 (0.4)	3.9 (0.0)	0.44
		L_1	3.75	0.16	244.5 (0.1)	3.6 (0.1)	0.07
		S-L_1	3.74	0.16	244.4 (0.1)	3.6 (0.1)	0.04
		WL1PCR	2.52	0.13	244.4 (0.2)	3.7 (0.1)	0.36
		S-WL1PCR	2.52	0.13	244.4 (0.2)	3.7 (0.1)	0.13
	EXP	SCAD	34.77	0.17	236.1 (0.3)	3.9 (0.0)	0.53
		S-SCAD	16.22	0.16	239.0 (0.5)	3.9 (0.0)	0.48
		MCP	36.90	0.17	235.9 (0.3)	3.9 (0.0)	0.52
		S-MCP	16.76	0.16	238.9 (0.4)	3.9 (0.0)	0.45
		L_1	3.31	0.16	244.2 (0.2)	3.7 (0.1)	0.08
		S-L_1	3.31	0.16	244.2 (0.2)	3.7 (0.1)	0.04
		WL1PCR	2.31	0.13	244.5 (0.2)	3.6 (0.1)	0.41
		S-WL1PCR	2.31	0.13	244.5 (0.2)	3.6 (0.1)	0.15
350 / 15000	DEV	SCAD	26.25	0.17	14987.4 (0.2)	3.8 (0.1)	15.68
		S-SCAD	22.36	0.18	14984.3 (0.3)	3.7 (0.1)	1.58
		MCP	25.94	0.17	14987.1 (0.2)	3.7 (0.1)	15.09
		S-MCP	23.22	0.18	14983.7 (0.3)	3.7 (0.1)	1.59
		L_1	4.29	0.17	14994.7 (0.1)	3.2 (0.1)	5.68
		S-L_1	4.28	0.17	14994.7 (0.1)	3.2 (0.1)	1.24
		WL1PCR	3.21	0.15	14993.8 (0.2)	3.4 (0.1)	31.59
		S-WL1PCR	3.21	0.15	14993.8 (0.2)	3.4 (0.1)	2.88
	EXP	SCAD	39.80	0.18	14987.6 (0.2)	3.6 (0.1)	15.68
		S-SCAD	34.63	0.18	14984.1 (0.3)	3.7 (0.1)	1.75
		MCP	36.67	0.18	14987.5 (0.2)	3.6 (0.1)	15.40
		S-MCP	35.25	0.18	14984.0 (0.3)	3.7 (0.1)	1.78
		L_1	3.98	0.17	14994.5 (0.2)	3.3 (0.1)	6.35
		S-L_1	4.03	0.17	14994.1 (0.5)	3.3 (0.1)	1.38
		WL1PCR	2.76	0.15	14994.6 (0.3)	3.0 (0.1)	35.48
		S-WL1PCR	2.76	0.15	14994.0 (0.5)	3.1 (0.1)	3.08

average number of misclassified cases in both sets. The results are summarized in Table 5.

From Table 5, we see that the two-step procedures are capable of identifying those relevant variables and obtaining a good estimation and prediction while considerably fewer computational resources are needed. Among all the two-step procedures, it turns out that the one using the L_1 penalty corresponds to the smallest number of misclassified cases for both the training and test sets using either the deviance or exponential loss. Although method WL1PCR without using the screening step performs slightly better than S-L_1, its computational time is much longer. Moreover, the choice of loss functions in the penalized-BD estimators has a relatively negligible impact on the classification performance.

TABLE 5

(Real data: Colon) *Average number of misclassified cases among* 45 *training samples and* 17 *test samples, average number of selected variables among all* 2000 *covariates and the average computation time in seconds, over* 100 *replications. Timing was carried out on an Intel* 3.60 *GHz processor.*

Loss	Method	# Error (training)	# Error (test)	# Selected	Time (sec)
DEV	SCAD	2.9	3.8	4.2	0.9
	S-SCAD	2.6	3.4	4.7	0.2
	MCP	2.9	3.8	4.2	0.8
	S-MCP	2.6	3.4	4.7	0.2
	L_1	1.0	3.8	16.1	1.7
	S-L_1	0.6	3.2	11.1	0.3
	WL1PCR	0.5	3.3	10.3	10.8
	S-WL1PCR	1.0	3.3	8.5	1.4
EXP	SCAD	3.8	3.8	4.3	1.0
	S-SCAD	3.1	3.8	4.6	0.3
	MCP	3.8	4.0	4.2	1.0
	S-MCP	3.2	3.7	4.6	0.3
	L_1	0.7	3.7	15.6	1.8
	S-L_1	0.7	3.3	10.5	0.3
	WL1PCR	0.4	3.2	9.6	11.6
	S-WL1PCR	1.0	3.4	8.3	1.5

Appendix: Proofs of Main Results

Notation. For notational brevity, let $\mathsf{X}_j = (1, X_j)^T$ and $\mathbf{b}_j = (\alpha_j, \beta_j)^T$ denote the two-dimensional covariate and parameter of the componentwise regression minimum-BD estimation in (3.3), respectively. Denote $\widehat{\mathbf{b}}_j^{\mathrm{CR}} = (\widehat{\alpha}_j^{\mathrm{CR}}, \widehat{\beta}_j^{\mathrm{CR}})^T$ and $\mathbf{b}_j^{\mathrm{CR}} = (\alpha_j^{\mathrm{CR}}, \beta_j^{\mathrm{CR}})^T$. Throughout this section, $\|\cdot\|_1$ is the L_1-norm, $\|\cdot\|_2$ is the Euclidean L_2-norm, and $\|\cdot\|_\infty$ is used to denote the L_∞-norm.

Condition. We have the following assumptions in which M, B, B' are sufficiently large constants. Those are not the weakest possible, but serve to facilitate the technical derivations.

A1. For all j, X_j are uniformly bounded, i.e. $\|\boldsymbol{X}\|_\infty \le M$. $\Sigma = \mathrm{var}(\boldsymbol{X})$ exists finitely and is nonsingular.

A2. $\mathrm{var}(Y \mid \boldsymbol{X}) > 0$, $E(Y^2) < \infty$ and the tail probability of Y satisfies that there exist some positive constants m_0 and m_1 such that for sufficiently large t, $\mathrm{P}(|Y| > t) \le m_0 \exp(-m_1 t)$.

A3. Assume that the quantities $\mathrm{q}_k(y;\theta) = (\partial^k/\partial\theta^k) Q\{y, F^{-1}(\theta)\}$, $k = 0, 1, \ldots$, exist finitely up to any order required. Suppose $\mathrm{q}_2(y;\theta) > 0$ for all $\theta \in \mathbb{R}$ and all y in the range of Y.

A4. $F(\cdot)$ is a bijection and F''' is continuous. Without loss of generality, assume $F'(\cdot) > 0$.

A5. For all j, $\mathbf{b}_j^{\mathrm{CR}}$ is an interior point of $\mathbb{R}_B^2 = \{(a,b) \in \mathbb{R}^2 : |a| \le B, |b| \le B\}$.

C. $\|\boldsymbol{\beta}_0^{(\mathrm{I})}\|_1 \le B'$.

D. Assume that the eigenvalues of $\mathbf{\Omega}_n$ and $\mathbf{H}_n$ are uniformly bounded away from 0; $\|\mathbf{H}_n^{-1}\mathbf{\Omega}_n\|_2$ is bounded away from ∞.

Proof of Theorem 1. Since $m(\boldsymbol{X}) = E(Y \mid \boldsymbol{X})$,

$$\begin{aligned}\operatorname{cov}\{E(Y \mid \boldsymbol{X}), X_j\} &= E\{E(Y \mid \boldsymbol{X})X_j\} - E\{E(Y \mid \boldsymbol{X})\}E(X_j)\\ &= E\{E(YX_j \mid \boldsymbol{X})\} - E(Y)E(X_j)\\ &= E(YX_j) - E(Y)E(X_j) = \operatorname{cov}(Y, X_j).\end{aligned}$$

It follows from Condition A3 that $Q\{y, F^{-1}(\theta)\}$ is strictly convex in θ. Therefore, the minimizer of (3.4) is the solution of the score equations of (3.4) which are given by

$$E\{\mathrm{q}_1(Y; \alpha_j^{\mathrm{CR}} + X_j\beta_j^{\mathrm{CR}})\} = E\Big\{\frac{(Y - \mu_j^{\mathrm{CR}})q''(\mu_j^{\mathrm{CR}})}{F'(\mu_j^{\mathrm{CR}})}\Big\} = 0,$$
$$E\{\mathrm{q}_1(Y; \alpha_j^{\mathrm{CR}} + X_j\beta_j^{\mathrm{CR}})X_j\} = E\Big\{\frac{X_j(Y - \mu_j^{\mathrm{CR}})q''(\mu_j^{\mathrm{CR}})}{F'(\mu_j^{\mathrm{CR}})}\Big\} = 0,$$

where $\mu_j^{\mathrm{CR}} = F^{-1}(\alpha_j^{\mathrm{CR}} + X_j\beta_j^{\mathrm{CR}})$.

We first show that if $\beta_j^{\mathrm{CR}} = 0$, then $\operatorname{cov}(Y, X_j) = 0$. When $\beta_j^{\mathrm{CR}} = 0$, μ_j^{CR} is a constant. Two score equations become

$$\mu_j^{\mathrm{CR}} = F^{-1}(\alpha_j^{\mathrm{CR}}) = E(Y), \quad E(X_jY) - \mu_j^{\mathrm{CR}}E(X_j) = 0,$$

which imply $\operatorname{cov}(Y, X_j) = 0$.

On the other side, if $\operatorname{cov}(Y, X_j) = 0$, it is easy to verify that $(F(E(Y)), 0)$ satisfies the score equations,

$$\begin{aligned}E\Big[\frac{\{Y - E(Y)\}q''(E(Y))}{F'(E(Y))}\Big] &= E\{Y - E(Y)\}\frac{q''(E(Y))}{F'(E(Y))} = 0,\\ E\Big[\frac{X_j\{Y - E(Y)\}q''(E(Y))}{F'(E(Y))}\Big] &= \operatorname{cov}(Y, X_j)\frac{q''(E(Y))}{F'(E(Y))} = 0.\end{aligned}$$

Therefore $\beta_j^{\mathrm{CR}} = 0$. □

Proof of Theorem 2. We first show part (i). The first two partial derivatives of $\ell_j^{\mathrm{CR}}(\alpha_j, \beta_j) = E[Q\{Y, F^{-1}(\alpha_j + X_j\beta_j)\}]$ with respect to α_j are given by

$$\begin{aligned}\frac{\partial \ell_j^{\mathrm{CR}}(\alpha_j, \beta_j)}{\partial \alpha_j} &= E\{\mathrm{q}_1(Y; \alpha_j + X_j\beta_j)\},\\ \frac{\partial^2 \ell_j^{\mathrm{CR}}(\alpha_j, \beta_j)}{\partial \alpha_j^2} &= E\{\mathrm{q}_2(Y; \alpha_j + X_j\beta_j)\}.\end{aligned}$$

By Condition A3, it follows that $\ell_j^{\mathrm{CR}}(\alpha_j, \beta_j)$ is convex in α_j. Then, for any given $\beta_j = b$, the minimizer of $\ell_j^{\mathrm{CR}}(\alpha_j, b)$ will be the solution to the following equation,

$$h_j(\alpha; b) = E\{\mathrm{q}_1(Y; \alpha + X_jb)\} = 0.$$

Denote by $\alpha_j(b)$ the solution of $h_j(\alpha; b) = 0$. Thus, $\alpha_j(b)$ is unique and a well-defined function of b.

Now we have an equivalent definition of β_j^{CR} given by

$$\beta_j^{\mathrm{CR}} = \arg\min_b \ell_j(b),$$

where $\ell_j(b) = E[Q\{Y, F^{-1}(\alpha_j(b) + X_j b)\}]$ and the first and second derivatives of $\ell_j(b)$ are given by

$$\begin{aligned} \ell_j'(b) = \frac{d\ell_j(b)}{db} &= E[\mathrm{q}_1\{Y; \alpha_j(b) + X_j b\}\{\alpha_j'(b) + X_j\}], \\ \ell_j''(b) = \frac{d^2\ell_j(b)}{db^2} &= E[\mathrm{q}_2\{Y; \alpha_j(b) + X_j b\}\{\alpha_j'(b) + X_j\}^2] > 0, \end{aligned}$$

which imply that $\ell_j(b)$ is convex in b and β_j^{CR} is unique and satisfies $\ell_j'(\beta_j^{\mathrm{CR}}) = 0$.

By the mean-value theorem,

$$\ell_j'(b) = \ell_j'(0) + b\ell_j''(b^*), \tag{6.1}$$

where b^* is between 0 and b. It can be shown that

$$\alpha_j(0) = F(E(Y)) \quad \text{for any } j = 1, \ldots, p_n.$$

Since $E\{\mathrm{q}_1(Y; \alpha_j(0))\} = 0$, we observe that

$$\ell_j'(0) = E\{\mathrm{q}_1(Y; \alpha_j(0))X_j\} = C_0 \operatorname{cov}(Y, X_j),$$

where $C_0 = q''(E(Y))/F'(E(Y))$. By Conditions A1 and A3, $|X_j| \le M$ and $\mathrm{q}_2(y; \theta) > 0$, we observe that for any b and $j = 1, \ldots, p_n$,

$$|\alpha_j'(b)| = \Big| - \frac{E\{\mathrm{q}_2(Y; \alpha_j(b) + X_j b)X_j\}}{E\{\mathrm{q}_2(Y; \alpha_j(b) + X_j b)\}} \Big| \le M.$$

Let $K_1 = \sup_{|\theta| \le (M+1)B} E\{\mathrm{q}_2(Y; \theta)\}$. Then for any $j = 1, \ldots, p_n$ and any $-B < b < B$,

$$\ell_j''(b) \le (2M)^2 K_1.$$

Let $c_1 = C_0/(4K_1M^2)$. By (6.1), for any $j = 1, \ldots, s_n$,

$$|\beta_j^{\mathrm{CR}}| \ge \frac{|C_0 \operatorname{cov}(Y, X_j)|}{4K_1M^2} = c_1|\operatorname{cov}(Y, X_j)| \ge c_1\mathcal{A}_n.$$

We now show part (ii). Let $K_2 = \inf_{|\theta| \le (M+1)B} E\{\mathrm{q}_2(Y; \theta)\}$. Similar to part (i), for any $j = 1, \ldots, p_n$ and any $-B < b < B$,

$$\ell_j''(b) \ge K_2 \operatorname{var}(X_j) = K_2.$$

Again by (6.1), for any $j = s_n + 1, \ldots, p_n$,

$$|\beta_j^{\mathrm{CR}}| \le |C_0 \operatorname{cov}(Y, X_j)|/(K_2\delta^2) = O(\mathcal{B}_n). \tag{6.2}$$

□

Proof of Theorem 3. To prove Theorem 3, the following lemma (which is Theorem 1 of [7]) will be needed.

Lemma 1. *Consider data* $\{(\boldsymbol{X}_i, Y_i) : i = 1, \ldots, n\}$ *which are* n i.i.d. *samples of* $(\boldsymbol{X}, Y) \in \mathcal{X} \times \mathcal{Y}$ *for some space* $\mathcal{X} \in \mathbb{R}^d$ *and* $\mathcal{Y} \in \mathbb{R}$. *A regression model for* $\boldsymbol{X}$ *and* Y *is assumed with loss function* $\ell(\boldsymbol{X}^T\boldsymbol{\beta}, Y)$. *Let*

$$\boldsymbol{\beta}_0 = \arg\min_{\boldsymbol{\beta}} E\{\ell(\boldsymbol{X}^T\boldsymbol{\beta}, Y)\}$$

be the population parameter. Assume that $\boldsymbol{\beta}_0$ *is an interior point of a sufficiently large, compact and convex set* $\mathbb{B} \in \mathbb{R}^d$. *Assume the following conditions on the model,*

(F1) *The matrix,*

$$I(\boldsymbol{\beta}) = E\Big[\Big\{\frac{\partial}{\partial \boldsymbol{\beta}}\ell(\boldsymbol{X}^T\boldsymbol{\beta}, Y)\Big\}\Big\{\frac{\partial}{\partial \boldsymbol{\beta}}\ell(\boldsymbol{X}^T\boldsymbol{\beta}, Y)\Big\}^T\Big],$$

exists finitely and is positive definite at $\boldsymbol{\beta} = \boldsymbol{\beta}_0$. *Moreover,*

$$\|I(\boldsymbol{\beta})\|_{\mathbb{B}} = \sup_{\boldsymbol{\beta}\in\mathbb{B}, \|\boldsymbol{x}\|_2=1} \|I(\boldsymbol{\beta})^{1/2}\boldsymbol{x}\|_2$$

exists.

(F2) *The function* $\ell(\boldsymbol{X}^T\boldsymbol{\beta}, Y)$ *satisfies the Lipschitz property with a positive constant* k_n,

$$|\ell(\boldsymbol{x}^T\boldsymbol{\beta}, y) - \ell(\boldsymbol{x}^T\boldsymbol{\beta}', y)|I_n(\boldsymbol{x}, y) \le k_n|\boldsymbol{x}^T\boldsymbol{\beta} - \boldsymbol{x}^T\boldsymbol{\beta}'|I_n(\boldsymbol{x}, y)$$

for any $\boldsymbol{\beta} \in \mathbb{B}$ *and* $\boldsymbol{\beta}' \in \mathbb{B}$, *where* $I_n(\boldsymbol{x}, y) = \mathrm{I}\{(\boldsymbol{x}, y) \in \Omega_n\}$ *with*

$$\Omega_n = \{(\boldsymbol{x}, y) : \|\boldsymbol{x}\|_\infty \le K_n,\ |y| \le K_n^*\}$$

for some sufficiently large positive constants K_n *and* K_n^*. *In addition, there exists a sufficiently large constant* C *such that with* $b_n = Ck_nV_n^{-1}(d/n)^{1/2}$ *and* V_n *given in Condition* (F3),

$$\sup_{\boldsymbol{\beta}\in\mathbb{B}, \|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|_2\le b_n} |E[\{\ell(\boldsymbol{X}^T\boldsymbol{\beta}, Y) - \ell(\boldsymbol{X}^T\boldsymbol{\beta}_0, Y)\}\{1 - I_n(\boldsymbol{X}, Y)\}]| \le o(d/n).$$

(F3) *The function* $\ell(\boldsymbol{X}^T\boldsymbol{\beta}, Y)$ *is convex in* $\boldsymbol{\beta}$, *satisfying*

$$E\{\ell(\boldsymbol{X}^T\boldsymbol{\beta}, Y) - \ell(\boldsymbol{X}^T\boldsymbol{\beta}_0, Y)\} \ge V_n\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2^2$$

for all $\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2 \le b_n$ *and some positive constant* V_n.

Then for any $t > 0$,

$$\mathrm{P}(\sqrt{n}\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2 \ge 16k_n(1 + t)/V_n) \le \exp(-2t^2/K_n^2) + n\mathrm{P}(\Omega_n^c).$$

We now prove Theorem 3. The main idea is to apply Lemma 1 by letting $d = 2$, $\boldsymbol{X} = \mathsf{X}_j$, $\boldsymbol{\beta} = \mathbf{b}_j$ and

$$\ell(\boldsymbol{X}^T\boldsymbol{\beta}, Y) = Q\{Y, F^{-1}(\mathsf{X}_j^T\mathbf{b}_j)\}.$$

So we need to show that the conditions (F1)–(F3) hold under our assumptions.

For condition (F1),

$$\begin{aligned} I_j(\mathbf{b}_j) &= E\{\mathrm{q}_1^2(Y; \mathsf{X}_j^T\mathbf{b}_j)\mathsf{X}_j\mathsf{X}_j^T\} \\ &= E\Big[\Big\{(Y-\mu_j)\frac{q''(\mu_j)}{F'(\mu_j)}\Big\}^2\mathsf{X}_j\mathsf{X}_j^T\Big], \end{aligned}$$

where $\mu_j = F^{-1}(\mathsf{X}_j^T\mathbf{b}_j)$. By assumptions A1 and A2, $I_j(\mathbf{b}_j)$ is bounded and positive definite at $\mathbf{b}_j = \mathbf{b}_j^{\mathrm{CR}}$.

For condition (F3), it suffices to show that $E\{\mathrm{q}_2(Y; \mathsf{x}_j^T\mathbf{b})\mathsf{X}_j\mathsf{X}_j^T\} \geq V\mathbf{I}_2$ for some $V > 0$, where $\mathbf{I}_2$ is the 2×2 identity matrix. Since $E(\mathsf{X}_j\mathsf{X}_j^T) = \mathbf{I}_2$ and $\mathrm{q}_2(Y; \mathsf{x}_j^T\mathbf{b}) > 0$, we only need to show that $E\{\mathrm{q}_2(Y; \mathsf{x}_j^T\mathbf{b})\} \geq V$ which is followed by

$$E\{\mathrm{q}_2(Y; \mathsf{x}_j^T\mathbf{b})\} \geq \mathrm{P}(|Y| \leq K)\xi,$$

where K is some sufficiently large positive constant such that $\mathrm{P}(|Y| \leq K) > 0$ and $\xi = \inf_{|y|\leq K, |\theta|\leq (M+1)B} \mathrm{q}_2(y;\theta)$.

Lastly for condition (F2), let $K_n = M$ and $K_n^* = \mathcal{A}_n^2 n$ and $V_n = V$. For $(\mathsf{x}, y) \in \Omega_n$, we have that, for any $\mathbf{b} \in \mathbb{B}$ and $\mathbf{b}' \in \mathbb{B}$,

$$\begin{aligned} &Q\{y, F^{-1}(\mathsf{x}^T\mathbf{b})\} - Q\{y, F^{-1}(\mathsf{x}^T\mathbf{b}')\} \\ &\quad = \{q(F^{-1}(\mathsf{x}^T\mathbf{b})) - q(F^{-1}(\mathsf{x}^T\mathbf{b}'))\} + y\{q'(F^{-1}(\mathsf{x}^T\mathbf{b})) - q'(F^{-1}(\mathsf{x}^T\mathbf{b}'))\} \\ &\qquad - \{F^{-1}(\mathsf{x}^T\mathbf{b})q'(F^{-1}(\mathsf{x}^T\mathbf{b})) - F^{-1}(\mathsf{x}^T\mathbf{b}')q'(F^{-1}(\mathsf{x}^T\mathbf{b}'))\} \\ &\quad = \{f_1(\mathsf{x}^T\mathbf{b}) - f_1(\mathsf{x}^T\mathbf{b}')\} + y\{f_2(\mathsf{x}^T\mathbf{b}) - f_2(\mathsf{x}^T\mathbf{b}')\} - \{f_3(\mathsf{x}^T\mathbf{b}) - f_3(\mathsf{x}^T\mathbf{b}')\}, \end{aligned}$$

where $f_1(t) = q(F^{-1}(t))$, $f_2(t) = q'(F^{-1}(t))$ and $f_3(t) = F^{-1}(t)q'(F^{-1}(t))$. Let

$$\begin{aligned} C_1 &= \sup_{|t|<(B+1)M} |f_1'(t)|, \\ C_2 &= \sup_{|t|<(B+1)M} |f_2'(t)| \quad \text{and} \\ C_3 &= \sup_{|t|<(B+1)M} |f_3'(t)|. \end{aligned}$$

Then

$$\begin{aligned} &|Q\{y, F^{-1}(\mathsf{x}^T\mathbf{b})\} - Q\{y, F^{-1}(\mathsf{x}^T\mathbf{b}')\}|I_n(\mathsf{x}, y) \\ &\quad \leq (C_1 + K_n^*C_2 + C_3)|\mathsf{x}^T\mathbf{b} - \mathsf{x}^T\mathbf{b}'|I_n(\mathsf{x}, y) \end{aligned}$$

which verifies the first part of condition (F2) with $k_n = C_1 + K_n^*C_2 + C_3$. For the second part of condition (F2), for all $j = 1, \ldots, p_n$,

$$|E[Q\{Y, F^{-1}(\mathsf{X}_j^T\mathbf{b})\} - Q\{Y, F^{-1}(\mathsf{X}_j^T\mathbf{b}_j^{\mathrm{CR}})\}\{1 - I_n(\mathsf{X}_j, Y)\}]|$$

$$\begin{aligned}
&\le E\{(C_1' + |Y|C_2' + C_3')\,\mathrm{I}(|Y| > K_n^*)\} \\
&\le \sqrt{E\{(C_1' + |Y|C_2' + C_3')^2\}}\sqrt{\mathrm{P}(|Y| > K_n^*)} \\
&\le O\{\exp(-m_1\mathcal{A}_n\sqrt{n}/2)\} = o(1/n),
\end{aligned}$$

where

$$\begin{aligned}
C_1' &= \sup_{|t|<(M+1)B} f_1(t) - \inf_{|t|<(M+1)B} f_1(t), \\
C_2' &= \sup_{|t|<(M+1)B} f_2(t) - \inf_{|t|<(M+1)B} f_2(t), \\
C_3' &= \sup_{|t|<(M+1)B} f_3(t) - \inf_{|t|<(M+1)B} f_3(t).
\end{aligned}$$

Then by Lemma 1, for any t and any $j = 1, \ldots, p_n$,

$$\mathrm{P}\{\sqrt{n}\|\widehat{\mathbf{b}}_j^{\mathrm{CR}} - \mathbf{b}_j^{\mathrm{CR}}\|_2 \ge 16k_n(1+t)/V_n\} \le \exp(-2t^2/K_n^2) + nm_0\exp(-m_1K_n^*).$$

Taking $(1+t) = \mathcal{A}_n\sqrt{n}V_n/(16k_n)$ and noting $K_n^* = \mathcal{A}_n^2 n$ yield

$$\mathrm{P}(\|\widehat{\mathbf{b}}_j^{\mathrm{CR}} - \mathbf{b}_j^{\mathrm{CR}}\|_2 \ge \mathcal{A}_n) \le \exp(-c_2\mathcal{A}_n^2 n) + nm_0\exp(-m_1\mathcal{A}_n^2 n),$$

where c_2 is a suitable positive constant. The desired result follows from using $\mathrm{P}(|\widehat{\beta}_j^{\mathrm{CR}} - \beta_j^{\mathrm{CR}}| \ge \mathcal{A}_n) \le \mathrm{P}(\|\widehat{\mathbf{b}}_j^{\mathrm{CR}} - \mathbf{b}_j^{\mathrm{CR}}\|_2 \ge \mathcal{A}_n)$ and Bonferroni inequality. □

Proof of Theorem 4. Define $\boldsymbol{\beta}^{\mathrm{CR}} = (\beta_1^{\mathrm{CR}}, \ldots, \beta_{p_n}^{\mathrm{CR}})^T$. We first prove that

$$\|\boldsymbol{\beta}^{\mathrm{CR}}\|_2^2 = \sum_{j=1}^{p_n} |\beta_j^{\mathrm{CR}}|^2 = O\{\lambda_{\max}(\Sigma)\}.$$

Let $C_4 = C_0/(K_2M^2)$. By (6.2), for all $j = 1, \ldots, p_n$,

$$\begin{aligned}
|\beta_j^{\mathrm{CR}}| &\le C_4|\mathrm{cov}(X_j, Y)| \\
&= C_4|E[\{X_j - E(X_j)\}E(Y \mid \boldsymbol{X})]| \\
&= C_4|E[\{X_j - E(X_j)\}F^{-1}(\beta_{0;0} + \boldsymbol{X}^T\boldsymbol{\beta}_0)]|.
\end{aligned}$$

On the other hand, we have

$$\begin{aligned}
&|\{X_j - E(X_j)\}\{F^{-1}(\beta_{0;0} + \boldsymbol{X}^T\boldsymbol{\beta}_0) - F^{-1}(\beta_{0;0} + E(\boldsymbol{X}^T\boldsymbol{\beta}_0))\}| \\
&\quad \le C_5|\{X_j - E(X_j)\}\{\boldsymbol{X} - E(\boldsymbol{X})\}^T\boldsymbol{\beta}_0|,
\end{aligned}$$

where $C_5 = \sup_{|t|<B'M+B}(F^{-1})'(t)$. Again, by taking the expectation on both sides and then putting it into the vector form, we have

$$\begin{aligned}
&\|E[\{\boldsymbol{X} - E(\boldsymbol{X})\}\{F^{-1}(\beta_{0;0} + \boldsymbol{X}^T\boldsymbol{\beta}_0)\}]\|_2^2 \\
&\quad = \|E[\{\boldsymbol{X} - E(\boldsymbol{X})\}\{F^{-1}(\beta_{0;0} + \boldsymbol{X}^T\boldsymbol{\beta}_0) - F^{-1}(\beta_{0;0} + E(\boldsymbol{X}^T\boldsymbol{\beta}_0))\}]\|_2^2 \\
&\quad \le C_5^2\|E[\{\boldsymbol{X} - E(\boldsymbol{X})\}\{\boldsymbol{X} - E(\boldsymbol{X})\}^T]\boldsymbol{\beta}_0\|_2^2 = C_5^2\|\Sigma\boldsymbol{\beta}_0\|_2^2 \\
&\quad \le C_5^2\lambda_{\max}(\Sigma)\|\Sigma^{1/2}\boldsymbol{\beta}_0\|_2^2.
\end{aligned}$$

Since $\|\Sigma^{1/2}\boldsymbol{\beta}_0\|_2^2 = \mathrm{var}(\boldsymbol{X}^T\boldsymbol{\beta}_0) \le M'$, it follows that

$$\|\boldsymbol{\beta}^{\mathrm{CR}}\|_2^2 \le C_4^2 C_5^2 M' \lambda_{\max}(\Sigma).$$

Finally by the above result, the cardinality of the set $\{j : |\beta_j^{\mathrm{CR}}| > \epsilon \mathcal{A}_n\}$ can not be bigger than $O(\mathcal{A}_n^{-2}\lambda_{\max})$ for any $\epsilon > 0$. The desired result can be easily seen from Theorem 3. □

Proof of Proposition 1. The oracle property was obtained by [23] for BD estimation when $s_n^5/n \to 0$ and $s_n(p_n - s_n) = o(n)$. If we can prove that when the number $p_n' = |\widehat{\mathcal{M}}|$ of variables selected in the screening step can be set appropriately, the event $\{\mathcal{M}_n \subseteq \widehat{\mathcal{M}}\}$ happens with probability approaching 1, then the conclusion should follow.

By Theorem 4, we have

$$\mathrm{P}(|\widehat{\mathcal{M}}_{\gamma_n}| \le O(\mathcal{A}_n^{-2}\lambda_{\max}(\Sigma))) = 1 - o(1).$$

Since we choose p_n' such that $\lambda_{\max}(\Sigma)/(p_n'\mathcal{A}_n^2) = o(1)$, it is equivalent to choosing another appropriate $\gamma_n' \le \gamma_n$. Thus, $\widehat{\mathcal{M}}_{\gamma_n} \subseteq \widehat{\mathcal{M}}$ and by Corollary 1,

$$\mathrm{P}(\mathcal{M}_n \subseteq \widehat{\mathcal{M}}) \ge \mathrm{P}(\mathcal{M}_n \subseteq \widehat{\mathcal{M}}_{\gamma_n}) = 1 - o(1).$$

The oracle property can be expressed as P(event O) $\to 1$ as $n \to \infty$. Since $s_n = o(n^{1/5})$ and $p_n' = o(n/s_n)$, we have

$$\mathrm{P}(\text{event O} \mid \mathcal{M}_n \subseteq \widehat{\mathcal{M}}) = 1 - o(1).$$

The desired result follows from

$$\mathrm{P}(\text{event O}) \ge \mathrm{P}(\text{event O} \mid \mathcal{M}_n \subseteq \widehat{\mathcal{M}})\mathrm{P}(\mathcal{M}_n \subseteq \widehat{\mathcal{M}}) = 1 - o(1).$$

□

Acknowledgements

The authors thank the Editor, George Michailidis, Associate Editor and anonymous referees for insightful comments. C. Zhang's research is supported by the NSF grants DMS–1308872 and DMS–1521761, and Wisconsin Alumni Research Foundation.

References

[1] Alon, U., Barkai, N., Notterman, D.A., Gish, K., Ybarra, S., Mack, D., and Levine, A.J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl. Acad. Sci. USA*, **96**, 6745–6750.

[2] Brègman, L. M. (1967). A relaxation method of finding a common point of convex sets and its application to the solution of problems in convex programming. *U.S.S.R. Comput. Math. and Math. Phys.*, **7**, 620–631. MR0215617

[3] Candes, E. and Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *Ann. Statist.*, **35**, 2313–2351. MR2382644

[4] Chang, J., Tang, C. Y. and Wu, Y. (2013). Marginal empirical likelihood and sure indpendence feature screening. *Ann. Statist.*, **41**, 2123–2148. MR3127860

[5] Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc. Ser. B*, **70**, 849–911. MR2530322

[6] Fan, J. (1997). Comment on "Wavelets in statistics: A review". *A. Antoniadis. J. Italian Statisit. Soc.*, **6**, 131–138.

[7] Fan, J. and Song, R. (2010). Sure independence screening in generalized linear models with NP-dimensionality. *Ann. Statist.*, **38**, 3567–3604. MR2766861

[8] Frank, I. E. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, **35**, 109–148.

[9] Friedman, J., Hastie, T., Höfling, H. and Tibshirani, R. (2007). Pathwise coordinate optimization. *Ann. Appl. Statist.*, **1**, 302–332. MR2415737

[10] Friedman, J., Hastie, T. and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, **33**, 1–22.

[11] Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfiel, C.D. and Lander, E.S. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–536.

[12] Hastie, T., Tibshirani, R. and Friedman, J. (2001). *The Elements of Statistical Learning*, Springer. MR1851606

[13] Li, G., Peng, H., Zhang, J. and Zhu, L. (2012). Robust rank correlation based screening. *Ann. Statist.*, **40**, 1846–1877. MR3015046

[14] Li, R., Zhong, W. and Zhu, L. (2012). Feature screening via distance correlation learning. *J. Amer. Statist. Assoc.*, **107**, 1129–1139. MR3010900

[15] McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models*, Chapman Hall CRC, Boca Raton. MR3223057

[16] McCullagh, P. (1983). Quasi-likelihood functions. *Ann. Statist.*, **11**, 59–67. MR0684863

[17] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, **58**, 267–288. MR1379242

[18] van de Geer, S. (2008). High-dimensional generalized linear models and the lasso. *Ann. Statist.*, **36**, 614–645. MR2396809

[19] Vapnik, V. (1996). *The Nature of Statistical Learning Theory*. Springer-Verlag, New York. MR1719582

[20] Wasserman, L. and Roeder, K. (2009). High-dimensional variable selection. *Ann. Statist.*, **37**, 2178–2201. MR2543689

[21] Zhang, C. H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.*, **38**, 894–942. MR2604701

[22] Zhang, C. M., Jiang, Y. and Shang, Z. (2009). New aspects of Bregman divergence in regression and classification with parametric and nonparametric estimation. *Canad. J. Statist.*, **37**, 119–139. MR2509465

[23] Zhang, C. M., Jiang, Y. and Chai, Y. (2010). Penalized Bregman divergence for large-dimensional regression and classification. *Biometrika*, **97**, 551–566. MR2672483

[24] Zhu, L., Li, L., Li, R. and Zhu, L. (2011). Model-free feature screening for ultrahigh-dimensional data. *J. Amer. Statist. Assoc.*, **106**, 1464–1475. MR2896849