Submitted to the Annals of Statistics

SURE INDEPENDENCE SCREENING IN GENERALIZED LINEAR MODELS WITH NP-DIMENSIONALITY*

By Jianqing Fan and Rui Song

Princeton University and Colorado State University

Ultrahigh dimensional variable selection plays an increasingly important role in contemporary scientific discoveries and statistical research. Among others, Fan and Lv (2008) propose an independent screening framework by ranking the marginal correlations. They showed that the correlation ranking procedure possesses a sure independence screening property within the context of the linear model with Gaussian covariates and responses. In this paper, we propose a more general version of the independent learning with ranking the maximum marginal likelihood estimates or the maximum marginal likelihood itself in generalized linear models. We show that the proposed methods, with Fan and Lv (2008) as a very special case, also possess the sure screening property with vanishing false selection rate. The conditions under which that the independence learning possesses a sure screening is surprisingly simple. This justifies the applicability of such a simple method in a wide spectrum. We quantify explicitly the extent to which the dimensionality can be reduced by independence screening, which depends on the interactions of the covariance matrix of covariates and true parameters. Simulation studies are used to illustrate the utility of the proposed approaches. In addition, we

*Supported in part by Grant NSF grants DMS-0714554 and DMS-0704337. The bulk of the work was conducted when Rui Song was a postdoctoral research fellow at Princeton University. The authors would like to thank the associate editor and two referees for their constructive comments that improve the presentation and the results of the paper.

AMS 2000 subject classifications: Primary 68Q32, 62J12; secondary 62E99, 60F10

Keywords and phrases: generalized linear models, independent learning, sure independent screening, variable selection.

establish an exponential inequality for the quasi-maximum likelihood estimator which is useful for high-dimensional statistical learning.

1. Introduction. Ultrahigh dimensional regression problem is a significant feature in many areas of modern scientific research using quantitative measurements such as microarrays, genomics, proteomics, brain images and genetic data. For example, in studying the associations between phenotypes such as height and cholesterol level and genotypes, it can involve millions of SNPs; in disease classification using microarray data, it can use thousands of expression profiles; and dimensionality grows rapidly when interactions are considered. Such a demand from applications brings a lot of challenge to statistical inference, as the dimension p can grow much faster than the sample size n such that many models are not even identifiable. By non-polynomial dimensionality or simply NP-dimensionality, we mean $\log p = O(n^a)$ for some $a > 0$. We will also loosely refer it to as an ultrahigh dimensionality. The phenomenon of noise accumulation in high-dimensional regression has also been observed by statisticians and computer scientists. See Fan and Lv (2008) and Fan and Fan (2008) for a comprehensive review and references therein. When dimension p is ultrahigh, it is often assumed that only a small number of variables among predictors $X_1, \ldots, X_p$ contribute to the response, which leads to the sparsity of the parameter vector $\boldsymbol{\beta}$. As a consequence, variable selection plays a prominent role in high dimensional statistical modeling.

Many variable selection techniques for various high dimensional statistical models have been proposed. Most of them are based on the penalized pseudo-likelihood approach, to name a few, the bridge regression in Frank and Friedman (1993), the LASSO in Tibshirani (1996), the SCAD and other folded-concave penalty in Fan and Li (2001), the Dantzig selector in Candes and Tao (2007) and their related methods (Zou, 2006; Zou and Li, 2008). Theoretical studies of these methods concentrate on the persistency (Greenshtein and Ritov, 2004; van de Geer, 2008), consistency and oracle properties (Fan and Li, 2001; Zou, 2006). However, in ultrahigh dimensional statistical learning problems, these methods may not perform well due to the simultaneous challenges of computational expediency, statistical accuracy and algorithmic stability (Fan et al., 2009).

Fan and Lv (2008) proposed a sure independent screening (SIS) method to select important variables in ultrahigh dimensional linear models. Their proposed two-stage procedure can deal with the aforementioned three challenges better than other methods. See also Huang et al. (2008) for a related study based on a marginal bridge regression. Fan and Lv (2008) showed that the correlation ranking of features possesses a sure independence screening (SIS) property under certain conditions, that is, with probability very close to 1, the independence screening technique retains all of the important variables in the model. However, the SIS procedure in Fan and Lv (2008) only restricts to the ordinary linear models and their technical arguments depend heavily on the joint normality assumptions and can not easily be extended even within the context of a linear model. This limits significantly its use in practice which excludes categorical variables. Huang et al. (2008) also investigate the marginal bridge regression in the ordinary linear model and their arguments depend also heavily on the explicit expressions of the least-square estimator and bridge regression. This calls for research on SIS procedures in more general models and under less restrictive assumptions.

In this paper, we consider an independence learning by ranking the maximum marginal likelihood estimator (MMLE) or maximum marginal likelihood itself for generalized linear models. That is, we fit p marginal regressions by maximizing the marginal likelihood with response Y and the marginal covariate $X_i,\ i = 1, \ldots, p$ (and the intercept) each time. The magnitude of the absolute values of the MMLE can preserve the non-sparsity information of the joint regression models, provided that the true values of the marginal likelihood preserve the non-sparsity of the joint regression models and that the MMLE estimates the true values of the marginal likelihood uniformly well. The former holds under a surprisingly simple condition, whereas the latter requires a development of uniform convergence over NP-dimensional marginal likelihoods. Hall et al. (2009) used a different marginal utility, derived from an empirical likelihood point of view. Hall and Miller (2009) proposed a generalized correlation ranking, which allows nonlinear regression. Both papers proposed an interesting bootstrap method to assess the authority of the selected features.

As the MMLE or maximum likelihood ranking is equivalent to the marginal correlation ranking in the ordinary linear models, our work can thus be considered as an important extension of SIS in Fan and Lv (2008), where the joint normality of the response and covariates is imposed. Moreover, our results improve over these in Fan and Lv (2008) in at least three aspects. Firstly, we establish a new framework for having SIS properties, which does not build on the normality assumption even in the linear model setting. Secondly, while it is not obvious (and could be hard) to generalize the proof of Fan and Lv (2008) to more complicated models, in the current framework, the SIS procedure can be applied to the generalized linear models and possibly other models. Thirdly, our results can easily be applied to the generalized correlation ranking (Hall and Miller, 2009) and other rankings based on a group of marginal variables.

Fitting marginal models to a joint regression is a type of model misspecification (White, 1982), since we drop out most covariates from the model fitting. In this paper, we establish a nonasymptotic tail probability bound for the MMLE under model misspecifications, which is beyond the traditional asymptotic framework of model misspecification and of interest in its own right. As a practical screening method, independent screening can miss variables that are marginally weakly correlated with the response variables, but jointly highly important to the response variables, and also rank some jointly unimportant variables too high by using marginal methods. Fan and Lv (2008) and Fan et al. (2009) develop iteratively conditional screening and selection methods to make the procedures robust and practical. The former focuses on ordinary linear models and the latter improves the idea in the former and expands significantly the scope of applicability, including generalized linear models.

The SIS property can be achieved as long as the surrogate, in this case, the marginal utility, can preserve the non-sparsity of the true parameter values. With a similar idea, Fan et al. (2009) proposed a SIS procedure for generalized linear models, by sorting the maximum likelihood functions, which is a type of "marginal likelihood ratio" ranking, whereas the MMLE can be viewed as a Wald type of statistic. The two methods are equivalent in terms of sure screening properties in our proposed framework. This will be demonstrated in our paper. The key technical challenge in the maximum marginal likelihood ranking is that the signal can even be weaker than the noise. We overcome this technical difficulty by using the invariance property of ranking under monotonic transforms.

The rest of the paper is organized as follows. In Section 2, we briefly introduce the setups of the generalized linear models. The SIS procedure is presented in Section 3. In Section 4, we provide an exponential bound for quasi maximum likelihood estimator. The SIS properties of the MMLE learning are presented in Section 5. In Section 6, we formulate the marginal likelihood screening and show the SIS property. Some simulation results are presented in Section 7. A summary of our findings and discussions is in Section 8. The detailed proofs are relegated to Section 9.

2. Generalized Linear Models. Assume that the random scalar Y is from an exponential family with the probability density function taking the canonical form

$$f_Y(y;\theta) = \exp\left\{y\theta - b(\theta) + c(y)\right\}, \tag{1}$$

for some known functions $b(\cdot)$, $c(\cdot)$ and unknown function θ. Here we do not consider the dispersion parameter as we only model the mean regression. We can easily introduce a dispersion parameter in (1) and the results continue to hold. The function θ is usually called the canonical or natural parameter. The mean response is $b'(\theta)$, the first derivative of $b(\theta)$ with respect to θ. We consider the problem of estimating a $(p+1)$-vector of parameter $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)$ from the following generalized linear model

$$E(Y|\mathbf{X} = \mathbf{x}) = b'(\theta(\mathbf{x})) = g^{-1}(\sum_{j=0}^{p} \beta_j x_j), \tag{2}$$

where $\mathbf{x} = \{x_0, x_1, \ldots, x_p\}^T$ is a (p+1)-dimensional covariate and $x_0 = 1$ represents the intercept. If g is the canonical link, i.e., $g = (b')^{-1}$, then $\theta(x) = \sum_{j=0}^{p} \beta_j x_j$. We focus on the canonical link function in this paper for simplicity of presentation.

Assume that the observed data $\{(\mathbf{X}_i, Y_i), i = 1, \ldots, n\}$ are i.i.d. copies of $(\mathbf{X}, Y)$, where the covariate $\mathbf{X} = (X_0, X_1, \ldots, X_p)$ is a $(p+1)$-dimensional random vector and $X_0 = 1$. We allow p to grow with n and denote it as p_n whenever needed.

We note that the ordinary linear model $Y = \mathbf{X}^T\boldsymbol{\beta} + \varepsilon$ with $\varepsilon \sim N(0,1)$ is a special case of model (2), by taking $g(\mu) = \mu$ and $b(\theta) = \theta^2/2$. When the design matrix $\mathbf{X}$ is standardized, the ranking by the magnitude of the marginal correlation is in fact the same as the ranking by the magnitude of the maximum marginal likelihood estimator (MMLE). Next we propose an independence screening method to GLIM based on the MMLE. We also assume that the covariates are standardized to have mean zero and standard deviation one:

$$EX_j = 0, \quad \text{and} \quad EX_j^2 = 1, \quad j = 1, \cdots, p_n.$$

3. Independence Screening with MMLE. Let $\mathcal{M}_\star = \{1 \le j \le p_n : \beta_j^\star \ne 0\}$ be the true sparse model with non-sparsity size $s_n = |\mathcal{M}_\star|$, where $\boldsymbol{\beta}^\star = (\beta_0^\star, \beta_1^\star, \ldots, \beta_{p_n}^\star)$ denotes the true value. In this paper, we refer to marginal models as fitting models with componentwise covariates. The maximum marginal likelihood estimator (MMLE) $\hat{\boldsymbol{\beta}}_j^M$, for $j = 1, \ldots, p_n$, is defined as the minimizer of the componentwise regression:

$$\hat{\boldsymbol{\beta}}_j^M = (\hat{\beta}_{j,0}^M, \hat{\beta}_j^M) = \text{argmin}_{\beta_0,\beta_j} \mathbb{P}_n l(\beta_0 + \beta_j X_j, Y),$$

where $l(Y;\theta) = -\left[\theta Y - b(\theta) - \log c(Y)\right]$ and $\mathbb{P}_n f(X,Y) = n^{-1}\sum_{i=1}^n f(X_i, Y_i)$ is the empirical measure. This can be rapidly computed and its implementation is robust, avoiding numerical instability in NP-dimensional problems. We correspondingly define the population version of the minimizer of the componentwise regression,

$$\boldsymbol{\beta}_j^M = (\beta_{j,0}^M, \beta_j^M) = \text{argmin}_{\beta_0,\beta_j} El(\beta_0 + \beta_j X_j, Y), \text{ for } j = 1, \ldots, p_n,$$

where E denotes the expectation under the true model.

We select a set of variables

$$\widehat{\mathcal{M}}_{\gamma_n} = \{1 \le j \le p_n : |\hat{\beta}_j^M| \ge \gamma_n\}, \tag{3}$$

where γ_n is a predefined threshold value. Such an independence learning ranks the importance of features according to their magnitude of marginal regression coefficients. With an independence learning, we dramatically decrease the dimension of the parameter space from p_n (possibly hundreds of thousands) to a much smaller number by choosing a large γ_n, hence the computation is much more feasible. Although the interpretations and implications of the marginal models are biased from the joint model, the non-sparse information about the joint model can be passed along to the marginal model under a mild condition. Hence it is suitable for the purpose of variable screening. Next we will show under certain conditions that the sure screening property holds, i.e., the set $\mathcal{M}_\star$ belongs to $\widehat{\mathcal{M}}_{\gamma_n}$ with probability one asymptotically, for an appropriate choice of γ_n. To accomplish this, we need the following technical device.

4. An exponential bound for QMLE.

In this section, we obtain an exponential bound for the quasi-MLE (QMLE), which will be used in the next section. Since this result holds under very general conditions and is of self-interest, in the following we make a more general description of the model and its conditions.

Consider data $\{\mathbf{X}_i, Y_i\}$, $i = 1, ..., n$ are n i.i.d. samples of $(\mathbf{X}, Y) \in \mathcal{X} \times \mathcal{Y}$ for some space $\mathcal{X}$ and $\mathcal{Y}$. A regression model for $\mathbf{X}$ and Y is assumed with quasi-likelihood function $-l(\mathbf{X}^T\boldsymbol{\beta}, Y)$. Here Y and $\mathbf{X} = (X_1, \ldots, X_q)^T$ represent the response and the q–dimensional covariate vector, which may include both discrete and continuous components and the dimensionality can also depend on n. Let

$$\boldsymbol{\beta}_0 = \text{argmin}_{\boldsymbol{\beta}} El(\mathbf{X}^T\boldsymbol{\beta}, Y)$$

be the population parameter. Assume that $\boldsymbol{\beta}_0$ is an interior point of a sufficiently large, compact and convex set $\mathbf{B} \in \mathbf{R}^q$. The following conditions on the model are needed:

A. The Fisher information

$$I(\boldsymbol{\beta}) = E\left\{\left[\frac{\partial}{\partial\boldsymbol{\beta}} l(\mathbf{X}^T\boldsymbol{\beta}, Y)\right]\left[\frac{\partial}{\partial\boldsymbol{\beta}} l(\mathbf{X}^T\boldsymbol{\beta}, Y)\right]^T\right\}$$

is finite and positive definite at $\boldsymbol{\beta} = \boldsymbol{\beta}_0$. Moreover, $\|I(\boldsymbol{\beta})\|_{\mathbf{B}} = \sup_{\boldsymbol{\beta}\in\mathbf{B},\ \|\mathbf{x}\|=1} \|I(\boldsymbol{\beta})^{1/2}\mathbf{x}\|$ exists, where $\|\cdot\|$ is the Euclidean norm.

B. The function $l(\mathbf{x}^T\boldsymbol{\beta}, y)$ satisfies the Lipschitz property with positive constant k_n:

$$|l(\mathbf{x}^T\boldsymbol{\beta}, y) - l(\mathbf{x}^T\boldsymbol{\beta}', y)|I_n(\mathbf{x}, y) \leq k_n|\mathbf{x}^T\boldsymbol{\beta} - \mathbf{x}^T\boldsymbol{\beta}'|I_n(\mathbf{x}, y),$$

for $\boldsymbol{\beta}, \boldsymbol{\beta}' \in \mathbf{B}$, where $I_n(\mathbf{x}, y) = I((\mathbf{x}, y) \in \Omega_n)$ with

$$\Omega_n = \{(\mathbf{x}, y) :\ \|\mathbf{x}\|_\infty \leq K_n,\ |y| \leq K_n^\star\},$$

for some sufficiently large positive constants K_n and $K_n^\star$, and $\|\cdot\|_\infty$ being the supremum norm. In addition, there exists a sufficiently large constant C such that with $b_n = Ck_nV_n^{-1}(q/n)^{1/2}$ and V_n given in Condition C

$$\sup_{\boldsymbol{\beta}\in\mathbf{B},\ \|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|\leq b_n} |E[l(\mathbf{X}^T\boldsymbol{\beta}, Y) - l(\mathbf{X}^T\boldsymbol{\beta}_0, Y)](1 - I_n(\mathbf{X}, Y))| \leq o(q/n).$$

where V_n is the constant given in Condition C.

C. The function $l(\mathbf{X}^T\boldsymbol{\beta}, Y)$ is convex in $\boldsymbol{\beta}$, satisfying

$$E(l(\mathbf{X}^T\boldsymbol{\beta}, Y) - l(\mathbf{X}^T\boldsymbol{\beta}_0, Y)) \geq V_n\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|^2,$$

for all $\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| \leq b_n$ and some positive constants V_n.

Condition A is analogous to assumption A6(b) of White (1982) and assumption R_s in Fahrmeir and Kaufmann (1986). It ensures the identifiability and the existence of the QMLE and is satisfied for many examples of generalized linear models. Conditions A and C are overlapped but not the same.

We now establish an exponential bound for the tail probability of the QMLE:

$$\hat{\boldsymbol{\beta}} = \operatorname{argmin}_{\boldsymbol{\beta}} \mathbb{P}_n \, l(\mathbf{X}^T \boldsymbol{\beta}, Y)$$

The idea of the proof is to connect $\sqrt{n}\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|$ to the tail of certain empirical processes and utilize the convexity and Lipschitz continuities.

THEOREM 1. *Under Conditions A–C, it holds that for any $t > 0$,*

$$P\left(\sqrt{n}\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| \geq 16k_n(1+t)/V_n\right) \leq \exp(-2t^2/K_n^2) + nP(\Omega_n^c).$$

5. Sure Screening Properties with MMLE. In this section, we introduce a new framework for establishing the sure screening property with MMLE in the canonical exponential family (1). We divide into three sections to present our findings.

5.1. *Population Aspect.* As fitting marginal regressions to a joint regression is a type of model misspecification, an important question would be: at what level the model information is preserved. Specifically for screening purpose, we are interested in the preservation of the non-sparsity from the joint regression to the marginal regression. This can be summarized into the following two questions. First, for the sure screening purpose, if a variable X_j is jointly important ($\beta_j^\star \neq 0$), will (and under what conditions) it still be marginally important ($\beta_j^M \neq 0$)? Second, for the model selection consistency purpose, if a variable X_j is jointly unimportant ($\beta_j^\star = 0$), will it still be marginally unimportant ($\beta_j^M = 0$)? We aim to answer these two questions in this section.

The following theorem reveals that the marginal regression parameter is in fact a measurement of the correlation between the marginal covariate and the mean response function.

THEOREM 2. *For $j = 1, \ldots, p_n$, the marginal regression parameters $\beta_j^M = 0$ if and only if $cov(b'(\mathbf{X}^T \boldsymbol{\beta}^\star), X_j) = 0$.*

By using the fact that that $\mathbf{X}^T \boldsymbol{\beta}^\star = \beta_0^\star + \sum_{j \in \mathcal{M}_\star} X_j \beta_j^\star$, we can easily show the following corollary.

COROLLARY 1. *If the partial orthogonality condition holds, i.e., $\{X_j,\ j \notin \mathcal{M}_\star\}$ is independent of $\{X_i,\ i \in \mathcal{M}_\star\}$, then $\beta_j^M = 0$, for $j \notin \mathcal{M}_\star$.*

This partial orthogonality condition is essentially the assumption made in Huang et al. (2008) who showed the model selection consistency in the special case with the ordinary linear model and bridge regression. Note that $\mathrm{cov}(b'(\mathbf{X}^T \boldsymbol{\beta}^\star), X_j) = \mathrm{cov}(Y, X_j)$. A necessary condition for sure screening is that the important variables X_j with $\beta_j^\star \neq 0$ are correlated with the response, which usually holds. When they are correlated with the response, by Theorem 2, $\beta_j^M \neq 0$, for $j \in \mathcal{M}_\star$. In other words, the marginal model pertains the information about the important variables in the joint model. This is the theoretical basis for the sure independence screening. On the other hand, if the partial orthogonality condition in Corollary 1 holds, then $\beta_j^M = 0$ for $j \notin \mathcal{M}_\star$. In this case, there exists a threshold γ_n such that the marginally selected model is model selection consistent:

$$\min_{j \in \mathcal{M}_\star} |\beta_j^M| \geq \gamma_n, \qquad \max_{j \notin \mathcal{M}_\star} |\beta_j^M| = 0.$$

To have a sure screening property based on the sample version (3), we need

$$\min_{j \in \mathcal{M}_\star} |\beta_j^M| \geq O(n^{-\kappa}),$$

for some $\kappa < 1/2$ so that the marginal signals are stronger than the stochastic noise. The following theorem shows that this is possible.

THEOREM 3. *If* $|cov(b'(\mathbf{X}^T\boldsymbol{\beta}^\star), X_j)| \geq c_1 n^{-\kappa}$ *for* $j \in \mathcal{M}_\star$ *and a positive constant* $c_1 > 0$*, then there exists a positive constant* c_2 *such that*

$$\min_{j \in \mathcal{M}_\star} |\beta_j^M| \geq c_2 n^{-\kappa},$$

provided that $b''(\cdot)$ *is bounded or*

$$EG(a|X_j|)|X_j|I(|X_j| \geq n^\eta) \leq dn^{-\kappa}, \text{ for some } 0 < \eta < \kappa,$$

and some sufficiently small positive constants a *and* d*, where* $G(|x|) = \sup_{|u| \leq |x|} |b'(u)|$.

Note that for the normal and Bernoulli distribution, $b''(\cdot)$ is bounded, whereas for the Poisson distribution, $G(|x|) = \exp(|x|)$ and Theorem 3 requires the tails of X_j to be light. Under some additional conditions, we will show in the proof of Theorem 5 that

$$\sum_{j=1}^p |\beta_j^M|^2 = O(\|\mathbf{\Sigma}\boldsymbol{\beta}^\star\|^2) = O(\lambda_{\max}(\mathbf{\Sigma})),$$

where $\mathbf{\Sigma} = \mathrm{var}(\mathbf{X})$, and $\lambda_{\max}(\mathbf{\Sigma})$ is its maximum eigenvalue. The first equality requires some efforts to prove, whereas the second equality follows easily from the assumption

$$\mathrm{var}(\mathbf{X}^T\boldsymbol{\beta}^\star) = \boldsymbol{\beta}^{\star T}\mathbf{\Sigma}\boldsymbol{\beta}^\star = O(1).$$

The implication of this result is that there can not be too many variables that have marginal coefficient $|\beta_j^M|$ that exceeds certain thresholding level. That achieves the sparsity in final selected model.

When the covariates are jointly normally distributed, the condition of Theorem 3 can be further simplified.

PROPOSITION 1. *Suppose that X and Z are jointly normal with mean zero and standard deviation 1. For a strictly monotonic function $f(\cdot)$, $cov(X, Z) = 0$ if and only if $cov(X, f(Z)) = 0$, provided the latter covariance exists. In addition,*

$$|cov(X, f(Z))| \geq |\rho| \inf_{|x| \leq c|\rho|} |g'(x)| \; EX^2 I(|X| \leq c),$$

for any $c > 0$, where $\rho = EXZ$, $g(x) = Ef(x + \varepsilon)$ with $\varepsilon \sim N(0, 1 - \rho^2)$.

The above proposition shows that the covariance of X and $f(Z)$ can be bounded from below by the covariance between X and Z, namely

$$|cov(X, f(Z))| \geq d|\rho|, \qquad d = \inf_{|x| \leq c} |g'(x)| \; EX^2 I(|X| \leq c),$$

in which $d > 0$ for a sufficiently small c. The first part of the proposition actually holds when the conditional density $f(z|x)$ of Z given X is a monotonic likelihood family (Bickel and Doksum, 2001) when x is regarded as a parameter. By taking $Z = \mathbf{X}^T \boldsymbol{\beta}^\star$, a direct application of Theorem 2 is that $\beta_j^M = 0$ if and only if

$$\mathrm{cov}(\mathbf{X}^T \boldsymbol{\beta}^\star, X_j) = 0,$$

provided that $\mathbf{X}$ is jointly normal, since $b'(\cdot)$ is an increasing function. Furthermore, if

$$|\mathrm{cov}(\mathbf{X}^T \boldsymbol{\beta}^\star, X_j)| \geq c_0 n^{-\kappa}, \qquad \kappa < 1/2, \tag{4}$$

for some positive constant c_0, a minimum condition required even for the least-squares model (Fan and Lv, 2008), then by the second part of Proposition 1, we have

$$|\mathrm{cov}(b'(\mathbf{X}^T\boldsymbol{\beta}^\star), X_j)| \geq c_1 n^{-\kappa},$$

for some constant c_1. Therefore, by Theorem 2, there exists a positive constant c_2 such that

$$|\beta_j^M| \geq c_2 n^{-\kappa}.$$

In other words, (4) suffices to have marginal signals that are above the maximum noise level.

5.2. *Uniform convergence and sure screening.* To establish the SIS property of MMLE, a key point is to establish the uniform convergence of the MMLEs. That is, to control the maximum noise level relative to the signal. Next we establish the uniform convergence rate for the MMLEs and sure screening property of the method in (3). The former will be useful in controlling the size of the selected set.

Let $\boldsymbol{\beta}_j = (\beta_{j,0}, \beta_j)^T$ denote the two-dimensional parameter and $\mathbf{X}_j = (1, X_j)^T$. Due to the concavity of the log-likelihood in GLIM with the canonical link, $El(\mathbf{X}_j^T\boldsymbol{\beta}_j, Y)$ has a unique minimum over $\boldsymbol{\beta}_j \in \mathcal{B}$ at an interior point $\boldsymbol{\beta}_j^M = (\beta_{j,0}^M, \beta_j^M)^T$, where $\mathcal{B} = \{|\beta_{j,0}^M| \leq B, |\beta_j^M| \leq B\}$ is a square with the width B over which the marginal likelihood is maximized. The following is an updated version of Conditions A, B and C for each marginal regression and two additional conditions for the covariates and the population parameters:

A'. The marginal Fisher information: $I_j(\boldsymbol{\beta}_j) = E\left\{b''(\mathbf{X}_j^T\boldsymbol{\beta}_j)\mathbf{X}_j\mathbf{X}_j^T\right\}$ is finite and positive definite at $\boldsymbol{\beta}_j = \boldsymbol{\beta}_j^M$, for $j = 1, \ldots, p_n$. Moreover, $\|I_j(\boldsymbol{\beta}_j)\|_{\mathcal{B}}$ is bounded from above.

B'. The second derivative of $b(\theta)$ is continuous and positive. There exists an $\varepsilon_1 > 0$ such that for all $j = 1, \dots, p_n$,

$$\sup_{\boldsymbol{\beta}\in\mathcal{B},\ \|\boldsymbol{\beta}-\boldsymbol{\beta}_j^M\|\le\varepsilon_1} |Eb(\mathbf{X}_j^T\boldsymbol{\beta})I(|X_j| > K_n)| \le o(n^{-1}).$$

C'. For all $\boldsymbol{\beta}_j \in \mathcal{B}$, we have $E(l(\mathbf{X}_j^T\boldsymbol{\beta}_j, Y) - l(\mathbf{X}_j^T\boldsymbol{\beta}_j^M, Y)) \ge V\|\boldsymbol{\beta}_j - \boldsymbol{\beta}_j^M\|^2$, for some positive V, bounded from below uniformly over $j = 1, \dots, p_n$.

D. There exists some positive constants m_0, m_1, s_0, s_1 and α, such that for sufficiently large t,

$$P(|X_j| > t) \le (m_1 - s_1)\exp\{-m_0 t^\alpha\}, \quad \text{for } j = 1, \dots, p_n,$$

and that

$$E\exp(b(\mathbf{X}^T\boldsymbol{\beta}^\star + s_0) - b(\mathbf{X}^T\boldsymbol{\beta}^\star)) + E\exp(b(\mathbf{X}^T\boldsymbol{\beta}^\star - s_0) - b(\mathbf{X}^T\boldsymbol{\beta}^\star)) \le s_1.$$

E. The conditions in Theorem 3 hold.

Conditions A'–C' are satisfied in a lot of examples of generalized linear models, such as linear regression, logistic regression and Poisson regression. Note that the second part of Condition D ensures the tail of the response variable Y to be exponentially light, as shown in the following lemma:

LEMMA 1. *If Condition D holds, for any $t > 0$,*

$$P(|Y| \ge m_0 t^\alpha / s_0) \le s_1 \exp(-m_0 t^\alpha).$$

Let $k_n = b'(K_n B + B) + m_0 K_n^\alpha / s_0$. Then Condition B holds for exponential family (1) with $K_n^\star = m_0 K_n^\alpha / s_0$. The Lipschitz constant k_n is bounded for the logistic regression, since Y and $b'(\cdot)$ are bounded. The following theorem gives a uniform convergence result of MMLEs and a sure screening property. Interestingly, the sure screening property does not directly depend on the property of the covariance matrix of the covariates such as the growth of its operator norm. This is an advantage over using the full likelihood.

THEOREM 4. *Suppose that Conditions A', B', C' and D hold.*

(i) If $n^{1-2\kappa}/(k_n^2 K_n^2) \to \infty$, *then for any* $c_3 > 0$, *there exists a positive constant* c_4 *such that*

$$P\left(\max_{1\le j\le p_n} |\hat{\beta}_j^M - \beta_j^M| \ge c_3 n^{-\kappa}\right)$$
$$\le \quad p_n\{\exp(-c_4 n^{1-2\kappa}/(k_n K_n)^2) + nm_1 \exp(-m_0 K_n^{\alpha})\}.$$

(ii) If, in addition, Condition E holds, then by taking $\gamma_n = c_5 n^{-\kappa}$ *with* $c_5 \le c_2/2$, *we have*

$$P\Big(\mathcal{M}_\star \subset \widehat{\mathcal{M}}_{\gamma_n}\Big) \ge 1 - s_n\{\exp(-c_4 n^{1-2\kappa}/(k_n K_n)^2) + nm_1 \exp(-m_0 K_n^{\alpha})\},$$

where $s_n = |\mathcal{M}_\star|$, *the size of non-sparse elements.*

REMARK 1. *If we assume that* $\min_{j\in\mathcal{M}_*} |cov(b'(\mathbf{X}^T\boldsymbol{\beta}^\star), X_j)| \ge c_1 n^{-\kappa+\delta}$ *for any* $\delta > 0$, *then one can take* $\gamma_n = cn^{-\kappa+\delta/2}$ *for any* $c > 0$ *in Theorem 4. This is essentially the thresholding used in Fan and Lv (2008).*

Note that when $b'(\cdot)$ is bounded as the Bernoulli model, k_n is a finite constant. In this case, by balancing the two terms in the upper bound of Theorem 4(i), the optimal order of K_n is given by

$$K_n = n^{(1-2\kappa)/(\alpha+2)},$$

and

$$P\left(\max_{1\le j\le p_n} |\hat{\beta}_j^M - \beta_j^M| \ge c_3 n^{-\kappa}\right) = O\left\{p_n \exp(-c_4 n^{(1-2\kappa)\alpha/(\alpha+2)})\right\},$$

for a positive constant c_4. When the covariates X_j are bounded, then k_n and K_n can be taken as finite constants. In this case,

$$P\left(\max_{1\le j\le p_n}|\hat{\beta}_j^M-\beta_j^M|\ge c_3 n^{-\kappa}\right)\le O\left\{p_n\exp(-c_4 n^{1-2\kappa})\right\}.$$

In both aforementioned cases, the tail probability in Theorem 4 is exponentially small. In other words, we can handle the NP-dimensionality:

$$\log p_n = o\left(n^{(1-2\kappa)\alpha/(\alpha+2)}\right),$$

with $\alpha=\infty$ for the case of bounded covariates.

For the ordinary linear model, $k_n = B(K_n+1)+K_n^\alpha/(2s_0)$ and by taking the optimal order of $K_n = n^{(1-2\kappa)/A}$ with $A=\max(\alpha+4, 3\alpha+2)$, we have

$$P\left(\max_{1\le j\le p_n}|\hat{\beta}_j^M-\beta_j^M|> c_3 n^{-\kappa}\right)= O\left\{p_n\exp(-c_4 n^{(1-2\kappa)\alpha/A})\right\}.$$

When the covariates are normal, $\alpha=2$ and our result is weaker than that given in Fan and Lv (2008) who permits $\log p_n = o(n^{1-2\kappa})$ whereas Theorem 4 can only handle $\log p_n = o(n^{(1-2\kappa)/4})$. However, we allow non-normal covariate and other error distributions.

The above discussion applies to the sure screening property given in Theorem 4(ii). It is only the size of non-sparse elements s_n that matters for the purpose of sure screening, not the dimensionality p_n.

5.3. *Controlling false selection rates.* After applying the variable screening procedure, the question arrives naturally how large the set $\widehat{\mathcal{M}}_{\gamma_n}$ is. In other words, has the number of variables been actually reduced by the independence learning? In this section, we aim to answer this question.

A simple answer to this question is the ideal case in which

$$\mathrm{cov}(b'(\mathbf{X}^T\boldsymbol{\beta}^\star), X_j) = o(n^{-\kappa}), \qquad \text{for } j\notin\mathcal{M}_\star.$$

In this case, under some mild conditions, we can show (see the proof of Theorem 3) that

$$\max_{j \notin \mathcal{M}_\star} |\beta_j^M| = o(n^{-\kappa}).$$

This, together with Theorem 4(i) shows that

$$\max_{j \notin \mathcal{M}_\star} |\hat{\beta}_j^M| \leq c_3 n^{-\kappa}, \qquad \text{for any } c_3 > 0,$$

with probability tending to one if the probability in Theorem 4(i) tends to zero. Hence, by the choice of γ_n as in Theorem 4(ii), we can achieve model selection consistency:

$$P(\widehat{\mathcal{M}}_{\gamma_n} = \mathcal{M}_\star) = 1 - o(1).$$

This kind of condition was indeed implied by the condition in Huang et al. (2008) in the special case with ordinary linear model using the bridge regression who draw a similar conclusion.

We now deal with the more general case. The idea is to bound the size of the selected set (3) by using the fact $\text{var}(Y)$ is bounded. This usually implies $\text{var}(\mathbf{X}^T\boldsymbol{\beta}^\star) = \boldsymbol{\beta}^{\star T}\boldsymbol{\Sigma}\boldsymbol{\beta}^\star = O(1)$. We need the following additional conditions:

F. The variance $\text{var}(\mathbf{X}^T\boldsymbol{\beta}^\star)$ is bounded from above and below.

G. Either $b''(\cdot)$ is bounded or $\mathbf{X}_M = (X_1, \cdots, X_{p_n})^T$ follows an elliptically contoured distribution, i.e.,

$$\mathbf{X}_M = \boldsymbol{\Sigma}_1^{1/2} R \mathbf{U},$$

and $|Eb'(\mathbf{X}^T\boldsymbol{\beta}^\star)(\mathbf{X}^T\boldsymbol{\beta}^\star - \beta_0^\star)|$ is bounded, where $\mathbf{U}$ is uniformly distributed on the unit sphere in p-dimensional Euclidean space, independent of the nonnegative random variable R, and $\boldsymbol{\Sigma}_1 = \text{var}(\mathbf{X}_M)$.

Note that $\boldsymbol{\Sigma} = \text{diag}(0, \boldsymbol{\Sigma}_1)$ in Condition G', since the covariance matrices differ only in the intercept term. Hence, $\lambda_{\max}(\boldsymbol{\Sigma}) = \lambda_{\max}(\boldsymbol{\Sigma}_1)$. The following result is about the size of $\widehat{\mathcal{M}}_{\gamma_n}$.

THEOREM 5. *Under Conditions A', B', C', D, F and G, we have for any $\gamma_n = c_5 n^{-2\kappa}$, there exists a c_4 such that*

$$\begin{aligned} & P[|\widehat{\mathcal{M}}_{\gamma_n}| \le O\{n^{2\kappa}\lambda_{\max}(\mathbf{\Sigma})\}] \\ \ge \quad & 1 - p_n\{\exp(-c_4 n^{1-2\kappa}/(k_n K_n)^2) + n m_1 \exp(-m_0 K_n^{\alpha})\}. \end{aligned}$$

The right hand side probability has been explained in Section 5.2. From the proof of Theorem 5, we actually show that the number of selected variables is of order $\|\mathbf{\Sigma}\boldsymbol{\beta}^{\star}\|^2/\gamma_n^2$, which is further bounded by $O\{n^{2\kappa}\lambda_{\max}(\mathbf{\Sigma})\}$ using $\mathrm{var}(\mathbf{X}^T\boldsymbol{\beta}^{\star}) = O(1)$. Interestingly, while the sure screening property does not depend on the behavior of $\mathbf{\Sigma}$, the number of selected variables is affected by how correlated the covariates are. When $n^{2\kappa}\lambda_{\max}(\mathbf{\Sigma})/p \to 0$, the number of selected variables are indeed negligible comparing to the original size. In this case, the percent of falsely discovered variables is of course negligible. In particular, when $\lambda_{\max}(\mathbf{\Sigma}) = O(n^{\tau})$, the size of selected variable is of order $O(n^{2\kappa+\tau})$. This is of the same order as in Fan and Lv (2008) for the multiple regression model with the Gaussian data who needs additional condition that $2\kappa + \tau < 1$. Our result is an extension of Fan and Lv (2008) even in this very specific case without the condition $2\kappa+\tau < 1$. In addition, our result is more intuitive: the number of selected variables is related to $\lambda_{\max}(\mathbf{\Sigma})$, or more precisely $\|\mathbf{\Sigma}\boldsymbol{\beta}^{\star}\|^2$ and the thresholding parameter γ_n.

6. A likelihood ratio screening. In a similar variable screening problem with generalized linear models, Fan et al. (2009) suggest to screen the variables by sorting the marginal likelihood. This method can be viewed as a marginal likelihood ratio screening, as it builds on the increments of the log-likelihood. In this section we show that the likelihood ratio screening is equivalent to the MMLE screening in the sense that they both possess the sure screening property and that the number of selected variables of the two methods are of the same order of magnitude.

We first formulate the marginal likelihood screening procedure. Let

$$L_{j,n} = \mathbb{P}_n\Big\{l(\hat{\beta}_0^M, Y) - l(\mathbf{X}_j^T\hat{\boldsymbol{\beta}}_j^M, Y)\Big\},\ j = 1, \dots, p_n,$$

and $\mathbf{L}_n = (L_{1,n}, \dots, L_{p_n,n})^T$, where $\hat{\beta}_0^M = \mathrm{argmin}_{\beta_0}\mathbb{P}_n l(\beta_0, Y)$. Correspondingly, let

$$L_j^\star = E\Big\{l(\beta_0^M, Y) - l(\mathbf{X}_j^T\boldsymbol{\beta}_j^M, Y)\Big\},\ j = 1, \dots, p_n,$$

and $\mathbf{L}^\star = (L_1^\star, \dots, L_{p_n}^\star)^T$, where $\beta_0^M = \mathrm{argmin}_{\beta_0} El(\beta_0, Y)$. It can be shown that $EY = b'(\beta_0^M)$ and that $\overline{Y} = b'(\hat{\beta}_0^M)$, where $\overline{Y}$ is the sample average. We sort the vector $\mathbf{L}_n$ in a descent order and select a set of variables

$$\widehat{\mathcal{N}}_{\nu_n} = \{1 \le j \le p_n : L_{j,n} \ge \nu_n\},$$

where ν_n is a predefined threshold value. Such an independence learning ranks the importance of features according to their marginal contributions to the magnitudes of the likelihood function. The marginal likelihood screening and the MMLE screening share a common computation procedure as solving p_n optimization problems over a two dimensional parameter space. Hence the computation is much more feasible than traditional variable selection methods.

Compared with MMLE screening, where the information utilized is only the magnitudes of the estimators, the marginal likelihood screening incorporates the whole contributions of the features to the likelihood increments: both the magnitudes of the estimators and their associated variation. Under current condition (Condition C'), the variance of the MMLEs are at a comparable level (through the magnitude of V, an implication of the convexity of the objective functions), and the two screening methods are equivalent. Otherwise, if V depends on n, the marginal likelihood screening can still preserve the non-sparsity structure, while the MMLE screening may need some corresponding adjustments, which we will not discuss in detail as it is beyond the scope of the current paper.

Next we will show that the sure screening property holds under certain conditions. Similarly to the MMLE screening, we first build the theoretical foundation of the marginal likelihood screening. That is, the marginal likelihood increment is also a measurement of the correlation between the marginal covariate and the mean response function.

THEOREM 6. *For $j = 1, \dots, p_n$, the marginal likelihood increment $L_j^\star = 0$ if and only if $cov(b'(\mathbf{X}^T\boldsymbol{\beta}^\star), X_j) = 0$.*

As a direct corollary of Theorem 1, we can easily show the following corollary for the purpose of model selection consistency.

COROLLARY 2. *If the partial orthogonality condition in Corollary 1 holds, then $L_j^\star = 0$, for $j \notin \mathcal{M}_\star$.*

We can also strengthen the result of minimum signals as follows. On the other hand, we also show that the total signals can not be too large. That is, there can not be too many signals that exceed certain threshold.

THEOREM 7. *Under the conditions in Theorem 3 and the Condition C', we have*

$$\min_{j\in\mathcal{M}_\star} |L_j^\star| \geq c_6 n^{-2\kappa},$$

for some positive constant c_6, provided that $|cov(b'(\mathbf{X}^T\boldsymbol{\beta}^\star), X_j)| \geq c_1 n^{-\kappa}$ for

$j \in \mathcal{M}_\star$. If, in addition, Conditions F and G hold, then

$$\|\mathbf{L}^\star\| = O(\|\boldsymbol{\beta}^M\|^2) = O(\|\boldsymbol{\Sigma}\boldsymbol{\beta}^\star\|^2) = O(\lambda_{\max}(\boldsymbol{\Sigma})).$$

The technical challenge is that the stochastic noise $\|\mathbf{L}_n - \mathbf{L}^\star\|_\infty$ is usually of the order of $O(n^{-2\kappa} + n^{-1/2}\log p_n)$, which can be an order of magnitude larger than the signals given in Theorem 7, unless $\kappa < 1/4$. Nevertheless, by a different trick that utilizes the fact that ranking is invariant under a strict monotonic transform, we are able to demonstrate the sure screening independence property for $\kappa < 1/2$.

THEOREM 8. *Suppose that Conditions A', B', C' and D, E and F hold. Then, by taking $\nu_n = c_7 n^{-2\kappa}$ for a sufficiently small $c_7 > 0$, there exists a $c_8 > 0$ such that*

$$P\Big(\mathcal{M}_\star \subset \widehat{\mathcal{N}}_{\nu_n}\Big) \geq 1 - s_n\{\exp(-c_8 n^{1-2\kappa}/(k_n K_n)^2) + nm_1 \exp(-m_0 K_n^\alpha)\}.$$

Similarly to the MMLE screening, we can control the size of $\widehat{\mathcal{N}}_{\nu_n}$ as follows. For simplicity of the technical argument, we focus only on the case where $b''(\cdot)$ is bounded.

THEOREM 9. *Under Conditions A', B', C', D, F and G, if $b''(\cdot)$ is bounded, then we have*

$$\begin{aligned} & P[|\widehat{\mathcal{N}}_{\nu_n}| \leq O\{n^{2\kappa}\lambda_{\max}(\boldsymbol{\Sigma})\}] \\ \geq \quad & 1 - p_n\{\exp(-c_8 n^{1-2\kappa}/(k_n K_n)^2) + nm_1 \exp(-m_0 K_n^\alpha)\}. \end{aligned}$$

7. Numerical Results. In this section, we present several simulation examples to evaluate the performance of SIS procedure with generalized linear models. It was demonstrated in Fan and Lv (2008) and Fan et al. (2009)

that independent screening is a fast but crude method of reducing the dimensionality to a more moderate size. Some methodological extensions include iterative SIS (ISIS) and multi-stage procedures, such as SIS-SCAD and SIS-LASSO, can be applied to perform the final variable selection and parameter estimation simultaneously. Extensive simulations on these procedures were also presented in Fan et al. (2009). To avoid repetition, in this paper, we focus on the vanilla SIS, with the aim to evaluate the sure screening property and to demonstrate some factors influencing the false selection rate. We vary the sample size from 80 to 600 for different scenarios to gauge the difficulties of the simulation models. The following three configurations with $p = 2000$, 5000 and 40000 predictor variables are considered for generating the covariates $\mathbf{X} = (X_1, \cdots, X_p)^T$:

S1. The covariates are generated according to

$$X_j = \frac{\varepsilon_j + a_j \varepsilon}{\sqrt{1 + a_j^2}}, \tag{5}$$

where ε and $\{\varepsilon_j\}_{j=1}^{[p/3]}$ are i.i.d. standard normal random variables, $\{\varepsilon_j\}_{j=[p/3]+1}^{[2p/3]}$ are i.i.d. and follow a double exponential distributions with location parameter zero and scale parameter one, and $\{\varepsilon_j\}_{j=[2p/3]+1}^{p}$ are i.i.d. and follow a mixture normal distribution with two components $N(-1, 1)$, $N(1, 0.5)$ and equal mixture proportion. The covariates are standardized to be mean zero and variance one. The constants $\{a_j\}_{j=1}^{q}$ are the same and chosen such that the correlation $\rho = \mathrm{corr}(X_i, X_j) = 0,\ 0.2,\ 0.4,\ 0.6$ and 0.8, among the first q variables, and $a_j = 0$ for $j > q$. The parameter q is also related to the overall correlation in the covariance matrix. We will present the numerical results with $q = 15$ for this setting.

S2. The covariates are also generated from (5), except that $\{a_j\}_{j=1}^{q}$ are

Table 1

The median of the 200 empirical maximum eigenvalues, with its robust estimate of SD in the parenthesis, of the corresponding sample covariance matrices of covariates based 200 simulations with partial combinations of $n = 80, 300, 600$, $p = 2000, 5000, 40000$ *and* $q = 15, 50$ *in the first two settings (S1 and S2).*

(p, n)	Setting	ρ 0	0.2	0.4	0.6	0.8
(40000,80)	S1 (q=15)	549.9(1.4)	550.1(1.4)	550.1(1.3)	550.1(1.3)	550.1(1.4)
(40000,80)	S2 (q=50)	550.0(1.4)	550.1(1.4)	550.4(1.5)	552.9(1.8)	558.5(2.4)
(40000,300)	S1 (q=15)	157.3(0.4)	157.4(0.4)	157.4(0.4)	157.4(0.3)	157.7(0.4)
(40000,300)	S2 (q=50)	157.4(0.4)	157.5(0.4)	160.9(1.2)	168.2(1.0)	176.9(1.0)
(5000,300)	S1 (q=15)	25.68(0.2)	25.68(0.2)	26.18(0.2)	27.99(0.4)	30.28(0.4)
(5000,300)	S1 (q=50)	25.69(0.1)	29.06(0.5)	37.98(0.7)	47.49(0.7)	57.17(0.5)
(2000,600)	S1 (q=15)	7.92(0.07)	8.32(0.15)	10.5(0.3)	13.09(0.3)	15.79(0.2)
(2000,600)	S1 (q=50)	7.93(0.07)	14.62(0.40)	23.95(0.7)	33.90(0.6)	43.56(0.5)
(2000,600)	S2 (q=50)	7.93(0.07)	14.62(0.40)	23.95(0.7)	33.90(0.6)	43.56(0.5)

i.i.d. normal random variables with mean a and variance 1 and $a_j = 0$ for $j > q$. The value of a is taken such that $E\text{corr}(X_i, X_j) = 0,\ 0.2,\ 0.4,\ 0.6$ and 0.8, among the first q variables. The simulation results to be presented for this setting use $q = 50$.

S3. Let $\{X_j\}_{j=1}^{p-50}$ be i.i.d. standard normal random variables and

$$X_k = \sum_{j=1}^{s} X_j(-1)^{j+1}/5 + \sqrt{25-s}/5\varepsilon_k, \qquad k = p-49, \cdots, p,$$

where $\{\varepsilon_k\}_{k=p-49}^{p}$ are standard normally distributed.

Table 1 summarizes the median of the empirical maximum eigenvalues of the covariance matrix and its robust estimate of the standard deviation (RSD) based 200 simulations in the first two settings (S1 and S2) with partial combinations of sample size $n = 80, 300, 600$, $p = 2000, 5000, 40000$ and $q = 15, 50$. RSD is the interquantile range (IQR) divided by 1.34. The empirical maximum eigenvalues are always larger than their population version, depending on the realizations of the design matrix. The empirical minimum eigenvalue is always zero, and the empirical condition numbers for the sample covariance matrix are infinite, since $p > n$. Generally, the empirical maximum eigenvalues increase as the correlation parameters ρ, q, the numbers of covariates p increase, and/or the sample sizes n decrease.

With these three settings, we aim to illustrate the behaviors of the two SIS procedures under different correlation structures. For each simulation and each model, we apply the two SIS procedures, the marginal MLE and the marginal likelihood ratio methods, to screen variables. The minimum model size (MMS) required for each method to have a sure screening, i.e. to contain the true model $\mathcal{M}_\star$, is used as a measure of the effectiveness of a screening method. This avoids the issues of choosing the thresholding parameter. To gauge the difficulty of the problem, we also include the LASSO and the SCAD as references for comparison when $p = 2000$ and 5000. The smaller p is used due to the computation burden of the LASSO and the SCAD. In addition, as demonstrated in our simulation results, they do not perform well when p is large. Our initial intension is to demonstrate that the simple SIS does not perform much worse than the far more complicated procedures like the LASSO and the SCAD. To our surprise, the SIS can even outperform those more complicated methods in terms of variable screening. Again, we record the MMS for the LASSO and the SCAD for each simulation and each model, which does not depend on the choice of regularization parameters. When the LASSO or the SCAD can not recover the true model even with the smallest regularization parameter, we average the model size with the smallest regularization parameter and p. These interpolated MMS' are presented with italic font in Tables 3–5 and 9 to distinct from the real MMS. Results for logistic regressions and linear regressions are presented in the following two subsections.

7.1. *Logistic Regressions.* The generated data $(\mathbf{X}_1^T, Y_1), \ldots, (\mathbf{X}_n^T, Y_n)$ are n i.i.d. copies of a pair $(\mathbf{X}^T, Y)$, in which the conditional distribution of the response Y given $\mathbf{X} = \mathbf{x}$ is binomial distribution with probability of success $p(\mathbf{x}) = \exp(\mathbf{x}^T \boldsymbol{\beta}^\star)/[1 + \exp(\mathbf{x}^T \boldsymbol{\beta}^\star)]$. We vary the size of the nonsparse set of coefficients as $s = 3,\ 6,\ 12,\ 15$ and 24. For each simulation, we evaluate each method by summarizing the median minimum model size (MMMS) of the selected models as well as its associated RSD, which is the associated interquartile range (IQR) divided by 1.34. The results, based on 200 simulations for each scenario are recorded in the second and third panel of Table 2 and the second panel of Tables 3–5. Specifically, Table 2 records the MMMS and the associated RSD for SIS under the first two settings when $p = 40000$, while Table 3–5 record these results for SIS, the LASSO and the SCAD when $p = 2000$ and 5000 under Settings 1, 2 and 3 respectively. The true parameters are also recorded in each corresponding table.

To demonstrate the difficulty of our simulated models, we depict the distribution, among 200 simulations, of the minimum $|t|$-statistics of s estimated regression coefficients in the oracle model in which the statistician does not know that all variables are statistically significant. This shows the difficulty to recover all significant variables even in the oracle model with the minimum model size s. The distribution was computed for each setting and scenario but only a few selected settings are shown presented in Figure 1. In fact, the distributions under Setting 1 are very similar to those under Setting 2 when the same q value is taken. It can be seen that the magnitude of the minimum $|t|$-statistics is reasonably small and getting smaller as the correlation within covariates (measured by ρ and q) increases, sometimes achieving three decimals. Given such small signal-to-noise ratio in the oracle models, the difficulty of our simulation models is a self-evident even if

TABLE 2
The MMMS and the associated RSD (in the parenthesis) of the simulated examples for logistic regressions in the first two settings (S1 and S2) when $p = 40000$.

ρ	n	SIS-MLR	SIS-MMLE	n	SIS-MLR	SIS-MMLE
	Setting 1, $q = 15$					
	$s = 3$, $\boldsymbol{\beta}^\star = (1, 1.3, 1)^T$			$s = 6$, $\boldsymbol{\beta}^\star = (1, 1, 3, 1, \ldots)^T$		
0	300	87.5(381)	89(375)	300	47(164)	50(170)
0.2	200	3(0)	3(0)	300	6(0)	6(0)
0.4	200	3(0)	3(0)	300	7(1)	7(1)
0.6	200	3(1)	3(1)	300	8(1)	8(2)
0.8	200	4(1)	4(1)	300	9(3)	9(3)
	$s = 12$, $\boldsymbol{\beta}^\star = (1, 1.3, \ldots)^T$			$s = 15$, $\boldsymbol{\beta}^\star = (1, 1.3, \ldots)^T$		
0	500	297(589)	302.5(597)	600	350(607)	359.5(612)
0.2	300	13(1)	13(1)	300	15(0)	15(0)
0.4	300	14(1)	14(1)	300	15(0)	15(0)
0.6	300	14(1)	14(1)	300	15(0)	15(0)
0.8	300	14(1)	14(1)	300	15(0)	15(0)
	Setting 2, $q = 50$					
	$s = 3$, $\boldsymbol{\beta}^\star = (1, 1.3, 1)^T$			$s = 6$, $\boldsymbol{\beta}^\star = (1, 1.3, 1, \ldots)^T$		
0	300	84.5(376)	88.5(383)	500	6(1)	6(1)
0.2	300	3(0)	3(0)	500	6(0)	6(0)
0.4	300	3(0)	3(0)	500	6(1)	6(1)
0.6	300	3(1)	3(1)	500	8.5(4)	9(5)
0.8	300	5(4)	5(4)	500	13.5(8)	14(8)
	$s = 12$, $\boldsymbol{\beta}^\star = (1, 1.3, \ldots)^T$			$s = 15$, $\boldsymbol{\beta}^\star = (1, 1.3, \ldots)^T$		
0	600	77(114)	78.5(118)	800	46(82)	47(83)
0.2	500	18(7)	18(7)	500	26(6)	26(6)
0.4	500	25(8)	25(10)	500	34(7)	33(8)
0.6	500	32(9)	31(8)	500	39(7)	38(7)
0.8	500	36(8)	35(9)	500	40(6)	42(7)

the signals seem not that small.

TABLE 3
The MMMS and the associated RSD (in the parenthesis) of the simulated examples for logistic regressions in Setting 1 (S1) when p=5000 and q=15. The values with italic font indicate that the LASSO or the SCAD can not recover the true model even with smallest regularization parameter and are estimated.

ρ	n	SIS-MLR	SIS-MMLE	LASSO	SCAD
		$s = 3$, $\boldsymbol{\beta}^\star = (1, 1.3, 1)^T$			
0	300	3(0)	3(0)	3(1)	3(1)
0.2	300	3(0)	3(0)	3(0)	3(0)
0.4	300	3(0)	3(0)	3(0)	3(0)
0.6	300	3(0)	3(0)	3(0)	3(1)
0.8	300	3(1)	3(1)	4(1)	4(1)
		$s = 6$, $\boldsymbol{\beta}^\star = (1, 1.3, 1, 1.3, 1, 1.3)^T$			
0	200	8(6)	9(7)	7(1)	7(1)
0.2	200	18(38)	20(39)	9(4)	9(2)
0.4	200	51(77)	64.5(76)	20(10)	16.5(6)
0.6	300	77.5(139)	77.5(132)	20(13)	19(9)
0.8	400	306.5(347)	313(336)	86(40)	70.5(35)
		$s = 12$, $\boldsymbol{\beta}^\star = (1, 1.3, \dots)^T$			
0	300	297.5(359)	300(361)	72.5(3704)	12(0)
0.2	300	13(1)	13(1)	12(1)	12(0)
0.4	300	14(1)	14(1)	14(1861)	13(1865)
0.6	300	14(1)	14(1)	*2552(85)*	12(3721)
0.8	300	14(1)	14(1)	*2556(10)*	12(3722)
		$s = 15$, $\boldsymbol{\beta}^\star = (3, 4, \dots)^T$			
0	300	479(622)	482(615)	69.5(68)	15(0)
0.2	300	15(0)	15(0)	16(13)	15(0)
0.4	300	15(0)	15(0)	38(3719)	15(3720)
0.6	300	15(0)	15(0)	*2555(87)*	15(1472)
0.8	300	15(0)	15(0)	*2552(8)*	15(1322)

The MMMS and RSD with fixed correlation (S1) and random correlation (S2) are comparable under the same q. As the correlation increases and/or the nonsparse set size increases, the MMMS and the associated RSD usually increase for all SIS, the LASSO and the SCAD. Among all the designed scenarios of Settings 1 and 2, SIS performs well, while the LASSO and the SCAD occasionally fail under very high correlations and relatively large nonsparse set size (s=12, 15 and 24). Interestingly, correlation within covariates can sometimes help SIS reduce the false selection rate, as it can

TABLE 4
The MMMS and the associated RSD (in the parenthesis) of the simulated examples for logistic regressions in Setting 2 (S2) when p=2000 and q=50. The values with italic font have the same meaning as Table 2.

ρ	n	SIS-MLR	SIS-MMLE	LASSO	SCAD
			$s = 3$, $\boldsymbol{\beta}^\star = (3, 4, 3)^T$		
0	200	3(0)	3(0)	3(0)	3(0)
0.2	200	3(0)	3(0)	3(0)	3(0)
0.4	200	3(0)	3(0)	3(0)	3(1)
0.6	200	3(1)	3(1)	3(1)	3(1)
0.8	200	5(5)	5.5(5)	6(4)	6(4)
			$s = 6$, $\boldsymbol{\beta}^\star = (3, -3, 3, -3, 3, -3)^T$		
0	200	8(6)	9(7)	7(1)	7(1)
0.2	200	18(38)	20(39)	9(4)	9(2)
0.4	200	51(77)	64.5(76)	20(10)	16.5(6)
0.6	300	77.5(139)	77.5(132)	20(13)	19(9)
0.8	400	306.5(347)	313(336)	86(40)	70.5(35)
			$s = 12$, $\boldsymbol{\beta}^\star = (3, 4, \ldots)^T$		
0	600	13(6)	13(7)	12(0)	12(0)
0.2	600	19(6)	19(6)	13(1)	13(2)
0.4	600	32(10)	30(10)	18(3)	17(4)
0.6	600	38(9)	38(10)	22(3)	22(4)
0.8	600	38(7)	39(8)	*1071(6)*	*1042(34)*
			$s = 24$, $\boldsymbol{\beta}^\star = (3, 4, \ldots)^T$		
0	600	180(240)	182(238)	35(9)	31(10)
0.2	600	45(4)	45(4)	35(27)	32(24)
0.4	600	46(3)	47(2)	*1099(17)*	1093(1456)
0.6	600	48(2)	48(2)	*1078(5)*	*1065(23)*
0.8	600	48(1)	48(1)	*1072(4)*	*1067(13)*

increase the marginal signals. It is notable that the LASSO and the SCAD usually can not select the important variables in the third setting, due to the violation of the irrepresentable condition for $s = 6, 12$ and 24, while SIS perform reasonably well.

7.2. *Linear Models.* The generated data $(\mathbf{X}_1^T, Y_1), \ldots, (\mathbf{X}_n^T, Y_n)$ are n i.i.d. copies of a pair $(\mathbf{X}^T, Y)$, in which the response Y follows a linear model with $Y = \mathbf{X}^T\boldsymbol{\beta}^\star + \varepsilon$, where the random error ε is standard normally distributed. The covariates are generated in the same manner as the logistic regression settings. We take the same true coefficients and correlation

TABLE 5
The MMMS and the associated RSD (in the parenthesis) of the simulated examples for logistic regressions in Setting 3 (S3) when p=2000 and n=600. The values with italic font have the same meaning as Table 2. M-$\lambda_{\max}$ and its RSD have the same meaning as Table 1.

s	M-$\lambda_{\max}$(RSD)	SIS-MLR	SIS-MMLE	LASSO	SCAD
3	8.47(0.17)	3(0)	3(0)	3(1)	3(0)
6	10.36(0.26)	56(0)	56(0)	*1227(7)*	*1142(64)*
12	14.69(0.39)	63(6)	63(6)	*1148(8)*	*1093(59)*
24	23.70(0.14)	214.5(93)	208.5(82)	*1120(5)*	*1087(24)*

structures for part of the scenarios ($p = 40000$) as the logistic regression examples, while vary the true coefficients for other scenarios, to gauge the difficulty of the problem. The sample size for each scenario is correspondingly decreased to reflect the fact that the linear model is more informative. The results are recorded in Tables 6–9 respectively. The trend of the MMMS and the associated RSD of SIS, the LASSO and the SCAD varying with the correlation and/or the nonsparse set size are similar to these in the logistic regression examples, but their magnitudes are usually smaller in the linear regression examples, as the model is more informative. Overall, the SIS does a very reasonable job in screening irrelevant variables and sometimes outperforms the LASSO and the SCAD.

8. Conclusion Remarks. In this paper, we propose two independent screening methods by ranking the maximum marginal likelihood estimators and the maximum marginal likelihood in generalized linear models. With Fan and Lv (2008) as a special case, the proposed method is shown to possess the sure independence screening property. The success of the marginal screening embarks the idea that any surrogates screening, besides the marginal utility screening introduced in this paper, as long as which can preserve the non-sparsity structure of the true model and is feasible in computation, can be a good option for population variable screening. It also

Table 6
The MMMS and the associated RSD (in the parenthesis) of the simulated examples in the first two settings (S1 and S2) for linear regressions when $p = 40000$.

ρ	n	SIS-MLR	SIS-MMLE	n	SIS-MLR	SIS-MMLE
	Setting 1, $q = 15$					
	$s = 3$, $\boldsymbol{\beta}^{\star} = (1, 1.3, 1)^T$			$s = 6$, $\boldsymbol{\beta}^{\star} = (1, 1, 3, 1, \ldots)^T$		
0	80	12(18)	12(18)	150	42(157)	42(157)
0.2	80	3(0)	3(0)	150	6(0)	6(0)
0.4	80	3(0)	3(0)	150	6.5(1)	6.5(1)
0.6	80	3(0)	3(0)	150	6(1)	6(1)
0.8	80	3(0)	3(0)	150	7(1)	7(1)
	$s = 12$, $\boldsymbol{\beta}^{\star} = (1, 1.3, \ldots)^T$			$s = 15$, $\boldsymbol{\beta}^{\star} = (1, 1.3, \ldots)^T$		
0	300	143(282)	143(282)	400	135.5(167)	135.5(167)
0.2	200	13(1)	13(1)	200	15(0)	15(0)
0.4	200	13(1)	13(1)	200	15(0)	15(0)
0.6	200	13(1)	13(1)	200	15(0)	15(0)
0.8	200	13(1)	13(1)	200	15(0)	15(0)
	Setting 2, $q = 50$					
	$s = 3$, $\boldsymbol{\beta}^{\star} = (1, 1.3, 1)^T$			$s = 6$, $\boldsymbol{\beta}^{\star} = (1, 1.3, 1, \ldots)^T$		
0	100	3(2)	3(2)	200	7.5(7)	7.5(7)
0.2	100	3(0)	3(0)	200	6(1)	6(1)
0.4	100	3(0)	3(0)	200	7(1)	7(1)
0.6	100	3(0)	3(0)	200	7(2)	7(2)
0.8	100	3(1)	3(1)	200	8(4)	8(4)
	$s = 12$, $\boldsymbol{\beta}^{\star} = (1, 1.3, \ldots)^T$			$s = 15$, $\boldsymbol{\beta}^{\star} = (1, 1.3, \ldots)^T$		
0	400	22(27)	22(27)	500	35(52)	35(52)
0.2	300	16(5)	16(5)	300	24(7)	24(7)
0.4	300	19(8)	19(8)	300	30(10)	30(10)
0.6	300	25(8)	25(8)	300	33.5(7)	33.5(7)
0.8	300	24(7)	24(7)	300	35(8)	35(8)

Table 7
The MMMS and the RSD (in the parenthesis) of the simulated examples for linear regressions in Setting 1 (S1) when $p = 5000$ and $q = 15$.

ρ	n	SIS-MLR	SIS-MMLE	LASSO	SCAD
	$s = 3$, $\boldsymbol{\beta}^{\star} = (0.5, 0.67, 0.5)^T$				
0	100	12(40)	12(40)	3(1)	3(1)
0.2	100	3(1)	3(1)	3(0)	3(0)
0.4	100	3(0)	3(0)	3(0)	3(0)
0.6	100	3(1)	3(1)	5(7)	5(5)
0.8	100	4(2)	4(2)	4(1)	4(1)
	$s = 6$, $\boldsymbol{\beta}^{\star} = (0.5, 0.67, 0.5, 0.67, 0.5, 0.67)^T$				
0	100	210.5(422)	210.5(422)	33.5(651)	25(22)
0.2	100	7(2)	7(2)	6(1)	6(1)
0.4	100	7(2)	7(2)	6(1)	6(1)
0.6	100	8(2)	8(2)	7(1)	7(1)
0.8	100	9(3)	9(3)	7(2)	8(1)
	$s = 12$, $\boldsymbol{\beta}^{\star} = (0.5, 0.67, \ldots)^T$				
0	300	49(76)	49(76)	12(1)	12(0)
0.2	100	14(2)	14(2)	12(1)	12(1)
0.4	100	14(1)	14(1)	12(1)	12(1)
0.6	100	14(1)	14(1)	13(1)	13(1)
0.8	100	14(1)	14(1)	13(1)	13(1)
	$s = 15$, $\boldsymbol{\beta}^{\star} = (0.5, 0.67, \ldots)^T$				
0	300	199(251)	199(251)	17(2)	15(0)
0.2	100	17(5)	17(5)	15(1)	15(0)
0.4	100	15(0)	15(0)	15(0)	15(0)
0.6	100	15(0)	15(0)	15(0)	15(0)
0.8	100	15(0)	15(0)	15(0)	15(1)

TABLE 8

The MMMS and the RSD (in the parenthesis) of the simulated examples for linear regressions in Setting 2 (S2) when $p = 2000$ and $q = 50$.

ρ	n	SIS-MLR	SIS-MMLE	LASSO	SCAD
	$s = 3$, $\boldsymbol{\beta}^\star = (0.6, 0.8, 0.6)^T$				
0	100	5(14)	6(16)	4(4)	4(2)
0.2	100	3(1)	3(1)	3(0)	3(0)
0.4	100	3(1)	4(1)	3(1)	3(1)
0.6	100	5(3)	7(5)	4(1)	4(1)
0.8	100	7(7)	14(12)	5(57)	7(4)
	$s = 6$, $\boldsymbol{\beta}^\star = (3, -3, 3, -3, 3, -3)^T$				
0	100	15(43)	18(47)	6(0)	6(1)
0.2	100	42(116)	47(99)	7(1)	7(1)
0.4	100	143(207)	129(226)	12(4)	12(5)
0.6	200	47(93)	49(110)	7(1)	7(1)
0.8	200	360(470)	376.5(486)	54(32)	51(25)
	$s = 12$, $\boldsymbol{\beta}^\star = (0.6, 0.8, \ldots)^T$				
0	200	151(212)	140(207)	15(4)	15(4)
0.2	100	37.5(10)	36(12)	16(3)	16(4)
0.4	100	39(7)	40.5(8)	18(3)	17(2)
0.6	100	41(7)	42(6)	19(3)	18(3)
0.8	100	44(5)	46(6)	23(1478)	24(50)
	$s = 24$, $\boldsymbol{\beta}^\star = (3, 4, \ldots)^T$				
0	400	229(283)	227(279)	24(0)	25(0)
0.2	100	61(43)	67(46)	30(2)	30(2)
0.4	100	48(2)	47(2)	31(2)	30(1)
0.6	100	48(2)	49(2)	32(2)	32(3)
0.8	100	49(2)	49(1)	32(2)	32(2)

TABLE 9

The MMMS and the associated RSD (in the parenthesis) of the simulated examples for linear regressions in Setting 3 (S3), where p=2000 and n=600. The values with italic font have the same meaning as Table 2. M-$\lambda_{\max}$ and its RSD have the same meaning as Table 1.

s	M-$\lambda_{\max}$(RSD)	SIS-MLR	SIS-MMLE	LASSO	SCAD
3	8.47(0.17)	3(0)	3(0)	3(0)	3(0)
6	10.36(0.26)	56(0)	56(0)	47(4)	45(3)
12	14.69(0.39)	62(0)	62(0)	*1610(10)*	*1304(2)*
24	23.70(0.14)	81(19)	81(23)	*1637(14)*	*1303(1)*

paves the way for the sample variable screening, as long as the surrogate signals are uniformly distinguishable from the stochastic noise. Along this line, many statistics, such as R square statistics, marginal pseudo likelihood (least square estimation, for example), can be potential basis for the independence learning. Meanwhile the proposed properties: sure screening and vanishing false selection rate will be good criteria for evaluating ultrahigh dimensional variable selection methods.

As our current results only hold when the log-likelihood function is concave in the regression parameters, the proposed procedure does not cover all generalized linear models, such as some noncanonical link cases. This leaves space for future research.

Unlike Fan and Lv (2008), the main idea of our technical proofs is broadly applicable. We conjecture that our results should hold when the conditional distribution of the outcome Y given the covariates $\mathbf{X}$ depends only on $\mathbf{X}^T\boldsymbol{\beta}^\star$ and is arbitrary and unknown otherwise. Therefore, besides GLIM, the SIS method can be applied to a rich class of general regression models, including transformation models (Bickel and Doksum, 1981; Box and Cox, 1964), censored regression models (Cox, 1972; Kosorok et al., 2004; Zeng and Lin, 2007), projection pursuit regression (Friedman and Stuetzle, 1981). These are also interesting future research topics.

Another important extension is to generalize the concept of marginal regression to the marginal group regression, where the number of covariates m in each marginal regression is greater or equal to one. This leads to a new procedure called grouped variables screening. It is expected to improve the situation when the variables are highly correlated and jointly important, but marginally the correlation between each individual variable and the response is weak. The current theoretical studies for the componentwise marginal regression can be directly extended to group variable screening, with appropriate conditions and adjustments. This leads to another interesting topic of future research.

In practice, how to choose the tuning parameter γ_n is an interesting and important problem. As discussed in Fan and Lv (2008), for the first stage of the iterative SIS procedure, our preference is to select sufficiently many features, such that $|\mathcal{M}_{\gamma_n}| = n$ or $n/\log(n)$. The FDR-based methods in multiple comparison can also possibly employed. In the second or final stage, Bayes information type of criterion can be applied. In practice, some data-driven methods may also be welcome for choosing the tuning parameter γ_n. This is an interesting future research topic and is beyond the scope of the current paper.

9. Proofs. To establish Theorem 1, the following symmetrization theorem in van der Vaart and Wellner (1996), contraction theorem in Ledoux and Talagrand (1991) and concentration theorem in Massart (2000) will be needed. We reproduce them here for the sake of readability.

LEMMA 2. *(Symmetrization, Lemma 2.3.1, van der Vaart and Wellner, 1996) Let $Z_1, \ldots, Z_n$ be independent random variables with values in $\mathcal{Z}$ and $\mathcal{F}$ is a class of real valued functions on $\mathcal{Z}$. Then*

$$E\left\{\sup_{f\in\mathcal{F}} |(\mathbb{P}_n - P)f(Z)|\right\} \le 2E\left\{\sup_{f\in\mathcal{F}} |\mathbb{P}_n \varepsilon f(Z)|\right\},$$

where $\varepsilon_1, \ldots, \varepsilon_n$ be a Rademacher sequence (i.e., i.i.d. sequence taking values ± 1 with probability $1/2$) independent of $Z_1, \ldots, Z_n$ and $Pf(Z) = Ef(Z)$.

LEMMA 3. *(Contraction theorem, Ledoux and Talagrand, 1991) Let $z_1, \ldots, z_n$ be nonrandom elements of some space $\mathcal{Z}$ and let $\mathcal{F}$ be a class* *of real valued functions on $\mathcal{Z}$. Let $\varepsilon_1, \ldots, \varepsilon_n$ be a Rademacher sequence. Consider Lipschitz functions $\gamma_i : \mathbf{R} \mapsto \mathbf{R}$, that is,*

$$|\gamma_i(s) - \gamma_i(\tilde{s})| \leq |s - \tilde{s}|, \ \forall s, \tilde{s} \in \mathbf{R}.$$

Then for any function $\tilde{f} : \mathcal{Z} \mapsto \mathbf{R}$, we have

$$E\{\sup_{f \in \mathcal{F}} |\mathbb{P}_n \varepsilon(\gamma(f) - \gamma(\tilde{f}))|\} \leq 2E\{\sup_{f \in \mathcal{F}} |\mathbb{P}_n \varepsilon(f - \tilde{f})|\}.$$

LEMMA 4. *(Concentration theorem, Massart, 2000) Let $Z_1, \ldots, Z_n$ be independent random variables with values in some space $\mathcal{Z}$ and let $\gamma \in \Gamma$, a class of real valued functions on $\mathcal{Z}$. We assume that for some positive constants $l_{i,\gamma}$ and $u_{i,\gamma}$, $l_{i,\gamma} \leq \gamma(Z_i) \leq u_{i,\gamma}$ $\forall \gamma \in \Gamma$. Define*

$$L^2 = \sup_{\gamma \in \Gamma} \sum_{i=1}^{n} (u_{i,\gamma} - l_{i,\gamma})^2 / n, \ \text{and}$$

$$\mathbf{Z} = \sup_{\gamma \in \Gamma} |(\mathbb{P}_n - P)\gamma(Z)|,$$

then for any $t > 0$,

$$P(\mathbf{Z} \geq E\mathbf{Z} + t) \leq \exp\left(-\frac{nt^2}{2L^2}\right).$$

Let $N > 0$, define a set of $\boldsymbol{\beta}$:

$$\mathcal{B}(N) = \{\boldsymbol{\beta} \in \mathcal{B}, \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| \leq N\}.$$

Let

$$\mathbb{G}_1(N) = \sup_{\boldsymbol{\beta} \in \mathcal{B}(N)} |(\mathbb{P}_n - P)\{l(\mathbf{X}^T\boldsymbol{\beta}, Y) - l(\mathbf{X}^T\boldsymbol{\beta}_0, Y)\} I_n(\mathbf{X}, Y)|,$$

where $I_n(\mathbf{X}, Y)$ is defined in Condition B. The next result is about the upper bound of the tail probability for $\mathbb{G}_1(N)$ in the neighborhood of $\mathcal{B}(N)$.

LEMMA 5. *For all $t > 0$, it holds that*

$$P(\mathbb{G}_1(N) \geq 4Nk_n(q/n)^{1/2}(1+t)) \leq \exp(-2t^2/K_n^2).$$

Proof of Lemma 5. The main idea is to apply the concentration theorem (Lemma 4). To this end, we first show that the random variables involved are bounded. By Condition B and the Cauchy-Schwartz inequality, we have that on the set Ω_n,

$$|l(\mathbf{X}^T\boldsymbol{\beta}, Y) - l(\mathbf{X}^T\boldsymbol{\beta}_0, Y)| \leq k_n|\mathbf{X}^T(\boldsymbol{\beta} - \boldsymbol{\beta}_0)| \leq k_n\|\mathbf{X}\| \, \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|.$$

On the set Ω_n, by the definition of $\mathcal{B}(N)$, the above random variable is further bounded by $k_n q^{1/2} K_n N$. Hence, $L^2 = 4k_n^2 q K_n^2 N^2$, using the notation of Lemma 4.

We need to bound the expectation $E\mathbb{G}_1(N)$. An application of the symmetrization theorem (Lemma 2) yields that

$$E\mathbb{G}_1(N) \leq 2E\left[\sup_{\boldsymbol{\beta}\in\mathcal{B}(N)} \left|\mathbb{P}_n\varepsilon\left\{l(\mathbf{X}^T\boldsymbol{\beta}, Y) - l(\mathbf{X}^T\boldsymbol{\beta}_0, Y)\right\} I_n(\mathbf{X}, Y)\right|\right].$$

By the contraction theorem (Lemma 3), and the Lipschitz condition in Condition B, we can bound the right hand side of the above inequality further by

$$4k_n E\left\{\sup_{\boldsymbol{\beta}\in\mathcal{B}(N)} \left|\mathbb{P}_n\varepsilon\mathbf{X}^T(\boldsymbol{\beta} - \boldsymbol{\beta}_0)I_n(\mathbf{X}, Y)\right|\right\}. \tag{6}$$

By the Cauchy-Schwartz inequality, the expectation in (6) is controlled by

$$E\|\mathbb{P}_n\varepsilon\mathbf{X}I_n(\mathbf{X}, Y)\| \sup_{\boldsymbol{\beta}\in\mathcal{B}(N)} \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| \leq E\|\mathbb{P}_n\varepsilon\mathbf{X}I_n(\mathbf{X}, Y)\|N. \tag{7}$$

By Jensen's inequality, the expectation in (7) is bounded above by

$$\left(E\|\mathbb{P}_n\varepsilon\mathbf{X}I_n(\mathbf{X}, Y)\|^2\right)^{1/2} = \left(E\|\mathbf{X}\|^2 I_n(\mathbf{X}, Y)/n\right)^{1/2} \leq (q/n)^{1/2},$$

by noticing that

$$E\|\mathbf{X}\|^2 I_n(\mathbf{X}, Y) \leq E\|\mathbf{X}\|^2 = E(X_1^2 + \cdots + X_q^2) = q,$$

since $EX_j^2 = 1$. Combining these results, we conclude that

$$E\mathbb{G}_1(N) \leq 4Nk_n(q/n)^{1/2}.$$

An application of the concentration theorem (Lemma 4) yields that

$$\begin{aligned} P(\mathbb{G}_1(N) \geq 4Nk_n(q/n)^{1/2}(1+t)) &\leq \exp\left(-\frac{n\{4Nk_n(q/n)^{1/2}t\}^2}{8qK_n^2k_n^2N^2}\right) \\ &= \exp(-2t^2/K_n^2). \end{aligned}$$

This proves the lemma. □

Proof of Theorem 1. The proof takes two main steps: we first bound $\|\hat{\boldsymbol\beta} - \boldsymbol\beta_0\|$ by $\mathbb{G}(N)$ for a small N, where N chosen so that Conditions B and C hold, and then utilize Lemma 5 to conclude.

Following a similar idea in van de Geer (2002), we define a convex combination $\boldsymbol\beta_s = s\hat{\boldsymbol\beta} + (1-s)\boldsymbol\beta_0$ with

$$s = \left(1 + \|\hat{\boldsymbol\beta} - \boldsymbol\beta_0\|/N\right)^{-1}.$$

Then, by definition,

$$\|\boldsymbol\beta_s - \boldsymbol\beta_0\| = s\|\hat{\boldsymbol\beta} - \boldsymbol\beta_0\| \leq N,$$

namely, $\boldsymbol\beta_s \in \mathcal{B}(N)$. Due to the convexity, we have

$$\begin{aligned} \mathbb{P}_n l(\mathbf{X}^T\boldsymbol\beta_s, Y) &\leq s\mathbb{P}_n l(\mathbf{X}^T\hat{\boldsymbol\beta}, Y) + (1-s)\mathbb{P}_n l(\mathbf{X}^T\boldsymbol\beta_0, Y) \\ &\leq \mathbb{P}_n l(\mathbf{X}^T\boldsymbol\beta_0, Y). \end{aligned} \tag{8}$$

Since $\boldsymbol\beta_0$ is the minimizer, we have

$$E[l(\mathbf{X}^T\boldsymbol\beta_s, Y) - l(\mathbf{X}^T\boldsymbol\beta_0, Y)] \geq 0,$$

where $\boldsymbol{\beta}_s$ is regarded a parameter in the above expectation. Hence, it follows from (8) that

$$\begin{aligned} & E[l(\mathbf{X}^T\boldsymbol{\beta}_s, Y) - l(\mathbf{X}^T\boldsymbol{\beta}_0, Y)] \\ \leq\ & (E - \mathbb{P}_n)[l(\mathbf{X}^T\boldsymbol{\beta}_s, Y) - l(\mathbf{X}^T\boldsymbol{\beta}_0, Y)] \\ \leq\ & \mathbb{G}(N), \end{aligned}$$

where

$$\mathbb{G}(N) = \sup_{\boldsymbol{\beta}\in\mathcal{B}(N)} |(\mathbb{P}_n - P)\{l(\mathbf{X}^T\boldsymbol{\beta}, Y) - l(\mathbf{X}^T\boldsymbol{\beta}_0, Y)\}|.$$

By Condition C, it follows that

$$\|\boldsymbol{\beta}_s - \boldsymbol{\beta}_0\| \leq [\mathbb{G}(N)/V_n]^{1/2}. \tag{9}$$

We now use (9) to conclude the result. Note that for any x,

$$P(\|\boldsymbol{\beta}_s - \boldsymbol{\beta}_0\| \geq x) \leq P(\mathbb{G}(N) \geq V_n x^2).$$

Setting $x = N/2$, we have

$$P(\|\boldsymbol{\beta}_s - \boldsymbol{\beta}_0\| \geq N/2) \leq P\{\mathbb{G}(N) \geq V_n N^2/4\}.$$

Using the definition of $\boldsymbol{\beta}_s$, the left hand side is the same as $P\{\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| \geq N\}$. Now, by taking $N = 4a_n(1+t)/V_n$ with $a_n = 4k_n\sqrt{q/n}$, we have

$$\begin{aligned} P\{\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| \geq N\} &\leq P\{\mathbb{G}(N) \geq V_n N^2/4\} \\ &= P\{\mathbb{G}(N) \geq N a_n(1+t)\}. \end{aligned}$$

The last probability is bounded by

$$P\{\mathbb{G}(N) \geq N a_n(1+t), \Omega_{n,\star}\} + P\{\Omega_{n,\star}^c\}, \tag{10}$$

where $\Omega_{n,\star} = \{\|\mathbf{X}_i\| \leq K_n, |Y_i| \leq K_n^\star\}$.

On the set $\Omega_{n,\star}$, since

$$\sup_{\boldsymbol{\beta}\in\mathcal{B}(N)} \mathbb{P}_n\Big|l(\mathbf{X}^T\boldsymbol{\beta}, Y) - l(\mathbf{X}^T\boldsymbol{\beta}_0, Y)\Big|(1 - I_n(\mathbf{X}, Y)) = 0,$$

by the triangular inequality,

$$\mathbb{G}(N) \le \mathbb{G}_1(N) + \sup_{\boldsymbol{\beta}\in\mathcal{B}(N)}\Big|E[l(\mathbf{X}^T\boldsymbol{\beta}, Y) - l(\mathbf{X}^T\boldsymbol{\beta}_0, Y)](1 - I_n(\mathbf{X}, Y))\Big|.$$

It follows from Condition B that (10) is bounded by

$$P\{\mathbb{G}_1(N) \ge Na_n(1+t) + o(q/n)\} + nP\{(\mathbf{X}, Y) \in \Omega_n^c\}$$

The conclusion follows from Lemma 5. □

Proof of Theorem 2. First of all, the target function $El(\beta_0 + \beta_j X_j, Y)$ is a convex function in β_j. We first show that if $\mathrm{cov}(b'(\mathbf{X}^T\boldsymbol{\beta}^\star), X_j) = 0$, then β_j^M must be zero. Recall $EX_j = 0$. The score equation of the marginal regression at β_j^M takes the form:

$$E\left\{b'(\mathbf{X}_j^T\boldsymbol{\beta}_j^M)X_j\right\} = E(YX_j) = E\left\{b'(\mathbf{X}^T\boldsymbol{\beta}^\star)X_j\right\}.$$

It can be equivalently written as

$$\mathrm{cov}(b'(\mathbf{X}_j^T\boldsymbol{\beta}_j^M), X_j) = \mathrm{cov}(b'(\mathbf{X}^T\boldsymbol{\beta}^\star), X_j) = 0.$$

Since both functions $f(t) = b'(\beta_{j,0}^M + t)$ and $h(t) = t$ are strictly monotone in t, when $t \ne 0$,

$$\{f(t) - f(0)\}(t - 0) > 0.$$

If $\beta_j^M \ne 0$, let $t = \beta_j^M X_j$,

$$\beta_j^M \mathrm{cov}(f(\beta_j^M X_j), X_j) = E[E\{f(t) - f(0)\}(t-0)|X_j \ne 0] > 0,$$

which leads to a contradiction. Hence β_j^M must be zero.

On the other side, if $\beta_j^M = 0$, the score equations now take the form:

$$E\left\{b'(\beta_{j,0}^M)\right\} = E\left\{b'(\mathbf{X}^T\boldsymbol{\beta}^\star)\right\}, \text{ and} \tag{11}$$

$$E\left\{b'(\beta_{j,0}^M)X_j\right\} = E\left\{b'(\mathbf{X}^T\boldsymbol{\beta}^\star)X_j\right\}. \tag{12}$$

Since $b'(\beta_{j,0}^M)$ is a constant, we can get the desired result by plugging (11) into (12). □

Proof of Theorem 3. We first prove the case that $b''(\theta)$ is bounded. By the Lipschitz continuity of the function $b'(\cdot)$, we have

$$\left|\{b'(\beta_{j,0}^M + X_j\beta_j^M) - b'(\beta_{j,0}^M)\}X_j\right| \le D_1|\beta_j^M|X_j^2.$$

$D_1 = \sup_x b''(x)$. By taking the expectation on both sides, we have

$$\left|E\{b'(\beta_{j,0}^M + X_j\beta_j^M) - b'(\beta_{j,0}^M)\}X_j\right| \le D_1|\beta_j^M|,$$

namely,

$$D_1|\beta_j^M| \ge |\text{cov}(b'(\beta_{j,0}^M + X_j\beta_j^M), X_j)|. \tag{13}$$

Note that $\beta_{j,0}^M$ and β_j^M satisfy the score equation

$$E\{b'(\beta_{0,1}^M + \beta_j^M X_j) - b'(\mathbf{X}^T\boldsymbol{\beta}^\star)\}X_j = 0. \tag{14}$$

It follows from (13) and $EX_j = 0$ that

$$|\beta_j^M| \ge D_1^{-1}c_1 n^{-\kappa}.$$

The conclusion follows.

We now prove the second case. The result holds trivially if $|\beta_j^M| \ge cn^{-\kappa}$ for a sufficiently large universal constant c. Now suppose that $|\beta_j^M| \le c_9 n^{-\kappa}$, for some positive constant c_9. We will show later that $|\beta_{j,0}^M - \beta_0^M| \le c_{10}$ for some

$c_{10} > 0$, where β_0^M is such that $b'(\beta_0^M) = EY$. In this case, if $|X_j| \le n^\kappa$, then the points $\beta_{j,0}^M$ and $(\beta_{j,0}^M + X_j\beta_j^M)$ falls in the interval $\beta_0^M \pm h$, independent of j, where $h = c_9 + c_{10}$.

By the Lipschitz continuity of the function $b'(\cdot)$ in the neighborhood around β_0^M, we have for $|X_j| \le n^\kappa$,

$$|\{b'(\beta_{j,0}^M + X_j\beta_j^M) - b'(\beta_{j,0}^M)\}X_j| \le D_2|\beta_j^M|X_j^2.$$

where $D_2 = \max_{x\in[\beta_0^M - h, \beta_0^M + h]} b''(x)$. By taking the expectation on both sides, it follows that

$$|E\{b'(\beta_{j,0}^M + X_j\beta_j^M) - b'(\beta_{j,0}^M)\}X_j I(|X_j| \le n^\kappa)| \le D_2|\beta_j^M|. \tag{15}$$

By using (14) and $EX_j = 0$, we deduce from (15) that

$$D_2|\beta_j^M| \ge |\mathrm{cov}(\mathbf{X}^T\boldsymbol{\beta}^\star, X_j)| - A_0 - A_1. \tag{16}$$

where $A_m = E|b'(\beta_{j,0}^M + X_j^m\beta_j^M)X_j|\ I(|X_j| \ge n^\kappa)$ for $m = 0$ and 1. Since $|\beta_{j,0}^M + X_j\beta_j^M| \le a|X_j|$ for $|X_j| \ge n^\kappa$ for a sufficiently large n, independent of j, by the condition given in the theorem, we have

$$A_m \le EG(a|X_j|)^m|X_j|\ I(|X_j| \ge n^\kappa) \le dn^{-\kappa}, \qquad \text{for } m = 0 \text{ and } 1.$$

The conclusion now follows from (16).

It remains to show that when $|\beta_j^M| \le c_9 n^{-\kappa}$ we have $|\beta_{j,0}^M - \beta_0^M| \le c_{10}$. To this end, let

$$\ell(\beta_0) = E\{b(\beta_0 + \beta_j^M X_j) - Y(\beta_0 + \beta_j^M X_j)\}.$$

Then, it is easy to see that

$$\ell'(\beta_0) = Eb'(\beta_0 + \beta_j^M X_j) - b'(\beta_0^M).$$

Observe that

$$|Eb'(\beta_0 + \beta_j^M X_j) - b'(\beta_0)| \le R_1 + R_2, \tag{17}$$

where $R_1 = \sup_{|x| \le c_9 n^{\eta-\kappa}} |b'(\beta_0 + x) - b'(\beta_0)|$ and $R_2 = 2EG(a|X_j|)I(|X_j| > n^\eta)$. Now, $R_1 = o(1)$ due to the continuity of $b'(\cdot)$ and $R_2 = o(1)$ by the condition of the theorem. Consequently, by (17), we conclude that

$$\ell'(\beta_0) = b'(\beta_0) - b'(\beta_0^M) + o(1).$$

Since $b'(\cdot)$ is a strictly increasing function, it is now obvious that

$$\ell'(\beta_0^M - c_{10}) < 0, \qquad \ell'(\beta_0^M + c_{10}) > 0.$$

for any given $c_{10} > 0$. Hence, $|\beta_{j,0}^M - \beta_0^M| < c_{10}$. □

Proof of Proposition 1. Without loss of generality, assume that $g(\cdot)$ is strictly increasing and $\rho > 0$. Since X and Z are jointly normally distributed, Z can be expressed as

$$Z = \rho X + \varepsilon,$$

where $\rho = E(XZ)$ is the regression coefficient and X and ε are independent. Thus,

$$Ef(Z)X = Eg(\rho X)X = E[g(\rho X) - g(0)]X, \tag{18}$$

where $g(x) = Ef(x+\varepsilon)$ is a strictly increasing function. The right hand side of (18) is always nonnegative and is zero if and only if $\rho = 0$.

To prove the second part, we first note that the random variable on the right hand side of (18) is nonnegative. Thus, by the mean-value theorem, we have that

$$|Ef(Z)X| \ \ \ge \ \ \inf_{|x| \le c\rho} |g'(x)| \, \rho E X^2 I(|X| \le c).$$

Hence, the result follows. □

Proof of Lemma 1. By Chebyshev's inequality,

$$P(Y \geq u) \leq \exp(-s_0 u) E \exp(s_0 Y).$$

Let $\theta = \mathbf{X}^T \boldsymbol{\beta}^\star$. Since Y belongs to an exponential family, we have

$$E\{\exp(s_0 Y)|\theta\} \quad = \quad \exp(b(\theta + s_0) - b(\theta)).$$

Hence

$$P(Y \geq u) \leq \exp(-s_0 u) E \exp(b(\mathbf{X}^T \boldsymbol{\beta}^\star + s_0) - b(\mathbf{X}^T \boldsymbol{\beta}^\star)).$$

Similarly we can get

$$P(Y \leq -u) \leq \exp(-s_0 u) E \exp(b(\mathbf{X}^T \boldsymbol{\beta}^\star - s_0) - b(\mathbf{X}^T \boldsymbol{\beta}^\star)).$$

The desired result thus follows from Condition D by letting $u = m_0 t^\alpha / s_0$. □

Proof of Theorem 4. Note that Condition B is satisfied with k_n defined in Section 5.2. The tail part of Condition B can also be easily checked. In fact,

$$\begin{aligned} & E[l(\mathbf{X}_j^T \boldsymbol{\beta}_j, Y) - l(\boldsymbol{\beta}_j^M, Y)](1 - I_n(\mathbf{X}_j, Y)) \\ \leq \quad & |Eb(\mathbf{X}_j^T \boldsymbol{\beta}_j) I(|X_j| \geq K_n)| + |Eb(\mathbf{X}_j^T \boldsymbol{\beta}_j^M) I(|X_j| \geq K_n)| + B(\boldsymbol{\beta}_j) + B(\boldsymbol{\beta}_j^M), \end{aligned}$$

where $B(\boldsymbol{\beta}_j) = |EY\mathbf{X}_j^T \boldsymbol{\beta}_j (1 - I_n(\mathbf{X}_j, Y))|$. The first two terms is of order $o(1/n)$ by assumption and the last two terms can be bounded by the exponential tail conditions in Condition D and the Cauchy-Schwartz inequality.

By Theorem 1, we have for any $t > 0$,

$$P(\sqrt{n}|\hat{\beta}_j^M - \beta_j^M| \geq 16(1+t)k_n/V) \leq \exp(-2t^2/K_n^2) + nm_1 \exp(-m_0 K_n^\alpha).$$

By taking $1+t = c_3 V n^{1/2-\kappa}/(16k_n)$, it follows that

$$P(|\hat{\beta}_j^M - \beta_j^M| \geq c_3 n^{-\kappa}) \leq \exp(-c_4 n^{1-2\kappa}/(k_n K_n)^2) + n m_1 \exp(-m_0 K_n^{\alpha}).$$

The first result follows from the union bound of probability.

To prove the second part, note that on the event

$$A_n \equiv \{\max_{j\in\mathcal{M}_\star} |\hat{\beta}_j^M - \beta_j^M| \leq c_2 n^{-\kappa}/2\},$$

by Theorem 3, we have

$$|\hat{\beta}_j^M| \geq c_2 n^{-\kappa}/2, \quad \text{for all } j \in \mathcal{M}_\star. \tag{19}$$

Hence, by the choice of γ_n, we have $\mathcal{M}_\star \subset \widehat{\mathcal{M}}_{\gamma_n}$. The result now follows from a simple union bound:

$$P(A_n^c) \leq s_n\{\exp(-c_4 n^{1-2\kappa}/(k_n K_n)^2) + n m_1 \exp(-m_0 K_n^{\alpha})\}.$$

This completes the proof. □

Proof of Theorem 5. The key idea of the proof is to show that

$$\|\boldsymbol{\beta}^M\|^2 = O(\|\boldsymbol{\Sigma}\boldsymbol{\beta}^\star\|^2) = O\{\lambda_{\max}(\boldsymbol{\Sigma})\}. \tag{20}$$

If so, the number of $\{j : |\beta_j^M| > \varepsilon n^{-\kappa}\}$ can not exceed $O\{n^{2\kappa}\lambda_{\max}(\boldsymbol{\Sigma})\}$ for any $\varepsilon > 0$. Thus, on the set

$$B_n = \{\max_{1\leq j\leq p_n} |\hat{\beta}_j^M - \beta_j^M| \leq \varepsilon n^{-\kappa}\},$$

the number of $\{j : |\hat{\beta}_j^M| > 2\varepsilon n^{-\kappa}\}$ can not exceed the number of $\{j : |\beta_j^M| > \varepsilon n^{-\kappa}\}$, which is bounded by $O\{n^{2\kappa}\lambda_{\max}(\boldsymbol{\Sigma})\}$. By taking $\varepsilon = c_5/2$, we have

$$P[|\widehat{\mathcal{M}}_{\gamma_n}| \leq O\{n^{2\kappa}\lambda_{\max}(\boldsymbol{\Sigma})\}] \geq P(B_n).$$

The conclusion follows from Theorem 4(i).

It remains to prove (20). We first bound β_j^M. Since $b'(\cdot)$ is monotonically increasing, the function

$$\{b'(\beta_{j,0}^M + X_j\beta_j^M) - b'(\beta_{j,0}^M)\}X_j\beta_j^M$$

is always positive. By Taylor's expansion, we have

$$\{b'(\beta_{j,0}^M + X_j\beta_j^M) - b'(\beta_{j,0}^M)\}\beta_j^M X_j \geq D_3(\beta_j^M X_j)^2 I(|X_j| \leq K),$$

where $D_3 = \inf_{|x|\leq K(B+1)} b''(x)$, since $(\beta_{j,0}^M, \beta_j^M)$ is an interior point of the square $\mathcal{B}$ with length $2B$. By taking the expectation on both sides and using $EX_j = 0$, we have

$$Eb'(\beta_{j,0}^M + X_j\beta_j^M)\beta_j^M X_j \geq D_3 E(\beta_j^M X_j)^2 I(|X_j| \leq K).$$

Since $EX_j^2 I(|X_j| \leq K) = 1 - EX_j^2 I(|X_j| > K)$, it is unformly bounded from below for a sufficiently large K, due to the uniform exponential tail bound in Condition D. Thus, it follows from (12) that

$$|\beta_j^M|^2 \leq D_4|Eb'(\mathbf{X}^T\boldsymbol{\beta}^\star)X_j|, \tag{21}$$

for some $D_4 > 0$.

We now further bound from above the right hand side of (21) by using $\mathrm{var}(\mathbf{X}^T\boldsymbol{\beta}^\star) = O(1)$. We first show the case where $b''(\cdot)$ is bounded. By the Lipschitz continuity of the function $b'(\cdot)$, we have

$$\Big|\{b'(\mathbf{X}^T\boldsymbol{\beta}^\star) - b'(\beta_0^\star)\}X_j\Big| \leq D_5\Big|X_j\mathbf{X}_M^T\boldsymbol{\beta}_1^\star\Big|,$$

where $\mathbf{X}_M = (X_1, \cdots, X_{p_n})^T$ and $\boldsymbol{\beta}_1^\star = (\beta_1^\star, \ldots, \beta_{p_n}^\star)^T$.

By putting the above equation into the vector form and taking the expectation on both sides, we have

$$\begin{aligned} \Big\|E\{b'(\mathbf{X}^T\boldsymbol{\beta}^\star) - b'(\beta_0^\star)\}\mathbf{X}_M\Big\|^2 &\leq D_5^2\Big\|E\mathbf{X}_M\mathbf{X}_M^T\boldsymbol{\beta}_1^\star\Big\|^2 \\ &\leq D_5^2\lambda_{\max}(\boldsymbol{\Sigma})\,\Big\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}^\star\Big\|^2. \end{aligned} \tag{22}$$

Using $E\mathbf{X}_M = 0$ and $\mathrm{var}(\mathbf{X}^T\boldsymbol{\beta}^\star) = O(1)$, we conclude that

$$\left\| Eb'(\mathbf{X}^T\boldsymbol{\beta}^\star)\mathbf{X}_M \right\|^2 \le D_6 \lambda_{\max}(\boldsymbol{\Sigma}),$$

for some positive constant D_6. This together with (21) entail (20).

It remains to bound (21) for the second case. Since $\mathbf{X}_M = R\boldsymbol{\Sigma}_1^{1/2}\mathbf{U}$, it follows that

$$Eb'(\beta_0^\star + \mathbf{X}_M^T\boldsymbol{\beta}_1^\star)\mathbf{X}_M = Eb'(\beta_0^\star + \boldsymbol{\beta}_2^T R\mathbf{U})R\boldsymbol{\Sigma}_1^{1/2}\mathbf{U},$$

where $\boldsymbol{\beta}_2 = \boldsymbol{\Sigma}_1^{1/2}\boldsymbol{\beta}_1^\star$. By conditioning on $\boldsymbol{\beta}_2^T\mathbf{U}$, it can be computed that

$$E(\mathbf{U}|\boldsymbol{\beta}_2^T\mathbf{U}) = \boldsymbol{\beta}_2^T\mathbf{U}/\|\boldsymbol{\beta}_2\|^2\boldsymbol{\beta}_2.$$

Therefore,

$$\begin{aligned} Eb'(\beta_0^\star + \mathbf{X}_M^T\boldsymbol{\beta}_1^\star)\mathbf{X}_M &= Eb'(\beta_0^\star + \boldsymbol{\beta}_2^T R\mathbf{U})R\boldsymbol{\Sigma}_1^{1/2}\boldsymbol{\beta}_2^T\mathbf{U}/\|\boldsymbol{\beta}_2\|^2\boldsymbol{\beta}_2 \\ &= Eb'(\mathbf{X}^T\boldsymbol{\beta}^\star)(\mathbf{X}^T\boldsymbol{\beta}^\star - \beta_0^\star)\boldsymbol{\Sigma}_1^{1/2}\boldsymbol{\beta}_2/\|\boldsymbol{\beta}_2\|^2. \end{aligned}$$

This entails that

$$\|Eb'(\mathbf{X}^T\boldsymbol{\beta}^\star)\mathbf{X}_M\|^2 = |Eb'(\mathbf{X}^T\boldsymbol{\beta}^\star)(\mathbf{X}^T\boldsymbol{\beta}^\star - \boldsymbol{\beta}_0)|^2 \; \|\boldsymbol{\Sigma}_1^{1/2}\boldsymbol{\beta}_2\|^2/\|\boldsymbol{\beta}_2\|^4, \tag{23}$$

By Condition G, $|Eb'(\mathbf{X}^T\boldsymbol{\beta}^\star)(\mathbf{X}^T\boldsymbol{\beta}^\star - \beta_0^\star)| = O(1)$. We also observe the facts that $\|\boldsymbol{\Sigma}_1^{1/2}\boldsymbol{\beta}_2\| \le \lambda_{\max}^{1/2}(\boldsymbol{\Sigma})\|\boldsymbol{\beta}_2\|$ and that $\|\boldsymbol{\beta}_2\| = \|\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}^\star\|$ is bounded. This proves (20) for the second case by using (21) and completes the proof. □

Proof of Theorem 6. If $\mathrm{cov}(b'(\mathbf{X}^T\boldsymbol{\beta}^\star), X_j) = 0$, by Theorem 2, we have $\beta_j^M = 0$, hence by the model identifiability at $\boldsymbol{\beta}_0^M$, $\beta_{j,0}^M = \beta_0^M$. Hence, $L_j^\star = 0$. On the other hand, if $L_j^\star = 0$, by Condition C', it follows that $\boldsymbol{\beta}_j^M = \boldsymbol{\beta}_0^M$, that is, $\beta_{j,0}^M = \beta_0^M$ and $\beta_j^M = 0$. Hence by Theorem 2, $\mathrm{cov}(b'(\mathbf{X}^T\boldsymbol{\beta}^\star), X_j) = 0$. □

Proof of Theorem 7. If $|\mathrm{cov}(b'(\mathbf{X}^T\boldsymbol{\beta}^\star), X_j)| \geq c_1 n^{-\kappa}$, for $j \in \mathcal{M}_\star$, by Theorem 3, we have $\min_{j\in\mathcal{M}_\star} |\beta_j^M| \geq c_2 n^{-\kappa}$. The first result thus follows from Condition C'.

To prove the second result, we will bound $L_j^\star$. We first show the case where $b''(\cdot)$ is bounded. By definition, we have

$$0 \leq L_j^\star \leq E\Big\{l(\beta_{j,0}^M, Y) - l(\mathbf{X}_j^T\boldsymbol{\beta}_j^M, Y)\Big\}. \tag{24}$$

By Taylor's expansion of the right hand side of (24), we have that

$$E\Big\{l(\beta_{j,0}^M, Y) - l(\mathbf{X}_j^T\boldsymbol{\beta}_j^M, Y)\Big\} \leq D_5(\beta_j^M)^2, \text{ for some } \ D_5 > 0. \tag{25}$$

The desired result thus follows from (24), (25) and the proof in Theorem 5 that

$$\|\mathbf{L}^\star\| \leq O(\|\boldsymbol{\beta}^M\|^2) = O(\lambda_{\max}(\boldsymbol{\Sigma})).$$

Now we prove the second case. By the mean-value theorem,

$$E\Big\{l(\beta_{j,0}^M, Y) - l(\mathbf{X}_j^T\boldsymbol{\beta}_j^M, Y)\Big\} = E\Big\{Y - b'(\beta_{j,0}^M + sX_j\beta_j^M)\Big\}X_j\beta_j^M, \tag{26}$$

for some $0 < s < 1$. Since $EYX_j = Eb'(\mathbf{X}_j^T\boldsymbol{\beta}_j^M)X_j$, the last term is equal to

$$E\Big\{b'(\mathbf{X}_j^T\boldsymbol{\beta}_j^M) - b'(\beta_{j,0}^M + sX_j\beta_j^M)\Big\}X_j\beta_j^M. \tag{27}$$

By the monotonicity of $b'(\cdot)$, when $X_j\beta_j^M \geq 0$, both factors in (27) is non-negative, and hence

$$\Big\{b'(\mathbf{X}_j^T\boldsymbol{\beta}_j^M) - b'(\beta_{j,0}^M + sX_j\beta_j^M)\Big\}X_j\beta_j^M \leq \Big\{b'(\mathbf{X}_j^T\boldsymbol{\beta}_j^M) - b'(\beta_{j,0}^M)\Big\}X_j\beta_j^M. \tag{28}$$

When $X_j\beta_j^M < 0$, both factors in (28) are negative and (28) continues to hold. It follows from (26)—(28) and $EX_j = 0$, the right hand side of (26) is bounded by

$$Eb'(\mathbf{X}_j^T\boldsymbol{\beta}_j^M)X_j\beta_j^M = Eb'(\mathbf{X}^T\boldsymbol{\beta}^\star)X_j\beta_j^M. \tag{29}$$

Combining (24), (26) and (29), we can bound $\|\mathbf{L}^\star\|$ in the vector form by Cauchy-Schwartz inequality:

$$\|\mathbf{L}^\star\| \leq \Big\|Eb'(\mathbf{X}^T\boldsymbol{\beta}^\star)\mathbf{X}_M\Big\|\|\boldsymbol{\beta}^M\| = O(\|\boldsymbol{\Sigma}\boldsymbol{\beta}^\star\|\|\boldsymbol{\beta}^M\|),$$

where (23) is used in the last equality. The desired result thus follows from Theorem 5. □

Proof of Theorem 8. To prove the result, we first bound $L_{j,n}$ from below to show the strength of the signals. Let $\hat{\boldsymbol{\beta}}_0^M = (\hat{\beta}_0^M, 0)^T$. Then, by Taylor's expansion, we have

$$2L_{j,n} = (\hat{\boldsymbol{\beta}}_0^M - \hat{\boldsymbol{\beta}}_j^M)\ell_j''(\boldsymbol{\xi}_n)(\hat{\boldsymbol{\beta}}_0^M - \hat{\boldsymbol{\beta}}_j^M) \geq \lambda_{j,\min}(\hat{\beta}_j{}^M)^2, \tag{30}$$

where $\lambda_{j,\min}$ is the minimum eigenvalue of the Hessian matrix

$$\ell_j''(\boldsymbol{\xi}_n) = \mathbb{P}_n b''(\boldsymbol{\xi}_n^T\mathbf{X}_j)\mathbf{X}_j\mathbf{X}_j^T,$$

where $\boldsymbol{\xi}_n$ lies between $\hat{\boldsymbol{\beta}}_0^M$ and $\hat{\boldsymbol{\beta}}_j^M$. We will show

$$P\{\lambda_{j,\min} > c_{11}\} = 1 - O\{\exp(-c_{12}n^{1-\kappa})\}, \tag{31}$$

for some $c_{11} > 0$ and $c_{12} > 0$.

Suppose (31) holds. Then, by (19), we have

$$\begin{aligned} & P\{\min_{j\in\mathcal{M}_\star} |\hat{\beta}_j^M| \geq c_2 n^{-\kappa}/2\} \\ = \ & 1 - O\left(s_n\{\exp(-c_4n^{1-2\kappa}/(k_nK_n)^2) + nm_1\exp(-m_0K_n^\alpha)\}\right). \end{aligned}$$

This, together with (30) and (31), implies

$$\begin{aligned} & P\{\min_{j\in\mathcal{M}_\star} L_{j,n} \geq c_{11}c_2^2 n^{-2\kappa}/8\} \\ = \ & 1 - O\left(s_n\{\exp(-c_4n^{1-2\kappa}/(k_nK_n)^2) + nm_1\exp(-m_0K_n^\alpha)\}\right). \end{aligned}$$

Hence, by choosing the thresholding $\nu_n = c_7 n^{-2\kappa}$, for $c_7 < c_{11} c_2^2/8$, $\mathcal{M}_\star \subset \widehat{\mathcal{N}}_{\nu_n}$ with the probability tending to one exponentially fast, and the result follows.

We now prove (31). It is obvious that

$$\ell_j''(\boldsymbol{\xi}_n) \geq \min_{|x| \leq (B+1)K} b''(x)\, \mathbb{P}_n \mathbf{X}_j \mathbf{X}_j^T I(|X_j| \leq K),$$

for any given K. Since the random variable involved is uniformly bounded in j, it follows from the Hoeffding inequality (Hoeffding, 1963) that

$$P\{|(\mathbb{P}_n - P) X_j^k I(|X_j| \leq K)| > \varepsilon\} \leq \exp(-2n\varepsilon^2/(4K^{2k})), \tag{32}$$

for any $k \geq 0$ and $\varepsilon > 0$. By taking $\varepsilon = n^{-\kappa/2}$, we have

$$P\{|(\mathbb{P}_n - P) X_j^k I(|X_j| \leq K)| > n^{-\kappa/2}\} \leq \exp(-2n^{1-\kappa}/(4K^{2k})).$$

Consequently, with probability tending to one exponentially fast, we have

$$\ell_j''(\boldsymbol{\xi}_n) \geq \min_{|x| \leq (B+1)K} b''(x)\, E\mathbf{X}_j \mathbf{X}_j^T I(|X_j| \leq K)/2, \tag{33}$$

The minimum eigenvalue of the matrix $E\mathbf{X}_j \mathbf{X}_j^T I(|X_j| \leq K)$ is

$$\min_{|a| \leq 1} E(a^2 + 2a\sqrt{1-a^2} X_j + (1-a^2) X_j^2) I(|X_j| \leq K).$$

It is bounded from below by

$$\min_{|a| \leq 1} E\{a^2 + (1-a^2) X_j^2 I(|X_j| \leq K)\} - 2|EX_j I(|X_j| \leq K)| - K^{-2}, \tag{34}$$

where we used $P(|X_j| \geq K) \leq K^{-2}$. Since $EX_j = 0$ and $EX_j^2 = 1$,

$$|EX_j I(|X_j| \leq K)| = |EX_j I(|X_j| > K)| \leq K^{-1} EX_j^2 I(|X_j| > K) \leq K^{-1}.$$

Hence the quantity in (34) can be further bounded from below by

$$\begin{aligned} & EX_j^2 I(|X_j| \leq K) + \min_{|a| \leq 1} a^2 EX_j^2 I(|X_j| > K) - 2K^{-1} - K^{-2} \\ \geq\ & 1 - \sup_j EX_j^2 I(|X_j| > K) - 2K^{-1} - K^{-2}. \end{aligned}$$

The result follows from Condition G and (33).

Proof of Theorem 9. By (30), it can be easily seen that

$$2L_{j,n} \leq D_1 \lambda_{\max}(\mathbb{P}_n \mathbf{X}_j \mathbf{X}_j^T)\, \|\hat{\boldsymbol{\beta}}_0^M - \hat{\boldsymbol{\beta}}_j^M\|^2,$$

where $D_1 = \sup_x b''(x)$ as defined in the proof of Theorem 3. By (32), with the exception on a set with negligible probability, it follows that

$$\lambda_{\max}(\mathbb{P}_n \mathbf{X}_j \mathbf{X}_j^T) \leq 2\lambda_{\max}(E\mathbf{X}_j \mathbf{X}_j^T) = 2,$$

uniformly in j. Therefore, with probability tending to one exponentially fast, we have

$$L_{j,n} \leq D_7\{(\hat{\beta}_0^M - \hat{\beta}_{j,0}^M)^2 + (\hat{\beta}_j^M)^2\}, \tag{35}$$

for some $D_7 > 0$.

We now use (35) to show that if $L_{j,n} > c_7 n^{-2\kappa}$, then $|\hat{\beta}_j^M| \geq D_8 n^{-\kappa}$, with exception on a set with negligible probability, where $D_8 = \{c_8/(2D_7)\}^{1/2}$. This implies that

$$|\widehat{\mathcal{N}}_{\nu_n}| \leq |\widehat{\mathcal{M}}_{\gamma_n}|,$$

with $\gamma_n = D_8 n^{-\kappa}$. The conclusion then follows from Theorem 5.

We now show that $L_{j,n} > c_7 n^{-2\kappa}$ implies that $|\hat{\beta}_j^M| \geq D_8 n^{-\kappa}$, with exception on a set with negligible probability. Suppose that $|\hat{\beta}_j^M| < D_8 n^{-\kappa}$. From the likelihood equations, we have

$$b'(\hat{\beta}_0^M) = \bar{Y} = \mathbb{P}_n b'(\hat{\beta}_{j,0}^M + \hat{\beta}_j^M X_j). \tag{36}$$

From the proof of Theorem 4, with exception on a set with negligible probability, we have $|\hat{\beta}_0^M - \beta_0^M| \leq c_{13} n^{-\kappa}$ and $|\hat{\beta}_{j,0}^M - \beta_{j,0}^M| \leq c_{14} n^{-\kappa}$, for some constants c_{13} and c_{14}. Since $(\beta_0^M, 0)$ and $(\beta_{j,0}^M, \beta_j^M)$ are interior points of the square $\mathcal{B}$ with length $2B$, it follows that with exception on a set with negligible probability, $|\hat\beta_0^M| \le B$ and $|\hat\beta_{j,0}^M| \le B$. Recall $D_1 = \sup_x b''(x)$. By Taylor expansion, for some $0 < s < 1$, we have

$$\begin{aligned} |\mathbb{P}_n b'(\hat\beta_{j,0}^M + \hat\beta_j^M X_j) - b'(\hat\beta_{j,0}^M)| &= |b''(\hat\beta_{j,0}^M + s\hat\beta_j^M X_j)\hat\beta_j^M \mathbb{P}_n X_j| \\ &\le D_1 |\hat\beta_j^M \mathbb{P}_n X_j| = o_P(|\hat\beta_j^M|), \end{aligned} \tag{37}$$

where the last step follows from the facts that $EX_j = 0$ and consequently $\mathbb{P}_n X_j = o(1)$ with an exception on a set of negligible probability, by applying the Hoeffding inequality on Ω_n^c and considering the exponential tail property of X_j. Hence, by (36) and (37), we have

$$|b'(\hat\beta_0^M) - b'(\hat\beta_{j,0}^M)| = o_P(|\hat\beta_j^M|).$$

Let $D_9 = \inf_{|x|\le 2B} b''(x)$, with exception on a set with negligible probability, we have

$$|b'(\hat\beta_0^M) - b'(\hat\beta_{j,0}^M)| \ge D_9 |\hat\beta_0^M - \hat\beta_{j,0}^M|.$$

Therefore, we conclude that

$$|\hat\beta_0^M - \hat\beta_{j,0}^M| = o_P(|\hat\beta_j^M|).$$

By (35), we have $|\hat\beta_j^M| > D_8 n^{-\kappa}$. This is a contraction, except on a set that has a negligible probability. This completes the proof. □

REFERENCES

BICKEL, P. and DOKSUM, K. (2001). *Mathematical statistics : basic ideas and selected topics*, 2nd ed. *Upper Saddle River, NJ: Prentice Hall.*

BICKEL, P. J. and DOKSUM, K. A. (1981). An analysis of transformations revisited. *Journal of the American Statistical Association*, **76** 296–311.

BOX, G. E. P. and COX, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society, Series B*, **26** 211–246.

Candes, E. and Tao, T. (2007). The dantzig selector: statistical estimation when p is much larger than n (with discussion). *The Annals of Statistics*, **35** 2313–2404.

Cox, D. R. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society, Series B: Methodological*, **34** 187–220.

Fahrmeir, L. and Kaufmann, H. (1986). Asymptotic inference in discrete response models. *Statistische Hefte*, **27** 179–205.

Fan, J. and Fan, Y. (2008). High-dimensional classification using features annealed independence rules. *The Annals of Statistics*, **36** 2605–2637.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, **96** 1348–1360.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, **70** 849–911.

Fan, J., Samworth, R. and Wu, Y. (2009). Ultra-dimensional variable selection via independent learning: beyond the linear model. *Journal of Machine Learning Research*, **10** 1829–1853.

Frank, I. E. and Friedman, J. H. (1993). Astatistical view of some chemometrics regression tools (with discussion). *Technometrics*, **35** 109–148.

Friedman, J. H. and Stuetzle, W. (1981). Projection pursuit regression. *Journal of the American Statistical Association*, **76** 817–823.

Greenshtein, E. and Ritov, Y. (2004). Persistence in high-dimensional linear preditor selection and the virtue of overparametrization. *Bernoulli*, **10** 971–988.

Hall, P. and Miller, H. (2009). Using generalised correlation to effect variable selection in very high dimensional problems. *The Journal of Computational and Graphical Statistics* To appear.

Hall, P., Titterington, D. and Xue, J. (2009). Tilting methods for assessing the influence of components in a classifier. *Journal of the Royal Statistical Society, Series B: Statistical Methodology* To appear.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, **58** 13–30.

Huang, J., Horowitz, J. and Ma, S. (2008). Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *The Annals of Statistics*, **36** 587–613.

Kosorok, M. R., Lee, B. L. and Fine, J. P. (2004). Robust inference for univariate proportional hazards frailty regression models. *The Annals of Statistics*, **32** 1448–1491.

Ledoux, M. and Talagrand, M. (1991). *Probability in Banach Spaces: Isoperimetry and Processes*. Springer-Verlag Inc.

Massart, P. (2000). About the constants in talagrands concentration inequalities for empirical processes. *Annals of Probability*, **28** 863–884.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B: Methodological*, **58** 267–288.

van de Geer, S. (2002). M-estimation using penalties or sieves. *Journal of Statistical Planning and Inference*, **108** 55–69.

van de Geer, S. (2008). High-dimensional generalized linear modelsand the lasso. *The Annals of Statistics*, **36** 614–645.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer, New York.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, **50** 1–26.

Zeng, D. and Lin, D. Y. (2007). Maximum likelihood estimation in semiparametric regression models with censored data. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, **69** 507–564.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, **101** 1418–1429.

Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *The Annals of Statistics*, **36** 1509–1533.

Jianqing Fan
Department of Operations Research
and Financial Engineering,
Princeton University,
Princeton, NJ 08544, U.S.A.
E-mail: jqfan@princeton.edu

Rui Song
Department of Statistics
Colorado State University,
Fort Collins 80526, U.S.A.
E-mail: song@stat.colostate.edu

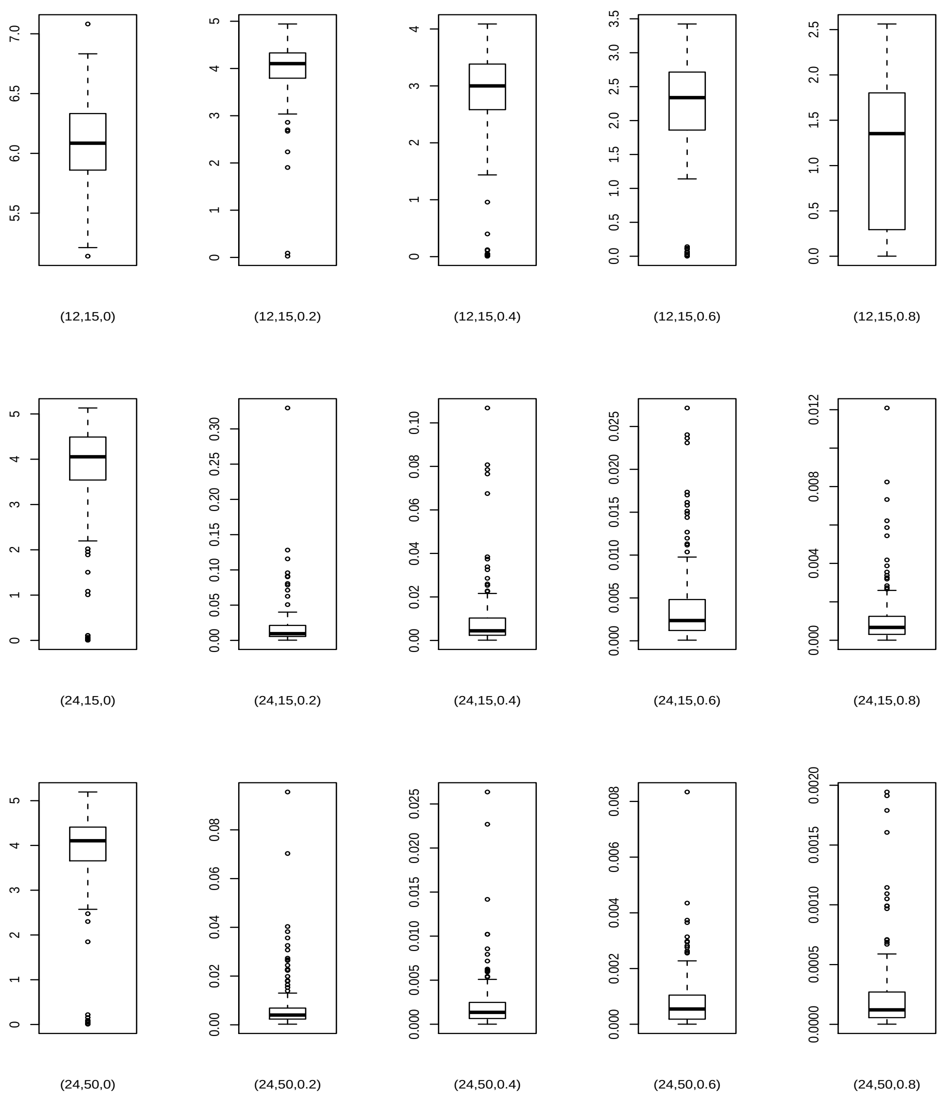

Fig 1. *The boxplots of the minimum* $|t|$*-statistics in the oracle models among 200 simulations for the first setting (S1) with logistic regression examples with* $\boldsymbol{\beta}^{\star} = (3, 4, \ldots)^T$ *when* $s = 12,\ 24$, $q = 15, 50$, $n = 600$ *and* $p = 2000$. *The triplets under each plot represent the corresponding values of* (s, q, ρ) *respectively.*