Published in final edited form as:
J Am Stat Assoc. 2014 ; 109(507): 1257–1269. doi:10.1080/01621459.2013.879531.

The Sparse MLE for Ultra-High-Dimensional Feature Screening

Chen Xu[*] **[Research Assistant]** and **Jiahua Chen [Canada Research Chair, Tier I]**
Department of Statistics, University of British Columbia, Vancouver, BC, Canada, V6T 1Z4.

Abstract

Feature selection is fundamental for modeling the high dimensional data, where the number of features can be huge and much larger than the sample size. Since the feature space is so large, many traditional procedures become numerically infeasible. It is hence essential to first remove most apparently non-influential features before any elaborative analysis. Recently, several procedures have been developed for this purpose, which include the sure-independent-screening (SIS) as a widely-used technique. To gain the computational efficiency, the SIS screens features based on their individual predicting power. In this paper, we propose a new screening method via the sparsity-restricted maximum likelihood estimator (SMLE). The new method naturally takes the joint effects of features in the screening process, which gives itself an edge to potentially outperform the existing methods. This conjecture is further supported by the simulation studies under a number of modeling settings. We show that the proposed method is screening consistent in the context of ultra-high-dimensional generalized linear models.

Keywords

Sure screening property; Hard-thresholding; Sparsity-constrained optimization; Ultra-high dimensionality; Penalized likelihood

1. INTRODUCTION

In genetics and other applications, there are often a huge number of potentially useful variables (features) for explaining the variability in a response of interest. For instance, modern geneticists routinely collect expression data over millions of genes in the hope of discovering those responsible to a disease. Financial analysts are eager to spot clues that lead to rise and fall of stock prices over oceans of information in business, economy and investment activities. Given a set of observations in these applications, statisticians are confronted with the task to identify a number of most influential features and build an interpretive model to link these features to the response variable. Because the feature space is so large, many traditional selection procedures become ineffective. For example, a best subset selection must evaluate a huge number of possibilities. The computationally less demanding stepwise-regression can also be unaffordable. Meanwhile, these procedures are often found unstable in the selection process. Recently, regularization approaches are found most favorable in such applications (Hastie et al. (2009), Donoho (2000), Fan and Lv

[*] (*chen.xu@stat.ubc.ca*).
Jiahua Chen (*jhchen@stat.ubc.ca*)

(2010)). In statistical modeling, these approaches are generally based on the penalized likelihood method (PLM), such as LASSO (Tibshirani (1996)), SCAD (Fan and Li (2001)), elastic-net (Zou and Hastie (2005)), and MCP (Zhang (2010)). They are demonstrated as computationally feasible alternatives for variable selection. Nevertheless, in some applications, the number of features p can be so large that the direct use of PLM becomes less satisfactory. While the LASSO-based methods suffer from selection inconsistency and prediction inaccuracy, the non-convex PLMs are still in search of efficient algorithms for its implementation. At the same time, the number of observations n often remains small or at most moderately sized. This raises the so-called large-p-small-n problem, which in addition makes the tuning of PLM a challenging task.

To meet the computational challenge caused by an exceedingly large p, Fan and Lv (2008) propose to first screen out a large number of non-influential features before a more elaborative selection. Specifically, their strategy is to quickly compute the marginal correlation on each of the p features with the response. Features with top absolute correlations are then retained for an elaborative second stage analysis. Under some model settings, this procedure is shown to retain important features with high probability (sure screening). At the same time, such an “independent” screening makes it computationally highly efficient in practice. This procedure is thereby referred to as the sure-independent-screening (SIS).

Since then, there have been further developments in SIS-based screening procedures. Recent examples include SIRS (Zhu et al. (2011)) and DC-SIS (Li et al. (2012)) that further relax the model flexibility. While these procedures greatly enhance the computational efficiency, they intrinsically ignore the joint effect of candidate features in the screening process. To overcome this shortcoming, Fan et al. (2009) propose to apply SIS iteratively (ISIS) to further improve its sure screening property. Wang (2009) re-vitalize the classical stepwise forward regression (FR) and show its consistency in the sense of Fan and Lv (2008). In this paper, we propose a new screening approach via the sparsity restricted maximum likelihood estimator (SMLE). The SMLE approximates the (ultra) high-dimensional model coefficients on a designated low-dimensional subspace and screens features with zero-estimated coefficients. Unlike the existing methods, the SMLE naturally accounts for the joint effects between features by jointly estimating their model coefficients. Thus, it has a good potential to provide more reliable screening results.

The basis of SMLE falls in a general class of sparsity constrained methods. In disciplines such as wavelet analysis, signal processing and compressed censoring, these methods are frequently used to construct parsimonious representations (approximation) of the high-resolution images/signals for the fast transmission and recovery (Donoho (2006), Candès et al. (2006), Blumensath and Davies (2009)). As for other sparsity constrained methods, a faithful implementation of the SMLE is usually computationally costly. Instead, we design an iterative hard-thresholding-based algorithm (IHT) to approximately solve the SMLE. Each iteration under this algorithm increases the value of the sparsity constrained likelihood via simple operations and thereby provides an improved sparse solution. The joint effects of features are naturally accounted at each iteration as a basis for the next update. A more competitive screening procedure is therefore anticipated.

We show that the proposed SMLE enjoys the sure screening property in the sense of Fan and Lv (2008) under the ultra-high dimensional generalized linear models (GLMs; McCullagh and Nelder (1989)). We establish the convergence of the accompanying IHT algorithm and further show its screening effectiveness via the LASSO-type initial settings. The high efficiency of the new method is illustrated through extensive simulation studies.

The rest of this paper is organized as follows. In Section 2, we briefly review the GLM and the SIS-based screening method. In Section 3, we investigate the use of SMLE for feature screening, discuss its asymptotic properties, and introduce the IHT algorithm. We discuss the screening-based PLM in Section 5 and assess the finite sample performance of the proposed method in Section 6. Concluding remarks are given in Section 7 and the proofs of theorems are provided in Appendix. We show additional numerical examples in the on-line supplemental file.

2. MODEL SETTING AND FEATURE SCREENING

Let Y be a random response depending on p explanatory variables (features) $\boldsymbol{x} = (x_1, \ldots, x_p)$. We postulate a generalized linear model (GLM) between Y and $\boldsymbol{x}$ as follows. Given $\boldsymbol{x}$, the density function of Y is in the form of

$$f(y;\theta) = exp\{\theta y - b(\theta) + c(y)\} \quad (1)$$

with respect to a σ-finite measure ν, where $b(\cdot)$ and $c(\cdot)$ are some known functions. Parameter θ is called the natural parameter and set $\Theta = \{\theta : \int f(y;\theta)\, d\nu < \infty\}$ is the natural parameter space. We assume that $\theta \in \bar{\Theta}$ with $\bar{\Theta}$ denoting a compact subspace of Θ and $b(.)$ is twice continuously differentiable. Under this model, $E(Y|\boldsymbol{x}) = b'(\theta)$ and $Var(Y|\boldsymbol{x}) = b''(\theta)$ with the primes being derivatives. Parameter θ is connected to $\boldsymbol{x}$ through a pre-specified link function

$$g(\mu) = \boldsymbol{x}^T \boldsymbol{\beta},$$

for some $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$. The canonical link $g = \mu^{-1}$ leads to a particular simple form with $\theta = \boldsymbol{x}^T\beta$. Classic GLMs with canonical link include the normal linear regression, the logistic regression and the Poisson regression among others.

Given n independent observations of $(Y, \boldsymbol{x})$, the statistical inference usually focuses on the regression coefficient β. A non-zero β_j characterizes the effect of feature x_j. When p is small, a well fitted GLM is often helpful to interpret the influence of $\boldsymbol{x}$ on the variability in Y. When p is very large, nevertheless, a full model with all $\beta_j \neq = 0$ has likely poor interpretive values. In this situation, it is desirable to build a model with many zero β_j's. This leads to a sparsity assumption on β in many applications. For instance, $\boldsymbol{x}$ may represent millions of genetic variations while Y denotes a disease status. Mostly likely, only a handful genetic features have meaningful influences on Y , which makes the feature selection essential in such analyses. However, as mentioned, the high-dimensionality posts serious challenges from both analytical and computational aspects.

For the convenience of presentation, we use s to denote an arbitrary subset of $\{1,\dots,p\}$, which amounts to a submodel with covariates $\boldsymbol{x}_s = \{x_j, j \in s\}$ and associated coefficients $\boldsymbol{\beta}_s = \{\beta_j, j \in s\}$. Also, we use $\|\cdot\|_0$ to denote the number of non-zero components of a vector (i.e. the l_0-norm), and use $\tau(s)$ to indicate the size of model s. In particular, we denote the true model by $s^* = \{j : \beta_j \neq 0\}$ with $\tau(s^*) = \|\boldsymbol{\beta}^*\|_0 = q$. The objective of feature selection boils down to identify s^* from $\{1,\dots,p\}$ through analyzing the data $\{(y_i, \boldsymbol{x}_i), i = 1,\dots,n\}$. Particularly, we are interested in the ultra-high dimensional situations where $p \gg n$.

When the dimension of $\boldsymbol{x}$ is ultra-high, screening features in $\boldsymbol{x}$ before an elaborative selection seems to be a computationally feasible strategy. To this end, Fan and Lv (2008) study a sure independent screening (SIS) procedure. In this approach, they first fit Y against a single feature x_j (standardized) and obtain the regression coefficient w_j for all j. With a pre-specified screening bound $k < n$, the SIS retains features in the set

$$\check{s} = \{j : |w_j| \quad \text{is among the} \quad k \quad \text{largests}\}.$$

More elaborate analysis will then be carried out by building a model between Y and $\boldsymbol{x}_{\check{s}}$.

Under the linear model, Fan and Lv (2008) prove that $P(s^* \subset \check{s}) \to 1$ as $n \to \infty$, even when p increases exponentially with n. This "sure screening property" together with its computational efficiency makes it widely adopted in the genetic studies (Efron (2007), Storey and Tibshirani (2003)). Fan and Song (2009) further extend their findings to the ultra-dimensional GLM with both discrete valued response and features.

Apparently, the SIS only considers the marginal effect of x_j one-at-a-time. This prompts an iterative SIS (ISIS) procedure (Fan et al. (2009)). The ISIS first applies SIS to select some k_1 features $\check{s}_1$. It then regresses Y against $\boldsymbol{x}_{\check{s}_1}$ via PLM such as SCAD to remove less significant features to get $\check{\omega}_1 \subset \check{s}_1$. Next, Y is regressed against $\boldsymbol{x}_{\check{\omega}_1}$ to obtain corresponding residuals, which are then treated as responses and the SIS is applied to $\check{\omega}_1^c$ to select k_2 additional features $\check{s}_2$. The iteration starts all over again on $\check{\omega}_1 \cup \check{s}_2$ to get $\check{\omega}_2$ and repeatedly until k features are recruited. Because the residuals from regressing Y against $\boldsymbol{x}_{\check{\omega}_i}$ are uncorrelated to $\boldsymbol{x}_{\check{\omega}_i}$, features correlated to Y through $\boldsymbol{x}_{\check{\omega}_i}$ are unlikely to be selected in the $(i + 1)$-th iteration. Such a consideration leaves room for features with weaker but still significant effects. The gain of ISIS is apparently built on higher computational cost and increased complexity. In the proposed new method, we hope to retain the insight of ISIS with a conceptually simpler and computationally cheaper procedure.

3. THE SPARSE-MLE AND IHT ALGORITHM

Under the canonical link and given a random sample of size n, the log-likelihood function of β is given by

$$\ell_n(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left\{ \left(\boldsymbol{x}_i^T \boldsymbol{\beta} \right) y_i - b\left(\boldsymbol{x}_i^T \boldsymbol{\beta} \right) \right\}. \quad (2)$$

Maximizing (2) with respect to β leads to its maximum likelihood estimator (MLE). Under some conditions and for a fixed p, the MLE is consistent as $n \to \infty$. This result, however, becomes not applicable when p is large (diverging with n), and even meaningless when $p \gg n$.

Suppose the effect of $\boldsymbol{x}$ is sparse. That is, the cardinality of the true model s^* is $\tau(s^*) = q \leq k$ for some known k. Then, it would be more sensible to search for the sparsity restricted MLE of β (SMLE) defined by

$$\hat{\boldsymbol{\beta}}_{[k]} = argmax_{\boldsymbol{\beta}} \quad \ell_n(\boldsymbol{\beta}) \quad \text{subject to} \quad \|\boldsymbol{\beta}\|_0 \leq k. \quad (3)$$

Let $\hat{s} = \left\{ j : \hat{\beta}_{[k]j} \neq 0 \right\}$ correspond to non-zero entries of $\hat{\boldsymbol{\beta}}_{[k]}$. If $\hat{\boldsymbol{\beta}}_{[k]}$ (and therefore $\hat{s}$) can be obtained at a relatively low computational cost, we then have a joint-likelihood-supported feature screening method, which naturally accounts for the joint effects between features.

There are two dilemmas for the use of SMLE, as it resembles the best-subset section procedure. As known, the best-subset methods for variable selection can be unstable and computationally expensive. For the screening purpose, however, the aim is to retain a relatively small number of features for the further selection. Hence, the instability might not be a serious concern as long as the screened set retains all relevant features. Computationally, a number of strategies have been developed to approximately solve the associated numerical problem. Examples include the matching pursuit algorithms (Mallat and Zhang (1993)) and the FOCUSS-based methods (Murray and Kreutz-Delgado (2001)). Among these, the hard-thresholding technique for linear models (Blumensath and Davies (2008)) is particularly helpful to our need.

We develop an iterative hard-thresholding algorithm (IHT) for GLM to approximately compute the SMLE. This algorithm starts with approximating $\ell_n(\cdot)$ at a generic β by

$$h_n(\boldsymbol{\gamma};\boldsymbol{\beta}) = \ell_n(\boldsymbol{\beta}) + (\boldsymbol{\gamma} - \boldsymbol{\beta})^T S_n(\boldsymbol{\beta}) - (u/2) \|\boldsymbol{\gamma} - \boldsymbol{\beta}\|_2^2, \quad (4)$$

for some scaling parameter $u > 0$, where $\|.\|_2$ denotes the L_2 norm and $S_n(\boldsymbol{\beta}) = \ell_n'(\boldsymbol{\beta})$ is the score function. The first two terms in (4) match the Taylor's expansion of $\ell_n(\cdot)$ at β, and $(u/2)\|\boldsymbol{\gamma} - \boldsymbol{\beta}\|_2^2$ is a regularization term. It is seen that $\ell_n(\boldsymbol{\beta}) = h_n(\boldsymbol{\beta};\boldsymbol{\beta})$, and that $h_n(\gamma; \beta)$ well approximates $\ell_n(\boldsymbol{\beta})$ for γ close to β. The additivity of $h_n(\gamma; \beta)$ in the components of γ makes the maximization of $h_n(\gamma; \beta)$ over γ computationally highly efficient.

Using (4), an approximate solution to (3) is obtained by the following iterative procedure

$$\boldsymbol{\beta}^{(t+1)} = arg\ \max_{\boldsymbol{\gamma}} h_n\left(\boldsymbol{\gamma};\boldsymbol{\beta}^{(t)}\right) \quad \text{subject to} \quad \|\boldsymbol{\gamma}\|_0 \leq k. \quad (5)$$

At each iteration, the regularization term in $h_n(\cdot; \cdot)$ prevents its maximizer moving far away from $\beta^{(t)}$. This feature makes $\beta^{(t+1)}$ a non-greedy update in searching for $\hat{\boldsymbol{\beta}}_{[k]}$. We start with an initial $\beta^{(0)}$ and carry out the iteration specified by (5) until $\|\beta^{(t+1)} - \beta^{(t)}\|_2$ falls below some tolerance level.

Let $\boldsymbol{y} = (y_1,..., y_n)^T$ and $\boldsymbol{X} = (\boldsymbol{x}_1,..., \boldsymbol{x}_n)^T$. Due to the additivity of $h_n(\gamma, \beta)$ in γ, the optimization in (5) takes a unified specific form as

$$\min_{\boldsymbol{\gamma}} \|\boldsymbol{\gamma} - \mathbf{u}^{-1}\left\{\mathbf{u}\boldsymbol{\beta} + \boldsymbol{X}^{\mathbf{T}}\boldsymbol{y} - \boldsymbol{X}^{\mathbf{T}}\mathbf{b}'(\boldsymbol{X}\boldsymbol{\beta})\right\}\|_2^2 \quad \text{subject to} \quad \|\boldsymbol{\gamma}\|_0 \leq k. \quad (6)$$

Clearly, without the sparsity constraint, the solution to (6) is the trivial $\tilde{\boldsymbol{\gamma}} = \boldsymbol{\beta} + \mathbf{u}^{-1}\boldsymbol{X}^{\mathbf{T}}\left\{\boldsymbol{y} - \mathbf{b}'(\boldsymbol{X}\boldsymbol{\beta})\right\}$. With the sparsity constraint, the analytical solution to (6) is then given by choosing the k largest components of $\tilde{\boldsymbol{\gamma}}$ in absolute value. Namely,

$$\hat{\boldsymbol{\gamma}} = \boldsymbol{H}(\tilde{\boldsymbol{\gamma}}; k) = \left[H(\tilde{\gamma}_1; r), \ldots, H\left(\tilde{\gamma}_p; r\right)\right]^T,$$

where r is the kth largest component of $|\tilde{\boldsymbol{\gamma}}|$ and $H(\gamma, r) = \gamma I(|\gamma| > r)$ is a hard thresholding function. Given the t-th sparse solution $\boldsymbol{\beta}^{(t)}$, the iteration (5) can be re-written as

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{H}\left(\boldsymbol{\beta}^{(t)} + u^{-1}\boldsymbol{X}^T\left\{\boldsymbol{y} - b'\left(\boldsymbol{X}\boldsymbol{\beta}^{(t)}\right)\right\}; k\right). \quad (7)$$

The thresholding-based iterative procedure involves no heavy-duty operations such as the matrix inversion. This advantage merits (7) particularly suitable for the large-p-small-n application. Meanwhile, the joint information carried in $\boldsymbol{X}$ is naturally accounted at each iteration as a basis for the next update, which further prompts the effectiveness of IHT. By Theorem 1, we show that the sequence $\{\boldsymbol{\beta}^{(t)}\}$ based on IHT increases the value of $\ell_n(\cdot)$ and necessarily converges to a local maximum of $\ell_n(\cdot)$.

Theorem 1. Following the settings and notations we introduced earlier, let $\{\boldsymbol{\beta}^{(t)}\}$ *be the sequence defined by (7). Denote by* ρ_1 *the largest eigenvalue of* $\boldsymbol{X}^T\boldsymbol{X}$ *and*

$$\rho^{(t)} = \max_{1 \leq i \leq n} \sup\left\{b''\left(\boldsymbol{x}_i^T\tilde{\boldsymbol{\beta}}\right) : \tilde{\boldsymbol{\beta}} = \alpha\boldsymbol{\beta}^{(t+1)} + (1-\alpha)\boldsymbol{\beta}^{(t)}, 0 \leq \alpha \leq 1\right\}.$$

If $u \geq \rho_1\rho^{(t)}$, *then*

$$\ell_n\left(\boldsymbol{\beta}^{(t+1)}\right) \geq \ell_n\left(\boldsymbol{\beta}^{(t)}\right).$$

Moreover, if $\sum_i \boldsymbol{x}_s\boldsymbol{x}_s^T > 0$ *for all s such that* $\|s\|_0 \leq k$, *then* $(\boldsymbol{\beta}^{(t)}\}$ *converges to a local maximum of* $\ell_n(\boldsymbol{\beta})$ *subject to* $\|\boldsymbol{\beta}\|_0 \leq k$. The proof is given in Appendix. Theorem 1 claims that, with an appropriate scale parameter u, the IHT necessarily improves the current estimate within the feasible region (i.e. $\|\boldsymbol{\beta}\|_0 \leq k$). It thereby serves as a refinement procedure for any initial value on $\boldsymbol{\beta}$, which leads to a sparse estimate better supported by the joint likelihood. The increment property of IHT features the SMLE a promising method for variable screening.

Regarding the setup for u, empirical studies suggest that a smaller value of u (larger step width) often leads to a faster convergence of IHT. Nevertheless, Theorem 1 indicates that only when u is large enough, the likelihood function is guaranteed to increase after each iteration. For normal linear regression, we have $b''(\theta) = 1$, and thus setting $u = \rho_1$ is an arbitrary choice that satisfies the theorem requirement. For logistic regression, we have $b''(\theta) \leq 1/4$, which implies $u = \rho_1/4$ is a reasonable choice. More generally, one may select u adaptively within each step of IHT, to guarantee the increment of $l_n(\beta^{(t)})$ and therefore the convergence of $\beta^{(t)}$. In simulations, we first initiate a tentative small size for u as $u = \rho_1$, after which we check whether or not $l_n(\beta^{(t+1)}) \geq l_n\, l_n(\beta^{(t)}$. If the equality is violated, we then double the current u value until $l_n(\beta^{(t+1)}) \geq l_n(\beta^{(t)}$ is satisfied. It is trivial that the adaptive procedure does not alter the convergence of IHT.

4. SURE SCREENING OF SMLE

We now provide some theoretical justifications for the SMLE-based feature screening. Recall that we use s^* to denote the true model and $\hat{s}$ is the screened model based on SMLE. In addition to the computational efficiency, a good screening approach is expected to retain all relevant features while removing the others. Such a desirable feature is referred to as the sure screening property (Fan and Lv (2008)). That is, as $n \to \infty$,

$$P\left(s^* \subset \hat{s}\right) \to 1. \quad (8)$$

To investigate whether the SMLE enjoys the property (8), we introduce some additional notations as follows. For any model s, let

$$\begin{aligned} S_n\left(\boldsymbol{\beta}_s\right) &= \frac{\partial \ell(\boldsymbol{\beta}_s)}{\partial \boldsymbol{\beta}_s} = \sum_{i=1}^{n} \left[y_i - b'\left(\boldsymbol{x}_{is}^T \boldsymbol{\beta}_s\right) \right] \boldsymbol{x}_{is}, \\ H_n\left(\boldsymbol{\beta}_s\right) &= -\frac{\partial^2 \ell(\boldsymbol{\beta}_s)}{\partial \boldsymbol{\beta}_s \partial \boldsymbol{\beta}_s^T} = \sum_{i=1}^{n} b''\left(\boldsymbol{x}_{is}^T \boldsymbol{\beta}_s\right) \boldsymbol{x}_{is} \boldsymbol{x}_{is}^T \end{aligned}$$

be the score function and the Hessian matrix of $\ell_n(\cdot)$ corresponding to β_s. Suppose a screening procedure retains k out of p features such that $\pi(s^*) = q < k$, we define

$$\boldsymbol{S}_+^k = \{s : s^* \subset s; \|s\|_0 \leq k\} \quad \text{and} \quad \boldsymbol{S}_-^k = \{s : s^* \not\subset s; \|s\|_0 \leq k\}$$

as the collections of the over-fitted models and the under-fitted models. We investigate the asymptotic properties of $\hat{\boldsymbol{\beta}}_{[k]}$ under the scenario where p, q, k and β^* vary along with the sample size n. Also, we assume the following conditions, some of which are purely technical and only serve to provide theoretical understanding of the new screening method. We do not intend to make these assumptions the weakest possible.

T1 $\log p = O(n^m)$ for some $0 \leq m < 1$.

T2 There exists $w_1, w_2 > 0$ and some non-negative constant τ_1, τ_2 such that

$$\min_{j\in s^*}|\beta_j^*| \geq w_1 n^{-\tau_1} \quad \text{and} \quad q<k \leq w_2 n^{\tau_2}.$$

T3 There exist constants $c_1 > 0$, $\delta_1 > 0$, such that for sufficiently large n,

$$\lambda_{min}\left[n^{-1}H_n(\boldsymbol{\beta}_s)\right] \geq c_1$$

for $\boldsymbol{\beta}_s \in \{\boldsymbol{\beta}_s : \|\boldsymbol{\beta}_s - \boldsymbol{\beta}_s^*\|_2 \leq \delta_1\}$ and $s \in \boldsymbol{S}_+^{2k}$, where $\lambda_{min}[\cdot]$ denotes the smallest eigenvalue of a matrix.

T4 Let $\sigma_i^2 = b''\left(\boldsymbol{x}_i^T\boldsymbol{\beta}^*\right)$. There exists positive constants c_2, c_3, such that $|x_{ij}| \leq c_2$, and when n is sufficiently large,

$$\max_{1\leq j\leq p} \max_{1\leq i\leq n}\left\{\frac{x_{ij}^2}{\sum_{i=1}^n x_{ij}^2\sigma_i^2}\right\} \leq c_3 \cdot n^{-1}.$$

By Condition T1, we assume that p diverges with n up to an exponential rate, which implies that the number of covariates can be substantially larger than the sample size. Condition T2 states a few requirements for establishing the sure screening of SMLE. The first one is the sparsity of β^* which makes the sure screening possible with $\tau(\hat{s}) = k > q$. Also, it requires the minimal component in β^* does not degenerate too fast, so that the signal is detectable in the asymptotic sequence. Meanwhile, together with T3, it confines an appropriate order of k that guarantees the identifiability of s^* over s for $\tau(s) \leq k$. Condition T3 corresponds to the UUP condition given by Candes and Tao (2007), which has been used as a relatively mild condition in the literature of high-dimensional methods (see e.g., Wang (2009), Chen and Chen (2012)). Condition T4 places restrictions on the observed values of $\boldsymbol{x}$. With a wide range of models, Condition T4 holds naturally for the random designs on $\boldsymbol{x}$ or for fixed designs with appropriate re-scaling operation.

We establish the sure screening property of SMLE by the following theorem.

Theorem 2. Suppose we have n independent observations with p candidate features from model (1) and conditions T1-T4 are satisfied with $\tau_1 + \tau_2 < (1 - m)/2$. Let $\hat{s}$ be the features obtained by the SMLE (3) of size k. We have

$$P(s^* \subset \hat{s}) \to 1, \quad as \quad n \to \infty.$$

The proof is given in Appendix. The sure screening of SMLE provides a necessary condition for correctly identifying s^* through a more elaborative selection based on $\hat{s}$. Nevertheless, Theorem 2 does not provide a concrete guidance on the choice of k for using SMLE in practice. In general, an appropriate k may depend on specific data structures as well as the nature of model sparsity. Apparently, a large k allows SMLE to retain more features in $\hat{s}$, although it brings more difficulties to the second stage analysis. However, an overly liberal k

tends to violate the identifiability of s^*, which in turn deteriorates the preference of SMLE. In applications such as risk investment and face cognition, prior knowledge on the model sparsity naturally provides a guidance for choosing a sensible k. In our simulations, we set $k = a\log(n)n^{1/3}$ with $a = (1, 2/3, 1/3)$ for linear models, Poisson models and logistic models respectively. These empirical choices are analogous to the recommended k values in Fan et al. (2009), which seem to work satisfactorily in our model settings. In fact, we find the performances of SMLE are robust to a wide choices of k, which facilitates the use of SMLE by avoiding an elaborative specification on k.

It is natural to question whether the proposed IHT algorithm necessarily leads to the SMLE $\hat{\boldsymbol{\beta}}_{[k]}$ (and therefore $\hat{s}$), so that the result of Theorem 2 becomes applicable. Unfortunately, due to the complexity of the optimization problem in (3), there is no guarantee that $\hat{\boldsymbol{\beta}}_{[k]}$ is the outcome of a single run of IHT. A good initial setup increases the chance of IHT to hit a local maximizer with the desired sure screening property that works as $\hat{\boldsymbol{\beta}}_{[k]}$. For the linear model, when the IHT starts with $\boldsymbol{\beta}^{(0)} = 0$, its first iteration corresponds to the SIS-based screening. This suggests that zero might be a reasonable choice of $\boldsymbol{\beta}^{(0)}$. At the same time, since the LASSO estimate of $\boldsymbol{\beta}$ is computationally convenient and estimation consistent, it is also an excellent choice of $\boldsymbol{\beta}^{(0)}$. Specifically, we show that, with an appropriate LASSO-type initial value, the IHT update $\boldsymbol{\beta}^{(t)}$ does enjoy the sure screening property within a finite number of iterations. This finding further enriches the proposed SMLE method from an implementary aspect.

To facilitate the proof, let $\|\cdot\|_1$ be the L_1 norm and s_c be the complement of s in $\{1,\ldots, p\}$. We introduce one more technical condition as follows:

T3' There exists a positive constant ν, such that for sufficiently large n,

$$\boldsymbol{\delta}^T H_n(\boldsymbol{\beta})\boldsymbol{\delta} \geq \nu \mathbf{n} \|\boldsymbol{\delta}_{\mathbf{s}^*}\|_{\mathbf{2}}^{\mathbf{2}}$$

for any $\boldsymbol{\delta} \neq 0$ satisfying $\|\boldsymbol{\delta}_{\boldsymbol{s}_c^*}\|_1 \leq 3\|\boldsymbol{\delta}_{\boldsymbol{s}^*}\|_1$, and for any $\boldsymbol{\beta}$ such that $\theta = \boldsymbol{x}^T\boldsymbol{\beta} \in \bar{\boldsymbol{\Theta}}$.

Condition T3' corresponds to the restricted eigenvalue condition given in Bickel et al. (2009). It ensures a desirable bound on the L_1 estimation loss of the LASSO estimator, which thereby serves as a foundation for the sure screening of the IHT iteration initiated by LASSO. This condition is slightly more stringent than T3 on the identifiability of s^*. We refer to Bickel et al. (2009) for more discussions and other similar conditions. The sure screening of the LASSO-initiated IHT procedure is stated as the following theorem.

Theorem 3. *Under the settings of Theorem 1, let* $\boldsymbol{\beta}^{(t)}$ be the t-th update of the IHT procedure (7) with $u \geq \xi kn$ for some $\xi > 0$ *and let* $s^{(t)} = \left\{j : \beta_j^{(t)} \neq 0\right\}$ *be its screened features. Suppose*

$$\boldsymbol{\beta}^{(0)} = arg\,\max_{\boldsymbol{\beta}} \{l_n(\boldsymbol{\beta}) - n\lambda\|\boldsymbol{\beta}\|_1\}$$

with λ *satisfying* $\lambda n^{(1-m)/2} \to \infty$ *and* $\lambda n^{\tau_1+\tau_2} \to 0$ *for* $0 \le \tau_1 + \tau_2 < (1-m)/2$. *Then, under T1-T2, T3' and T4,*

$$P\left(s^* \subset s^{(t)}\right) \to 1$$

for any finite $t \ge 1$, *as* $n \to \infty$.

The proof is given in Appendix. Clearly, Theorem 3 relies on the appropriate choice of tuning parameter λ, which controls the amount of L_1 regularization on $\beta^{(0)}$. Generally, one may choose λ via any tuning strategy that is well-developed for LASSO. In simulations, we find the λ that leads to $\|\beta^{(0)}\|_0 = n - 1$ is often an adequate choice for the IHT-implemented SMLE method.

Also note that, in Theorem 3, there are practically no restrictions on the size of ξ and consequently on u. We may therefore simply follow the requirement of Theorem 1 in deciding the size of u to ensure the convergence of IHT and the validity of Theorem 3. The theorem also implies that, with an appropriate $\beta^{(0)}$, the IHT-implemented SMLE can potentially retain all relevant features after any pre-given number of iterations.

5. THE SCREENING-BASED PLM SELECTION PROCEDURE

We have proposed a SMLE method for feature screening in ultra-high dimensional data analyses. While the relevant features are retained after screening, there are still many irrelevant ones survived in $\hat{s}$. To obtain a more interpretable model, it is therefore necessary to perform a more elaborative selection on the remaining features. In particular, a PLM can be conveniently used for this purpose.

Following the modeling settings and notations in Section 2, the SMLE-PLM procedure estimates β by $\hat{\boldsymbol{\beta}}_\lambda(\hat{s})$ through maximizing

$$Q(\boldsymbol{\beta}_{\hat{s}}) = \ell_n(\boldsymbol{\beta}_{\hat{s}}) - n\sum_{j\in\hat{s}} \phi_\lambda(|\beta_j|), \quad (9)$$

where $\hat{s}$ is the k-dimensional submodel obtained from the SMLE and $\varphi_\lambda(.)$ is a specified penalty function. As in typical PLMs, an appropriate $\varphi_\lambda(.)$ with spikes at origin leads to the sparse structure of $\hat{\boldsymbol{\beta}}_\lambda(\hat{s})$. We may tune the degree of regularization to achieve desired level of model sparsity and therefore select the final features.

Specifically, given a penalty function $\varphi_\lambda(\cdot)$, the sparsity of the final solution is determined by the tuning parameter λ. With different values of λ, $\hat{\boldsymbol{\beta}}_\lambda(\hat{s})$ leads to models of differing sparsity. A criterion is then needed to determine the final choice among these candidate models (and thereby the influential features). To this end, cross-validation (Stone (1974)) and information-based criteria such as AIC (Akaike (1973)) or BIC (Schwarz (1978)) are frequently used in the literature. The former criterion gauges the prediction accuracy of the resulting model, while the later ones balance the goodness of fit with the model complexity. In this paper, we recommend choosing a model sparsity via minimizing an extended Bayes

information criterion (Chen and Chen (2008)). More specifically, for a given $\varphi_\lambda(\cdot)$ (such as LASSO or SCAD), we determine the final model via minimizing

$$EBIC(\lambda) = -2\ell\left(\hat{\boldsymbol{\beta}}_\lambda(\hat{s})\right) + \|\hat{\boldsymbol{\beta}}_\lambda(\hat{s})\|_0 (\log n + 0.5\log p). \quad (10)$$

Compared with traditional BIC, the EBIC places more penalties on the model complexity by linking the regularization to the number of features p. This modification is particularly helpful to select an appropriate λ in SMLE-PLM that leads to a parsimonious model in the ultra-high dimensional applications where $p >> n$.

Under mild regularity conditions, Chen and Chen (2008) established the selection consistency of EBIC under the ulra-high dimensional GLMs. This result together with the sure screening of SMLE guarantees the consistency of the SMLE-PLM procedure as specified above. For a conceptually more focused presentation, we choose to skip the tedious details in this article.

6. NUMERICAL STUDIES

We assess the finite sample performance of the SMLE and its associated PLM procedure through simulation studies. There are many factors to be considered in order to obtained a relatively complete picture of the new method. In particular, the performance of screening-based methods is heavily dependent on the correlation structure between features, the number of relevant features, sample size, total number of candidate features, the error variance and others. We use several subsections to describe these aspects of the simulation settings.

6.1. General Settings

We examine the methods under three classes of models: linear regression, logistic regression and Poisson regression. Specifically, for the linear model, we compare five screening methods: SIS, ISIS, FR, LASSO and SMLE, while we do not include FR for the logistic and Poisson models due to the exceeding computational expenses.

For each class of models, we consider three correlation structures between candidate features. The first one is when the features are ideally independent with each other, under which the task of feature selection is the most straightforward. The second is when the features have an auto-correlation structure. The neighboring features are correlated but distant ones are virtually uncorrelated. This type of correlation is commonly used in modeling the data with a natural order. The last one set every feature either relevant or correlated with some relevant features. This structure makes it challenging to distinguish the relevant features from irrelevant features.

More specifically, the correlation structures under consideration are listed as follows.

S1: Candidate features are independently generated as a random sample from $N(0, 1)$.

S2: Candidate features $x_1,..., x_p$ are joint normal, marginally $N(0, 1)$, with $\text{cov}(x_j, x_{j-1}) = 2/3$, $\text{cov}(x_j, x_{j-2}) = 1/3$ for $j \geq 3$ and $\text{cov}(x_j, x_h) = 0$ for $|j - h| \geq 3$.

S3: Candidate features $x_1,..., x_p$ are joint normal, marginally $N(0, 1)$, with $\text{cov}(x_j, x_h) = 0.15$ for $j, h \in s^*$ and $\text{cov}(x_j, x_h) = 0.3$ for j or $h \in s_c^*$.

These structures are also chosen because they have been discussed in many papers (see, e.g., Fan et al. (2009), Wang (2009)). Thus, it prevents potential bias in favor of the new method.

6.2. Implementation Issues

All screening methods give potential users some flexibilities. Because of this, we need to make a few specific choices in the simulation study. In general, we follow the recommendation of authors for each method included in this simulation.

Specifically, we conduct all simulation studies by software R on an Unix server with 2.4 GHz CPUs. For SIS/ISIS, we use the R-function **GLMvanISISscad** of package **SIS**, where a maximum of five ISIS loops are carried out. The number of relevant features kept in each loop is decided by SCAD with the AIC-based tuning method; see Fan et al. (2009). We use the IHT algorithm to implement SMLE as proposed, where we set $\beta^{(0)}$ by the LASSO estimate with sparsity $n - 1$ and stop the iterations when $\|\beta^{(t)} - \beta^{(t-1)}\|_2 < 10^{-3}$. We also treat LASSO as a screening method (retaining the first k features in its solution path) and use R-function **glmnet** for its implementation.

As discussed, we set the screening bound $k = a\log(n)n^{1/3}$ with $a = (1, 2/3, 1/3)$ for linear models, Poisson models and logistic models respectively. In simulation setups, these choices also closely match the recommenced k values in Fan et al. (2009) for (I)SIS, which should make the comparison objective. Since the goal of our simulation is to compare different screening methods, we treat these model-based k values as common benchmarks for the comparison. In fact, similar simulation results can be obtained under a wide range of k values.

After the completion of screening, a more elaborative second stage analysis is conducted to choose the relevant features. Two commonly used PLMs, LASSO and SCAD, are readily recommended and we tune them by EBIC. R-packages **glmnet** and **SIS** are used for LASSO and SCAD respectively.

6.3. Summary Statistics

We assess the performances of screening methods based on $T = 500$ simulation replications. Specifically, let $\hat{s}_t$ denote the model selected in the tth replication by any specific method. We measure its retaining capacity (RC) of relevant features by

$$RC = T^{-1}\sum_{t=1}^{T} I\left(s^* \subset \hat{s}_t\right).$$

In particular, we compute the RC values for each feature screening method and its associated PLM procedures.

We characterize the model selectivity in terms of positive selection rate (PSR) and false discovery rate (FDR):

$$PSR=\frac{\sum_{t=1}^{T}\|s^*\cap\hat{s}_t\|_0}{T\|s^*\|_0},\quad FDR=\frac{\sum_{t=1}^{T}\|\hat{s}_t-s^*\|_0}{T\|\hat{s}_t\|_0}.$$

The PSR and FDR depict two different aspects of a selection result: a high PSR means most relevant features are identified, while a low FDR indicates few irrelevant features are miss-selected. As further references, we also report the averaged model size (AMS) of $\hat{s}_t$, averaged computational time for conducting each study (TIME, in seconds), as well as the proportion of times when $\hat{s}_t = s^*$ (correct selection rate; CSR).

For each model under study, we further specify its parameter settings in the following subsections, where the corresponding simulation results are also given.

6.4. Linear Regression

In this example, data $(y_i, \boldsymbol{x}_i)$ for $i = 1,..., n$ are generated as independent copies from $(Y, \boldsymbol{x})$ according to a linear model

$$Y=\boldsymbol{x}^T\boldsymbol{\beta}+\epsilon,$$

where ε is a $N(0, \sigma^2)$ distributed random error. The model parameters used in each of the three correlation setups are as follows

S1: s^* is a simple random sample of size 8 from $\{1,..., p\}$. β_{s^*} are independent samples from $(4\sqrt{n\,log\,n}+|Z|)\,U$ and $\boldsymbol{\beta}_{\boldsymbol{s}_c^*}=0$, where $P(U = 1) = 0.6$, $P(U = –1) = 0.4$ with $Z \sim N(0, 1)$. We set $(n, p, \sigma) = (200, 10000, 3)$.

S2: $s^* = \{1, 3, 5, 7, 9\}$ with effects $\beta_{s^*} = (5, 3.5, 2.8, 2.5, 2.2)^T$ and $\boldsymbol{\beta}_{\boldsymbol{s}_c^*}=0$. We set $(n, p, \sigma) = (120, 5000, 5)$.

S3: $s^* = \{1, 2, 3, 4\}$ with effects $\beta_{s^*} = (2.5, 2.5, 2.5, 2.5)^T$ and $\boldsymbol{\beta}_{\boldsymbol{s}_c^*}=0$. We set $(n, p, \sigma) = (100, 1000, 1)$.

Setup S1 is taken from Example 1 of Wang (2009). In S2, s^* is chosen to have non-negligible correlations between relevant features. For S3, the most challenging situation, we fix s^* to be the set of the first four features and set an equal coefficient for all the relevant feature at 2.5. The σ values are chosen after pilot studies to control an appropriate signal-to-noise ratio. These specifications were paired up the the covariance structures S1, S2 and S3 in section 6.1.

Under linear model, the prediction accuracy is a good performance measure for each model selected after screening and PLM. For this purpose, we generated a test data of size n from the true model s^*, $(\tilde{y}, \tilde{\boldsymbol{x}})$. For each $\hat{s}_t$ corresponding $\hat{\boldsymbol{\beta}}_t$, we calculated a relative prediction error by

$$\mathrm{P}.err=\frac{1}{T}\sum_{t=1}^{T}\left\{1-\frac{\sum_i\left(\tilde{y}_i-\tilde{\boldsymbol{x}}_i^T[s^*]\hat{\boldsymbol{\beta}}[s^*]\right)^2}{\sum_i\left(\tilde{y}_i-\tilde{\boldsymbol{x}}_i^T[\hat{s}_t]\hat{\boldsymbol{\beta}}_t[\hat{s}_t]\right)^2}\right\},$$

where $\hat{\boldsymbol{\beta}}[s^*]$ is the least squares estimate under s^*, and $\hat{\boldsymbol{\beta}}_t$ is the least squares estimate based on the selected features $\hat{s}_t$ through a screening method and its associated PLM procedure. The simulation results are summarized in Table 1, where symbol "-" denotes no method is further applied for the second stage selection.

Under setup S1, features are independent. All screening methods except for SIS have a high retaining capacity. The poor performance of SIS is likely attributable to its use of feature correlation in isolation. Clearly, the ISIS large improve SIS at some extra computational cost. Note that the high FDR for screening methods are not of concern, as they simply implies that more elaborated second stage analysis is further needed. The methods of ISIS, FR, LASSO and SMLE under this setup are comparable in most aspects, except for the computational time as shown in the last column. Considering the numerical cost, we find the combination of first stage LASSO and SMLE with the second stage SCAD are the best.

Under setup 2, the drawback of FR is clearly observed, where the features have an autocorrelation structure. If two features are highly correlated, they have similar correlation strength with the response variable. When FR is used for screening, the chance for retaining both of them is very small. Because, after retaining one of them, the other one could be weakly correlated with the residual. Other screening methods are not affected as much by this correlation. Likely as a consequence, the FR has a lower PSR and a higher FDR. Nevertheless, it should be noticed that the ultimate predication errors of FR-based methods are not worse compared with other methods. Our proposed SMLE remained satisfactorily but does not outperform in any obvious manner.

Under setup S3, the most challenging one, the strong collinearity among features badly deteriorates the performances of SIS and LASSO (screening method). The ISIS recoveries from the failure of SIS to a large degree. Nevertheless, FR and SMLE are not seriously affected. It is particularly noticeable that after SMLE and associated SCAD, the resulting selection outcomes outperform other combinations by a big margin. We find the SMLE-SCAD combination has near perfect retaining power, full positive selection, negligible false discovery rate and very low prediction error rate.

In summary, under linear model, the proposed method SMLE and the SMLE-SCAD combination is among the best performers under Setup S1 and S2. It outperforms by a big merging under Setup S3.

6.5. Logistic Regression

We now consider the situation when the response variable Y is binary and $\pi = P(Y = 1|\boldsymbol{x})$ satisfying $\text{logit}(\pi) = \boldsymbol{x}^T\beta$. That is, $(Y, \boldsymbol{x})$ satisfy a logistic regression model. The parameters are specified as follows:

S1: s^* is a simple random sample of size 8 from $\{1,..., p\}$. β_{s*} are independent samples from $U\left(4\log n/\sqrt{n}+|Z|/4\right)$ and $\boldsymbol{\beta}_{\boldsymbol{s}_c^*}=0$, where $P(U = 1) = 0.5$, $P(U = –1) = 0.5$ with $Z \sim N(0, 1)$.

S2: $s^* = \{1, 3, 5, 7, 9\}$ with effects $\beta_{s*} = (2, –1.8, 1.6, –1.4, 1.2)^T$ and $\boldsymbol{\beta}_{\boldsymbol{s}_c^*}=0$.

S3: $s^* = \{1, 2, 3, 4\}$ with effects $\beta_{s*} = (1.5, 1.5, 1.5, 1.5)^T$ and $\boldsymbol{\beta}_{\boldsymbol{s}_c^*}=0$.

Similar to setups under the linear model, these values match simulation settings used in other publications. We used $n = 400$ and $p = 1000$ for all three setups to allow appropriate parameter estimation.

Logistic model is often used to predict the outcome of a future observation with given $\boldsymbol{x}$. When $P(Y = 1|\boldsymbol{x}) \geq 0.5$ according to the fitted model, we predict the future outcome Y as 1 or 0 otherwise. The prediction error (P.err) of each selected model is evaluated by the proportion of incorrect predictions based on an independent testing data. We compute other performance measures in the same way as for linear models. The results of other methods are given in Table 2.

Similar to the linear model situation, under setup S1, we again notice that all screening methods have good performances in terms of retaining relevant features. The performances after second stage PLM, using either LASSO or SCAD, remain satisfactory. Except for ISIS, all other methods show their high computational efficiency for the screening purpose.

Under setup S2, the SIS and its associated PLM do not perform satisfactorily. Other three methods all work well for the feature screening, among which SMLE performs the best. For the second stage analysis, the merit of using SCAD is observed. We note that both ISIS-SCAD and SMLE-SCAD combination have satisfactory performances.

Under setup S3, the correlation structure is least favorable to feature selection. As observed in the linear model, the simulation results immediately put SIS and LASSO out of contest. The comparison between ISIS-SCAD and SMLE-SCAD are not sharp in terms of prediction error. Yet we notice significant improvement of SMLE-SCAD over ISIS-SCAD for a higher PSR and a lower FDR, with a lower computational cost.

6.6. Poisson Regression

We now move to Poisson regression model. We fix $n = 200$ and $p = 1000$ for all three cases with other parameters specified as follows:

S1: s^* is a simple random sample of size 8 from $\{1,..., p\}$. β_{s*} are independent samples from $U\left(\log n/\sqrt{n}+|Z|/8\right)$ and $\boldsymbol{\beta}_{\boldsymbol{s}_c^*}=0$, where $P(U = 1) = 0.8$, $P(U = –1) = 0.2$ with $Z \sim N(0, 1)$.

S2: $s^* = \{1, 3, 5, 7, 9\}$ with effects $\beta_{s*} = (2, –1.8, 1.6, –1.4, 1.2)^T$ and $\boldsymbol{\beta}_{\boldsymbol{s}_c^*}=0$.

S3: $s^* = \{1, 2, 3, 4\}$ with effects $\beta_{s*} = (0.7, 0.7, 0.7, 0.7)^T$ and $\boldsymbol{\beta}_{\boldsymbol{s}_c^*}=0$.

The notion of prediction error is less applicable under this model. We generate a test data set from the true model and obtain the log-likelihood function $\tilde{\ell}(\cdot)$ based on this data. Let $\hat{\boldsymbol{\beta}}[s^*]$ is the MLE of β based on true model s^* and $\hat{\boldsymbol{\beta}}_t$ be the MLE of β based on the selected model obtained after 2-stage analysis. We then compute to what degree the likelihood at $\hat{\boldsymbol{\beta}}_t$ matches the likelihood at $\hat{\boldsymbol{\beta}}[s^*]$. The following quantity is clearly not prediction error, but for notational consistency, we still define

$$\mathrm{P}.err = \frac{1}{T}\sum_{t=1}^{T}\left\{1 - \frac{\tilde{\ell}\left(\hat{\boldsymbol{\beta}}_t\right)}{\tilde{\ell}\left(\hat{\boldsymbol{\beta}}[s^*]\right)}\right\}.$$

The simulation results are organized in the same way and are given in Table 3. In this modeling context, we observe that the SIS even fails under setup S1, while the performance of LASSO is also unsatisfactory. In comparison, the ISIS and SMLE both achieve high retaining capability of relevant features. Yet, when the computational cost is considered, SMLE is the winner. For the second stage analysis, we observe that the new method SMLE-SCAD continues to outperform, which achieves a high PSR, a low FDR, and the highest CSR.

6.7. Other Simulation Examples

In order to fully assess the proposed method, we further conduct simulation studies under more distorted data structures, including the cases where the GLM model assumption is violated. The performance of SMLE remains satisfactory under the cases we tested. We exhibit these examples in the online supplemental file.

6.8. Real Data Example

We apply the SMLE-based screening method in a genetic example. In Singh et al. (2002), expression levels of 12600 genes were measured from prostate specimens of 52 prostate cancer patients and 50 healthy controls. One objective in this study is to build a gene expression-based classification rule to predict the identity of unknown prostate samples. Such a classification tool is helpful in early detection of prostate cancer, which provides a better opportunity for curative surgery. Identifying influential genes to the disease outcome also provides deeper understanding on prostate tumours from a genetic aspect.

By performing a permutation-based correlation test, Singh et al. (2002) suggested 456 potential genes that are likely differently expressed between tumour and normal samples. Based on the information of 6033 genes from the complete dataset, Efron (2009) further suggested 377 genes for prediction though an empirical Bayes approach, while Chen and Chen (2012) spotted 3 genes by the EBIC-based LASSO.

In this example, we reanalyze the dataset by building a logistic regression

$$logit\left\{P\left(Y=1|\boldsymbol{x}\right)\right\} = \boldsymbol{x}^T\boldsymbol{\beta},$$

where Y is the binary status of the prostate cancer (with $Y = 1$ for a tumor sample, $Y = 0$ for a normal sample) and $\boldsymbol{x}$ is the 12600 gene expression levels. Accordingly, we predict $Y = 1$ when $P(Y = 1|\boldsymbol{x})$ is estimated over 0.75 and predict $Y = 0$ otherwise. The SMLE-SCAD is used due to its decent performance in our simulation studies. Specifically, we randomly select a set of 10 subjects from each of the tumour and normal sample groups as the testing set, and treat the rest as the training set. We set the screening bound $k = 20$ for SMLE-SCAD used on the training set and assess the prediction accuracy of the selected model on the testing set. We summarize the assessment based on $T = 200$ replications in terms of sensitivity, specificity as well as the overall prediction error. For comparison, we also include the results of SIS, ISIS and LASSO followed by SCAD with $k = 20$. All tuning parameters are selected by EBIC as in the simulation examples.

From Table 4, we see that four methods performed comparably well by choosing a parsimonious model with relatively low prediction error. The models chosen by SMLE have higher sensitivity, which mistaken fewer cancer patients as healthy. Our analysis based on SMLE is consistent with Chen and Chen (2012) on the number of selected genes, but ours have a lower prediction error.

7. SUMMARY AND CONCLUSIONS

In this paper, we developed a new feature screening approach SMLE for the ultra-high dimensional regression analysis. The proposed method has a good potential to improve the existing marginal-information-based methods by taking the joint effects between features into consideration. We established the screening consistency of the new method and further developed an iterative hard-thresholding algorithm to facilitate its implementation.

Our simulation studies indicated that SMLE has an excellent capability of retaining relevant features under all cases that we consider. Meanwhile, the proposed procedure was observed to have a high computational efficiency that is comparable to SIS or LASSO. These characteristics all merit SMLE a promising approach in the analysis of ultra-high dimensional data.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors are grateful to the Editor, the AE, and two anonymous referees for the insightful comments and helpful advices, which lead to a substantially improved manuscript. This research was supported by Natural Science and Engineering Research Council of Canada (NSERC) and National Institute on Drug Abuse (NIDA) grant P50-DA10075. The content is solely the responsibility of the authors and does not necessarily represent the official views of NSERC or NIDA.

APPENDIX: PROOFS

We provide here proofs for the theorems in this paper.

Proof of Theorem 1 (convergence of IHT)

We first prove that $\ell_n\left(\boldsymbol{\beta}^{(t)}\right)$ increases after each iteration. Recall that the *h* function and the IHT algorithm are defined by

$$\begin{aligned} h_n\left(\boldsymbol{\gamma};\boldsymbol{\beta}\right) = &\quad \ell_n\left(\boldsymbol{\beta}\right) + \left(\boldsymbol{\gamma}-\boldsymbol{\beta}\right)^T S_n\left(\boldsymbol{\beta}\right) - \left(u/2\right)\left\|\boldsymbol{\gamma}-\boldsymbol{\beta}\right\|_2^2, \\ \boldsymbol{\beta}^{(t+1)} = &\quad arg\,\max_{\boldsymbol{\gamma}} h_n\left(\boldsymbol{\gamma};\boldsymbol{\beta}^{(\mathbf{t})}\right) \quad \text{subject to} \quad \left\|\boldsymbol{\gamma}\right\|_0 \le k. \end{aligned}$$

It is seen that

$$\begin{aligned} \ell\left(\boldsymbol{\beta}^{(t)}\right) =& h\left(\boldsymbol{\beta}^{(t)},\boldsymbol{\beta}^{(t)}\right) \le h\left(\boldsymbol{\beta}^{(t+1)},\boldsymbol{\beta}^{(t)}\right) \\ =& \ell_n\left(\boldsymbol{\beta}^{(t+1)}\right) - \left(u/2\right)\left\|\boldsymbol{\beta}^{(t+1)}-\boldsymbol{\beta}^{(t)}\right\|^2 \\ &+ \sum_{i=1}^{n}\left\{b\left(\boldsymbol{x}_i^T\boldsymbol{\beta}^{(t+1)}\right) - b\left(\boldsymbol{x}_i^T\boldsymbol{\beta}^{(t)}\right) - b'\left(\boldsymbol{x}_i^T\boldsymbol{\beta}^{(t)}\right)\left(\boldsymbol{x}_i^T\boldsymbol{\beta}^{(t+1)} - \boldsymbol{x}_i^T\boldsymbol{\beta}^{(t)}\right)\right\}. \end{aligned}$$

By the Taylor's expansion, there exists a $\tilde{\theta}$ between θ and θ_0 such that

$$b\left(\theta\right) - b\left(\theta_0\right) - b'\left(\theta_0\right)\left(\theta-\theta_0\right) = \left(1/2\right)b''\left(\tilde{\theta}\right)\left(\theta-\theta_0\right)^2.$$

Note that $\rho^{(t)}$ is the largest possible value of $b''\left(\tilde{\theta}\right)$ in the current context. Thus, we have

$$\begin{aligned} \ell\left(\boldsymbol{\beta}^{(t)}\right) \le& \ell\left(\boldsymbol{\beta}^{(t+1)}\right) \\ &- \left(u/2\right)\left\|\boldsymbol{\beta}^{(t+1)}-\boldsymbol{\beta}^{(t)}\right\|^2 \\ &+ \left(1/2\right)\rho^{(t)}\left\|\boldsymbol{X}\boldsymbol{\beta}^{(t+1)} - \boldsymbol{X}\boldsymbol{\beta}^{(t)}\right\|^2 \le \ell\left(\boldsymbol{\beta}^{(t+1)}\right) \\ &+ \left(1/2\right)\left(\rho^{(t)}\rho_1 - u\right)\left\|\boldsymbol{\beta}^{(t+1)}-\boldsymbol{\beta}^{(t)}\right\|^2 \le \ell\left(\boldsymbol{\beta}^{(t+1)}\right). \end{aligned}$$

The last inequality holds because *u* is chosen such that $u > \rho^{(t)}\rho_1$ in every iteration.

Apparently, the quality holds only if $\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)}$. This proves that $\ell_n\left(\boldsymbol{\beta}^{(t)}\right)$ increases after each iteration.

We next prove the convergence of $\{\boldsymbol{\beta}^{(t)}\}$. Under GLM, the summand $\left(\boldsymbol{x}_i^T\boldsymbol{\beta}\right)y_i - b\left(\boldsymbol{x}_i^T\boldsymbol{\beta}\right)$ of $\ell\left(\cdot\right)$ is strictly concave in $\boldsymbol{x}_i^T\boldsymbol{\beta}$. The property $\ell\left(\boldsymbol{\beta}^{(t)}\right) \le \ell\left(\boldsymbol{\beta}^{(t+1)}\right)$ at every iteration *t*, therefore, implies $\boldsymbol{x}_i^T\boldsymbol{\beta}^{(t)}$ is confined in a bounded (compact) region. Thus, $\rho^* = \max\rho^{(t)} < \infty$ and

$$\ell\left(\boldsymbol{\beta}^{(t+1)}\right) - \ell\left(\boldsymbol{\beta}^{(t)}\right) \ge \left(1/2\right)\left(u - \rho^*\rho_1\right)\left\|\boldsymbol{\beta}^{(t+1)}-\boldsymbol{\beta}^{(t)}\right\|^2.$$

Therefore, $\|\beta^{(t+1)} - \beta^{(t)}\| \to 0$ as $t \to \infty$ because $\ell(\cdot)$ is bounded.

Recall that $\|\beta^{(t)}\|_0 \le k$ for all t and the dimension of β is a finite p. There are only $\binom{p}{k}$ "sparsity patterns" for $\beta^{(t)}$. We can hence find a subsequence of $\{\beta^{(t)}\}$ with the same sparsity patten, denote it as s. By assumption, $\sum_i \boldsymbol{x}[s]\boldsymbol{x}^T[s] > 0$ for all s such that $\|s\| \le k$. Hence, $\ell(\cdot)$ confined on β with this sparsity pattern s is strictly convex. This further implies that this subsequence $\{\beta^{(t)}\}$ is confined in a compact region and has at least one limiting point, say, $\tilde{\boldsymbol{\beta}} = \left\{\tilde{\beta}_1, \cdots, \tilde{\beta}_p\right\}^T$. Let $\{t_m\}$ be a subsequence such that $\lim \boldsymbol{\beta}^{(t_m)} = \tilde{\boldsymbol{\beta}}$. Since $\|\beta^{(t+1)} - \beta^{(t)}\| \to 0$, we also have $\lim \boldsymbol{\beta}^{(t_m+1)} = \tilde{\boldsymbol{\beta}}$. This fact further implies that $\tilde{\boldsymbol{\beta}}$ is a local maximum of $\ell_n(\boldsymbol{\beta})$ subject to $\|\beta\|_0 \le k$ shown as follows. Note that

$$\boldsymbol{\beta}^{(t_m+1)} = arg\, \max_{\boldsymbol{\gamma}} \left\{ h\left(\boldsymbol{\gamma}; \boldsymbol{\beta}^{(\mathbf{t_m})}\right) : \|\boldsymbol{\gamma}\|_0 \le k \right\}$$

Letting $m \to \infty$, we get

$$\tilde{\boldsymbol{\beta}} = arg\, \max_{\boldsymbol{\gamma}} \left\{ h\left(\boldsymbol{\gamma}; \tilde{\boldsymbol{\beta}}\right) : \|\boldsymbol{\gamma}\|_0 \le k \right\}.$$

That is, $\tilde{\boldsymbol{\beta}}$ maximizes $h\left(\boldsymbol{\gamma}; \tilde{\boldsymbol{\beta}}\right)$ with respect to γ and sparsity pattern s. We show that it is a local maximum of $\ell(\cdot)$ under two possibilities in sparsity.

Case 1, $\|\tilde{\boldsymbol{\beta}}\|_0 < k$: Let $\tilde{\boldsymbol{\gamma}} = \tilde{\boldsymbol{\beta}} + u^{-1}\boldsymbol{X}^T\left\{\boldsymbol{y} - b'\left(\boldsymbol{X}\tilde{\boldsymbol{\beta}}\right)\right\}$, so that $\tilde{\boldsymbol{\beta}} = \boldsymbol{H}(\tilde{\boldsymbol{\gamma}}; k)$. The fact $\|\tilde{\boldsymbol{\beta}}\|_0 < k$ implies $\tilde{\boldsymbol{\gamma}}$ has fewer than k non-zero entries. By the definition of $\boldsymbol{H}$, it leads to $\tilde{\boldsymbol{\beta}} = \tilde{\boldsymbol{\gamma}}$, and hence $S\left(\tilde{\boldsymbol{\beta}}\right) = \boldsymbol{X}^T\left\{\boldsymbol{y} - b'\left(\boldsymbol{X}\tilde{\boldsymbol{\beta}}\right)\right\} = 0$. If so, $\tilde{\boldsymbol{\beta}}$ is an unconstrained maximum of $\ell(\cdot)$ satistying $\|\beta\|_0 \le k$.

Case 2, $\|\tilde{\boldsymbol{\beta}}\|_0 = k$: In this case, $\tilde{\boldsymbol{\beta}}$ being a local maximum leads to

$$\partial h\left(\boldsymbol{\gamma}, \tilde{\boldsymbol{\beta}}\right) / \partial \gamma_j \big|_{\boldsymbol{\gamma} = \tilde{\boldsymbol{\beta}}} = 0 \quad (11)$$

for every j such that $\tilde{\beta}_j \ne 0$. Note that we can write

$$h\left(\boldsymbol{\gamma}, \tilde{\boldsymbol{\beta}}\right) = \ell(\boldsymbol{\gamma}) + T\left(\boldsymbol{\gamma}, \tilde{\boldsymbol{\beta}}\right) \quad (12)$$

with

$$T\left(\boldsymbol{\gamma}, \tilde{\boldsymbol{\beta}}\right) = -(u/2)\|\boldsymbol{\gamma} - \tilde{\boldsymbol{\beta}}\|_2^2 + \sum_{i=1}^{n}\left\{b\left(\boldsymbol{x}_i^T\gamma_j\right) - b\left(\boldsymbol{x}_i^T\tilde{\beta}_j\right) - b'\left(\tilde{\beta}_j\right)\left(\boldsymbol{x}_i^T\gamma_j - \boldsymbol{x}_i^T\tilde{\beta}_j\right)\right\}.$$

It is seen thta $\partial T\left(\boldsymbol{\gamma},\tilde{\boldsymbol{\beta}}\right)/\partial\boldsymbol{\gamma}|_{\boldsymbol{\gamma}=\tilde{\boldsymbol{\beta}}}=0$. In view of (11), this fact implies $\partial l(\boldsymbol{\beta})/\partial\beta_j = 0$ is zero at $\boldsymbol{\beta}=\tilde{\boldsymbol{\beta}}$ for j such that $\tilde{\beta}_j \neq 0$. Hence, in the space of $\boldsymbol{\beta}$ of this sparsity pattern, $\tilde{\boldsymbol{\beta}}$ is the unique maximum because $\ell(\cdot)$ is strictly convex.

Based on the above analysis of two cases, we have shown that any limiting point of $\{\boldsymbol{\beta}^{(t)}\}$ is a local maximum satisfying sparsity constraint $\|\boldsymbol{\beta}\|_0 \leq k$. This also implies that $\ell(\cdot)$ has finite many local maxima. This conclusion is helpful to show that the entire series $\{\boldsymbol{\beta}^{(t)}\}$ converges.

Suppose $\boldsymbol{\beta}^{(t)}$ has two distinct limiting points, say $\tilde{\boldsymbol{\beta}}_1 \neq \tilde{\boldsymbol{\beta}}_2$. They are both local maxima of $\ell(\cdot)$. Let $\epsilon=\|\tilde{\boldsymbol{\beta}}_1 - \tilde{\boldsymbol{\beta}}_2\|$. Since $\|\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)}\| \to 0$, the distance between successive $\boldsymbol{\beta}^{(t)}$ goes to 0 as $t \to \infty$. Thus, there must be infinite many $\boldsymbol{\beta}^{(j)}$ which are at least $\varepsilon/3$ distance from these two limiting points. Yet we have seen that any subsequence of $\{\boldsymbol{\beta}^{(t)}\}$ must have a limiting point and this one $\tilde{\boldsymbol{\beta}}_3$ must be different from the previous ones. Applying this logic repeatedly implies $\ell(\cdot)$ has infinite many stationary points. This is in contradiction to the conclusion at the end of last paragraph. Hence, $\boldsymbol{\beta}^{(t)}$ can only have one limiting point and therefore converges.

Before getting into the detailed illustrations of Theorem 2, we first state one technical lemma as follows.

Lemma 1. *Let Y_i, i = 1,..., n be independent random variables from exponential family distributions with natural parameters $\theta_i \in \bar{\Theta}$. Let μ_i and σ_i^2* denote the mean and variance of Y_i respectively. Let t_{ni}, i = 1,..., *n be real numbers such that*

$$\sum_{i=1}^{n} t_{ni}^2\sigma_i^2=1, \quad \max_{1\leq i\leq n}\left\{t_{ni}^2\right\}=O\left(n^{-1}\right).$$

Then, for some positive sequence $d_n = o(n^{1/2})$ and a sufficiently large n,

$$P\left(\sum_{i=1}^{n} t_{ni}\left(Y_i-\mu_i\right)>d_n\right) \leq exp\left(-d_n^2/3\right).$$

Lemma 1 states an useful property of exponential family for the illustration of Theorem 2, the proof of which can be found in Chen and Chen (2012).

Proof of Theorem 2 (Sure screening of SMLE)

We illustrate the theorem by showing that asymptotically $\hat{s}$ falls into the collection of over-fitted models that contain s^* as a submodel. This is implied by the fact that the maximum likelihood score based on an over-fitted model is asymptotically greater than that of any under-fitted model with at least one feature in s^* excluded.

Let $\hat{\boldsymbol{\beta}}_s$ be the (unrestricted) MLE of β based on model s. The theorem is implied if $P\left\{\hat{s} \in \boldsymbol{S}_+^k\right\} \to 1$. Thus, it suffices to show that

$$P\left\{\max_{s \in \boldsymbol{S}_-^k} \ell_n\left(\hat{\boldsymbol{\beta}}_s\right) \geq \min_{s \in \boldsymbol{S}_+^k} \ell_n\left(\hat{\boldsymbol{\beta}}_s\right)\right\} \to 0, \quad (13)$$

as $n \to \infty$.

For any $s \in \boldsymbol{S}_-^k$, define $s' = s \cup s^* \in \boldsymbol{S}_+^{2k}$. Consider $\beta_{s'}$ close to $\boldsymbol{\beta}_{s'}^*$ such that $\|\boldsymbol{\beta}_{s'} - \boldsymbol{\beta}_{s'}^*\| = w_1 n^{-\tau_1}$ for some $w_1,\ \tau_1 > 0$. Clearly, when n is sufficiently large, $\beta_{s'}$ falls into a small neighborhood of $\boldsymbol{\beta}_{s'}^*$, so that condition T3 becomes applicable. Thus, by Taylor's expansion, we have

$$\begin{aligned}\ell_n\left(\boldsymbol{\beta}_{s'}\right) - \ell_n\left(\boldsymbol{\beta}_{s'}^*\right) &= \left[\boldsymbol{\beta}_{s'} - \boldsymbol{\beta}_{s'}^*\right]^T S_n\left(\boldsymbol{\beta}_{s'}^*\right) \\ &- (1/2)\left[\boldsymbol{\beta}_{s'} - \boldsymbol{\beta}_{s'}^*\right]^T H_n\left(\tilde{\boldsymbol{\beta}}_{s'}\right)\left[\boldsymbol{\beta}_{s'} - \boldsymbol{\beta}_{s'}^*\right] \leq \left[\boldsymbol{\beta}_{s'} - \boldsymbol{\beta}_{s'}^*\right]^T S_n\left(\boldsymbol{\beta}_{s'}^*\right) \\ &- (c_1/2)\, n\|\boldsymbol{\beta}_{s'} - \boldsymbol{\beta}_{s'}^*\|_2^2 \leq w_1 n^{-\tau_1}\|S_n\left(\boldsymbol{\beta}_{s'}^*\right)\|_2 \\ &- (c_1/2)\, w_1^2 n^{1-2\tau_1},\end{aligned} \quad (14)$$

where $\tilde{\boldsymbol{\beta}}_{s'}$ is an intermediate value between $\beta_{s'}$ and $\boldsymbol{\beta}_{s'}^*$. Thus, we have

$$P\left\{\ell_n\left(\boldsymbol{\beta}_{s'}\right) - \ell_n\left(\boldsymbol{\beta}_{s'}^*\right) \leq 0\right\} \leq P\left\{\|S_n\left(\boldsymbol{\beta}_{s'}^*\right)\|_2 \geq (c_1 w_1/2)\, n^{1-\tau_1}\right\} \leq \sum_{j \in s'} P\left\{S_{nj}^2\left(\boldsymbol{\beta}_{s'}^*\right) \geq k^{-1}(c_1 w_1/2)^2 n^{2-2\tau_1}\right\},$$

where

$$S_{nj}\left(\boldsymbol{\beta}_{s'}^*\right) = \sum_{i=1}^n \left[y_i - b'\left(\boldsymbol{x}_{is'}^T \boldsymbol{\beta}_{s'}^*\right)\right] x_{ij} = \sum_{i=1}^n \left[y_i - E\left(y_i|\boldsymbol{x}_i\right)\right] x_{ij}.$$

Let $t_{ni} = x_{ij}\left(\sum_{i=1}^n x_{ij}^2 \sigma_i^2\right)^{-1/2}$ such that $\sum_{i=1}^n t_{ni}^2 \sigma_i^2 = 1$. Under the model assumptions that $\theta = \boldsymbol{x}^T \boldsymbol{\beta} \in \bar{\boldsymbol{\Theta}}$ is continuous, we have $b''(\theta) < \sigma^2$ for some positive constant σ. Thus, by T4, we have $max_i\left\{t_{ni}^2\right\} \leq c_3 n^{-1}$ and $\sum_{i=1}^n x_{ij}^2 \sigma_i^2 \leq c_2^2 n \sigma^2$. Also, by T2, we have $k \leq w_2 n^{\tau_2}$. With these conditions, Lemma 1 gives the following probability inequality

$$P\left\{S_{nj}\left(\boldsymbol{\beta}_{s'}^*\right) \geq (c_1 w_1/2)\, k^{-(1/2)} n^{1-\tau_1}\right\} \leq P\left\{\sum_{i=1}^n t_{ni}\left[y_i - E\left(y_i|\boldsymbol{x}_i\right)\right] > c n^{0.5(1-2\tau_1-\tau_2)}\right\} \leq exp\left(-\left(c^2/3\right) n^{1-2\tau_1-\tau_2}\right), \quad (15)$$

where c denotes some generic positive constant. Also, by the same arguments, we have

$$P\left\{S_{nj}\left(\boldsymbol{\beta}_{s'}^*\right) \leq -(c_1 w_1/2)\, k^{-(1/2)} n^{1-\tau_1}\right\} \leq exp\left(-\left(c^2/3\right) n^{1-2\tau_1-\tau_2}\right). \quad (16)$$

The inequalities (15) and (16) imply that,

$$P\left\{\ell_n\left(\boldsymbol{\beta}_{s'}\right) \geq \ell_n\left(\boldsymbol{\beta}^*_{s'}\right)\right\} \leq 4k\, exp\left(-\left(c^2/3\right) n^{1-2\tau_1-\tau_2}\right).$$

Consequently, by Bonferroni inequality and condition $\tau_1 + \tau_2 < (1 - m)/2$, we have

$$\begin{aligned} P\left\{\max_{s\in s^k_-} \ell_n\left(\boldsymbol{\beta}_{s'}\right) \geq \ell_n\left(\boldsymbol{\beta}^*_{s'}\right)\right\} \leq & \sum_{s\in s^k_-} P\left\{\ell_n\left(\boldsymbol{\beta}_{s'}\right) \geq \ell_n\left(\boldsymbol{\beta}^*_{s'}\right)\right\} \\ \leq & 4kp^k\, exp\left(-\left(c^2/3\right) n^{1-2\tau_1-\tau_2}\right) \\ \leq & a_1\, exp\left(\tau_2\, log\, n + a_2 n^{\tau_2+m} - a_3 n^{1-2\tau_1-\tau_2}\right) = o\left(1\right) \end{aligned}$$

for some generic positive constants a_1, a_2 and a_3. Because $\ell_n\left(\boldsymbol{\beta}_{s'}\right)$ is concave in $\beta_{s'}$, above result holds for any $\beta_{s'}$ such that $\left\|\boldsymbol{\beta}_{s'} - \boldsymbol{\beta}^*_{s'}\right\|_2 \geq w_1 n^{-\tau_1}$.

For any $s \in \boldsymbol{S}^k_-$, let $\breve{\boldsymbol{\beta}}_{s'}$ be $\hat{\boldsymbol{\beta}}_s$ augmented with zeros corresponding to the elements in s'/s^*. By condition T2, it is seen that $\left\|\breve{\boldsymbol{\beta}}_{s'} - \boldsymbol{\beta}^*_{s'}\right\|_2 \geq \left\|\boldsymbol{\beta}^*_{s^*/s}\right\|_2 \geq w_1 n^{-\tau_1}$. Consequently,

$$P\left\{\max_{s\in \boldsymbol{S}^k_-} \ell_n\left(\hat{\boldsymbol{\beta}}_s\right) \geq \min_{s\in \boldsymbol{S}^k_+} \ell_n\left(\hat{\boldsymbol{\beta}}_s\right)\right\} \leq P\left\{\max_{s\in s^k_-} \ell_n\left(\breve{\boldsymbol{\beta}}_{s'}\right) \geq \ell_n\left(\boldsymbol{\beta}^*_{s'}\right)\right\} = o\left(1\right)$$

The theorem is proved.

The following lemma provides a desirable bound in the L_1 of LASSO, which is helpful for proving Theorem 3.

Lemma 2. *Let* $\beta_L = \arg\max_\beta\{l_n(\beta) - n\lambda\|\beta\|_1\}$ *for some* $\lambda > 0$. *Then, under the conditions of Theorem 3, as* $n \to \infty$,

$$P\left(\left\|\boldsymbol{\beta}_L - \boldsymbol{\beta}^*\right\|_1 \leq 16\nu^{-1}\lambda q\right) \to 1$$

with the constant ν defined in T3'.

Proof. Our proof follows the framework of Theorem 7.2 in Bickel et al. (2009). By definition of β_L,

$$l_n\left(\boldsymbol{\beta}_L\right) - l_n\left(\boldsymbol{\beta}^*\right) \geq n\lambda\left\|\boldsymbol{\beta}_L\right\|_1 - n\lambda\left\|\boldsymbol{\beta}^*\right\|_1. \quad (17)$$

Let $\delta = (\beta_L - \beta^*) = (\delta_1, ..., \delta_p)^T$. Expanding $l_n(\beta_L)$ at $\beta_L = \beta^*$, we get

$$l_n\left(\boldsymbol{\beta}_L\right) - l_n\left(\boldsymbol{\beta}^*\right) = \boldsymbol{\delta}^T S_n\left(\boldsymbol{\beta}^*\right) - (1/2)\,\boldsymbol{\delta}^T H_n\left(\tilde{\boldsymbol{\beta}}\right)\boldsymbol{\delta}$$

for some intermediate value $\tilde{\boldsymbol{\beta}}$ between β_L and β^*. Hence, we have

$$n^{-1}\boldsymbol{\delta}^T H_n\left(\tilde{\boldsymbol{\beta}}\right)\boldsymbol{\delta} \le \mathbf{2n^{-1}}|\boldsymbol{\delta}|^{\mathbf{T}}|\mathbf{S_n}\left(\boldsymbol{\beta}^*\right)|+\mathbf{2}\lambda\|\boldsymbol{\beta}^*\|_{\mathbf{1}} - \mathbf{2}\lambda\|\boldsymbol{\beta}_{\boldsymbol{L}}\|_{\mathbf{1}}. \quad (18)$$

We now search for a bound on S_n. Define

$$\mathscr{A}=\left\{\max_{1\le j\le p}|S^*_{nj}| \le n\lambda/2\right\},$$

where for each $j \in \{1,..., p\}$

$$S^*_{nj}=S_{nj}\left(\boldsymbol{\beta}^*\right)=\sum_{i=1}^{n}\left\{y_i - b'\left(\boldsymbol{x}_i^T\boldsymbol{\beta}^*\right)\right\}x_{ij}=\sum_{i=1}^{n}\left\{y_i - E\left(y_i|\boldsymbol{x}_i\right)\right\}x_{ij}$$

is the jth component of $S_n(\beta^*)$. Note that,

$$P\left(|S^*_{nj}|>n\lambda/2\right)=P\left(S^*_{nj}>n\lambda/2\right)+P\left(-S^*_{nj}>n\lambda/2\right) \quad (19)$$

Following the arguments in (15), we can similarly show that both probabilities on the left hand side of (19) are bounded by $\exp(-cn\lambda^2)$ for a generic constant $c > 0$. Thus, under conditions T1 and $\lambda n^{(1-m)/2} \to \infty$, we have

$$P\left(\mathscr{A}^c\right) \le \sum_{j=1}^{p}P\left(|S^*_{nj}|>n\lambda/2\right) \le 2p\,exp\left(-c\lambda^2 n\right) \le 2\,exp\left(an^m - cn\lambda^2\right) \to 0 \quad (20)$$

as $n \to \infty$, where a is another generic positive constant. This implies that $P\left(\mathscr{A}\right) \to 1$ and $\|S_n(\beta^*)\|_\infty = O_p(n\lambda)$. Clearly, under event $\mathscr{A}$, we have the left hand side of (18) bounded by

$$n^{-1}\boldsymbol{\delta}^T H_n\left(\tilde{\boldsymbol{\beta}}\right)\boldsymbol{\delta} \le \lambda\|\boldsymbol{\delta}\|_{\mathbf{1}}+\mathbf{2}\lambda\|\boldsymbol{\beta}^*\|_{\mathbf{1}} - \mathbf{2}\lambda\|\boldsymbol{\beta}_{\boldsymbol{L}}\|_{\mathbf{1}},$$

which further implies that

$$\begin{aligned} & n^{-1}\boldsymbol{\delta}^T H_n\left(\tilde{\boldsymbol{\beta}}\right)\boldsymbol{\delta} \\ & \quad +\lambda\|\boldsymbol{\delta}\|_1 \le 2\lambda\left(\sum_{j=1}^{p}|\hat{\beta}_{l_{1j}} - \beta^*_j|+|\beta^*_j| - |\hat{\beta}_{l_{1j}}|\right) \\ & \quad =2\lambda\left(\sum_{j\in s^*}|\hat{\beta}_{l_{1j}} - \beta^*_j|+|\beta^*_j| - |\hat{\beta}_{l_{1j}}|\right) \le 4\lambda\sum_{j\in s^*}|\delta_j| \\ & \quad =4\lambda\|\boldsymbol{\delta}_{s^*}\|_1. \end{aligned} \quad (21)$$

Since $\boldsymbol{\delta}^T H_n\left(\tilde{\boldsymbol{\beta}}\right)\boldsymbol{\delta} \ge \mathbf{0}$, inequality (21) implies that $\|\delta\|_1 \le 4\|\delta_{s*}\|_1$. Subsequently, $\|\boldsymbol{\delta}_{\boldsymbol{s}^*_c}\|_1 \le 3\|\boldsymbol{\delta}_{\boldsymbol{s}^*}\|_1$ which leads to condition T3': for sufficiently large n,

$$\boldsymbol{\delta}^T H_n\left(\tilde{\boldsymbol{\beta}}\right)\boldsymbol{\delta} \geq \mathbf{n}\nu\|\boldsymbol{\delta}_{\mathbf{s}^*}\|_{\mathbf{2}}^{\mathbf{2}}. \quad (22)$$

By Cauchy inequality, (22) and then (21), we get

$$\|\boldsymbol{\delta}_{s^*}\|_1^2 \leq q\|\boldsymbol{\delta}_{s^*}\|_2^2 \leq q\nu^{-1}\left[n^{-1}\boldsymbol{\delta}^T H_n\left(\tilde{\boldsymbol{\beta}}\right)\boldsymbol{\delta}\right] \leq 4\nu^{-1}\lambda q\|\boldsymbol{\delta}_{s^*}\|_1.$$

This leads to $\|\delta_{s*}\|_1 \leq 4\nu^{-1}\lambda q$. Subsequently, by $\|\boldsymbol{\delta}_{s_c^*}\|_1 \leq 3\|\boldsymbol{\delta}_{s^*}\|_1$, we get

$$\|\boldsymbol{\delta}\|_1 = \|\boldsymbol{\delta}_{s_c^*}\|_1 + \|\boldsymbol{\delta}_{s^*}\|_1 \leq 4\|\boldsymbol{\delta}_{s^*}\|_1 \leq 16\nu^{-1}\lambda q$$

under event $\mathscr{A}$. Since we have shown that $P(\mathscr{A}) \to 1$, the lemma is therefore proved.

Proof of Theorem 3 (Sure screening of IHT)

Let $\|\cdot\|_\infty$ denote the L_∞ norm of a vector (i.e. the maximum absolute component) and $d = min_{j\in s^*}|\beta_j^*|$. The theorem is established if, for any $t \geq 0$, $P\left\{\|\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^*\|_\infty < d/2\right\} \to 1$. Clearly, this is implied by

$$\|\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^*\|_\infty = o_p(d). \quad (23)$$

We use mathematical induction to prove (23) in two steps.

We first prove that (23) holds for $t = 0$. Recall that $\beta^{(0)}$ in the procedure responds to the LASSO estimate of. By Lemma 2,

$$P\left\{\|\boldsymbol{\beta}^{(0)} - \boldsymbol{\beta}^*\|_1 \leq 16\nu^{-1}\lambda q\right\} \to 1. \quad (24)$$

Also, by condition T2, $q = O(n^{\tau_2})$ and $d^{-1} = O(n^{\tau_1})$. These together with $\lambda = o(n^{-(\tau_1+\tau_2)})$ lead to $\lambda q = o(d)$. Hence, (24) implies (23) for $t = 0$.

We now carry out the second step of the mathematical induction. Suppose (23) is true for $t-1$, we show that it is also true for t with $t \geq 1$. Recall that $\beta^{(t)} = \boldsymbol{H}(\gamma^{(t)}; k)$ with $\gamma^{(t)} = \beta^{(t-1)} + u^{-1}\boldsymbol{X}^T\{\boldsymbol{y} - b'(\boldsymbol{X}\beta^{(t-1)})\}$ and $\boldsymbol{H}$ being the hard thresholding function. If $\|\gamma^{(t)} - \beta^*\|_\infty = o_p(d)$, we would have $\|\boldsymbol{\gamma}_{\boldsymbol{s}_c^*}^{(t)}\|_\infty = o_p(d)$ and $\|\boldsymbol{\gamma}_{\boldsymbol{s}^*}^{(t)}\|_\infty = O_p(d)$, where the latter one is in a strict sense. Hence, components in $\boldsymbol{\gamma}_{\boldsymbol{s}^*}^{(t)}$ are among the ones with the top k largest absolute values in probability. Consequently, we would have

$$\|\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^*\|_\infty = \|\boldsymbol{H}\left(\boldsymbol{\gamma}^{(t)}; k\right) - \boldsymbol{\beta}^*\|_\infty \leq \|\boldsymbol{\gamma}^{(t)} - \boldsymbol{\beta}^*\|_\infty = o_p(d).$$

Thus, the second step is completed by showing $\|\gamma^{(t)} - \beta^*\|_\infty = o_p(d)$.

Note that,

$$\|\boldsymbol{\gamma}^{(t)} - \boldsymbol{\beta}^*\|_\infty \leq \|\boldsymbol{\beta}^{(t-1)} - \boldsymbol{\beta}^*\|_\infty + u^{-1}\|\boldsymbol{X}^T\left\{\boldsymbol{y} - b'\left(\boldsymbol{X}\boldsymbol{\beta}^{(t-1)}\right)\right\}\|_\infty. \quad (25)$$

Under the induction assumption, $\|\boldsymbol{\beta}^{(t-1)} - \boldsymbol{\beta}^*\|_\infty = o_p(d)$. By Taylor's expansion, we have

$$\begin{aligned}\|\boldsymbol{X}^T\left\{\boldsymbol{y} - b'\left(\boldsymbol{X}\boldsymbol{\beta}^{(t-1)}\right)\right\}\|_\infty &\leq \|\boldsymbol{X}^T\left\{\left(\boldsymbol{y} - b'\left(\boldsymbol{X}\boldsymbol{\beta}^*\right)\right)\right\}\|_\infty \\ &+\|\boldsymbol{X}^T b''\left(\boldsymbol{X}\tilde{\boldsymbol{\beta}}\right)\boldsymbol{X}\left(\boldsymbol{\beta}^{(t-1)} - \boldsymbol{\beta}^*\right)\|_\infty \\ &=\|S_n\left(\boldsymbol{\beta}^*\right)\|_\infty + \|\psi\|_\infty,\end{aligned} \quad (26)$$

where $\psi = \boldsymbol{X}^{\mathbf{T}}\mathbf{b}''\left(\boldsymbol{X}\tilde{\boldsymbol{\beta}}\right)\boldsymbol{X}\left(\boldsymbol{\beta}^{(\mathbf{t}-\mathbf{1})} - \boldsymbol{\beta}^*\right)$ and $\tilde{\boldsymbol{\beta}}$ is some intermediate value between $\boldsymbol{\beta}^*$ and $\boldsymbol{\beta}^{(t-1)}$. By (20), $\|S_n(\boldsymbol{\beta}^*)\|_\infty = O_p(n\lambda)$. Under the conditions of this theorem, $\lambda n^{\tau_1+\tau_2} \to 0$ and $u > \xi nk$, we find $u^{-1}\|S_n(\boldsymbol{\beta}^*)\|_\infty = o_p(d)$. Meanwhile, note that $u \geq \xi kn$, $b''(\cdot) \leq \sigma^2$ for some constant $\sigma > 0$ under the model assumption. Under condition T4, we get

$$u^{-1}\|\psi\|_\infty \leq (\xi nk)^{-1}\sigma^2\|\boldsymbol{\beta}^{(t-1)} - \boldsymbol{\beta}^*\|_\infty \max_{1\leq j\leq p}\left\{\sum_{i=1}^{n}\sum_{h\in s_t}|x_{ih}x_{ij}|\right\} \leq 2\xi^{-1}c_2^2\sigma^2\|\boldsymbol{\beta}^{(t-1)} - \boldsymbol{\beta}^*\|_\infty = o_p(d),$$

where $s_t = s^{(t-1)} \cup s^*$ with $\pi(s_t) \leq 2k$. Thus, the second terms on the right hand side of (25) is $o_p(d)$ and the second step of the mathematical induction is completed. The theorem is therefore proved.

REFERENCES

Akaike H. Information theory and an extension of the maximum likelihood principle. In 2nd Internatinal Symposium on Information Theory. 1973; 1:267–281.

Bickel P, Ritov Y, Tsybakov A. Simultaneous analysis of lasso and dantzig selector. The Annals of Statistics. 2009; 37:1705–1732.

Blumensath T, Davies M. Iterative thresholding for sparse approximations. The Journal of Fourier Analysis and Applications. 2008; 14:629–654.

Blumensath T, Davies M. Iterative hard thresholding for compressed sensing. Applied and Computational Harmonic Analysis. 2009; 27:265–274.

Candes E, Tao T. The dantzig selector: statistial estimation when p is much larger than n. The Annals of Statistics. 2007; 35:2313–2351.

Candès EJ, Romberg J, Tao T. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. IEEE Transactions on information theory. 2006; 52:489–509.

Chen J, Chen Z. Extended bayisan information criterion for model selection with large model spaces. Biometrika. 2008; 95:759–771.

Chen J, Chen Z. Extended bic for small-n-large-p sparse glm. Statistica Sinica. 2012; 22:555–574.

Donoho D. High-dimentional data analysis: the curses and blessings of dimentionality. Aide-Memoire of a Lecture at AMS Conference on Math Challenges of the 21st Century. 2000

Donoho D. Compressed sensing. IEEE Transactions on information theory. 2006; 52:1289–1306.

Efron B. Correlation and large-scale simulations significance testing. Journal of the American Statistical Association. 2007; 102:93–103.

Efron B. Empirical bayes estimates for large-scale predition problem. Journal of the American Statistical Association. 2009; 104:1015–1028. [PubMed: 20333278]

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

Fan J, Lv J. Sure independence screening for ultrahigh dimentional feature space (with discussion). Journal of the Royal Statistical Society. Series B. 2008; 70:849–911. [PubMed: 19603084]

Fan J, Lv J. A selective overview of variable selection in high dimentional feature space. Statistica Sinica. 2010; 20:101–148. [PubMed: 21572976]

Fan J, Samworth R, Wu Y. Ultrahigh dimentional variable selection: beyond the linear model. Journal of Machine Learning Research. 2009; 10:1829–1853.

Fan J, Song R. Sure independence screening in generalized linear models with np-dimensionality. The Annals of Statistics. 2009; 38:3567–3604.

Hastie, T.; Tibshirani, R.; Fridman, J. The elements of statistical learning: data mining. 2 edition. Springer-Verlag; New York: 2009.

Li R, Zhong W, Zhu L. Feature screening via distance correlation learning. Journal of the American Statistical Association. 2012; 107:1129–1139. [PubMed: 25249709]

Mallat S, Zhang Z. Matching pursuits with time-frequency dictionaries. IEEE Transactions on Signal Processing. 1993; 41:3397–3415.

McCullagh, P.; Nelder, JA. Generalized Linear Models. 2 edition. Chapman and Hall; London: 1989.

Schwarz G. Estimating the dimension of a model. The Annals of Statistics. 1978; 6:461–464.

Singh D, Febbo PG, Ross K, Jackson DG, Manola J, Ladd C, Tamayo P, Renshaw AA, D'Amico AV, Richie JP, Lander ES, Loda M, Kanto PW, Golub TR, Sellers WR. Gene expression correlates of clinical prostate cancer behavior. Cancer Cell. 2002; 1:203–209. [PubMed: 12086878]

Stone M. Cross-validatory choice and assessment of statistical predictions (with discussion). Journal of the Royal Statistical Society. Series B. 1974; 39:111–147.

Storey J, Tibshirani R. Statistical significance for genome-wide studies. Proceedings of the National Academy of Sciences. 2003; 100:9440–9445.

Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B. 1996; 58:267–288.

Wang H. Forward regression for ultra-high dimentional varibale screening. Journal of the American Statistical Association. 2009; 104:1512–1524.

Zhang C. Nearly unbiased variable selection under minimax concave penalty. The Annals of Statistics. 2010; 38:894–942.

Zhu L-P, Li L, Li R, Zhu L-X. Model-free feature screening for ultrahigh-dimensional data. Journal of the American Statistical Association. 2011; 106:1464–1475. [PubMed: 22754050]

Zou H, Hastie T. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B. 2005; 67:301–320.

Table 1

Simulation results under linear regression

Setup	1st Stage	2nd Stage	RC	PSR	FDR	CSR	AMS	P.err	TIME
S1	SIS	–	.19	.83	.79	.00	31.0	.50	10.22
		LASSO	.18	.83	.11	.13	7.5	.62	10.45
		SCAD	.19	.83	.11	.17	7.6	.42	10.60
	ISIS	–	.95	.99	.74	.00	31.0	.41	57.48
		LASSO	.87	.98	.10	.41	8.9	.58	57.74
		SCAD	.94	.99	.07	.54	8.6	.12	58.69
	FR	–	1.00	1.00	.74	.00	31.0	.52	161.43
		LASSO	.99	.99	.07	.59	8.7	.44	161.66
		SCAD	.99	.99	.09	.40	8.8	.12	161.82
	LASSO	–	.98	.99	.72	.00	29.0	.35	.43
		LASSO	.85	.98	.11	.40	9.0	.59	.67
		SCAD	.96	.99	.07	.47	8.7	.20	1.80
	SMLE	–	.99	1.00	.74	.00	31.0	.46	2.68
		LASSO	.93	.99	.08	.52	8.8	.56	2.94
		SCAD	.98	.99	.07	.49	8.7	.12	3.16
S2	SIS	–	.58	.91	.81	.00	24.0	.34	4.67
		LASSO	.32	.83	.09	.23	4.6	.42	4.91
		SCAD	.26	.78	.15	.17	4.7	.26	5.12
	ISIS	–	.63	.92	.81	.00	24.0	.41	24.18
		LASSO	.38	.85	.10	.25	4.8	.41	24.41
		SCAD	.26	.78	.20	.14	5.0	.30	24.76
	FR	–	.34	.78	.84	.00	24.0	.59	46.66
		LASSO	.29	.76	.23	.17	5.1	.33	46.85
		SCAD	.19	.73	.28	.08	5.2	.25	47.25
	LASSO	–	.89	.97	.78	.00	22.4	.34	.14
		LASSO	.44	.86	.09	.26	4.8	.41	.36
		SCAD	.29	.80	.17	.17	5.0	.31	.94
	SMLE	–	.77	.95	.80	.00	24.0	.44	1.25
		LASSO	.47	.87	.09	.28	4.9	.39	1.47
		SCAD	.25	.79	.20	.12	5.0	.28	2.04
S3	SIS	–	.01	.32	.94	.00	21.0	.89	.70
		LASSO	.01	.23	.91	.00	9.8	.95	.89
		SCAD	.01	.24	.89	.00	9.7	.94	.93
	ISIS	–	.70	.83	.84	.00	21.0	.58	5.15
		LASSO	.51	.69	.66	.01	8.9	.87	5.36
		SCAD	.70	.77	.32	.49	6.1	.29	5.67
	FR	–	.88	.89	.84	.00	21.0	.59	6.14
		LASSO	.79	.86	.53	.02	7.9	.83	6.34

Setup	1st Stage	2nd Stage	RC	PSR	FDR	CSR	AMS	P.err	TIME
		SCAD	.88	.89	.19	.57	5.1	.11	6.87
	LASSO	–	.25	.62	.87	.00	19.7	.71	.02
		LASSO	.03	.32	.86	.00	9.8	.95	.21
		SCAD	.25	.49	.66	.24	8.4	.71	.25
	SMLE	–	.99	1.00	.82	.00	21.0	.49	.30
		LASSO	.67	.89	.55	.02	8.5	.91	.52
		SCAD	.99	.99	.07	.71	4.4	.01	.85

Table 2

Simulation results under logistic regression

Setup	1st Stage	2nd Stage	RC	PSR	FDR	CSR	AMS	P.err	TIME
S1	SIS	–	.92	.99	.47	.00	15.0	.16	3.33
		LASSO	.92	.99	.01	.83	8.0	.16	3.74
		SCAD	.92	.99	.01	.82	8.0	.14	4.41
	ISIS	–	1.00	1.00	.47	.00	15.0	.17	15.18
		LASSO	.99	1.00	.03	.81	8.2	.16	15.52
		SCAD	.99	1.00	.05	.63	8.5	.14	16.41
	LASSO	–	.99	1.00	.42	.00	14.0	.16	.09
		LASSO	.99	.99	.02	.84	8.2	.16	.58
		SCAD	.99	1.00	.04	.69	8.4	.14	1.34
	SMLE	–	.99	1.00	.47	.00	15.0	.17	3.61
		LASSO	.99	.99	.02	.82	8.2	.16	3.95
		SCAD	.99	1.00	.04	.70	8.4	.14	4.87
S2	SIS	–	.09	.73	.76	.00	15.0	.29	3.18
		LASSO	.07	.60	.24	.00	4.1	.29	3.45
		SCAD	.08	.60	.27	.02	4.2	.27	3.73
	ISIS	–	.86	.96	.68	.00	15.0	.24	23.93
		LASSO	.75	.93	.22	.21	6.2	.26	24.22
		SCAD	.82	.95	.13	.47	5.6	.20	24.76
	LASSO	–	.82	.96	.64	.00	13.4	.23	.08
		LASSO	.58	.88	.29	.05	6.4	.26	.34
		SCAD	.78	.95	.19	.18	6.0	.21	.82
	SMLE	–	.97	.99	.67	.00	15.0	.23	3.17
		LASSO	.87	.96	.18	.30	6.2	.25	3.44
		SCAD	.90	.98	.11	.51	5.6	.20	3.93
S3	SIS	–	.01	.43	.89	.00	15.0	.21	3.17
		LASSO	.01	.32	.78	.00	6.1	.26	3.43
		SCAD	.01	.34	.78	.00	6.3	.25	3.58
	ISIS	–	.61	.82	.78	.00	15.0	.20	21.45
		LASSO	.37	.65	.65	.00	7.4	.22	21.77
		SCAD	.58	.76	.54	.06	7.2	.19	22.22
	LASSO	–	.14	.56	.84	.00	13.1	.20	.04
		LASSO	.03	.36	.77	.00	6.4	.25	.29
		SCAD	.08	.42	.74	.01	6.7	.24	.42
	SMLE	–	.77	.92	.78	.00	15.0	.19	6.38
		LASSO	.58	.85	.50	.01	7.1	.21	6.73
		SCAD	.76	.91	.39	.13	6.4	.17	7.31

Table 3

Simulation results under Poisson regression model

Setup	1st Stage	2nd Stage	RC	PSR	FDR	CSR	AMS	P.err	TIME
S1	SIS	–	.08	.76	.70	.00	21.0	.34	3.78
		LASSO	.06	.74	.31	.00	8.9	.46	4.54
		SCAD	.08	.76	.25	.02	8.3	.31	8.38
	ISIS	–	.95	.99	.62	.00	21.0	.16	19.69
		LASSO	.86	.97	.21	.12	10.1	.29	20.36
		SCAD	.95	.99	.08	.54	8.7	.04	24.96
	LASSO	–	.76	.96	.61	.00	20.0	.13	.03
		LASSO	.53	.91	.32	.03	11.1	.39	.69
		SCAD	.76	.96	.14	.31	9.1	.09	3.45
	SMLE	–	.94	.99	.62	.00	21.0	.21	.48
		LASSO	.92	.99	.11	.35	9.0	.25	1.12
		SCAD	.94	.99	.05	.66	8.4	.04	5.96
S2	SIS	–	.01	.53	.86	.00	21.0	.46	3.52
		LASSO	.01	.47	.64	.00	7.2	.46	3.84
		SCAD	.01	.46	.65	.00	7.2	.42	4.99
	ISIS	–	.88	.96	.77	.00	21.0	.18	24.03
		LASSO	.64	.89	.43	.03	8.1	.34	24.38
		SCAD	.87	.95	.22	.27	6.5	.05	26.41
	LASSO	–	.48	.86	.77	.00	19.4	.20	.03
		LASSO	.10	.66	.57	.00	8.1	.43	.36
		SCAD	.47	.83	.38	.07	7.1	.15	1.84
	SMLE	–	.93	.98	.76	.00	21.0	.23	.52
		LASSO	.85	.96	.30	.12	7.3	.29	.90
		SCAD	.92	.98	.14	.41	5.8	.03	3.05
S3	SIS	–	.00	.21	.96	.00	21.0	.61	3.54
		LASSO	.00	.14	.93	.00	9.0	.70	4.15
		SCAD	.00	.17	.92	.00	8.3	.62	6.03
	ISIS	–	.59	.74	.86	.00	21.0	.34	24.04
		LASSO	.41	.61	.69	.02	8.8	.54	24.68
		SCAD	.58	.72	.53	.09	7.4	.26	28.01
	LASSO	–	.01	.30	.93	.00	19.0	.57	.03
		LASSO	.00	.19	.92	.00	9.3	.70	.62
		SCAD	.00	.27	.86	.00	8.3	.60	2.52
	SMLE	–	.93	.98	.81	.00	21.0	.27	.73
		LASSO	.59	.85	.58	.01	8.6	.52	1.29
		SCAD	.90	.96	.34	.13	6.4	.08	5.23

Table 4

Results for analyzing the prostate data.

Screening	AMS	Sensitivity	Specificity	P.err
SIS	2.2	.71	.96	.17
ISIS	2.3	.68	.96	.18
LASSO	2.7	.71	.94	.17
SMLE	2.7	.77	.94	.14