ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistical Methodology
Series B

J. R. Statist. Soc. B (2016)

Variable selection for support vector machines in moderately high dimensions

Xiang Zhang and Yichao Wu,

North Carolina State University, Raleigh, USA

Lan Wang

University of Minnesota, Minneapolis, USA

and Runze Li

Pennsylvania State University, University Park, USA

[Received April 2013. Final revision September 2014]

Summary. The support vector machine (SVM) is a powerful binary classification tool with high accuracy and great flexibility. It has achieved great success, but its performance can be seriously impaired if many redundant covariates are included. Some efforts have been devoted to studying variable selection for SVMs, but asymptotic properties, such as variable selection consistency, are largely unknown when the number of predictors diverges to ∞. We establish a unified theory for a general class of non-convex penalized SVMs. We first prove that, in ultrahigh dimensions, there is one local minimizer to the objective function of non-convex penalized SVMs having the desired oracle property. We further address the problem of non-unique local minimizers by showing that the local linear approximation algorithm is guaranteed to converge to the oracle estimator even in the ultrahigh dimensional setting if an appropriate initial estimator is available. This condition on the initial estimator is verified to be automatically valid as long as the dimensions are moderately high. Numerical examples provide supportive evidence.

Keywords: Local linear approximation; Non-convex penalty; Oracle property; Support vector machines; Ultrahigh dimensions; Variable selection

1. Introduction

Owing to the recent advent of new technologies for data acquisition and storage, we have seen an explosive growth of data complexity in a variety of research areas such as genomics, imaging and finance. As a result, the number of predictors becomes huge. However, there are only a moderate number of instances that are available for study (Donoho, 2000). For example, in tumour classification using genomic data, expression values of tens of thousands of genes are available, but the number of arrays is typically of the order of tens. Classification of high dimensional data poses many statistical challenges and calls for new methods and theories. In this paper we consider high dimensional classification where the number of covariates diverges with the sample size and can be potentially much larger than the sample size.

The support vector machine (SVM) (Vapnik, 1996) is a powerful binary classification tool with high accuracy and great flexibility. It has achieved success in many applications. However,

Address for correspondence: Yichao Wu, Department of Statistics, North Carolina State University, Raleigh, NC 27695-8203, USA.
E-mail: wu@stat.ncsu.edu

 1369–7412/16/78000

one serious drawback of the standard SVM is that its performance can be adversely affected if many redundant variables are included in building the decision rule (Friedman *et al.*, 2001); see the evidence in the numerical results of Section 5.1. Classification using all features has been shown to be as poor as random guessing because of the accumulation of noise in high dimensional space (Fan and Fan, 2008). Many methods have been proposed to remedy this problem, such as recursive feature elimination suggested by Guyon *et al.* (2002). In particular, superior performance can be achieved with a unified method, namely achieving variable selection and prediction simultaneously (Fan and Li, 2001) by using an appropriate sparsity penalty. It is well known that the standard SVM can fit in the regularization framework of *loss* plus *penalty* by using hinge loss and the L_2-penalty. Based on this, several attempts have been made to achieve variable selection for the SVM by replacing the L_2-penalty with other forms of penalty. Bradley and Mangasarian (1998), Zhu *et al.* (2004) and Wegkamp and Yuan (2011) considered the L_1-penalized SVM; Zou and Yuan (2008) proposed to use the F_∞-norm SVM to select groups of predictors; Wang *et al.* (2006, 2008) suggested the elastic net penalty for the SVM; Zou (2007) proposed to penalize the SVM with the adaptive lasso penalty; Zhang *et al.* (2006), Becker *et al.* (2011) and Park *et al.* (2012) studied smoothly clipped absolute deviation (SCAD) (Fan and Li, 2001) penalized SVM. Recently Park *et al.* (2012) studied the oracle property of the SCAD-penalized SVM with a fixed number of predictors. Yet, to the best of our knowledge, the theory of variable selection consistency of sparse SVMs in high dimensions or ultrahigh dimensions (Fan and Lv, 2008) has not been studied so far.

In this paper, we study the variable selection consistency of sparse SVMs. Instead of using the L_2-penalty, we consider the penalized SVM with a general class of non-convex penalties, such as the SCAD penalty or the minimax concave penalty (MCP) (Zhang, 2010). Though the convex L_1-penalty can also induce sparsity, it is well known that its variable selection consistency in linear regression relies on the stringent 'irrepresentability condition' on the design matrix. This condition, however, can easily be violated in practice; see the examples in Zou (2006) and Meinshausen and Yu (2009). Moreover, the regularization parameter for model selection consistency in this case is not optimal for prediction accuracy (Meinshausen and Bühlmann, 2006; Zhao and Yu, 2007). For the non-convex penalty, Kim *et al.* (2008) investigated the oracle property of SCAD-penalized least squares regression in the high dimensions. However, a different set of proving techniques is needed for the non-convex penalized SVMs because the hinge loss in the SVM is not a smooth function. The Karush–Kuhn–Tucker local optimality condition is generally not sufficient for the set-up of a non-smooth loss plus a non-convex penalty. A new sufficient optimality condition based on subgradient calculation is used in the technical proof in this paper. We prove that under some general conditions, with probability tending to 1, the oracle estimator is a local minimizer of the non-convex penalized SVM objective function where the number of variables may grow exponentially with the sample size. By oracle estimator, we mean an estimator obtained by minimizing the empirical hinge loss with only relevant covariates. As one referee pointed out, with a finite sample, the empirical hinge loss may have multiple minimizers because the objective function is piecewise linear. This issue will vanish asymptotically because we assume that the population hinge loss has a unique minimizer. Such an assumption on the population hinge loss has been made in the existing literature (Koo *et al.*, 2008).

Even though the non-convex penalized SVMs are shown to enjoy the aforementioned local oracle property, it is largely unknown whether numerical algorithms can identify this local minimizer since the objective function involved is non-convex and typically multiple local minimizers exist. Existing methods rely heavily on conditions that guarantee that the local minimizer is unique. In general, when the convexity of the hinge loss function dominates the concavity of

the penalty, the non-convex penalized SVMs actually have a unique minimizer due to global convexity. Recently Kim and Kwon (2012) gave sufficient conditions for a unique minimizer of the non-convex penalized least square regression when global convexity is not satisfied. However, for ultrahigh dimensional cases, it would be unrealistic to assume the existence of a unique local minimizer. See Wang *et al*. (2013) for relevant discussion and a possible solution to non-convex penalized regression.

In this paper, we further address the non-uniqueness issue of local minimizers by verifying that, with probability tending to 1, the local linear approximation (LLA) algorithm (Zou and Li, 2008) is guaranteed to yield an estimator with the desired oracle property in merely two iterations under the localizability condition (Fan *et al*., 2014). This convergence result extends the work of Fan *et al*. (2014) by relaxing the differentiability assumption of the loss function and holds in the ultrahigh dimensional setting with $p = o\{\exp(n^{\delta})\}$ for some positive constant δ. We further show that the localizability condition is automatically valid for the moderately high dimensional setting with $p = o(\sqrt{n})$. To the best of our knowledge, this is the first result on the convergence of the LLA algorithm in the set-up of a non-smooth loss function with a non-convex penalty.

The rest of this paper is organized as follows. Section 2 introduces the methodology of non-convex penalized SVMs. Section 3 contains the main results of the properties of non-convex penalized SVMs. The implementation procedure is summarized in Section 4. Simulation studies and a real data example are provided in Section 5, followed by a discussion in Section 6. Technical proofs are presented in Appendix A. A file containing R demonstration code for one simulation example and the real data example is available from `http://www4.stat.ncsu.edu/~wu/soft/VarSelforSVMbyZhangWuWangLi.zip`.

2. Non-convex penalized support vector machines

We begin with the basic set-up and notation. In binary classification, we are typically given a random sample $\{(Y_i, \mathbf{X}_i)\}_{i=1}^n$ from an unknown population distribution $P(\mathbf{X}, Y)$. Here $Y_i \in \{1, -1\}$ denotes the categorical label and $\mathbf{X}_i = (X_{i0}, X_{i1}, \ldots, X_{ip})^{\mathrm{T}} = (X_{i0}, (\mathbf{X}_i^*)^{\mathrm{T}})^{\mathrm{T}}$ denotes the input covariates with $X_{i0} = 1$ corresponding to the intercept term. The goal is to estimate a classification rule that can be used to predict output labels for future observations with input covariates only. With potentially varying misclassification cost specified by weight $W_i = w$ if $Y_i = 1$ and $W_i = 1 - w$ if $Y_i = -1$ for some $0 < w < 1$, the linear weighted SVM (Lin *et al*., 2002) estimates the classification boundary by solving

$$\min_{\boldsymbol{\beta}} n^{-1} \sum_{i=1}^{n} W_i (1 - Y_i \mathbf{X}_i^{\mathrm{T}} \boldsymbol{\beta})_+ + \lambda \boldsymbol{\beta}^{*\mathrm{T}} \boldsymbol{\beta}^*,$$

where $(1-u)_+ = \max\{1-u, 0\}$ denotes the hinge loss, $\lambda > 0$ is a regularization parameter and $\boldsymbol{\beta} = (\beta_0, (\boldsymbol{\beta}^*)^{\mathrm{T}})^{\mathrm{T}}$ with $\boldsymbol{\beta}^* = (\beta_1, \beta_2, \ldots, \beta_p)^{\mathrm{T}}$. The standard SVM is a special case of the weighted SVM with weight parameter $w = 0.5$. In this paper, we consider the weighted SVM for more generality. In general, the corresponding decision rule, $\mathrm{sgn}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta})$, uses all covariates and is not capable of selecting relevant covariates.

Towards variable selection for the linear weighted SVM, we consider the population linear weighted hinge loss $\mathbb{E}\{W(1 - Y\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta})_+\}$. Let $\boldsymbol{\beta}_0 = (\beta_{00}, \beta_{01}, \ldots, \beta_{0p})^{\mathrm{T}} = (\beta_{00}, (\boldsymbol{\beta}_0^*)^{\mathrm{T}})^{\mathrm{T}}$ denote the true parameter value, which is defined as the minimizer of the population weighted hinge loss, namely

$$\boldsymbol{\beta}_0 = \arg\min_{\boldsymbol{\beta}} \mathbb{E}\{W(1 - Y\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta})_+\}. \tag{1}$$

The number of covariates $p = p_n$ is allowed to increase with the sample size n. It is even possible that p_n is much larger than n. In this paper we assume that the true parameter β_0 is sparse. Let $A = \{1 \leqslant j \leqslant p_n; \beta_{0j} \neq 0\}$ be the index set of the non-zero coefficients. Let $q = q_n = |A|$ be the cardinality of set A, which is also allowed to increase with n. Without loss of generality, we assume that the last $p_n - q_n$ components of β_0 are 0, i.e. $\beta_0^{\mathrm{T}} = (\beta_{01}^{\mathrm{T}}, \mathbf{0}^{\mathrm{T}})$. Correspondingly, we write $\mathbf{X}_i^{\mathrm{T}} = (\mathbf{Z}_i^{\mathrm{T}}, \mathbf{R}_i^{\mathrm{T}})$, where $\mathbf{Z}_i = (X_{i0}, X_{i1}, \ldots, X_{iq})^{\mathrm{T}} = (1, (\mathbf{Z}_i^*)^{\mathrm{T}})^{\mathrm{T}}$ and $\mathbf{R}_i = (X_{i[q+1]}, \ldots, X_{ip})^{\mathrm{T}}$. Further we denote π_+ and π_- respectively to be the marginal probability of the label $Y = 1$ and $Y = -1$.

To facilitate our theoretical analysis, we introduce the gradient vector and Hessian matrix of the population linear weighted hinge loss. Let $L(\beta_1) = \mathbb{E}\{W(1 - Y\mathbf{Z}^{\mathrm{T}}\beta_1)_+\}$ be the population linear weighted hinge loss by using only relevant covariates. Define $S(\beta_1) = (S(\beta_1)_j)$ to be the $(q_n + 1)$-dimensional vector given by

$$S(\beta_1) = -\mathbb{E}\{I(1 - Y\mathbf{Z}^{\mathrm{T}}\beta_1 \geqslant 0)WY\mathbf{Z}\},$$

where $I(\cdot)$ denotes the indicator function. Also define $H(\beta_1) = (H(\beta_1)_{jk})$ to be the $(q_n + 1) \times (q_n + 1)$ matrix given by

$$H(\beta_1) = \mathbb{E}\{\delta(1 - Y\mathbf{Z}^{\mathrm{T}}\beta_1)W\mathbf{Z}\mathbf{Z}^{\mathrm{T}}\},$$

where $\delta(\cdot)$ denotes the Dirac delta function. It can be shown that, if well defined, $S(\beta_1)$ and $H(\beta_1)$ can be considered to be the gradient vector and Hessian matrix of $L(\beta_1)$ respectively. See lemma 2 of Koo *et al.* (2008) for details.

2.1. Non-convex penalized support vector machines

By acting as if the true sparsity structure is known in advance, the oracle estimator is defined as $\hat{\beta} = (\hat{\beta}_1^{\mathrm{T}}, \mathbf{0}^{\mathrm{T}})^{\mathrm{T}}$, where

$$\hat{\beta}_1 = \arg\min_{\beta_1} n^{-1} \sum_{i=1}^{n} W_i(1 - Y_i\mathbf{Z}_i^{\mathrm{T}}\beta_1)_+. \tag{2}$$

Here the objective function is piecewise linear. With a finite sample, it may have multiple minimizers. In that case, $\hat{\beta}_1$ can be chosen to be any minimizer. Our forthcoming theoretical results still hold. In the limit as $n \to \infty$, $\hat{\beta}_1$ minimizes the population version of the objective function $\mathbb{E}\{W(1 - Y\mathbf{Z}^{\mathrm{T}}\beta_1)_+\}$. Lin (2002) showed that, when the misclassification costs are equal, the minimizer of $\mathbb{E}\{(1 - Y f(\mathbf{Z}))_+\}$ over measurable $f(\mathbf{Z})$ is the Bayes rule $\mathrm{sgn}\{p(\mathbf{Z}) - \frac{1}{2}\}$, where $p(\mathbf{z}) = P(Y = 1 | \mathbf{Z} = \mathbf{z})$. This suggests that the oracle estimator is aiming at approximating the Bayes rule. In practice, achieving an estimator with the desired oracle property is very challenging, because the sparsity structure of the true parameter β_0 is largely unknown. Later we shall show that, under some regularity conditions, our proposed algorithm can find an estimator with oracle property and we claim convergence with high probability. Indeed, the numerical examples in Section 5.1 demonstrate that the estimator selected by our proposed algorithm has performance that is close to that of the Bayes rule. Note that the Bayes rule is unattainable here because we assume no knowledge on the high dimensional conditional density $P(\mathbf{X}|Y)$.

In this paper, we consider the non-convex penalized hinge loss objective function

$$Q(\beta) = n^{-1} \sum_{i=1}^{n} W_i(1 - Y_i\mathbf{X}_i^{\mathrm{T}}\beta)_+ + \sum_{j=1}^{p_n} p_{\lambda_n}(|\beta_j|), \tag{3}$$

where $p_{\lambda_n}(\cdot)$ is a symmetric penalty function with tuning parameter λ_n. Let $p'_{\lambda_n}(t)$ be the deriva-

tive of $p_{\lambda_n}(t)$ with respect to t. We consider a general class of non-convex penalties that satisfy the following conditions.

Assumption 1. The symmetric penalty $p_{\lambda_n}(t)$ is assumed to be non-decreasing and concave for $t \in [0, \infty)$, with a continuous derivative $p'_{\lambda_n}(t)$ on $(0, \infty)$ and $p_{\lambda_n}(0) = 0$.

Assumption 2. There exists $a > 1$ such that $\lim_{t \to 0+} p'_{\lambda_n}(t) = \lambda_n$, $p'_{\lambda_n}(t) \geqslant \lambda_n - t/a$ for $0 < t < a\lambda$ and $p'_{\lambda_n}(t) = 0$ for $t \geqslant a\lambda$.

The motivation for such a non-convex penalty is that the convex L_1-penalty lacks the oracle property owing to the overpenalization of large coefficients in the model selected. Consequently it is undesirable to use the L_1-penalty when the purpose of the data analysis is to select the relevant covariates among potentially high dimensional candidates in classification. Note that p, q, λ and other related quantities are allowed to depend on n, and we suppress the subscript n whenever there is no confusion.

Two commonly used non-convex penalties that satisfy assumptions 1 and 2 are the SCAD penalty and the MCP. The SCAD penalty (Fan and Li, 2001) is defined by

$$p_\lambda(|\beta|) = \lambda|\beta| \, I(0 \leqslant |\beta| < \lambda) + \frac{a\lambda|\beta| - (\beta^2 + \lambda^2)/2}{a - 1} I(\lambda \leqslant |\beta| \leqslant a\lambda) + \frac{(a+1)\lambda^2}{2} I(|\beta| > a\lambda)$$

for some $a > 2$. The MCP (Zhang, 2010) is defined by

$$p_\lambda(|\beta|) = \lambda\left(|\beta| - \frac{\beta^2}{2a\lambda}\right) I(0 \leqslant |\beta| < a\lambda) + \frac{a\lambda^2}{2} I(|\beta| \geqslant a\lambda) \qquad \text{for some } a > 1.$$

3. Oracle property

3.1. Regularity conditions

To facilitate our technical proofs, we impose the following regularity conditions.

Condition 1. The densities of $\mathbf{Z}^*$ given $Y = 1$ and $Y = -1$ are continuous and have common support in $\mathcal{R}^q$.

Condition 2. $\mathbb{E}(X_j^2) < \infty$ for $1 \leqslant j \leqslant q$.

Condition 3. The true parameter $\boldsymbol{\beta}_0$ is unique and a non-zero vector.

Condition 4. $q_n = O(n^{c_1})$, namely $\lim_{n \to \infty} q_n / n^{c_1} < \infty$, for some $0 \leqslant c_1 < \frac{1}{2}$.

Condition 5. There is a constant $M_1 > 0$ such that $\lambda_{\max}(n^{-1}\mathbf{X}_A^{\mathrm{T}}\mathbf{X}_A) \leqslant M_1$, where $\mathbf{X}_A$ is the first $q_n + 1$ columns of the design matrix and $\lambda_{\max}$ denotes the largest eigenvalue. It is further assumed that $\max_{1 \leqslant i \leqslant n} \|\mathbf{Z}_i\| = O_p\{\sqrt{q_n} \log(n)\}$, $(\mathbf{Z}_i, Y_i)$ are in general position (Koenker (2005), section 2.2) and X_{ij} are sub-Gaussian random variables for $1 \leqslant i \leqslant n, q_n + 1 \leqslant j \leqslant p_n$.

Condition 6. $\lambda_{\min}\{H(\boldsymbol{\beta}_{01})\} \geqslant M_2$ for some constant $M_2 > 0$, where $\lambda_{\min}$ denotes the smallest eigenvalue.

Condition 7. $n^{(1-c_2)/2} \min_{1 \leqslant j \leqslant q_n} |\beta_{0j}| \geqslant M_3$ for some constant $M_3 > 0$ and $2c_1 < c_2 \leqslant 1$.

Condition 8. Denote the conditional density of $\mathbf{Z}^{\mathrm{T}}\boldsymbol{\beta}_{01}$ given $Y = 1$ and $Y = -1$ as f and g respectively. It is assumed that f is uniformly bounded away from 0 and ∞ in a neighbourhood of 1 and g is uniformly bounded away from 0 and ∞ in a neighbourhood of -1.

Remark 1. Conditions 1–3 and 6 were also assumed for fixed p in Koo *et al.* (2008). We need these assumptions to ensure that the oracle estimator is consistent in the scenario of diverging p. Condition 3 states that the optimal classification decision function is not constant, which is required to ensure that $\mathbf{S}(\boldsymbol{\beta})$ and $\mathbf{H}(\boldsymbol{\beta})$ are a well-defined gradient vector and Hessian matrix of the hinge loss; see lemma 2 and lemma 3 of Koo *et al.* (2008). Conditions 4 and 7 are common in the literature of high dimensional inference (Kim *et al.*, 2008). More specifically, condition 4 states that the divergence rate of the number of non-zero coefficients cannot be faster than $\sqrt{n}$ and condition 7 simply states that the signals cannot decay too quickly. The condition on the largest eigenvalues of the design matrix in condition 5 is similar to the sparse Riesz condition and was also assumed in Zhang and Huang (2008), Yuan (2010) and Zhang (2010). Note that the bound on the smallest eigenvalue is not specified. The condition on the maximum norm in assumption 5 holds when $\mathbf{Z}^*$ given Y follows a multivariate normal distribution. $(\mathbf{Z}_i, Y_i)$ are in general position if with probability 1 there are exactly $q_n + 1$ elements in $\mathbf{D} = \{i : 1 - Y_i \mathbf{Z}_i^{\mathrm{T}} \hat{\boldsymbol{\beta}}_1 = 0\}$ (Koenker (2005), section 2.2). The condition for general position is true with probability 1 with respect to Lebesgue measure. Condition 8 requires that there is enough information around the non-differentiable point of the hinge loss, similarly to condition (C5) in Wang *et al.* (2012) for quantile regression.

For illustrative examples that satisfy all the above conditions, assume that $0 < \pi_+ = 1 - \pi_- < 1$ and let the number of signals be fixed. The first example is that the conditional distributions of $\mathbf{X}^*$ given Y have unbounded support $\mathcal{R}^p$ with sub-Gaussian tails. It can be easily seen that the Fisher discriminant analysis is one special case when $\mathbf{X}^*$ given Y are Gaussian. Conditions 1–4 and 7 are trivial. Condition 5 holds by the properties of sub-Gaussian random variables. Koo *et al.* (2008) showed that condition 6 holds if the supports of the conditional densities of $\mathbf{Z}^*$ given Y are convex, which are naturally satisfied for $\mathcal{R}^q$. Condition 8 is trivially satisfied by the unbounded support of the conditional distribution of $\mathbf{Z}^*$ given Y. Another example is the probit model that $\mathbf{X}^*$ has unbounded support $\mathcal{R}^p$ with sub-Gaussian tails and $\Pr(Y = 1 | \mathbf{X}^*) = \Phi(\mathbf{X}^{\mathrm{T}} \boldsymbol{\beta})$ for some $\boldsymbol{\beta} \neq \mathbf{0}$. It can be easily checked that the conditional distributions of $\mathbf{X}^*$ given Y also have unbounded supports $\mathcal{R}^p$ and hence all the conditions are satisfied.

3.2. Local oracle property

In this subsection, we establish the theory of the local oracle property for the non-convex penalized SVMs, namely the oracle estimator is one of the local minimizers of the objective function $Q(\boldsymbol{\beta})$ defined in equation (3). We start with the following lemma on the consistency of the oracle estimator, which can be viewed as an extension of the consistency result in Koo *et al.* (2008) to the diverging p scenario.

Lemma 1. Assume that conditions 1–7 are satisfied. The oracle estimator $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_1^{\mathrm{T}}, \mathbf{0}^{\mathrm{T}})^{\mathrm{T}}$ satisfies $\|\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{01}\| = O_p\{\sqrt{(q_n/n)}\}$ when $n \to \infty$.

Though the convexity of the non-convex penalized hinge loss objective function $Q(\boldsymbol{\beta})$ is not guaranteed, it can be written as the difference between two convex functions:

$$Q(\boldsymbol{\beta}) = g(\boldsymbol{\beta}) - h(\boldsymbol{\beta}), \tag{4}$$

where

$$g(\boldsymbol{\beta}) = n^{-1} \sum_{i=1}^{n} W_i (1 - Y_i \mathbf{X}_i^{\mathrm{T}} \boldsymbol{\beta})_+ + \lambda_n \sum_{j=1}^{p} |\beta_j|$$

and

$$h(\boldsymbol{\beta}) = \lambda_n \sum_{j=1}^{p} |\beta_j| - \sum_{j=1}^{p} p_{\lambda_n}(|\beta_j|) = \sum_{j=1}^{p} H_{\lambda_n}(\beta_j).$$

The form of $H_\lambda(\beta_j)$ depends on the penalty function. For the SCAD penalty, we have

$$H_\lambda(\beta_j) = \frac{\beta_j^2 - 2\lambda|\beta_j| + \lambda^2}{2(a-1)} I(\lambda \leqslant |\beta_j| \leqslant a\lambda) + \left\{ \lambda|\beta_j| - \frac{(a+1)\lambda^2}{2} \right\} I(|\beta_j| > a\lambda),$$

whereas, for the MCP, we have $H_\lambda(\beta_j) = \{\beta_j^2/(2a)\} I(0 \leqslant |\beta_j| < a\lambda) + (\lambda|\beta_j| - a\lambda^2/2) I(|\beta_j| \geqslant a\lambda)$. This decomposition is useful, as it naturally satisfies the form of the difference of convex functions algorithm (An and Tao, 2005).

To prove the oracle property of the non-convex penalized SVMs, we shall use a sufficient local optimality condition for the difference convex programming that was first presented in Tao and An (1997). This sufficient condition is based on subgradient calculus. The subgradient can be viewed as an extension of the gradient of the smooth convex function to the non-smooth convex function. Let $\mathrm{dom}(g) = \{\mathbf{x} : g(\mathbf{x}) < \infty\}$ be the effective domain of a convex function g. The subgradient of $g(\mathbf{x})$ at a point $\mathbf{x}_0$ is defined as $\partial g(\mathbf{x}_0) = \{\mathbf{t} : g(\mathbf{x}) \geqslant g(\mathbf{x}_0) + (\mathbf{x} - \mathbf{x}_0)^{\mathrm{T}}\mathbf{t}\}$. Note that, at the non-differentiable point, the subgradient contains a collection of vectors. One can easily check that the subgradient of the hinge loss function at the oracle estimator is the collection of vectors $s(\hat{\boldsymbol{\beta}}) = (s_0(\hat{\boldsymbol{\beta}}), \ldots, s_p(\hat{\boldsymbol{\beta}}))^{\mathrm{T}}$ with

$$s_j(\hat{\boldsymbol{\beta}}) = -n^{-1} \sum_{i=1}^{n} W_i Y_i \mathbf{X}_{ij} I(1 - Y_i \mathbf{X}_i^{\mathrm{T}} \hat{\boldsymbol{\beta}} > 0) - n^{-1} \sum_{i=1}^{n} W_i Y_i \mathbf{X}_{ij} v_j, \tag{5}$$

where $-1 \leqslant v_i \leqslant 0$ if $1 - Y_i X_i^{\mathrm{T}} \hat{\boldsymbol{\beta}} = 0$ and $v_i = 0$ otherwise, $j = 0, \ldots, p$. Under some regularity conditions, we can study the asymptotic behaviours of the subgradient at the oracle estimator. The results are summarized in the following theorem.

Theorem 1. Suppose that conditions 1–8 hold, and the tuning parameter satisfies $\lambda = o(n^{-(1-c_2)/2})$ and $\log(p)q\log(n)n^{-1/2} = o(\lambda)$. For the oracle estimator $\hat{\boldsymbol{\beta}}$, there exists v_i^* which satisfies $v_i^* = 0$ if $1 - Y_i X_i^{\mathrm{T}} \hat{\boldsymbol{\beta}} \neq 0$ and $v_i^* \in [-1, 0]$ if $1 - Y_i X_i^{\mathrm{T}} \hat{\boldsymbol{\beta}} = 0$, such that, for $s_j(\hat{\boldsymbol{\beta}})$ with $v_i = v_i^*$, with probability approaching 1, we have

$$\begin{aligned} s_j(\hat{\boldsymbol{\beta}}) &= 0, && j = 0, 1, \ldots, q, \\ |\hat{\beta}_j| &\geqslant (a + \tfrac{1}{2})\lambda, && j = 1, \ldots, q, \\ |s_j(\hat{\boldsymbol{\beta}})| &\leqslant \lambda \text{ and } |\hat{\beta}_j| = 0, && j = q+1, \ldots, p, \end{aligned}$$

Theorem 1 characterizes the subgradients of the hinge loss at the oracle estimator. It basically says that in a regular setting, with probability arbitrarily close to 1, those components of the subgradients corresponding to the relevant covariates are exactly 0 and those corresponding to irrelevant covariates are not far from 0.

We now present the sufficient optimality condition based on subgradient calculation. Corollary 1 of Tao and An (1997) states that, if there is a neighbourhood U around the point $\mathbf{x}^*$ such that $\partial h(\mathbf{x}) \cap \partial g(\mathbf{x}^*) \neq \emptyset, \forall\, \mathbf{x} \in U \cap \mathrm{dom}(g)$, then $\mathbf{x}^*$ is a local minimizer of $g(\mathbf{x}) - h(\mathbf{x})$. To verify this local sufficiency condition, we study the asymptotic behaviours of subgradients of the two convex functions in the aforementioned decomposition (4) of $Q(\boldsymbol{\beta})$. Note that, based on equation (5), the subgradient function of $g(\boldsymbol{\beta})$ at $\boldsymbol{\beta}$ can be shown to be the following collection of vectors:

$$\partial g(\boldsymbol{\beta}) = \left\{ \boldsymbol{\xi} = (\xi_0, \ldots, \xi_p)^{\mathrm{T}} \in \mathcal{R}^{p+1} : \xi_j = -n^{-1} \sum_{i=1}^{n} W_i Y_i \mathbf{X}_{ij} I(1 - Y_i \mathbf{X}_i^{\mathrm{T}} \hat{\boldsymbol{\beta}} > 0) \right.$$
$$\left. - n^{-1} \sum_{i=1}^{n} W_i Y_i \mathbf{X}_{ij} v_j + \lambda l_j, j = 0, \ldots, p \right\},$$

where $l_0 = 0$, $l_j = \operatorname{sgn}(\beta_j)$ if $\beta_j \neq 0$ and $l_j \in [-1, 1]$ otherwise for $1 \leqslant j \leqslant p$, and $-1 \leqslant v_i \leqslant 0$ if $1 - Y_i X_i^{\mathrm{T}} \hat{\boldsymbol{\beta}} = 0$ and $v_i = 0$ otherwise for $1 \leqslant i \leqslant n$. Furthermore, by assumption 2 of the class of non-convex penalty functions, $\lim_{t \to 0+} H'_\lambda(t) = \lim_{t \to 0-} H'_\lambda(t) = \lambda \operatorname{sgn}(t) - \lambda \operatorname{sgn}(t) = 0$. Thus $h(\boldsymbol{\beta})$ is differentiable everywhere. Consequently the subgradient of $h(\boldsymbol{\beta})$ at point $\boldsymbol{\beta}$ is a singleton:

$$\partial h(\boldsymbol{\beta}) = \left\{ \boldsymbol{\mu} = (\mu_0, \ldots, \mu_p) \in \mathcal{R}^{p+1} : \mu_j = \frac{\partial h(\boldsymbol{\beta})}{\partial \beta_j}, j = 0, \ldots, p \right\}.$$

For the class of non-convex penalty functions under consideration, $\partial h(\boldsymbol{\beta})/\partial \beta_j = 0$ for $j = 0$. For $1 \leqslant j \leqslant p$,

$$\frac{\partial h(\boldsymbol{\beta})}{\partial \beta_j} = \frac{\beta_j - \lambda \operatorname{sgn}(\beta_j)}{a - 1} I(\lambda \leqslant |\beta_j| \leqslant a\lambda) + \lambda \operatorname{sgn}(\beta_j) I(|\beta_j| > a\lambda)$$

for the SCAD penalty, and

$$\frac{\partial h(\boldsymbol{\beta})}{\partial \beta_j} = \frac{\beta_j}{a} I(0 \leqslant |\beta_j| < a\lambda) + \lambda \operatorname{sgn}(\beta_j) I(|\beta_j| \geqslant a\lambda)$$

for the MCP.

Combining this with theorem 1, we shall prove that with probability tending to 1, for any $\boldsymbol{\beta}$ in a ball in $\mathcal{R}^{p+1}$ with the centre $\hat{\boldsymbol{\beta}}$ and radius $\lambda/2$, there is a subgradient $\boldsymbol{\xi} = (\xi_0, \ldots, \xi_p)^{\mathrm{T}} \in \partial g(\hat{\boldsymbol{\beta}})$ such that $h(\boldsymbol{\beta})/\partial \beta_j = \xi_j$, $j = 0, 1, \ldots, p$. Consequently the oracle estimator $\hat{\boldsymbol{\beta}}$ is itself a local minimizer of equation (3). This is summarized in the following theorem.

Theorem 2. Assume that conditions 1–8 hold. Let $B_n(\lambda)$ be the set of local minimizers of the objective function $Q(\boldsymbol{\beta})$ with regularization parameter λ. The oracle estimator $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_1^{\mathrm{T}}, \mathbf{0}^{\mathrm{T}})^{\mathrm{T}}$ satisfies

$$\Pr\{\hat{\boldsymbol{\beta}} \in B_n(\lambda)\} \to 1$$

as $n \to \infty$, if $\lambda = o(n^{-(1-c_2)/2})$, and $\log(p) q \log(n) n^{-1/2} = o(\lambda)$.

It can be shown that, if we take $\lambda = n^{-1/2+\delta}$ for some $c_1 < \delta < c_2/2$, then the oracle property holds even for $p = o\{\exp(n^{(\delta - c_1)/2})\}$. Therefore, the local oracle property holds for the non-convex penalized SVM even when the number of covariates grows exponentially with the sample size.

3.3. An algorithm with provable convergence to the oracle estimator

Theorem 2 indicates that one of the local minimizers has the oracle property. However, there can potentially be multiple local minimizers and it remains challenging to identify the oracle estimator. In the high dimensional setting, assuming that the local minimizer is unique would not be realistic.

In this paper, instead of assuming the uniqueness of solutions, we work directly on the conditions under which the oracle estimator can be identified by some numerical algorithms that

solve the non-convex penalized SVM objective function. One possible algorithm is the LLA algorithm that was proposed by Zou and Li (2008). We focus on theoretical development first in this section and delay the detailed LLA algorithm for the non-convex penalized SVMs to Section 4. Recently the LLA has been shown to be capable of identifying the oracle estimator in the set-up of folded concave penalized estimation with a differentiable loss function (Wang *et al.*, 2013; Fan *et al.*, 2014). We generalize their results to non-differentiable loss functions, so that they can fit in the framework of the non-convex penalized SVMs. Similarly to their work, the main condition required is the existence of an appropriate initial estimator inputted in the iterations of the LLA algorithm. Denote the initial estimator as $\tilde{\boldsymbol{\beta}}^{(0)}$. Intuitively, if the initial estimator $\tilde{\boldsymbol{\beta}}^{(0)}$ lies in a small neighbourhood of the true value $\boldsymbol{\beta}_0$, the algorithm should converge to the good local minimizer around $\boldsymbol{\beta}_0$. This localizability will be formalized in terms of L_∞-distance later. With such an appropriate initial estimator, under the aforementioned regularity conditions, one can prove that the LLA algorithm converges to the oracle estimator with probability tending to 1 even in ultrahigh dimensions.

Let $\tilde{\boldsymbol{\beta}}^{(0)} = (\tilde{\beta}_0^{(0)}, \ldots, \tilde{\beta}_p^{(0)})^{\mathrm{T}}$. Consider the following events:

(a) $F_{n1} = \{|\tilde{\beta}_j^{(0)} - \beta_{0j}| > \lambda, \text{ for some } 1 \leqslant j \leqslant p\}$;
(b) $F_{n2} = \{|\beta_{0j}| < (a+1)\lambda, \text{ for some } 1 \leqslant j \leqslant q\}$;
(c) $F_{n3} = \{\text{for all subgradients } s(\hat{\boldsymbol{\beta}}), |s_j(\hat{\boldsymbol{\beta}})| > (1 - 1/a)\lambda \text{ for some } q+1 \leqslant j \leqslant p \text{ or } |s_j(\hat{\boldsymbol{\beta}})| \neq 0 \text{ for some } 0 \leqslant j \leqslant q\}$;
(d) $F_{n4} = \{|\hat{\beta}_j| < a\lambda, \text{ for some } 1 \leqslant j \leqslant q\}$.

Denote the corresponding probability as $P_{ni} = \Pr(F_{ni}), i = 1, 2, 3, 4$. P_{n1} represents the localizability of the problem. When we have an appropriate initial estimator, we expect P_{n1} to converge to 0 as $n \to \infty$. P_{n2} is the probability that the true signal is too small to be detected by any method. P_{n3} describes the behaviour of the subgradients at the oracle estimator. As stated in theorem 1, there is a subgradient such that its components corresponding to irrelevant variables are near 0 and those corresponding to relevant variables are exactly 0, so P_{n3} cannot be too large. P_{n4} is concerned with the magnitude of the oracle estimator on relevant variables. Under regularity conditions, the oracle estimator will detect the true signals and hence P_{n4} will be very small.

Now we provide conditions for the LLA algorithm to find the oracle estimator $\hat{\boldsymbol{\beta}}$ in the non-convex penalized SVMs based on P_{n1}, P_{n2}, P_{n3} and P_{n4}.

Theorem 3. With probability at least $1 - P_{n1} - P_{n2} - P_{n3} - P_{n4}$, the LLA algorithm initiated by $\tilde{\boldsymbol{\beta}}^{(0)}$ finds the oracle estimator $\hat{\boldsymbol{\beta}}$ after two iterations. Furthermore, if conditions 1–8 hold, $\lambda = o(n^{-(1-c_2)/2})$ and $\log(p)q\log(n)n^{-1/2} = o(\lambda)$, then $P_{n2} \to 0$, $P_{n3} \to 0$ and $P_{n4} \to 0$ as $n \to \infty$.

The first part of theorem 3 provides a non-asymptotic lower bound on the probability that the LLA algorithm converges to the oracle estimator. As we shall show in Appendix A, if none of the events F_{ni} happen, the LLA algorithm initiated with $\tilde{\boldsymbol{\beta}}^{(0)}$ will find the oracle estimator in the first iteration, and in the second iteration it will find the oracle estimator again and thus claim convergence. Only a single correction is required in the first iteration and the second iteration is needed to stop the algorithm. Therefore, the LLA algorithm can identify the oracle estimator after two iterations and this result holds generally without conditions 1–8.

The second part of theorem 3 indicates that, under conditions 1–8, the lower bound is determined only by the limiting behaviour of the initial estimator. As long as an appropriate initial estimator is available, the problem of selecting the oracle estimator from potential multiple local minimizers is addressed. Let $\hat{\boldsymbol{\beta}}^{L_1}$ be the solution to the L_1-penalized SVM. When the initial

estimator $\tilde{\boldsymbol{\beta}}^{(0)}$ is taken to be $\hat{\boldsymbol{\beta}}^{L_1}$ and the following condition 9 holds, by theorem 3 the oracle estimator can be identified even in the ultrahigh dimensional setting. The result is summarized in the following corollary.

Condition 9. $\Pr(|\hat{\beta}_j^{L_1} - \beta_{0j}| > \lambda, \text{for some } 1 \leqslant j \leqslant p) \to 0$ as $n \to \infty$.

Corollary 1. Let $\hat{\boldsymbol{\beta}}(\lambda)$ be the solution found by the LLA algorithm initiated by $\hat{\boldsymbol{\beta}}^{L_1}$ after two iterations. Assume that the same conditions in theorem 3 and condition 9 hold; then

$$\Pr\{\hat{\boldsymbol{\beta}}(\lambda) = \hat{\boldsymbol{\beta}}\} \to 1 \qquad \text{as } n \to \infty.$$

In the ultrahigh dimensional case, we may require more stringent conditions to guarantee condition 9. For the non-convex penalized least square regression, one can use the lasso solution (Tibshirani, 1996) as the initial estimator and condition 9 holds if one can further assume the restricted eigenvalue condition of the design matrix (Bickel *et al.*, 2009). However, it is still largely unknown whether this conclusion also applies to the setting where both the loss and the penalty are non-differentiable. Without imposing any new regularity conditions, we next prove that, in the moderately high dimensions with $p = o(\sqrt{n})$, the solution to the L_1-penalized SVM satisfies condition 9 under conditions that are quite similar to 1–8.

The following regularity conditions are modified from 1–8. Conditions 3 and 7 and 8 are the same as aforementioned.

Condition 1′. The densities of $\mathbf{X}^*$ given $Y = 1$ and $Y = -1$ are continuous and have a common support in $\mathcal{R}^p$.

Condition 2′. $\mathbb{E}[X_j^2] < \infty$ for $1 \leqslant j \leqslant p$.

Condition 3′. $p_n = O(n^{c_1})$ for some $0 \leqslant c_1 < \frac{1}{2}$.

Condition 5′. There is a constant $M_1 > 0$ such that $\lambda_{\max}(n^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{X}) \leqslant M_1$. It is further assumed that $\max_{1 \leqslant i \leqslant n} \|\mathbf{X}_i\| = O_p\{\sqrt{p_n}\log(n)\}$, $(\mathbf{X}_i, Y_i)$ are in general position (Koenker (2005), section 2.2) and X_{ij} are sub-Gaussian random variables for $1 \leqslant i \leqslant n, q_n + 1 \leqslant j \leqslant p_n$.

Condition 6′. $\lambda_{\min}\{H(\boldsymbol{\beta}_0)\} \geqslant M_3$ for some constant $M_3 > 0$.

Under the new regularity conditions, we can conclude that the solution to the L_1-penalized SVM is an appropriate initial estimator. Combined with theorem 3, the LLA algorithm initiated with a zero vector can identify the oracle estimator with one more iteration. The results are summarized in the following theorem.

Theorem 4. Assume that $\hat{\boldsymbol{\beta}}^{L_1}$ is the solution to the L_1-penalized SVM with tuning parameter c_n. If the modified conditions hold, $\lambda = o(n^{-(1-c_2)/2})$, $p\log(n)n^{-1/2} = o(\lambda)$ and $c_n = o(n^{-1/2})$, then we have $\Pr(|\hat{\beta}_j^{L_1} - \beta_{0j}| > \lambda, \text{ for some } 1 \leqslant j \leqslant p) \to 0$ as $n \to \infty$. Further, the LLA algorithm initiated by $\hat{\boldsymbol{\beta}}^{L_1}$ finds the oracle estimator in two iterations with probability tending to 1. i.e. $\Pr\{\hat{\boldsymbol{\beta}}(\lambda) = \hat{\boldsymbol{\beta}}\} \to 1$ as $n \to \infty$.

Theorem 4 can guarantee that the LLA algorithm initialized by the $\hat{\boldsymbol{\beta}}^{L_1}$ identifies the oracle estimator with high probability only when $p = o(\sqrt{n})$. However, our empirical studies suggest that, even for cases with p much larger than n, the LLA algorithm initiated by $\hat{\boldsymbol{\beta}}^{L_1}$ usually converges within two iterations and the local minimizer identified has acceptable performance.

4. Implementation and tuning

To solve the non-convex penalized SVMs, we use the LLA algorithm. More explicitly, we start with an initial value $\{\tilde{\boldsymbol{\beta}}^{(0)} : \tilde{\beta}_j^{(0)} = 0, j = 1, 2, \ldots, p\}$. At each step $t \geqslant 1$, we update by solving

$$\min_{\beta}\left\{n^{-1}\sum_{i=1}^{n} W_i(1-Y_i\mathbf{X}_i^{\mathrm{T}}\beta)_+ + \sum_{j=1}^{p} p'_\lambda(|\tilde{\beta}_j^{(t-1)}|)|\beta_j|\right\}, \tag{6}$$

where $p'_\lambda(\cdot)$ denotes the derivative of $p_\lambda(\cdot)$. Following the literature, when $\tilde{\beta}_j^{(t-1)}=0$, we take $p'_\lambda(0)$ as $p'_\lambda(0+)=\lambda$. The LLA algorithm is an instance of the majorize–minimize algorithm and converges to a local minimizer of the non-convex objective function.

With slack variables, the convex optimization problem (6) can be easily recast as a linear programming problem

$$\min_{\xi,\eta,\beta}\left\{n^{-1}\sum_{i=1}^{n} W_i\xi_i + \sum_{j=1}^{p} p'_\lambda(|\tilde{\beta}_j^{(t-1)}|)\eta_j\right\}$$

subject to

$$\begin{aligned} \xi_i &\geqslant 0, & i&=1,2,\ldots,n,\\ \xi_i &\geqslant 1-Y_i\mathbf{X}_i^{\mathrm{T}}\beta, & i&=1,2,\ldots,n,\\ \eta_j &\geqslant \beta_j,\ \eta_j \geqslant -\beta_j, & j&=1,2,\ldots,p. \end{aligned}$$

We propose to use the stopping rule that $p'_\lambda(|\tilde{\beta}_j^{(t-1)}|)$ stabilizes for $j=1,2,\ldots,p$, namely when $\Sigma_{j=1}^{p}\{p'_\lambda(|\tilde{\beta}_j^{(t-1)}|)-p'_\lambda(|\tilde{\beta}_j^{(t)}|)\}^2$ is sufficiently small.

For the choice of tuning parameter λ, Claeskens *et al.* (2008) suggested the SVM information criterion SVMIC which, for a subset S of $\{1,2,\ldots,p\}$, is defined as

$$\mathrm{SVMIC}(S)=\sum_{i=1}^{n}\xi_i+\log(n)|S|,$$

where $|S|$ is the cardinality of S and ξ_i, $i=1,2,\ldots,n$, denote the corresponding optimal slack variables. This criterion directly follows the spirit of the Bayesian information criterion BIC by Schwarz (1978). Chen and Chen (2008) showed that BIC can be too liberal when the model space is large and proposed the extended BIC

$$\mathrm{EBIC}_\gamma(S)=-2\,\mathrm{logLikelihood}+\log(n)|S|+2\gamma\binom{p}{|S|},\qquad 0\leqslant\gamma\leqslant 1.$$

By combining these ideas, we suggest the SVM-extended BIC

$$\mathrm{SVMIC}_\gamma(S)=\sum_{i=1}^{n}2W_i\xi_i+\log(n)|S|+2\gamma\binom{p}{|S|},\qquad 0\leqslant\gamma\leqslant 1.$$

Note that SVMIC_γ reduces to SVMIC when $\gamma=0$ and $w=0.5$. We use $\gamma=0.5$ as suggested by Chen and Chen (2008) and choose the λ that minimizes SVMIC_γ.

5. Simulation and real data examples

We carry out Monte Carlo studies to evaluate the finite sample performance of the non-convex penalized SVMs. We compare the performance of the SCAD-penalized SVM, MCP-penalized SVM, standard L_2-SVM, L_1-penalized SVM, adaptively weighted L_1-penalized SVM (Zou, 2007) and hybrid Huberized SVM (Wang *et al.*, 2008) (denoted by SCAD-svm, MCP-svm, L_2-svm, L_1-svm, Adap L_1-svm and Hybrid-svm respectively) with weight parameter $w=0.5$. The main interest here is the ability to identify the relevant covariates and the control of test error when $p>n$.

5.1. Simulation study

We consider two data generation processes. The first, which is adapted from Park *et al.* (2012), is essentially a standard linear discriminant analysis setting. The second is related to probit regression.

(a) Model 1: $\Pr(Y=1)=\Pr(Y=-1)=0.5$, $\mathbf{X}^*|(Y=1)\sim \mathrm{MN}(\boldsymbol{\mu},\boldsymbol{\Sigma})$, $\mathbf{X}^*|(Y=-1)\sim \mathrm{MN}(-\boldsymbol{\mu},\boldsymbol{\Sigma})$, $q=5$, $\boldsymbol{\mu}=(0.1,0.2,0.3,0.4,0.5,0,\ldots,0)^{\mathrm{T}}\in\mathbf{R}^p$, $\boldsymbol{\Sigma}=(\sigma_{ij})$ with non-zero elements $\sigma_{ii}=1$ for $i=1,2,\ldots,p$ and $\sigma_{ij}=\rho=-0.2$ for $1\leqslant i\neq j\leqslant q$. The Bayes rule is $\mathrm{sgn}(2.67\,X_1+2.83X_2+3X_3+3.17X_4+3.33X_5)$ with Bayes error 6.3%.

(b) Model 2: $\mathbf{X}^*\sim \mathrm{MN}(\mathbf{0}_p,\boldsymbol{\Sigma})$, $\boldsymbol{\Sigma}=(\sigma_{ij})$ with non-zero elements $\sigma_{ii}=1$ for $i=1,2,\ldots,p$ and $\sigma_{ij}=0.4^{|i-j|}$ for $1\leqslant i\neq j\leqslant p$, $\Pr(Y=1|\mathbf{X}^*)=\Phi\{(\mathbf{X}^*)^{\mathrm{T}}\boldsymbol{\beta}^*\}$ where $\Phi(\cdot)$ is the cumulative density function of the standard normal distribution, $\boldsymbol{\beta}^*=(1.1,1.1,1.1,1.1,0,\ldots,0)^{\mathrm{T}}$ and $q=4$. The Bayes rule is $\mathrm{sgn}(X_1+X_2+X_3+X_4)$ with Bayes error 10.4%.

We consider various (n,p) settings for each data generation process with p much larger than n. Similarly to Mazumder *et al.* (2011), an independent tuning data set of size $10n$ is generated to tune any regularization parameter for all methods by minimizing the estimated prediction error calculated over the tuning data set. We also report the performance of the SCAD- and MCP-penalized SVMs by using SVMIC_γ to select the tuning parameter λ. Tuning by a large independent tuning data set of $10n$ approximates the ideal 'population tuning', which is usually not available in practice. By giving all the other methods the best possible tuning, we are controlling the effect of tuning parameter selection and being conservative about the performance of the non-convex penalized SVMs tuned by SVMIC_γ. As we shall see later, the results of SCAD- and MCP-penalized SVMs by using the independent tuning data set are slightly better than the corresponding results by using SVMIC_γ tuning; and all other methods have no ability to select the correct model exactly, even with an unrealistically good tuning parameter. The range of λ is $\{2^{-6},\ldots,2^3\}$. We use $a=3.7$ for the SCAD penalty and $a=3$ for the MCP as suggested in the literature. We generate an independent test data set of size n to report the estimated test error. The columns 'Signal' and 'Noise' summarize the average number of selected relevant and irrelevant covariates respectively. The numbers in the 'Correct' column summarize the percentages of selecting the exactly true model over replications.

Table 1 shows the results for model 1 for various (n,p) settings. The numbers in parentheses are the corresponding standard errors based on 100 replications. When tuned by using an independent tuning set of size $10n$, both SCAD- and MCP-penalized SVMs identify more relevant variables than any other methods and they also reduce the number of falsely selected variables dramatically. When tuned by SVMIC_γ, SCAD- and MCP-penalized SVMs select slightly fewer signals when $n=100$, but this is based on the fact that other methods select a much larger model without proper control of noise. A large proportion of the missed relevant covariates are from X_1 as it has the weakest signal. Note that SVMIC_γ performs almost the same as population tuning when n is relatively large. In general, the non-convex penalized SVMs have an overwhelmingly high probability of selecting the exact true mode as n and p increase, whereas other methods show very weak, if any, ability to recover the exact true model. This is consistent with our theory of the asymptotic oracle property of non-convex penalized SVMs. The test errors of SCAD- and MCP-penalized SVMs are uniformly smaller than those of any other method in all settings, even in the settings with a small sample size $n=100$ and tuned by SVMIC_γ, where they select slightly fewer signals. This is because, in high dimensional classification problems, a large number of falsely selected variables will greatly blur the prediction power of the relevant variables.

Table 1. Simulation results for model 1

Method	*n*	*p*	*Signal*	*Noise*	*Correct (%)*	*Test error (%)*
SCAD-svm	100	400	4.94 (0.03)	0.89 (0.19)	64	8.71 (0.4)
	100	800	4.93 (0.03)	0.93 (0.14)	51	9.39 (0.4)
	200	800	5.00 (0.00)	0.09 (0.05)	96	7.20 (0.2)
	200	1600	5.00 (0.00)	0.07 (0.04)	96	7.24 (0.2)
MCP-svm	100	400	4.90 (0.04)	0.88 (0.17)	53	8.96 (0.4)
	100	800	4.92 (0.03)	1.37 (0.20)	40	10.59 (0.5)
	200	800	5.00 (0.00)	0.06 (0.04)	97	7.30 (0.2)
	200	1600	5.00 (0.00)	0.09 (0.03)	92	6.79 (0.2)
SCAD-svm$^{(\mathrm{SVMIC}_\gamma)}$	100	400	4.64 (0.08)	0.48 (0.11)	64	10.32 (0.6)
	100	800	4.63 (0.09)	0.57 (0.09)	52	11.68 (0.7)
	200	800	5.00 (0.00)	0.03 (0.02)	97	7.24 (0.2)
	200	1600	4.99 (0.01)	0.05 (0.03)	95	7.23 (0.2)
MCP-svm$^{(\mathrm{SVMIC}_\gamma)}$	100	400	4.46 (0.10)	0.44 (0.08)	45	11.81 (0.6)
	100	800	4.34 (0.11)	0.68 (0.11)	38	13.13 (0.7)
	200	800	5.00 (0.00)	0.09 (0.03)	92	7.34 (0.2)
	200	1600	5.00 (0.00)	0.06 (0.03)	95	7.19 (0.2)
L_1-svm	100	400	4.87 (0.05)	32.97 (1.47)	0	16.08 (0.5)
	100	800	4.63 (0.07)	44.34 (2.18)	0	19.71 (0.6)
	200	800	5.00 (0.00)	21.33 (1.70)	0	9.59 (0.3)
	200	1600	4.99 (0.01)	33.37 (0.96)	0	10.88 (0.3)
Hybrid-svm	100	400	4.78 (0.05)	24.74 (1.37)	0	16.34 (0.5)
	100	800	4.62 (0.06)	27.16 (1.30)	0	19.93 (0.6)
	200	800	5.00 (0.00)	12.86 (0.99)	0	9.93 (0.2)
	200	1600	4.99 (0.01)	10.85 (0.98)	0	10.53 (0.3)
Adap L_1-svm	100	400	4.39 (0.08)	13.14 (0.90)	0	16.76 (0.5)
	100	800	3.99 (0.08)	12.50 (0.69)	0	20.19 (0.6)
	200	800	4.86 (0.04)	3.93 (0.25)	1	10.04 (0.3)
	200	1600	4.49 (0.06)	1.01 (0.09)	4	13.43 (0.4)
L_2-svm	100	400	5.00 (0.00)	395.00 (0.00)	0	39.23 (0.5)
	100	800	5.00 (0.00)	795.00 (0.00)	0	42.99 (0.5)
	200	800	5.00 (0.00)	795.00 (0.00)	0	39.22 (0.3)
	200	1600	5.00 (0.00)	1595.00 (0.00)	0	42.50 (0.4)

Table 2 shows the results for model 2 for $n = 250$ and $p = 800$. The numbers in parentheses are the corresponding standard errors based on 200 replications. We observe similar performance patterns in terms of both variable selection and prediction error. Because of the higher correlation between signal and noise, in model 2 it is generally more difficult to select the relevant covariates. Both SCAD- and MCP-penalized SVMs still have a reasonable performance in identifying the underlying true model and result in more accurate prediction. Note that under this data generation process the adaptively weighted L_1-penalized SVM behaves similarly to non-convex penalized SVMs, though its oracle property is largely unknown.

5.2. Real data application

We next use a real data set to illustrate the performance of the non-convex penalized SVM. This data set is part of the 'MicroArray quality control II' project, which is available from the gene expression omnibus database with accession number GSE20194. It contains 278 patient samples from two classes: 164 have positive oestrogen receptor status and 114 have negative oestrogen receptor status. Each sample is described by 22283 genes.

Table 2. Simulation results for model 2 with $n = 250$ and $p = 800$

Method	*Signal*	*Noise*	*Correct (%)*	*Test error (%)*
SCAD-svm	3.99 (0.01)	0.26 (0.08)	92.5	11.4 (0.1)
MCP-svm	3.99 (0.01)	0.17 (0.07)	93.5	11.3 (0.1)
SCAD-svm$^{(\text{SVMIC}_\gamma)}$	3.96 (0.02)	0.05 (0.02)	94	11.5 (0.1)
MCP-svm$^{(\text{SVMIC}_\gamma)}$	3.98 (0.01)	0.07 (0.02)	92.5	11.4 (0.1)
L_1-svm	4.00 (0.00)	6.84 (0.42)	7.5	12.4 (0.1)
Hybrid-svm	4.00 (0.00)	4.03 (0.41)	10.5	11.9 (0.1)
Adap L_1-svm	4.00 (0.00)	2.90 (0.28)	38	11.8 (0.1)
L_2-svm	4.00 (0.00)	796.00 (0.00)	0	32.5 (0.2)

Table 3. Classification error of the gene data set

Method	*Test error (%)*	*Genes*
SCAD-svm	9.8 (0.2)	2.06 (0.43)
MCP-svm	9.6 (0.2)	1.04 (0.02)
L_1-svm	10.9 (0.2)	28.74 (1.36)
Adap L_1-svm	13.1 (0.2)	34.30 (1.03)
Hybrid-svm	10.0 (0.1)	1391.60 (94.86)
L_2-svm	10.8 (0.2)	3000.00 (0.00)

The original data have been standardized for each predictor. To reduce the computational burden, only the 3000 genes with largest absolute values of the two sample t-statistics are used. Such simplification has been considered in Cai and Liu (2011). Though only 3000 genes are used, the classification result is satisfactory. We randomly split the data into an equally balanced training set with 50 samples with positive oestrogen receptor status and 50 samples with negative oestrogen receptor status, and the rest were designated as the test set. As in the simulation study, we use $a = 3.7$ for the SCAD penalty and $a = 3$ for the MCP penalty. To obtain a fair comparison, a fivefold cross-validation is implemented on the training set to select a tuning parameter by a grid search over $\{2^{-15}, \ldots, 2^3\}$ for all methods and the test error is calculated on the test data. This procedure was repeated 100 times.

Table 3 summarizes the average classification error and number of genes selected. The numbers in parentheses are the corresponding standard errors based on 100 replications. Non-convex penalized SVMs achieve a significantly lower test error than all the other methods except for the doubly penalized hybrid SVM. Although the doubly penalized hybrid SVM performs similarly to SCAD- and MCP-penalized SVMs in terms of test error, it selects a much more complex model in general. In addition, the number of genes selected by non-convex penalized SVMs is stable, whereas the model size that is selected by hybrid SVMs ranges from 102 genes to 2576 genes across the 100 replications. Such stability is desirable, so the procedure is robust to the random partition of the data. The numerical results confirm that SCAD- and MCP-penalized SVMs can achieve both promising prediction power and excellent gene selection ability.

6. Discussion

In this paper we study the non-convex penalized SVMs with a diverging number of covariates

in terms of variable selection. When the true model is sparse, under some regularity conditions, we prove that it enjoys the oracle property, i.e. one of the local minimizers of the non-convex penalized SVM behaves like the oracle estimator as if the true sparsity is known in advance and only the relevant variables are used to form the decision boundary. We also show that, as long as we have an appropriate initial estimator, we can identify the oracle estimator with probability tending to 1.

6.1. Connection to Bayes rule

In this paper, the true model and the oracle property are built on $\boldsymbol{\beta}_0$, which is the minimizer of the population version of the hinge loss. This definition has a strong connection to the Bayes rule, which is theoretically optimal if the underlying distribution is known. In the equal weight case ($w = \frac{1}{2}$), the Bayes rule is given by $\operatorname{sgn}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_{\mathrm{Bayes}})$ with $\boldsymbol{\beta}_{\mathrm{Bayes}} = \arg\min_{\boldsymbol{\beta}} \mathbb{E}[I\{\operatorname{sgn}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}) \neq Y\}]$. To appreciate the connection, we first note that $\boldsymbol{\beta}_{\mathrm{Bayes}}$ and $\boldsymbol{\beta}_0$ are equivalent to each other in the important special case of Fisher linear discriminant analysis. Indeed, consider an informative example setting with $\pi_+ = \pi_- = \frac{1}{2}$, $\mathbf{X}^*|(Y=1) \sim N(\boldsymbol{\mu}_+, \boldsymbol{\Sigma})$ and $\mathbf{X}^*|(Y=-1) \sim N(\boldsymbol{\mu}_-, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu}_+$ and $\boldsymbol{\mu}_-$ denote different mean vectors for two classes and $\boldsymbol{\Sigma}$ a same variance–covariance matrix. It is known that in this case the Bayes rule boundary is given by

$$(\boldsymbol{\mu}_+ - \boldsymbol{\mu}_-)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\{\mathbf{x}^* - \tfrac{1}{2}(\boldsymbol{\mu}_+ + \boldsymbol{\mu}_-)\} = 0.$$

$\boldsymbol{\beta}_0$ as the minimizer of the population hinge loss satisfies the gradient condition

$$\mathbf{S}(\boldsymbol{\beta}_0) = -\mathbb{E}\{I(1 - Y\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0 \geqslant 0)Y\mathbf{X}\} = 0,$$

which is equivalent to the equations

$$\begin{aligned} \Pr(1 - \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0 \geqslant 0 | Y = 1) &= \Pr(1 + \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0 \geqslant 0 | Y = -1), \\ \mathbb{E}\{I(1 - \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0 \geqslant 0)\mathbf{X}^* | Y = 1\} &= \mathbb{E}\{I(1 + \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0 \geqslant 0)\mathbf{X}^* | Y = -1\}. \end{aligned} \tag{7}$$

For any $\boldsymbol{\beta}^*_{0,\perp}$ that satisfies $(\boldsymbol{\beta}^*_0)^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\beta}^*_{0,\perp} = 0$, $(\mathbf{X}^*)^{\mathrm{T}}\boldsymbol{\beta}^*_0$ and $(\mathbf{X}^*)^{\mathrm{T}}\boldsymbol{\beta}^*_{0,\perp}$ are conditionally independent given Y and thus we can decompose the conditional expectation in equation (7) into two parts. It can be seen from equation (7) that

$$\begin{aligned} \beta_{00} &= -\tfrac{1}{2}\boldsymbol{\beta}^{*\mathrm{T}}_0(\boldsymbol{\mu}_+ + \boldsymbol{\mu}_-), \\ (\boldsymbol{\mu}_+ - \boldsymbol{\mu}_-)^{\mathrm{T}}\boldsymbol{\beta}^*_{0,\perp} &= 0, \qquad \forall\, \boldsymbol{\beta}^*_{0,\perp} \text{satisfying } \boldsymbol{\beta}^{*\mathrm{T}}_0\boldsymbol{\Sigma}\boldsymbol{\beta}^*_{0,\perp} = 0, \end{aligned}$$

i.e. $\boldsymbol{\mu}_+ - \boldsymbol{\mu}_-$ lies in the space that is spanned by $\boldsymbol{\Sigma}\boldsymbol{\beta}^*_0$. The decision boundary defined by the true value is then

$$\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0 \equiv C(\boldsymbol{\mu}_+ - \boldsymbol{\mu}_-)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\{\mathbf{x}^* - \tfrac{1}{2}(\boldsymbol{\mu}_+ + \boldsymbol{\mu}_-)\} = 0$$

for some constant C. Therefore, the Bayes rule is equivalent to $\boldsymbol{\beta}_0$.

In more general settings, $\boldsymbol{\beta}_{\mathrm{Bayes}}$ and $\boldsymbol{\beta}_0$ may not be the same. However, Lin (2000) showed that the non-linear SVM approaches the Bayes rule in a direct fashion, and its expected misclassification rate quickly converges to that of the Bayes rule even though its extension to a linear SVM is largely unknown. Furthermore, denote $R(f)$ and $R_0(f)$ to be the risk in terms of the 0–1 loss and hinge loss respectively, for any measurable f, i.e. $R(f) = \mathbb{E}(I[\operatorname{sgn}\{f(\mathbf{X})\} \neq Y])$ and $R_0(f) = \mathbb{E}\{(1 - Y f(\mathbf{X}))_+\}$. It is known that minimizing $R(f)$ directly is very difficult because minimizing the empirical 0–1 loss is infeasible in practice (Bartlett *et al.*, 2006). Instead, we can always shift the target from the 0–1 loss to a convex surrogate such as the hinge loss. Assume that the minimizers of $R(f)$ and $R_0(f)$ are both linear functions, and by definition they are $\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_{\mathrm{Bayes}}$

and $\mathbf{X}^{\mathrm{T}}\beta_0$ respectively. By theorem 1 of Bartlett *et al*. (2006), we have the optimal excess risk upper bound

$$R(\mathbf{X}^{\mathrm{T}}\beta) - R(\mathbf{X}^{\mathrm{T}}\beta_{\mathrm{Bayes}}) \leqslant R_0(\mathbf{X}^{\mathrm{T}}\beta) - R_0(\mathbf{X}^{\mathrm{T}}\beta_0)$$

for any β. Hence pursuing the oracle property on β_0 has the potential to control the excess risk efficiently. As can be seen in this paper, the main advantages of working with the hinge loss instead of the 0–1 loss are the theoretical tractability and convenience in practical implementation.

6.2. Other issues

As one referee pointed out, the objective function (2) in the definition of our oracle estimator is piecewise linear and may have multiple minimizers. The same issue applies to the L_1-penalized SVM and the non-convex penalized SVM. On the basis of our theoretical development, non-uniqueness of the minimizer of function (2) is not essential. When the minimizer is not unique, our theoretical results still hold for any particular minimizer. In this sense, we can first use the non-convex penalized SVM to identify important predictors. In the next step, to obtain a unique classifier, a refitting can be applied by using the standard L_2-penalized SVM on those identified important predictors. For model 1 in Section 5.1, we considered this refitting. This additional refitting step does not lead to much improvement: it reduces the average test errors in some settings but not in others. Thus the refitting result is not reported here.

An alternative approach to deal with this non-uniqueness is to consider a joint penalty by using both a non-convex penalty and a standard L_2-penalty. The objective function then becomes

$$n^{-1}\sum_{i=1}^{n} W_i(1 - Y_i\mathbf{X}_i^{\mathrm{T}}\beta)_+ + \sum_{j=1}^{p_n} p_{\lambda_{1n}}(|\beta_j|) + \sum_{j=1}^{p_n} \lambda_{2n}\beta_j^2$$

for two different tuning parameters λ_{1n} and λ_{2n}. The corresponding oracle estimator is then defined as the minimizer of the objective function for the standard L_2-SVM by using only the relevant covariates. One advantage of this joint penalty formulation over the method that is proposed in this paper is that the uniqueness of the oracle estimator is guaranteed in the finite sample case. However, it involves simultaneously selecting two tuning parameters, and this may not be convenient in practice. We conduct a simple numerical experiment using model 1 in Section 5.1 with $n = 200$ and $p = 600$ or $p = 800$. The simulation results are summarized in Table 4. As shown in Table 4, our numerical example suggests that the performance of this joint penalty method is similar to the approach that is proposed in this paper.

Several issues remain unsolved. In this paper we study only the SVMs in non-separable cases in the limit. Although the non-separable cases are important in practical applications,

Table 4. Comparision between SCAD and joint penalized SVMs by using model 1

Method	*p*	*Signal*	*Noise*	*Correct (%)*	*Test error (%)*
SCAD-svm	600	5.00 (0.00)	0.17 (0.07)	93	7.04 (0.2)
	800	5.00 (0.00)	0.13 (0.06)	93	7.25 (0.2)
Joint SCAD + L_2-svm	600	5.00 (0.00)	1.22 (0.28)	65	7.12 (0.2)
	800	5.00 (0.00)	2.64 (0.62)	50	7.10 (0.2)

it would be interesting to show similar results for separable cases. The asymptotic analysis of separable cases requires the positiveness of the limit of the regularization term, which is different from the analysis in this paper. Another issue is the availability of an appropriate initial estimator in ultrahigh dimensions. Our empirical studies suggest that the L_1-penalized SVM provides a reasonable initial estimator and the LLA algorithm converges very quickly even for cases with $p \gg n$. However, it still lacks theoretical justification since our theorem 4 provides theoretical support in only moderately high dimensions with $p = o(\sqrt{n})$. One could try to extend the work of Bickel *et al.* (2009) by assuming similar types of restricted eigenvalues conditions. This extension would require new techniques because both the loss function and the penalty are non-differentiable and the non-smooth locations are different in L_1-penalized SVMs, whereas the set-up in Bickel *et al.* (2009) is a smooth loss function with a non-smooth penalty.

Acknowledgements

We thank the Joint Editors, the Associate Editor and three referees for very constructive comments and suggestions which have improved the presentation of the paper. We also thank Amanda Applegate for her help in the presentation of this paper. The research is partially supported by National Science Foundation grants DMS-1055210 (Wu) and DMS-1308960 (Wang) and National Institutes of Health grants R01-CA149569 (Zhang and Wu), P01-CA142538 (Wu), P50-DA10075 (Li) and P50-DA036107 (Li).

Appendix A

We first prove lemma 1.

A.1. Proof of lemma 1

Let $l(\boldsymbol{\beta}_1) = n^{-1}\sum_{i=1}^n W_i(1 - Y_i\mathbf{Z}_i^{\mathrm{T}}\boldsymbol{\beta}_1)_+$. Note that $\hat{\boldsymbol{\beta}}_1 = \arg\min_{\beta_1} l(\boldsymbol{\beta}_1)$. We shall show that, when $\forall\,\eta > 0$, there is a constant Δ such that, for all n sufficiently large, $\Pr[\inf_{||\mathbf{u}||=\Delta} l\{\boldsymbol{\beta}_{01} + \sqrt{(q/n)}\mathbf{u}\} > l(\boldsymbol{\beta}_{01})] \geqslant 1 - \eta$. Because $l(\boldsymbol{\beta}_1)$ is convex, with probability at least $1 - \eta$, $\boldsymbol{\beta}_1$ is in the ball $\{\boldsymbol{\beta}_1 : ||\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{01}|| \leqslant \Delta\sqrt{(q/n)}\}$. Denote $\Lambda_n(\mathbf{u}) = nq^{-1}[l\{\boldsymbol{\beta}_{01} + \sqrt{(q/n)}\mathbf{u}\} - l(\boldsymbol{\beta}_{01})]$. Observe that $\mathbb{E}\{\Lambda_n(\mathbf{u})\} = nq^{-1}[L\{\boldsymbol{\beta}_{01} + \sqrt{(q/n)}\mathbf{u}\} - L(\boldsymbol{\beta}_{01})]$. Recall also that $\boldsymbol{\beta}_0 = \arg\min_{\beta} \mathbb{E}\{W(1 - Y\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta})\}$. If we restrict the last $p - q$ elements to be 0, it can be easily seen that $\boldsymbol{\beta}_{01} = \arg\min_{\beta_1} \mathbb{E}\{W(1 - Y\mathbf{Z}^{\mathrm{T}}\boldsymbol{\beta}_1)\} = \arg\min_{\beta_1} L(\boldsymbol{\beta}_1)$; thus $S(\boldsymbol{\beta}_{01}) = 0$. By Taylor series expansion of $L(\boldsymbol{\beta}_1)$ around $\boldsymbol{\beta}_{01}$, we have $\mathbb{E}\{\Lambda_n(\mathbf{u})\} = \frac{1}{2}\mathbf{u}^{\mathrm{T}}H(\tilde{\boldsymbol{\beta}})\mathbf{u} + o_p(1)$, where $\tilde{\boldsymbol{\beta}} = \boldsymbol{\beta}_{01} + \sqrt{(q/n)}t\mathbf{u}$ for some $0 < t < 1$. As shown in Koo *et al.* (2008), for $0 \leqslant j, k \leqslant q$, the (j, k)th element of the Hessian matrix $H(\boldsymbol{\beta}_{01})$ is continuous given condition 1 and 2; thus $H(\boldsymbol{\beta})$ is continuous. By continuity of $H(\boldsymbol{\beta})$ at $\boldsymbol{\beta}_{01}$, then $\frac{1}{2}\mathbf{u}^{\mathrm{T}}H(\tilde{\boldsymbol{\beta}})\mathbf{u} = \frac{1}{2}\mathbf{u}^{\mathrm{T}}H(\boldsymbol{\beta}_{01})\mathbf{u} + o(1)$ as $n \to \infty$. Define $\mathbf{W}_n = -\Sigma_{i=1}^n \zeta_i W_i Y_i \mathbf{Z}_i$ where $\zeta_i = I(1 - Y_i\mathbf{Z}_i^{\mathrm{T}}\boldsymbol{\beta}_{01} \geqslant 0)$. Recall that $S(\boldsymbol{\beta}_{01}) = -\mathbb{E}(\zeta_i W_i Y_i \mathbf{Z}_i) = 0$. If we define

$$R_{i,n}(\mathbf{u}) = W_i\left(1 - Y_i\mathbf{Z}_i^{\mathrm{T}}\left(\boldsymbol{\beta}_{01} + \frac{\sqrt{q}}{\sqrt{n}}\mathbf{u}\right)\right)_+ - W_i(1 - Y_i\mathbf{Z}_i^{\mathrm{T}}\boldsymbol{\beta}_{01})_+ + \zeta_i W_i Y_i \mathbf{Z}_i^{\mathrm{T}}\sqrt{(q/n)}\mathbf{u}$$

then we have

$$\Lambda_n(\mathbf{u}) = \mathbb{E}\{\Lambda_n(\mathbf{u})\} + \mathbf{W}_n^{\mathrm{T}}\mathbf{u}/\sqrt{(qn)} + q^{-1}\sum_{i=1}^n [R_{i,n}(\mathbf{u}) - \mathbb{E}\{R_{i,n}(\mathbf{u})\}]. \tag{8}$$

Then similarly to equation (28) in Koo *et al.* (2008) we have

$$q^{-2}\sum_{i=1}^n E[|R_{i,n}(\mathbf{u}) - E\{R_{i,n}(\mathbf{u})\}|^2] \leqslant C\Delta^2\, E[q^{-1}(1 + \|\mathbf{Z}\|^2)\, U\{\sqrt{(1 + \|\mathbf{Z}\|^2)}\Delta\sqrt{(q/n)}\}],$$

where $U(t) = I(|1 - Y_i\mathbf{Z}_i^{\mathrm{T}}\boldsymbol{\beta}_{01}| < t)$. Condition 2 implies that $E\{q^{-1}(1 + \|\mathbf{Z}\|^2)\} < \infty$. Hence, for any $\varepsilon > 0$, we can choose a positive constant C such that $E[q^{-1}(1 + \|\mathbf{Z}\|^2)\, I\{q^{-1}(1 + \|\mathbf{Z}\|^2) > C\}] < \varepsilon/2$; then

$$E[q^{-1}(1+\|\mathbf{Z}\|^2)U\{\sqrt{(1+\|\mathbf{Z}\|^2)}\Delta\sqrt{(q/n)}\}] \leqslant E[q^{-1}(1+\|\mathbf{Z}\|^2)I\{q^{-1}(1+\|\mathbf{Z}\|^2)>C\}] + C\Pr\{|1-Y_i\mathbf{Z}_i^{\mathrm{T}}\boldsymbol{\beta}_{01}|<C\Delta\sqrt{(q/n)}\}.$$

We can take a large N such that $\Pr\{|1-Y_i\mathbf{Z}_i^{\mathrm{T}}\boldsymbol{\beta}_{01}|<C\Delta\sqrt{(q/n)}\}<\varepsilon/(2C)$ for all $n>N$ by condition 4. This proves that $q^{-2}\Sigma_{i=1}^n E[|R_{i,n}(\mathbf{u})-E\{R_{i,n}(\mathbf{u})\}|^2]\to 0$ as $n\to\infty$. Observe that $E\{\mathbf{W}_n^{\mathrm{T}}\mathbf{u}/\sqrt{(qn)}\}=0$, and

$$\operatorname{var}\{\mathbf{W}_n^{\mathrm{T}}\mathbf{u}/\sqrt{(qn)}\} \leqslant Cn^{-1}q^{-1}\sum_{i=1}^n(\mathbf{Z}_i^{\mathrm{T}}\mathbf{u})^2 \leqslant Cq^{-1}\lambda_{\max}(n^{-1}\mathbf{X}_A^{\mathrm{T}}\mathbf{X}_A)\|\mathbf{u}\|^2\to 0$$

as $n\to\infty$. Therefore, the first term of equation (8) will dominate other terms as $n\to\infty$. By condition 6 we have $\frac{1}{2}\mathbf{u}^{\mathrm{T}}H(\boldsymbol{\beta}_{01})\mathbf{u}>0$. Thus we can choose a sufficiently large Δ such that $\Lambda_n(\mathbf{u})>0$ with probability $1-\eta$ for $\|\mathbf{u}\|=\Delta$ and all sufficiently large n.

The proof of theorem 1 relies on the following lemmas.

Lemma 2.

$$\Pr\left\{\max_{q+1\leqslant j\leqslant p} n^{-1}\left|\sum_{i=1}^n W_iY_iX_{ij}\,I(1-Y_i\mathbf{Z}_i^{\mathrm{T}}\boldsymbol{\beta}_{01}\geqslant 0)\right|>\lambda/2\right\}\to 0 \qquad \text{as } n\to\infty.$$

Proof. Recall that $\mathbb{E}\{W_iY_iX_{ij}\,I(1-Y_i\mathbf{Z}_i^{\mathrm{T}}\boldsymbol{\beta}_{01}\geqslant 0)\}=0$. By condition 5 and lemma 14.9 of Bühlmann and Van De Geer (2011), we have $\Pr\{n^{-1}|\Sigma_{i=1}^n W_iY_iX_{ij}\,I(1-Y_i\mathbf{Z}_i^{\mathrm{T}}\boldsymbol{\beta}_{01}\geqslant 0)|>\lambda/2\}\leqslant\exp(-Cn\lambda^2)$. Note that

$$\Pr\left\{\max_{q+1\leqslant j\leqslant p} n^{-1}\left|\sum_{i=1}^n W_iY_iX_{ij}\,I(1-Y_i\mathbf{Z}_i^{\mathrm{T}}\boldsymbol{\beta}_{01}\geqslant 0)\right|>\lambda/2\right\} = \Pr\left\{\bigcup_{q+1\leqslant j\leqslant p}\left\{n^{-1}\left|\sum_{i=1}^n W_iY_iX_{ij}\,I(1-Y_i\mathbf{Z}_i^{\mathrm{T}}\boldsymbol{\beta}_{01}\geqslant 0)\right|>\lambda/2\right\}\right\}\leqslant p\exp(-Cn\lambda^2)\to 0$$

as $n\to\infty$ by the fact that $\log(p)=o(n\lambda^2)$.

Lemma 3. For any $\Delta>0$,

$$\Pr\left[\max_{q+1\leqslant j\leqslant p}\sup_{\|\boldsymbol{\beta}_1-\boldsymbol{\beta}_{01}\|\leqslant\Delta\sqrt{(q/n)}}\left|\sum_{i=1}^n W_iY_iX_{ij}\{I(1-Y_i\mathbf{Z}_i^{\mathrm{T}}\boldsymbol{\beta}_1\geqslant 0)-I(1-Y_i\mathbf{Z}_i^{\mathrm{T}}\boldsymbol{\beta}_{01}\geqslant 0) - \Pr(1-Y_i\mathbf{Z}_i^{\mathrm{T}}\boldsymbol{\beta}_1\geqslant 0)+\Pr(1-Y_i\mathbf{Z}_i^{\mathrm{T}}\boldsymbol{\beta}_{01}\geqslant 0)\}\right|>n\lambda\right]\to 0$$

as $n\to\infty$.

Proof. We generalize an approach by Welsh (1989). We cover the ball $\{\boldsymbol{\beta}_1:\|\boldsymbol{\beta}_1-\boldsymbol{\beta}_{01}\|\leqslant\Delta\sqrt{(q/n)}\}$ with a net of balls with radius $\Delta\sqrt{(q/n^5)}$. It can be shown that this net can be constructed with cardinality $N\leqslant dn^{4q}$ for some $d>0$. Denote the N balls by $B(\mathbf{t}_1),\ldots,B(\mathbf{t}_N)$, where $\mathbf{t}_k$, $k=1,\ldots,N$, are the centres. Denote $\kappa_i(\boldsymbol{\beta}_1)=1-Y_i\mathbf{Z}_i^{\mathrm{T}}\boldsymbol{\beta}_1$, and

$$J_{nj1}=\sum_{k=1}^N\Pr\left(\left|\sum_{i=1}^n W_iY_iX_{ij}[I\{\kappa_i(\mathbf{t}_k)\geqslant 0\}-I\{\kappa_i(\boldsymbol{\beta}_{01})\geqslant 0\}-\Pr\{\kappa_i(\mathbf{t}_k)\geqslant 0\}+\Pr\{\kappa_i(\boldsymbol{\beta}_{01})\geqslant 0\}]\right|>n\lambda/2\right),$$

$$J_{nj2}=\sum_{k=1}^N\Pr\left(\sup_{\tilde{\boldsymbol{\beta}}_1\in B(\mathbf{t}_k)}\left|\sum_{i=1}^n W_iY_iX_{ij}[I\{\kappa_i(\tilde{\boldsymbol{\beta}}_1)\geqslant 0\}-I\{\kappa_i(\mathbf{t}_k)\geqslant 0\}-\Pr\{\kappa_i(\tilde{\boldsymbol{\beta}}_1)\geqslant 0\}+\Pr\{\kappa_i(\mathbf{t}_k)\geqslant 0\}]\right|>n\lambda/2\right).$$

Then, by condition 5,

$$\Pr\left(\sup_{\|\boldsymbol{\beta}_1-\boldsymbol{\beta}_{01}\|\leqslant\Delta\sqrt{(q/n)}}\left|\sum_{i=1}^n W_iY_iX_{ij}[I\{\kappa_i(\boldsymbol{\beta}_1)\geqslant 0\}-I\{\kappa_i(\boldsymbol{\beta}_{01})\geqslant 0\} - \Pr\{\kappa_i(\boldsymbol{\beta}_1)\geqslant 0\}+\Pr\{\kappa_i(\boldsymbol{\beta}_{01})\geqslant 0\}]\right|>n\lambda\right)\leqslant J_{nj1}+J_{nj2}.$$

To evaluate J_{nj1}, let $U_i=W_iY_iX_{ij}[I\{\kappa_i(\mathbf{t}_k)\geqslant 0\}-I\{\kappa_i(\boldsymbol{\beta}_{01})\geqslant 0\}-\Pr\{\kappa_i(\mathbf{t}_k)\geqslant 0\}+\Pr\{\kappa_i(\boldsymbol{\beta}_{01})\geqslant 0\}]$. The

U_i are independent mean 0 random variables, and $\text{var}(U_i) = \mathbb{E}(U_i^2) = \mathbb{E}(U_i^2|Y_i = 1)\ \Pr(Y_i = 1) + \mathbb{E}(U_i^2|Y_i = -1)\Pr(Y_i = -1)$. Denote F and G the cumulative density function of the conditional distribution of $\mathbf{Z}^{\mathrm{T}}\boldsymbol{\beta}_{01}$ given $Y = 1$ and $Y = -1$. Observe that

$$\begin{aligned}\mathbb{E}(U_i^2|Y_i = 1) &\leqslant C(F_i\{1 + \mathbf{Z}_i^{\mathrm{T}}(\boldsymbol{\beta}_{01} - \mathbf{t}_k)\}[1 - F_i\{1 + \mathbf{Z}_i^{\mathrm{T}}(\boldsymbol{\beta}_{01} - \mathbf{t}_k)\}] + F_i(1)\{1 - F_i(1)\} \\ &\quad - 2F_i[\min\{1 + \mathbf{Z}_i^{\mathrm{T}}(\boldsymbol{\beta}_{01} - \mathbf{t}_k), 1\}] + 2\,F_i(1)F_i\{1 + \mathbf{Z}_i^{\mathrm{T}}(\boldsymbol{\beta}_{01} - \mathbf{t}_k)\}) \\ &\leqslant C|\mathbf{Z}_i^{\mathrm{T}}(\mathbf{t}_k - \boldsymbol{\beta}_{01})|,\end{aligned}$$

and it follows by condition 8 that

$$\begin{aligned}\mathbb{E}(U_i^2|Y_i = -1) &\leqslant C\{G_i\{-1 + \mathbf{Z}_i^{\mathrm{T}}(\boldsymbol{\beta}_{01} - \mathbf{t}_k)\}[1 - G_i\{-1 + \mathbf{Z}_i^{\mathrm{T}}(\boldsymbol{\beta}_{01} - \mathbf{t}_k)\}] + G_i(-1)\{1 - G_i(-1)\} \\ &\quad - 2(1 - G_i[\max\{-1 + \mathbf{Z}_i^{\mathrm{T}}(\boldsymbol{\beta}_{01} - \mathbf{t}_k), -1\}]) + 2\{1 - G_i(-1)\}[1 - G_i\{-1 + \mathbf{Z}_i^{\mathrm{T}}(\boldsymbol{\beta}_{01} - \mathbf{t}_k)\}]\} \\ &\leqslant C|\mathbf{Z}_i^{\mathrm{T}}(\mathbf{t}_k - \boldsymbol{\beta}_{01})|.\end{aligned}$$

Thus we have

$$\sum_{i=1}^{n} \text{var}(U_i) \leqslant nC \max_i \|\mathbf{Z}_i\|\|\mathbf{t}_k - \boldsymbol{\beta}_{01}\| = n\,O\{\sqrt{q}\log(n)\}O\{\sqrt{(q/n)}\} = O\{\sqrt{nq}\,\log(n)\}.$$

Applying lemma 14.9 of Bühlmann and Van De Geer (2011), for some positive constants C_1 and C_2 under the assumptions on the rate of λ,

$$J_{nj1} \leqslant 2N\,\exp\left\{-\frac{n^2\lambda^2/4}{C_1\sqrt{nq}\,\log(n) + C_2 n\lambda}\right\} \leqslant C\exp\{4q\log(n) - Cn\lambda\}. \tag{9}$$

To evaluate J_{nj2}, note that $I(x \geqslant s)$ is decreasing in s. Denote

$$V_i = [I\{\kappa_i(\tilde{\boldsymbol{\beta}}_1) \geqslant 0\} - I\{\kappa_i(\mathbf{t}_k) \geqslant 0\} - \Pr\{\kappa_i(\tilde{\boldsymbol{\beta}}_1) \geqslant 0\} + \Pr\{\kappa_i(\mathbf{t}_k) \geqslant 0\}].$$

We have $-B_i \leqslant V_i \leqslant A_i$ for any $\tilde{\boldsymbol{\beta}}_1 \in B(\mathbf{t}_k)$, where

$$\begin{aligned}A_i &= I\{\kappa_i(\mathbf{t}_k) \geqslant -\Delta\sqrt{(q/n^5)}\} - I\{\kappa_i(\mathbf{t}_k) \geqslant 0\} - \Pr\{\kappa_i(\mathbf{t}_k) \geqslant \Delta\sqrt{(q/n^5)}\} + \Pr\{\kappa_i(\mathbf{t}_k) \geqslant 0\}, \\ B_i &= I\{\kappa_i(\mathbf{t}_k) \geqslant 0\} - I\{\kappa_i(\mathbf{t}_k) \geqslant \Delta\sqrt{(q/n^5)}\} - \Pr\{\kappa_i(\mathbf{t}_k) \geqslant 0\} + \Pr\{\kappa_i(\mathbf{t}_k) \geqslant -\Delta\sqrt{(q/n^5)}\}.\end{aligned}$$

Therefore, we have

$$\begin{aligned}&\Pr\left(\sup_{\tilde{\boldsymbol{\beta}}_1 \in B(\mathbf{t}_k)} \left|\sum_{i=1}^{n} W_i Y_i X_{ij}[I\{\kappa_i(\tilde{\boldsymbol{\beta}}_1) \geqslant 0\} - I\{\kappa_i(\mathbf{t}_k) \geqslant 0\} - \Pr\{\kappa_i(\tilde{\boldsymbol{\beta}}_1) \geqslant 0\} + \Pr\{\kappa_i(\mathbf{t}_k) \geqslant 0\}]\right| > n\lambda/2\right) \\ &\quad \leqslant \Pr\left(C\max_i |X_{ij}| \sup_{\tilde{\boldsymbol{\beta}}_1 \in B(\mathbf{t}_k)} \left|\sum_{i=1}^{n} V_i\right| > n\lambda/2\right) \leqslant \Pr\left\{C\max_i |X_{ij}| \max\left(\sum_{i=1}^{n} A_i, \sum_{i=1}^{n} B_i\right) > n\lambda/2\right\}\end{aligned}$$

by the fact that $A_i > 0$ and $B_i > 0$. Note that

$$\begin{aligned}\sum_{i=1}^{n} A_i &= \sum_{i=1}^{n} [I\{\kappa_i(\mathbf{t}_k) \geqslant -\Delta\sqrt{(q/n^5)}\} - I\{\kappa_i(\mathbf{t}_k) \geqslant 0\} - \Pr\{\kappa_i(\mathbf{t}_k) \geqslant -\Delta\sqrt{(q/n^5)}\} + \Pr\{\kappa_i(\mathbf{t}_k) \geqslant 0\}] \\ &\quad + \sum_{i=1}^{n} [\Pr\{\kappa_i(\mathbf{t}_k) \geqslant -\Delta\sqrt{(q/n^5)}\} - \Pr\{\kappa_i(\mathbf{t}_k) \geqslant \Delta\sqrt{(q/n^5)}\}]\end{aligned}$$

and

$$\begin{aligned}&\sum_{i=1}^{n} [\Pr\{\kappa_i(\mathbf{t}_k) \geqslant -\Delta\sqrt{(q/n^5)}\} - \Pr\{\kappa_i(\mathbf{t}_k) \geqslant \Delta\sqrt{(q/n^5)}\}] \\ &\quad = [F_i\{1 + \Delta\sqrt{(q/n^5)} - \mathbf{Z}_i^{\mathrm{T}}(\boldsymbol{\beta}_{01} - \mathbf{t}_k)\} - F_i\{1 - \Delta\sqrt{(q/n^5)} - \mathbf{Z}_i^{\mathrm{T}}(\boldsymbol{\beta}_{01} - \mathbf{t}_k)\}]\Pr(Y_i = 1) \\ &\qquad + [G_i\{-1 + \Delta\sqrt{(q/n^5)} - \mathbf{Z}_i^{\mathrm{T}}(\boldsymbol{\beta}_{01} - \mathbf{t}_k)\} - G_i\{-1 - \Delta\sqrt{(q/n^5)} - \mathbf{Z}_i^{\mathrm{T}}(\boldsymbol{\beta}_{01} - \mathbf{t}_k)\}]\Pr(Y_i = -1) \\ &\quad \leqslant Cn\,\log(q)\sqrt{(q/n^5)}\sqrt{q} = C\,\log(q)qn^{-3/2}\end{aligned}$$

by condition 8. Denote

$$O_i = [I\{\kappa_i(\mathbf{t}_k) \geqslant -\Delta\sqrt{(q/n^5)}\} - I\{\kappa_i(\mathbf{t}_k) \geqslant 0\} - \Pr\{\kappa_i(\mathbf{t}_k) \geqslant -\Delta\sqrt{(q/n^5)}\} + \Pr\{\kappa_i(\mathbf{t}_k) \geqslant 0\}].$$

Thus, for sufficiently large n by $\lambda = o(n^{-(1-c_2)/2})$ and condition 7, we have

$$\sum_{k=1}^{N} \Pr\left(C \sum_{i=1}^{n} A_i > n\lambda/2 \right) \leqslant \sum_{k=1}^{N} \Pr\left\{ C \sum_{i=1}^{n} O_i > n\lambda/2 - C \log(q) q n^{-3/2} \right\} \leqslant \sum_{k=1}^{N} \Pr\left(C \sum_{i=1}^{n} O_i > n\lambda/4 \right).$$

Note that O_i are independent mean 0 random variables, and

$$\mathbb{E}(O_i^2) = \mathbb{E}[I\{\kappa_i(\mathbf{t}_k) \geqslant -\Delta\sqrt{(q/n^5)}\} - I\{\kappa_i(\mathbf{t}_k) \geqslant 0\}]^2 \leqslant \sqrt{(q/n^5)} \max_i \|\mathbf{Z}_i\| = Cq \log(n) n^{-5/2},$$

using a similar idea to deriving the upper bound of $\mathbb{E}(U_i^2)$. Applying Bernstein's inequality and the fact that $\max_i |X_{ij}| = O_p\{\sqrt{\log(n)}\}$ for sub-Gaussian random variables, for some positive constant C_1 and C_2,

$$\sum_{k=1}^{N} \Pr\left(C \max_i |X_{ij}| \sum_{i=1}^{n} A_i > \frac{n\lambda}{2} \right) \leqslant N \exp\left\{ - \frac{n^2\lambda^2/4}{C_1 q n^{-3/2} \log(n)^{3/2} + C_2 n\lambda} \right\} \leqslant C \exp\{4q \log(n) - Cn\lambda\}.$$

Similarly, we can prove that $\Sigma_{k=1}^{N} \Pr(C \max_i |X_{ij}| \Sigma_{i=1}^{n} B_i > n\lambda/2) \leqslant C \exp\{4q \log(n) - Cn\lambda\}$. Therefore, we have

$$J_{nj2} \leqslant C \exp\{4q \log(n) - Cn\lambda\}. \tag{10}$$

Using inequalities (9) and (10), then the probability of lemma 3 is bounded by

$$\sum_{j=q+1}^{p} (J_{nj1} + J_{nj2}) \leqslant C \exp\{\log(p) + 4q \log(n) - Cn\lambda\} \to 0 \tag{11}$$

which completes the proof.

Now we prove Theorem 1.

A.2. Proof of theorem 1

The unpenalized hinge loss objective function is convex. By the convex optimization theorem, there exists v_i^* such that $s_j(\hat{\boldsymbol{\beta}}) = 0$, $j = 0, 1, \ldots, q$, with $v_i = v_i^*$.

Note that $\min_{1 \leqslant j \leqslant q} |\hat{\beta}_j| \geqslant \min_{1 \leqslant j \leqslant q} |\beta_{0j}| - \max_{1 \leqslant j \leqslant q} |\hat{\beta}_j - \beta_{0j}|$. By condition 7 we have $n^{(1-c_2)/2} \times \min_{1 \leqslant j \leqslant q_n} |\beta_{0j}| \geqslant M_1$, and $\max_{1 \leqslant j \leqslant q} |\hat{\beta}_j - \beta_{0j}| = O_p\{\sqrt{(q/n)}\}$ by theorem 1. Thus we have $\min_{1 \leqslant j \leqslant q} |\hat{\beta}_j| = O_p(n^{-(1-c_2)/2})$. By $\lambda = o(n^{-(1-c_2)/2})$, we have $\Pr\{|\hat{\beta}_j| \geqslant (a + \frac{1}{2})\lambda\} \to 1$ for $j = 0, 1, \ldots, q$.

By the definition of the oracle estimator, we have $|\hat{\beta}_j| = 0$, $j = q+1, \ldots, p$. It suffices to show that $\Pr\{|s_j(\hat{\boldsymbol{\beta}})| > \lambda, \text{for some} j = q+1, \ldots, p\} \to 0$. Let $\mathbf{D} = \{i : 1 - Y_i \mathbf{Z}_i^{\mathrm{T}} \hat{\boldsymbol{\beta}}_1 = 0\}$; then, for $j = q+1, \ldots, p$, we have

$$s_j(\hat{\boldsymbol{\beta}}) = -n^{-1} \sum_{i=1}^{n} W_i Y_i X_{ij} I(1 - Y_i \mathbf{Z}_i^{\mathrm{T}} \hat{\boldsymbol{\beta}}_1 \geqslant 0) - n^{-1} \sum_{i \in \mathbf{D}} W_i Y_i X_{ij} (v_j - 1),$$

where $-1 \leqslant v_i \leqslant 0$ if $i \in \mathbf{D}$ and $v_i = 0$ otherwise. By condition 5 $(\mathbf{Z}_i, Y_i)$ are in general positions; with probability 1 there are exactly $q+1$ elements in $\mathbf{D}$. Then by condition 4, with probability 1 $|n^{-1} \Sigma_{i \in \mathbf{D}} W_i Y_i X_{ij} (v_j - 1)| = O\{qn^{-1} \log(q)\} = o(\lambda)$. Thus we need to show only that $\Pr\{\max_{q+1 \leqslant j \leqslant p} |n^{-1} \Sigma_{i=1}^{n} W_i Y_i X_{ij} I(1 - Y_i \mathbf{Z}_i^{\mathrm{T}} \hat{\boldsymbol{\beta}}_1 \geqslant 0)| > \lambda\} \to 0$. Observe that

$$\begin{aligned}
&\Pr\left\{ \max_{q+1 \leqslant j \leqslant p} \left| n^{-1} \sum_{i=1}^{n} W_i Y_i X_{ij} I(1 - Y_i \mathbf{Z}_i^{\mathrm{T}} \hat{\boldsymbol{\beta}}_1 \geqslant 0) \right| > \lambda \right\} \\
&\quad \leqslant \Pr\left[\max_{q+1 \leqslant j \leqslant p} \left| n^{-1} \sum_{i=1}^{n} W_i Y_i X_{ij} \{ I(1 - Y_i \mathbf{Z}_i^{\mathrm{T}} \hat{\boldsymbol{\beta}}_1 \geqslant 0) - I(1 - Y_i \mathbf{Z}_i^{\mathrm{T}} \boldsymbol{\beta}_{01} \geqslant 0) \} \right| > \lambda/2 \right] \\
&\qquad + \Pr\left\{ \max_{q+1 \leqslant j \leqslant p} \left| n^{-1} \sum_{i=1}^{n} W_i Y_i X_{ij} I(1 - Y_i \mathbf{Z}_i^{\mathrm{T}} \boldsymbol{\beta}_{01} \geqslant 0) \right| > \lambda/2 \right\}.
\end{aligned} \tag{12}$$

By lemma 1 the second term of inequality (12) is $o_p(1)$. From lemma 1, the first term of inequality (12) is bounded by

$$\Pr\left[\max_{q+1\leqslant j\leqslant p}\left|n^{-1}\sum_{i=1}^{n}W_iY_iX_{ij}\{I(1-Y_i\mathbf{Z}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}_1\geqslant 0)-I(1-Y_i\mathbf{Z}_i^{\mathrm{T}}\boldsymbol{\beta}_{01}\geqslant 0)\}\right|>\lambda/2\right]$$
$$\leqslant \Pr\left[\max_{q+1\leqslant j\leqslant p}\sup_{\|\boldsymbol{\beta}_1-\boldsymbol{\beta}_{01}\|\leqslant\Delta\sqrt{(q/n)}}\left|n^{-1}\sum_{i=1}^{n}W_iY_iX_{ij}\{I(1-Y_i\mathbf{Z}_i^{\mathrm{T}}\boldsymbol{\beta}_1\geqslant 0)-I(1-Y_i\mathbf{Z}_i^{\mathrm{T}}\boldsymbol{\beta}_{01}\geqslant 0)\right.\right.$$
$$\left.\left.-\Pr(1-Y_i\mathbf{Z}_i^{\mathrm{T}}\boldsymbol{\beta}_1\geqslant 0)+\Pr(1-Y_i\mathbf{Z}_i^{\mathrm{T}}\boldsymbol{\beta}_{01}\geqslant 0)\}\right|>\lambda/4\right]$$
$$+\Pr\left[\max_{q+1\leqslant j\leqslant p}\sup_{\|\boldsymbol{\beta}_1-\boldsymbol{\beta}_{01}\|\leqslant\Delta\sqrt{(q/n)}}\left|n^{-1}\sum_{i=1}^{n}W_iY_iX_{ij}\{\Pr(1-Y_i\mathbf{Z}_i^{\mathrm{T}}\boldsymbol{\beta}_1\geqslant 0)\right.\right.$$
$$\left.\left.-\Pr(1-Y_i\mathbf{Z}_i^{\mathrm{T}}\boldsymbol{\beta}_{01}\geqslant 0)\}\right|>\lambda/4\right]. \tag{13}$$

By lemma 3, the first term of inequality (13) is $o_p(1)$. Thus we need to bound only the second term of inequality (13). Note that

$$|\Pr(1-Y_i\mathbf{Z}_i^{\mathrm{T}}\boldsymbol{\beta}_1\geqslant 0)-\Pr(1-Y_i\mathbf{Z}_i^{\mathrm{T}}\boldsymbol{\beta}_{01}\geqslant 0)|\leqslant|F_i\{1+\mathbf{Z}_i^{\mathrm{T}}(\boldsymbol{\beta}_1-\boldsymbol{\beta}_{01})\}-F_i(1)|\Pr(Y_i=1)$$
$$+|G_i\{-1+\mathbf{Z}_i^{\mathrm{T}}(\boldsymbol{\beta}_1-\boldsymbol{\beta}_{01})\}-G_i(-1)|\Pr(Y_i=-1).$$

Then we have

$$\max_{q+1\leqslant j\leqslant p}\sup_{\|\boldsymbol{\beta}_1-\boldsymbol{\beta}_{01}\|\leqslant\Delta\sqrt{(q/n)}}\left|n^{-1}\sum_{i=1}^{n}W_iY_iX_{ij}\{\Pr(1-Y_i\mathbf{Z}_i^{\mathrm{T}}\boldsymbol{\beta}_1\geqslant 0)-\Pr(1-Y_i\mathbf{Z}_i^{\mathrm{T}}\boldsymbol{\beta}_{01}\geqslant 0)\}\right|$$
$$\leqslant C\max_{i,j}|X_{ij}|\sup_{\|\boldsymbol{\beta}_1-\boldsymbol{\beta}_{01}\|\leqslant\Delta\sqrt{(q/n)}}n^{-1}\sum_{i=1}^{n}\|\mathbf{Z}_i\|\|\boldsymbol{\beta}_1-\boldsymbol{\beta}_{01}\|=O_p\{\sqrt{\log(pn)}\}\,O\{\sqrt{(q/n)}\}\,O_p\{\sqrt{q}\log(n)\}$$
$$=o_p(\lambda).$$

Thus

$$\Pr\left[\max_{q+1\leqslant j\leqslant p}\sup_{\|\boldsymbol{\beta}_1-\boldsymbol{\beta}_{01}\|\leqslant\Delta\sqrt{(q/n)}}\left|n^{-1}\sum_{i=1}^{n}W_iY_iX_{ij}\{\Pr(1-Y_i\mathbf{Z}_i^{\mathrm{T}}\boldsymbol{\beta}_1\geqslant 0)-\Pr(1-Y_i\mathbf{Z}_i^{\mathrm{T}}\boldsymbol{\beta}_{01}\geqslant 0)\}\right|>\lambda/4\right]=o_p(1),$$

which completes the proof.

Now we prove theorem 2.

A.3. Proof of theorem 2

We shall show that $\hat{\boldsymbol{\beta}}$ is a local minimizer of $Q(\boldsymbol{\beta})$ by writing $Q(\boldsymbol{\beta})$ as $g(\boldsymbol{\beta})-h(\boldsymbol{\beta})$.

By theorem 1, we have $\Pr\{\mathcal{G}\subseteq\partial g(\hat{\boldsymbol{\beta}})\}\to 1$, where

$$\mathcal{G}=\{\boldsymbol{\xi}=(\xi_0,\ldots,\xi_p):\xi_0=0;\xi_j=\lambda\,\mathrm{sgn}(\hat{\boldsymbol{\beta}})_j,\,j=1,\ldots,q;\xi_j=s_j(\boldsymbol{\beta})+\lambda l_j,\,j=q+1,\ldots,p\},$$

where $l_j\in[-1,1]$, $j=q+1,\ldots,p$.

Consider any $\boldsymbol{\beta}$ in $\mathbf{R}^{p+1}$ with centre $\hat{\boldsymbol{\beta}}$ and radius $\lambda/2$. It suffices to show that there exist $\boldsymbol{\xi}^*\in\mathcal{G}$ such that $\Pr\{\xi_j^*=\partial h(\boldsymbol{\beta})/\partial\beta_j\}\to 1$ as $n\to\infty$.

Since $\partial h(\boldsymbol{\beta})/\partial\beta_0=0$, we have $\xi_0^*=\partial h(\boldsymbol{\beta})/\partial\beta_0$.

For $j=1,\ldots,q$, we have $\min_{1\leqslant j\leqslant q}|\beta_j|\geqslant\min_{1\leqslant j\leqslant q}|\hat{\beta}_j|-\max_{1\leqslant j\leqslant q}|\hat{\beta}_j-\beta_j|\geqslant(a+\frac{1}{2})\lambda-\lambda/2=a\lambda$ with probability 1 by theorem 1. Therefore by assumption 2 of the class of penalties $\Pr\{\partial h(\boldsymbol{\beta})/\partial\beta_j=\lambda\,\mathrm{sgn}(\beta_j)\}\to 1$ for $j=1,\ldots,q$. For sufficently large n, $\mathrm{sgn}(\beta_j)=\mathrm{sgn}(\hat{\beta}_j)$. Thus we have $\Pr\{\xi_j^*=\partial h(\boldsymbol{\beta})/\partial\beta_j\}\to 1$ as $n\to\infty$ for $j=1,\ldots,q$.

For $j=q+1,\ldots,p$, we have $\Pr\{|\beta_j|\leqslant|\hat{\beta}_j|+|\beta_j-\hat{\beta}_j|\leqslant\lambda\}\to 1$ by theorem 1. Therefore we have $\Pr\{\partial h(\boldsymbol{\beta})/\partial\beta_j=0\}\to 1$ for SCAD and $\Pr\{\partial h(\boldsymbol{\beta})/\partial\beta_j=-\beta_j/a\}\to 1$ for the MCP. Observe that by assumption 2 we have $\Pr\{|\partial h(\boldsymbol{\beta})/\partial\beta_j|\leqslant\lambda\}\to 1$ for the class of penalties. By lemma 1 we have $\Pr\{|s_j(\hat{\beta}_j)|\leqslant\lambda\}\to 1$ for $j=q+1,\ldots,p$. We can always find $l_j\in[-1,1]$ such that $\Pr\{\xi_j^*=s_j(\hat{\beta})+\lambda l_j=\partial h(\boldsymbol{\beta})/\partial\beta_j\}\to 1$ for $j=1,\ldots,q$, for both penalties. This completes the proof.

The proof of theorem 3 consists of two parts. First we shall show that the LLA algorithm initiated by $\tilde{\boldsymbol{\beta}}^{(0)}$ gives the oracle estimator after one iteration. Then we shall show that, once the LLA algorithm

has found the oracle estimator $\hat{\boldsymbol{\beta}}$, the LLA algorithm will find it again in the next iteration, i.e. the LLA algorithm will converge.

A.4. Proof of theorem 3

Assume that none of the events F_{ni} is true, for $i = 1, \ldots, 4$. The probability that none of these event is true is at least $1 - P_{n1} - P_{n2} - P_{n3} - P_{n4}$. Then we have

$$|\tilde{\beta}_j^{(0)}| = |\tilde{\beta}_j^{(0)} - \beta_{0j}| \leqslant \lambda, \qquad q+1 \leqslant j \leqslant p,$$
$$|\tilde{\beta}_j^{(0)}| \geqslant |\beta_{0j}| - |\tilde{\beta}_j^{(0)} - \beta_{0j}| \geqslant a\lambda, \qquad 1 \leqslant j \leqslant q.$$

By assumption 2 of the class of non-convex penalties, we have $p'_\lambda(|\tilde{\beta}_j^{(0)}|) = 0$ for $1 \leqslant j \leqslant q$. Therefore the solution of the next iteration of $\tilde{\boldsymbol{\beta}}^{(1)}$ is the solution to the convex optimization

$$\tilde{\boldsymbol{\beta}}^{(1)} = \arg\min_{\boldsymbol{\beta}} n^{-1} \sum_{i=1}^{n} W_i (1 - Y_i \mathbf{X}_i^{\mathrm{T}} \boldsymbol{\beta})_+ + \sum_{q+1 \leqslant j \leqslant p} p'_\lambda(|\tilde{\beta}_j^{(0)}|)\, |\beta_j|. \tag{14}$$

By the fact the F_{n3} is not true, there are some subgradients of oracle estimator $s(\hat{\boldsymbol{\beta}})$ such that $s_j(\hat{\boldsymbol{\beta}}) = 0$ for $0 \leqslant j \leqslant q$ and $|s_j(\hat{\boldsymbol{\beta}})| < (1 - 1/a)\lambda$ for $q+1 \leqslant j \leqslant p$. By the definition of subgradient, we have

$$\begin{aligned} n^{-1} \sum_{i=1}^{n} W_i (1 - Y_i \mathbf{X}_i^{\mathrm{T}} \boldsymbol{\beta})_+ &\geqslant n^{-1} \sum_{i=1}^{n} W_i (1 - Y_i \mathbf{X}_i^{\mathrm{T}} \hat{\boldsymbol{\beta}})_+ + \sum_{0 \leqslant j \leqslant p} s_j(\hat{\boldsymbol{\beta}})(\beta_j - \hat{\beta}_j) \\ &= n^{-1} \sum_{i=1}^{n} W_i (1 - Y_i \mathbf{X}_i^{\mathrm{T}} \hat{\boldsymbol{\beta}})_+ + \sum_{q+1 \leqslant j \leqslant p} s_j(\hat{\boldsymbol{\beta}})(\beta_j - \hat{\beta}_j). \end{aligned}$$

Then we have for any $\boldsymbol{\beta}$

$$\begin{aligned} &\left\{ n^{-1} \sum_{i=1}^{n} W_i (1 - Y_i \mathbf{X}_i^{\mathrm{T}} \boldsymbol{\beta})_+ + \sum_{q+1 \leqslant j \leqslant p} p'_\lambda(|\tilde{\beta}_j^{(0)}|) |\beta_j| \right\} - \left\{ n^{-1} \sum_{i=1}^{n} W_i (1 - Y_i \mathbf{X}_i^{\mathrm{T}} \hat{\boldsymbol{\beta}})_+ + \sum_{q+1 \leqslant j \leqslant p} p'_\lambda(|\tilde{\beta}_j^{(0)}|) |\hat{\beta}_j| \right\} \\ &\quad \geqslant \sum_{q+1 \leqslant j \leqslant p} \{ p'_\lambda(|\tilde{\beta}_j^{(0)}|) - s_j(\hat{\boldsymbol{\beta}}) \operatorname{sgn}(\beta_j) \} |\beta_j| \geqslant \sum_{q+1 \leqslant j \leqslant p} \{ (1 - 1/a)\lambda - s_j(\hat{\boldsymbol{\beta}}) \operatorname{sgn}(\beta_j) \} |\beta_j| \geqslant 0. \end{aligned}$$

The strict inequality holds unless $\beta_j = 0$ for all $q+1 \leqslant j \leqslant p$. Since we consider the non-separable case that the oracle estimator is unique, we know that the oracle estimator is the unique minimizer of problem (14) and hence $\tilde{\boldsymbol{\beta}}^{(1)} = \hat{\boldsymbol{\beta}}$. This proves that the LLA algorithm finds the oracle estimator after one iteration.

In the case that F_{n2} is not true, we have $|\hat{\beta}_j| > a\lambda$ for all $1 \leqslant j \leqslant q$. Hence by assumption 2 of the class of penalties $p'_\lambda(|\hat{\beta}_j|) = 0$ for all $1 \leqslant j \leqslant q$ and $p'_\lambda(|\hat{\beta}_j|) = p'_\lambda(0) = \lambda$ for all $q+1 \leqslant j \leqslant p$. Once the LLA algorithm has found $\hat{\boldsymbol{\beta}}$, the solution to the next LLA iteration $\tilde{\boldsymbol{\beta}}^{(2)}$ is the minimizer of the convex optimization problem

$$\tilde{\boldsymbol{\beta}}^{(2)} = \arg\min_{\boldsymbol{\beta}} n^{-1} \sum_{i=1}^{n} W_i (1 - Y_i \mathbf{X}_i^{\mathrm{T}} \boldsymbol{\beta})_+ + \sum_{q+1 \leqslant j \leqslant p} \lambda |\beta_j|. \tag{15}$$

Then we have for any $\boldsymbol{\beta}$

$$\begin{aligned} &\left\{ n^{-1} \sum_{i=1}^{n} W_i (1 - Y_i \mathbf{X}_i^{\mathrm{T}} \boldsymbol{\beta})_+ + \sum_{q+1 \leqslant j \leqslant p} \lambda |\beta_j| \right\} - \left\{ n^{-1} \sum_{i=1}^{n} W_i (1 - Y_i \mathbf{X}_i^{\mathrm{T}} \hat{\boldsymbol{\beta}})_+ + \sum_{q+1 \leqslant j \leqslant p} \lambda |\hat{\beta}_j| \right\} \\ &\qquad \geqslant \sum_{q+1 \leqslant j \leqslant p} \{ \lambda - s_j(\hat{\boldsymbol{\beta}}) \operatorname{sgn}(\beta_j) \} |\beta_j| \geqslant 0, \end{aligned}$$

and hence $\tilde{\boldsymbol{\beta}}^{(2)} = \hat{\boldsymbol{\beta}}$ is the unique minimizer of problem (15), i.e. the LLA algorithm finds the oracle estimator again and stops.

As $n \to \infty$, by theorem 1 we have $P_{n2} \to 0$ and $P_{n4} \to 0$. The proof for $P_{n3} \to 0$ is similar to the proof for theorem 1 by changing the constant to $1 - 1/a$.

Now we prove theorem 4.

A.5. Proof of theorem 4

Let $\|\cdot\|_1$ be the L_1-norm of a vector. Denote $l_n(\boldsymbol{\beta}) = n^{-1}\sum_{i=1}^n W_i(1 - Y_i\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta})_+ + c_n\|\boldsymbol{\beta}\|_1$. Note that

$$E(np^{-1}[l_n\{\boldsymbol{\beta}_0 + \sqrt{(p/n)}\mathbf{u}\} - l_n(\boldsymbol{\beta}_0)]) = E[np^{-1}\{W(1 - Y\mathbf{X}^{\mathrm{T}}\{\boldsymbol{\beta}_0 + \sqrt{(p/n)}\mathbf{u}\})_+ - W(1 - Y\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)_+\}] + np^{-1}c_n\{\|\boldsymbol{\beta}_0 + \sqrt{(p/n)}\mathbf{u}\|_1 - \|\boldsymbol{\beta}_0\|_1\}$$

for some constant Δ that $\|\mathbf{u}\| = \Delta$. Observe that $\|\boldsymbol{\beta}_0 + \sqrt{(p/n)}\mathbf{u}\|_1 - \|\boldsymbol{\beta}_0\|_1 \leqslant \|\sqrt{(p/n)}\mathbf{u}\|_1 = \sqrt{(p/n)}\|\mathbf{u}\|_1$. By the fact that $c_n = o(n^{-1/2})$, we have $np^{-1}c_n\{\|\boldsymbol{\beta}_0 + \sqrt{(p/n)}\mathbf{u}\|_1 - \|\boldsymbol{\beta}_0\|_1\} \to 0$ as $n \to \infty$. Then, similarly to the proof of lemma 1, we can show that the expectation is dominated by $\frac{1}{2}u^{\mathrm{T}}H(\boldsymbol{\beta}_0)\mathbf{u} > 0$ and $\Pr[\inf_{\|\mathbf{u}\|_1 \leqslant \Delta} l_n\{\boldsymbol{\beta}_0 + \sqrt{(p/n)}\mathbf{u}\} > l_n(\boldsymbol{\beta}_0)] \geqslant 1 - \eta$. Hence $\|\hat{\boldsymbol{\beta}}^{L_1} - \boldsymbol{\beta}_0\| = O_p\{\sqrt{(p/n)}\}$. Because $pn^{-1/2} = o(\lambda)$, $\Pr(|\hat{\beta}_j^{L_1} - \beta_{0j}| > \lambda$, for some$1 \leqslant j \leqslant p) \to 0$ as $n \to \infty$. Then using theorem 1 and corollary 1 we have $\Pr\{\hat{\boldsymbol{\beta}}(\lambda) = \hat{\boldsymbol{\beta}}\} \to 1$, which completes the proof.

References

An, L. T. H. and Tao, P. D. (2005) The DC (difference of convex functions) programming and DCA revisited with DC models of real world nonconvex optimization problems. *Ann. Ops Res.*, **133**, 23–46.

Bartlett, P. L., Jordan, M. I. and McAuliffe, J. D. (2006) Convexity, classification and risk bounds. *J. Am. Statist. Ass.*, **101**, 138–156.

Becker, N., Toedt, G., Lichter, P. and Benner, A. (2011) Elastic scad as a novel penalization method for svm classification tasks in high-dimensional data. *BMC Bioinform.*, **12**, article 138.

Bickel, P., Ritov, Y. and Tsybakov, A. (2009) Simultaneous analysis of Lasso and Dantzig selector. *Ann. Statist.*, **37**, 1705–1732.

Bradley, P. and Mangasarian, O. (1998) Feature selection via concave minimization and support vector machines. In *Proc. 15th Int. Conf. Machine Learning*, pp. 82–90. San Francisco: Morgan Kaufmann.

Bühlmann, P. and Van De Geer, S. (2011) *Statistics for High-dimensional Data: Methods, Theory and Applications*. Berlin: Springer.

Cai, T. and Liu, W. (2011) A direct estimation approach to sparse linear discriminant analysis. *J. Am. Statist. Ass.*, **106**, 1566–1577.

Chen, J. and Chen, Z. (2008) Extended bayesian information criteria for model selection with large model spaces. *Biometrika*, **95**, 759–771.

Claeskens, G., Croux, C. and Van Kerckhoven, J. (2008) An information criterion for variable selection in support vector machines. *J. Mach. Learn. Res.*, **9**, 541–558.

Donoho, D. (2000) High-dimensional data analysis: the curses and blessings of dimensionality. *Math Challenges Lecture*, pp. 1–32. American Mathematical Society.

Fan, J. and Fan, Y. (2008) High dimensional classification using features annealed independence rules. *Ann. Statist.*, **36**, 2605–2637.

Fan, J. and Li, R. (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Statist. Ass.*, **96**, 1348–1360.

Fan, J. and Lv, J. (2008) Sure independence screening for ultrahigh dimensional feature space (with discussion). *J. R. Statist. Soc.* B, **70**, 849–911.

Fan, J., Xue, L. and Zou, H. (2014) Strong oracle optimality of folded concave penalized estimation. *Ann. Statist.*, **42**, 819–849.

Friedman, J., Hastie, T. and Tibshirani, R. (2001) *The Elements of Statistical Learning*, vol. 1. New York: Springer.

Guyon, I., Weston, J., Barnhill, S. and Vapnik, V. (2002) Gene selection for cancer classification using support vector machines. *Mach. Learn.*, **46**, 389–422.

Kim, Y., Choi, H. and Oh, H. (2008) Smoothly clipped absolute deviation on high dimensions. *J. Am. Statist. Ass.*, **103**, 1665–1673.

Kim, Y. and Kwon, S. (2012) Global optimality of nonconvex penalized estimators. *Biometrika*, **99**, 315–325.

Koenker, R. (2005) *Quantile Regression*. Cambridge: Cambridge University Press.

Koo, J., Lee, Y., Kim, Y. and Park, C. (2008) A Bahadur representation of the linear support vector machine. *J. Mach. Learn. Res.*, **9**, 1343–1368.

Lin, Y. (2000) Some asymptotic properties of the support vector machine. *Technical Report 1029*. Department of Statistics, University of Wisconsin—Madison, Madison.

Lin, Y. (2002) Support vector machines and the bayes rule in classification. *Data Minng Knowl. Discov.*, **6**, 259–275.

Lin, Y., Lee, Y. and Wahba, G. (2002) Support vector machines for classification in nonstandard situations. *Mach. Learn.*, **46**, 191–202.

Mazumder, R., Friedman, J. and Hastie, T. (2011) Sparsenet: coordinate descent with nonconvex penalties. *J. Am. Statist. Ass.*, **106**, 1125–1138.

Meinshausen, N. and Bühlmann, P. (2006) High-dimensional graphs and variable selection with the lasso. *Ann. Statist.*, **34**, 1436–1462.

Meinshausen, N. and Yu, B. (2009) Lasso-type recovery of sparse representations for high-dimensional data. *Ann. Statist.*, **37**, 246–270.
Park, C., Kim, K.-R., Myung, R. and Koo, J.-Y. (2012) Oracle properties of scad-penalized support vector machine. *J. Statist. Planng Inf.*, **142**, 2257–2270.
Schwarz, G. (1978) Estimating the dimension of a model. *Ann. Statist.*, **6**, 461–464.
Tao, P. and An, L. (1997) Convex analysis approach to D.C. programming: theory, algorithms and applications. *Acta Math. Vietnam.*, **22**, 289–355.
Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *J. R. Statist. Soc.* B, **58**, 267–288.
Vapnik, V. (1996) *The Nature of Statistical Learning Theory*. New York: Springer.
Wang, L., Kim, Y. and Li, R. (2013) Calibrating non-convex penalized regression in ultra-high dimension. *Ann. Statist.*, **41**, 2505–2536.
Wang, L., Wu, Y. and Li, R. (2012) Quantile regression for analyzing heterogeneity in ultra-high dimension. *J. Am. Statist. Ass.*, **107**, 214–222.
Wang, L., Zhu, J. and Zou, H. (2006) The doubly regularized support vector machine. *Statist. Sin.*, **16**, 589–615.
Wang, L., Zhu, J. and Zou, H. (2008) Hybrid huberized support vector machines for microarray classification and gene selection. *Bioinformatics*, **24**, 412–419.
Wegkamp, M. and Yuan, M. (2011) Support vector machines with a reject option. *Bernoulli*, **17**, 1368–1385.
Welsh, A. (1989) On *m*-processes and *m*-estimation. *Ann. Statist.*, **17**, 337–361.
Yuan, M. (2010) High dimensional inverse covariance matrix estimation via linear programming. *J. Mach. Learn. Res.*, **99**, 2261–2286.
Zhang, C. (2010) Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.*, **38**, 894–942.
Zhang, H., Ahn, J., Lin, X. and Park, C. (2006) Gene selection using support vector machines with non-convex penalty. *Bioinformatics*, **22**, 88–95.
Zhang, C.-H. and Huang, J. (2008) The sparsity and bias of the lasso selection in high-dimensional linear regression. *Ann. Statist.*, **36**, 1567–1594.
Zhao, P. and Yu, B. (2007) On model selection consistency of lasso. *J. Mach. Learn. Res.*, **7**, 2541–2563.
Zhu, J., Rosset, S., Hastie, T. and Tibshirani, R. (2004) 1-norm support vector machines. *Adv. Neur. Inform. Process. Syst.*, **16**, 49–56.
Zou, H. (2006) The adaptive lasso and its oracle properties. *J. Am. Statist. Ass.* **101**, 1418–1429.
Zou, H. (2007) An improved 1-norm svm for simultaneous classification and variable selection. *J. Mach. Learn. Res.*, 675–681.
Zou, H. and Li, R. (2008) One-step sparse estimates in nonconcave penalized likelihood models. *Ann. Statist.*, **36**, 1509–1533.
Zou, H. and Yuan, M. (2008) The f-infinity norm support vector machine. *Statist. Sin.*, **18**, 379–398.