Biometrika (2014), **101**, 1, *pp.* 57–70
Printed in Great Britain
doi: 10.1093/biomet/ast047
Advance Access publication 13 December 2013

Asymptotic properties for combined L_1 and concave regularization

BY YINGYING FAN AND JINCHI LV

Data Sciences and Operations Department, University of Southern California, Los Angeles, California 90089, U.S.A.

fanyingy@marshall.usc.edu jinchilv@marshall.usc.edu

SUMMARY

Two important goals of high-dimensional modelling are prediction and variable selection. In this article, we consider regularization with combined L_1 and concave penalties, and study the sampling properties of the global optimum of the suggested method in ultrahigh-dimensional settings. The L_1 penalty provides the minimum regularization needed for removing noise variables in order to achieve oracle prediction risk, while a concave penalty imposes additional regularization to control model sparsity. In the linear model setting, we prove that the global optimum of our method enjoys the same oracle inequalities as the lasso estimator and admits an explicit bound on the false sign rate, which can be asymptotically vanishing. Moreover, we establish oracle risk inequalities for the method and the sampling properties of computable solutions. Numerical studies suggest that our method yields more stable estimates than using a concave penalty alone.

Some key words: Concave penalty; Global optimum; Lasso penalty; Prediction; Variable selection.

1. INTRODUCTION

Prediction and variable selection are two important goals in many contemporary large-scale problems. Many regularization methods in the context of penalized empirical risk minimization have been proposed to select important covariates. See, for example, Fan & Lv (2010) for a review of some recent developments in high-dimensional variable selection. Penalized empirical risk minimization has two components: empirical risk for a chosen loss function for prediction, and a penalty function on the magnitude of parameters for reducing model complexity. The loss function is often chosen to be convex. The inclusion of the regularization term helps prevent overfitting when the number of covariates p is comparable to or exceeds the number of observations n.

Generally speaking, two classes of penalty functions have been proposed in the literature: convex ones and concave ones. When a convex penalty such as the lasso penalty (Tibshirani, 1996) is used, the resulting estimator is a well-defined global optimizer. For the properties of L_1 regularization methods, see, for example, Chen et al. (1999), Efron et al. (2004), Zou (2006), Candès & Tao (2007), Rosset & Zhu (2007), and Bickel et al. (2009). In particular, Bickel et al. (2009) proved that using the L_1 penalty leads to estimators satisfying the oracle inequalities under the prediction loss and L_q loss, with $1 \leqslant q \leqslant 2$, in high-dimensional nonparametric regression models. An oracle inequality means that with an overwhelming probability, the loss of the regularized estimator is within a logarithmic factor, a power of $\log p$, of that of the oracle estimator, with the power depending on the chosen estimation loss. Despite these nice properties, the L_1

penalty tends to yield a larger model than the true one for optimizing predictions, and many of the selected variables may be insignificant, showing that the resulting method may not be ideal for variable selection. The relatively large model size also reduces the interpretability of the selected model.

Concave penalties, on the other hand, have been shown to lead to nice variable selection properties. The oracle property was introduced in Fan & Li (2001) to characterize the performance of concave regularization methods, in relation to the oracle procedure knowing the true sparse model in advance. In fixed dimensions, concave regularization has been shown to have the oracle property, recovering the true model with asymptotic probability one. This work has been extended to higher dimensions in different contexts, and the key message is the same. See, for example, Lv & Fan (2009), Zhang (2010), and Fan & Lv (2011). In particular, the weak oracle property, a surrogate of the oracle property, was introduced in Lv & Fan (2009). When $p > n$, it is generally difficult to study the properties of the global optimizer for concave regularization methods. Thus, most studies have focused on some local optimizer that has appealing properties in high-dimensional settings. The sampling properties of the global optimizers for these methods are less well understood in high dimensions.

In this article, we characterize theoretically the global optimizer of the regularization method with the combined L_1 and concave penalty, in the setting of the high-dimensional linear model. We prove that the resulting estimator combines the prediction power of the L_1 penalty and the variable selection power of the concave penalty. On the practical side, the L_1 penalty contributes the minimum amount of regularization necessary to remove noise variables for achieving oracle prediction risk, while the concave penalty incorporates additional regularization to control model sparsity. On the theoretical side, the use of an L_1 penalty helps us to study the various properties of the global optimizer. Specifically, we prove that the global optimizer enjoys the oracle inequalities under the prediction loss and L_q loss, with $1 \leqslant q \leqslant 2$, as well as an asymptotically vanishing bound on the false sign rate. We also establish its oracle risk inequalities under various losses, as well as the sampling properties of computable solutions. In addition, we show that the refitted least-squares estimator can enjoy the oracle property, in the context of Fan & Li (2001). These results are also closely related to those in Zhang & Zhang (2012). Our work complements theirs in three important respects. First, the bound on the number of false positives in Zhang & Zhang (2012) is generally of the same order as the true model size, while our bound on the stronger measure of the rate of false signs can be asymptotically vanishing. Second, our estimation and prediction bounds depend only on the universal regularization parameter for the L_1 component and are free of the regularization parameter λ for the concave component, whereas the bounds in Zhang & Zhang (2012) generally depend on λ alone. Third, our oracle risk inequalities are new and stronger than those for losses, since the risks involve the expectations of losses and thus provide a more complete view of the stability of the method. It is unclear whether the concave method alone would enjoy similar risk bounds.

Our proposal shares a similar spirit to that of Liu & Wu (2007), who proposed a combination of L_0 and L_1 penalties for variable selection and studied its properties in linear regression with fixed dimensionality. Their new penalty yields more stable variable selection results than the L_0 penalty, and outperforms both L_0 and L_1 penalties in terms of variable selection, while maintaining good prediction accuracy. Our theoretical results and numerical study reveal that this advantage persists in high dimensions and for more general concave penalties. Our work differs from theirs in two main respects: we provide more complete and unified theory in ultra high-dimensional settings, and we consider a large class of concave penalties with only mild conditions on their shape. The idea of combining strengths of different penalties has also been exploited in, for example, Zou & Zhang (2009).

2. Model setting

Consider the linear regression model

$$y = X\beta + \varepsilon, \tag{1}$$

where $y = (Y_1, \ldots, Y_n)^{\mathrm{T}}$ is an n-dimensional vector of responses, $X = (x_1, \ldots, x_p)$ is an $n \times p$ design matrix, $\beta = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$ is an unknown p-dimensional vector of regression coefficients, and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^{\mathrm{T}}$ is an n-dimensional vector of noise variates. We are interested in variable selection when the true regression coefficient vector $\beta_0 = (\beta_{0,1}, \ldots, \beta_{0,p})^{\mathrm{T}}$ has many zero components. The main goal is to effectively identify the true underlying sparse model, that is, the support $\mathrm{supp}(\beta_0) = \{j = 1, \ldots, p : \beta_{0,j} \neq 0\}$, with asymptotic probability one, and to efficiently estimate the nonzero regression coefficients $\beta_{0,j}$. A popular approach to estimating sparse β_0 is penalized least squares, which regularizes the conventional least-squares estimation by penalizing the magnitude of parameters $|\beta_j|$. A zero component of the resulting estimate indicates that the corresponding covariate x_j is screened from the model.

Penalized least-squares estimation minimizes the objective function

$$(2n)^{-1}\|y - X\beta\|_2^2 + \|p_\lambda(\beta)\|_1$$

over $\beta \in \mathbb{R}^p$, where we use the compact notation $p_\lambda(\beta) = p_\lambda(|\beta|) = (p_\lambda(|\beta_1|), \ldots, p_\lambda(|\beta_p|))^{\mathrm{T}}$ with $|\beta| = (|\beta_1|, \ldots, |\beta_p|)^{\mathrm{T}}$, and $p_\lambda(t)$, $t \in [0, \infty)$, is a penalty function indexed by the regularization parameter $\lambda \geqslant 0$. The lasso (Tibshirani, 1996) corresponds to the L_1 penalty $p_\lambda(t) = \lambda t$. As shown in Bickel et al. (2009), the lasso enjoys the oracle inequalities for prediction and estimation, but it tends to yield large models. Concave penalties have received much attention due to their oracle properties. Yet, as discussed in § 1, the sampling properties of the global optimizer for concave regularization methods are relatively less well understood in high dimensions. To overcome these difficulties, we suggest combining the L_1 penalty $\lambda_0 t$ with a concave penalty $p_\lambda(t)$, and study the resulting regularization problem

$$\min_{\beta \in \mathbb{R}^p} \{(2n)^{-1}\|y - X\beta\|_2^2 + \lambda_0\|\beta\|_1 + \|p_\lambda(\beta)\|_1\}, \tag{2}$$

where $\lambda_0 = c\{(\log p)/n\}^{1/2}$ for some positive constant c. Throughout the paper, we fix such a choice of the universal regularization parameter for the L_1 penalty, and the minimizer of (2) is implicitly referred to as the global minimizer. The L_1 component $\lambda_0\|\beta\|_1$ helps us to study the global minimizer of (2), and reflects the minimum amount of regularization for removing the noise in prediction. The concave component $\|p_\lambda(\beta)\|_1$ serves to adapt the model sparsity for variable selection.

3. Main results

3·1. *Hard-thresholding property*

To understand why the combination of L_1 and concave penalties can yield better variable selection than can the L_1 penalty alone, we consider the hard-thresholding penalty $p_{\mathrm{H},\lambda}(t) = 2^{-1}\{\lambda^2 - (\lambda - t)_+^2\}$, $t \geqslant 0$. Assume that each covariate x_j is rescaled to have an L_2-norm $n^{1/2}$. Let $\hat\beta = (\hat\beta_1, \ldots, \hat\beta_p)^{\mathrm{T}}$ be the global minimizer of (2) with $p_\lambda(t) = p_{\mathrm{H},\lambda}(t)$. The global optimality of $\hat\beta$ entails that each $\hat\beta_j$ is the global minimizer of the corresponding univariate penalized least-squares problem along the jth coordinate. All these univariate problems share a common form,

with generally different scalars z,

$$\hat{\beta}(z) = \arg\min_{\beta \in \mathbb{R}} \{2^{-1}(z-\beta)^2 + \lambda_0|\beta| + p_{\mathrm{H},\lambda}(|\beta|)\}, \tag{3}$$

since all covariates have L_2-norm $n^{1/2}$. Simple calculus shows that the solution in (3) is

$$\hat{\beta}(z) = \mathrm{sgn}(z)(|z| - \lambda_0)1_{\{|z|>\lambda+\lambda_0\}}, \tag{4}$$

so the resulting estimator has the same feature as the hard-thresholded estimator: each component is either zero or of magnitude larger than λ. This provides an appealing distinction between insignificant covariates, whose coefficients are zero and should be estimated as such, and significant covariates, whose coefficients are significantly nonzero and should be estimated as nonzero, improving the variable selection performance of soft-thresholding by L_1 penalty.

The hard-thresholding feature is shared by many other penalty functions, as now shown.

PROPOSITION 1. *Assume that $p_\lambda(t)$ $(t \geqslant 0)$ is increasing and concave with $p_\lambda(t) \geqslant p_{\mathrm{H},\lambda}(t)$ on $[0, \lambda]$, $p'_\lambda\{(1-c_1)\lambda\} \leqslant c_1\lambda$ for some $c_1 \in [0, 1)$, and $-p''_\lambda(t)$ is decreasing on $[0, (1-c_1)\lambda]$. Then any local minimizer of (2) that is a global minimizer in each coordinate has the hard-thresholding feature that each component is either zero or of magnitude larger than $(1-c_1)\lambda$.*

Although we used the derivatives $p'_\lambda(t)$ and $p''_\lambda(t)$ in the above proposition, the results continue to hold if we replace $-p'_\lambda(t)$ with the subdifferential of $-p_\lambda(t)$, and $-p''_\lambda(t)$ with the local concavity of $p_\lambda(t)$ at point t, when the penalty function is nondifferentiable at t (Lv & Fan, 2009). The hard-thresholding penalty $p_{\mathrm{H},\lambda}(t)$ satisfies the conditions of Proposition 1, with $c_1 = 0$. This class of penalty functions also includes, for example, the L_0 penalty and the smooth integration of counting and absolute deviation penalty (Lv & Fan, 2009), with suitably chosen $c_1 \in [0, 1)$ and tuning parameters.

3·2. *Technical conditions*

We consider a wide range of error distributions for the linear model (1). Throughout this paper, we make the following assumption on the distribution of the model error ε:

$$\mathrm{pr}(\|n^{-1}X^{\mathrm{T}}\varepsilon\|_\infty > \lambda_0/2) = O(p^{-c_0}), \tag{5}$$

where c_0 is some arbitrarily large, positive constant depending only on c, the constant defining λ_0. This condition was imposed in Fan & Lv (2011), who showed for independent $\varepsilon_1, \ldots, \varepsilon_n$ that Gaussian errors and bounded errors satisfy (5) without any extra assumption, and that light-tailed error distributions satisfy (5) with additional mild assumptions on the design matrix X.

Without loss of generality, we assume that only the first s components of β_0 are nonzero, where the true model size s can diverge with the sample size n. Write the true regression coefficient vector as $\beta_0 = (\tilde{\beta}_{0,1}^{\mathrm{T}}, \tilde{\beta}_{0,2}^{\mathrm{T}})^{\mathrm{T}}$ with $\tilde{\beta}_{0,1} = (\beta_{0,1}, \ldots, \beta_{0,s})^{\mathrm{T}} \in \mathbb{R}^s$ the subvector of all nonzero coefficients and $\tilde{\beta}_{0,2} = 0$, and let $p_\lambda(\infty) = \lim_{t\to\infty} p_\lambda(t)$. We impose the following conditions on the design matrix and penalty function, respectively.

Condition 1. For some positive constant κ_0, $\min_{\|\delta\|_2=1,\ \|\delta\|_0<2s} n^{-1/2}\|X\delta\|_2 \geqslant \kappa_0$ and

$$\kappa = \kappa(s, 7) = \min_{\delta \neq 0,\ \|\tilde{\delta}_2\|_1 \leqslant 7\|\tilde{\delta}_1\|_1} \{n^{-1/2}\|X\delta\|_2/(\|\tilde{\delta}_1\|_2 \vee \|\tilde{\delta}_2'\|_2)\} > 0, \tag{6}$$

where $\delta = (\tilde{\delta}_1^{\mathrm{T}}, \tilde{\delta}_2^{\mathrm{T}})^{\mathrm{T}}$ with $\tilde{\delta}_1 \in \mathbb{R}^s$ and $\tilde{\delta}_2'$ the subvector of $\tilde{\delta}_2$ consisting of the components with the s largest absolute values.

Condition 2. The penalty $p_\lambda(t)$ satisfies the conditions of Proposition 1 with $p_\lambda'\{(1-c_1)\lambda\} \leqslant \lambda_0/4$, and $\min_{j=1,\ldots,s} |\beta_{0,j}| > \max\{(1-c_1)\lambda, 2\kappa_0^{-1} p_\lambda^{1/2}(\infty)\}$.

The first part of Condition 1 is a mild sparse eigenvalue condition, and the second part combines the restricted eigenvalue assumptions in Bickel et al. (2009), which were introduced for studying the oracle inequalities for the lasso estimator and Dantzig selector (Candès & Tao, 2007). To see the intuition for (6), recall that ordinary least-squares estimation requires that the Gram matrix $X^{\mathrm{T}}X$ be positive definite, that is,

$$\min_{0 \neq \delta \in \mathbb{R}^p} \{n^{-1/2}\|X\delta\|_2 / \|\delta\|_2\} > 0. \tag{7}$$

In the high-dimensional setting where $p > n$, condition (7) is always violated. Condition 1 replaces the norm $\|\delta\|_2$ in the denominator of (7) with the L_2-norm of only a subvector of δ. Condition 1 also has an additional bound involving $\|\tilde{\delta}_2'\|_2$. This is needed only when dealing with the L_q loss with $q \in (1, 2]$. For other losses, the bound can be relaxed to

$$\kappa = \kappa(s, 7) = \min_{\delta \neq 0,\ \|\tilde{\delta}_2\|_1 \leqslant 7\|\tilde{\delta}_1\|_1} \{n^{-1/2}\|X\delta\|_2 / \|\tilde{\delta}_1\|_2\} > 0.$$

For simplicity, we use the same notation κ in these bounds.

In view of the basic constraint (A7), the restricted eigenvalue assumptions in (6) can be weakened to other conditions such as the compatibility factor or the cone invertibility factor (Zhang & Zhang, 2012). We adopt the assumptions in Bickel et al. (2009) to simplify our presentation.

Condition 2 ensures that the concave penalty $p_\lambda(t)$ satisfies the hard-thresholding property, requires that its tail grow relatively slowly, and puts a constraint on the minimum signal strength.

3·3. *Asymptotic properties of the global optimum*

In this section, we study the sampling properties of the global minimizer $\hat{\beta}$ of (2) with p implicitly understood as $\max(n, p)$ in all bounds. To evaluate the variable selection performance, we consider the number of falsely discovered signs,

$$\mathrm{FS}(\hat{\beta}) = |\{j = 1, \ldots, p : \mathrm{sgn}(\hat{\beta}_j) \neq \mathrm{sgn}(\beta_{0,j})\}|,$$

which is a stronger measure than the total number of false positives and false negatives.

THEOREM 1. *Assume that Conditions 1–2 and the deviation probability bound (5) hold, and that $p_\lambda(t)$ is continuously differentiable. Then the global minimizer $\hat{\beta}$ of (2) has the hard-thresholding property stated in Proposition 1 and, with probability $1 - O(p^{-c_0})$, satisfies simultaneously that*

$$n^{-1/2}\|X(\hat{\beta} - \beta_0)\|_2 = O(\kappa^{-1}\lambda_0 s^{1/2}), \tag{8}$$

$$\|\hat{\beta} - \beta_0\|_q = O(\kappa^{-2}\lambda_0 s^{1/q}), \quad q \in [1, 2], \tag{9}$$

$$\mathrm{FS}(\hat{\beta}) = O\{\kappa^{-4}(\lambda_0/\lambda)^2 s\}.$$

If in addition $\lambda \geqslant 56(1-c_1)^{-1}\kappa^{-2}\lambda_0 s^{1/2}$, *then with probability* $1-O(p^{-c_0})$ *we also have that* $\mathrm{sgn}(\hat{\beta}) = \mathrm{sgn}(\beta_0)$ *and* $\|\hat{\beta}-\beta_0\|_\infty = O\{\lambda_0\|(n^{-1}X_1^{\mathrm{T}}X_1)^{-1}\|_\infty\}$, *where* X_1 *is the* $n \times s$ *submatrix of* X *corresponding to* s *nonzero regression coefficients* $\beta_{0,j}$.

From Theorem 1, we see that if λ is chosen such that $\lambda_0/\lambda \to 0$, then the number of falsely discovered signs $\mathrm{FS}(\hat{\beta})$ is of order $o(s)$ and thus the false sign rate $\mathrm{FS}(\hat{\beta})/s$ is asymptotically vanishing. In contrast, Bickel et al. (2009) showed that under the restricted eigenvalue assumptions, the lasso estimator, with the L_1 component $\lambda_0\|\beta\|_1$ alone, generally gives a sparse model with size of order $O(\phi_{\max}s)$, where $\phi_{\max}$ is the largest eigenvalue of the Gram matrix $n^{-1}X^{\mathrm{T}}X$. This implies that the false sign rate for the lasso estimator can be of order $O(\phi_{\max})$, which does not vanish asymptotically. Similarly, Zhang & Zhang (2012) proved that the number of false positives of the concave regularized estimator is generally of order $O(s)$, which means that the false sign rate can be asymptotically nonvanishing.

The convergence rates in the oracle inequalities (8)–(9), involving both sample size n and dimensionality p, are the same as those for the L_1 component alone in Bickel et al. (2009), and are consistent with those for the concave component alone in Zhang & Zhang (2012). A distinctive feature is that our estimation and prediction bounds in (8)–(9) depend only on the universal regularization parameter $\lambda_0 = c\{(\log p)/n\}^{1/2}$ for the L_1 component, and are independent of the regularization parameter λ for the concave component. In contrast, the bounds in Zhang & Zhang (2012) generally depend on λ alone. The logarithmic factor $\log p$ reflects the general price one needs to pay to search for important variables in high dimensions. In addition, when the signal strength is stronger and the regularization parameter λ is chosen suitably, with the aid of the concave component we have a stronger variable selection result of sign consistency than from using the L_1 penalty alone, in addition to the oracle inequality. Thanks to the inclusion of the L_1 component, another nice feature is that our theory analyses the sampling properties on the whole parameter space $\mathbb{R}^p$, the full space of all possible models, rather than the restriction to the union of lower-dimensional coordinate subspaces such as in Fan & Lv (2011).

The bound on the L_∞ estimation loss in Theorem 1 involves $\|(n^{-1}X_1^{\mathrm{T}}X_1)^{-1}\|_\infty$, which is bounded from above by $s^{1/2}\|(n^{-1}X_1^{\mathrm{T}}X_1)^{-1}\|_2 \leqslant s^{1/2}\kappa_0^{-2}$. The former bound is in general tighter than the latter one. To see this, let us consider the special case where all column vectors of the $n \times s$ subdesign matrix X_1 have equal pairwise correlation $\rho \in [0,1)$. Then the Gram matrix takes the form $n^{-1}X_1^{\mathrm{T}}X_1 = (1-\rho)I_s + \rho 1_s 1_s^{\mathrm{T}}$. By the Sherman–Morrison–Woodbury formula, we have $(n^{-1}X_1^{\mathrm{T}}X_1)^{-1} = (1-\rho)^{-1}I_s - \rho(1-\rho)^{-1}\{1+(s-1)\rho\}^{-1}1_s 1_s^{\mathrm{T}}$, which gives

$$\|(n^{-1}X_1^{\mathrm{T}}X_1)^{-1}\|_\infty = (1-\rho)^{-1}[1+\rho(s-2)\{1+(s-1)\rho\}^{-1}] \leqslant 2(1-\rho)^{-1}.$$

It is interesting to observe that the above matrix ∞-norm has a dimension-free upper bound. Thus in this case, the bound on the L_∞ estimation loss becomes $O[\{(\log p)/n\}^{1/2}]$.

Due to the presence of the L_1 penalty in (2), the resulting global minimizer $\hat{\beta}$ characterized in Theorem 1 may not have the oracle property in the context of Fan & Li (2001). This issue can be resolved using the refitted least-squares estimator on the support $\mathrm{supp}(\hat{\beta})$.

COROLLARY 1. *Assume that all conditions of Theorem 1 hold, and let* $\tilde{\beta}$ *be the refitted least-squares estimator given by covariates in* $\mathrm{supp}(\hat{\beta})$, *with* $\hat{\beta}$ *the estimator in Theorem 1. Then with probability* $1-O(p^{-c_0})$, $\tilde{\beta}$ *equals the oracle estimator, and has the oracle property if the oracle estimator is asymptotically normal.*

Corollary 1 follows immediately from the second part of Theorem 1. Additional regularity conditions ensuring the asymptotic normality of the oracle estimator can be found in, for example, Theorem 4 in Fan & Lv (2011).

THEOREM 2. *Assume that the conditions of Theorem 1 hold, with $\varepsilon_1, \ldots, \varepsilon_n$ independent and identically distributed as ε_0. Then the regularized estimator $\hat{\beta}$ in Theorem 1 satisfies that for any $\tau > 0$,*

$$E\{n^{-1}\|X(\hat{\beta} - \beta_0)\|_2^2\} = O(\kappa^{-2}\lambda_0^2 s + m_{2,\tau} + \gamma\lambda_0 p^{-c_0}), \tag{10}$$

$$E(\|\hat{\beta} - \beta_0\|_q^q) = O[\kappa^{-2q}\lambda_0^q s + (2-q)\lambda_0^{-1} m_{2,\tau} + (q-1)\lambda_0^{-2} m_{4,\tau} + \{(2-q)\gamma + (q-1)\gamma^2\}p^{-c_0}], \quad q \in [1,2], \tag{11}$$

$$E\{\mathrm{FS}(\hat{\beta})\} = O\{\kappa^{-4}(\lambda_0/\lambda)^2 s + \lambda^{-2} m_{2,\tau} + (\gamma\lambda_0/\lambda^2 + s)p^{-c_0}\}, \tag{12}$$

where $m_{q,\tau} = E(|\varepsilon_0|^q 1_{\{|\varepsilon_0|>\tau\}})$ denotes the tail moment and $\gamma = \|\beta_0\|_1 + s\lambda_0^{-1} p_\lambda(\infty) + \tau^2\lambda_0^{-1}$. If in addition $\lambda \geqslant 56(1-c_1)^{-1}\kappa^{-2}\lambda_0 s^{1/2}$, then we also have that $E\{\mathrm{FS}(\hat{\beta})\} = O\{\lambda^{-2} m_{2,\tau} + (\gamma\lambda_0/\lambda^2 + s)p^{-c_0}\}$ and $E(\|\hat{\beta} - \beta_0\|_\infty) = O\{\lambda_0\|(n^{-1}X_1^{\mathrm{T}}X_1)^{-1}\|_\infty + \lambda_0^{-1} m_{2,\tau} + \gamma p^{-c_0}\}$.

Observe that λ_0 enters all bounds for the oracle risk inequalities, whereas λ enters only the risk bound for the variable selection loss. This again reflects the different roles played by the L_1 penalty and concave penalty in prediction and variable selection. The estimation and prediction risk bounds in (10)–(11) as well as the variable selection risk bound in (12) can have leading orders given in their first terms. To understand this, note that each of these first terms is independent of τ and p^{-c_0}, and the remainders in each upper bound can be made sufficiently small, since τ and c_0 can be chosen arbitrarily large. In fact, for bounded error ε_i with range $[-b, b]$, taking $\tau = b$ makes the tail moments $m_{q,\tau}$ vanish. For Gaussian error $\varepsilon_i \sim N(0, \sigma^2)$, by the Gaussian tail probability bound, we can show that $m_{q,\tau} = O[\tau^{q-1}\exp\{-\tau^2/(2\sigma^2)\}]$ for positive integer q. In general, the tail moments can have sufficiently small order by taking a sufficiently large τ diverging with n. All terms involving p^{-c_0} can also be of sufficiently small order by taking a sufficiently large positive constant c in λ_0; see (5).

Our new oracle risk inequalities complement the common results on the oracle inequalities for losses. The inclusion of the L_1 component $\lambda_0 t$ stabilizes prediction and variable selection, and leads to oracle risk bounds. It is, however, unclear whether the concave method alone can enjoy similar risk bounds.

3·4. *Asymptotic properties of computable solutions*

In § 3·3 we have shown that the global minimizer for combined L_1 and concave regularization can enjoy appealing asymptotic properties. Such a global minimizer, however, is not guaranteed to be found by a computational algorithm due to the general nonconvexity of the objective function in (2). Thus a natural question is whether these nice properties can be shared by the computable solution by any algorithm, where a computable solution is typically a local minimizer. Zhang & Zhang (2012) showed that under regularity conditions, any two sparse local solutions can be close to each other. This result, along with the sparsity of the global minimizer in Theorem 1, entails that any sparse computable solution, in the sense of being a local minimizer, can be close to the global minimizer, and thus can enjoy properties similar to the global minimizer. The following theorem establishes these results for sparse computable solutions.

THEOREM 3. *Let $\hat{\beta}$ be a computable local minimizer of (2) that is a global minimizer in each coordinate produced by any algorithm satisfying $\|\hat{\beta}\|_0 \leqslant c_2 s$ and $\|n^{-1}X^{\mathrm{T}}(y - X\hat{\beta})\|_\infty = O(\lambda_0)$, $\lambda \geqslant c_3\lambda_0$, and $\min_{\|\delta\|_2=1,\ \|\delta\|_0\leqslant c_4 s} n^{-1/2}\|X\delta\|_2 \geqslant \kappa_0$ for some positive constants c_2, c_3, κ_0 and sufficiently large positive constant c_4. Then under the conditions of Theorem 1, $\hat{\beta}$ has the same asymptotic properties as the global minimizer in Theorem 1.*

For practical implementation of the method in (2), we employ the path-following coordinate optimization algorithm (Fan & Lv, 2011; Mazumder et al., 2011) and choose the initial estimate to be the lasso estimator $\hat{\beta}_{\text{lasso}}$ with the regularization parameter tuned to minimize the cross-validated prediction error. An analysis of the convergence properties of such an algorithm was presented by Lin & Lv (2013). The use of the lasso estimator as the initial value has also been exploited in, for example, Zhang & Zhang (2012). With the coordinate optimization algorithm, one can obtain a path of sparse computable solutions that are global minimizers in each coordinate. Theorem 3 suggests that a sufficiently sparse computable solution with small correlation between the residual vector and all covariates can enjoy desirable properties.

4. SIMULATION STUDY

We simulated 100 datasets from the linear regression model (1) with $\varepsilon \sim N(0, \sigma^2 I_n)$ and $\sigma = 0{\cdot}25$. For each simulated dataset, the rows of X were sampled as independent and identically distributed copies from $N(0, \Sigma_0)$ with $\Sigma_0 = (0{\cdot}5^{|i-j|})$. We considered $(n, p) = (80, 1000)$ and $(160, 4000)$, and set β as $\beta_0 = (1, -0{\cdot}5, 0{\cdot}7, -1{\cdot}2, -0{\cdot}9, 0{\cdot}3, 0{\cdot}55, 0, \ldots, 0)^{\mathrm{T}}$. For each dataset, we employed the lasso, combined L_1 and smoothly clipped absolute deviation (Fan & Li, 2001), combined L_1 and hard-thresholding, and combined L_1 and smooth integration of counting and absolute deviation penalties to produce a sparse estimate. The minimax concave penalty in Zhang (2010) performed very similarly to the smoothly clipped absolute deviation penalty, so we omit its results to save space. The tuning parameters were selected using BIC.

We considered six performance measures for the estimate $\hat{\beta}$: the prediction error, L_2 loss, L_1 loss, L_∞ loss, the number of false positives, and the number of false negatives. The prediction error is defined as $E(Y - x^{\mathrm{T}}\hat{\beta})^2$, with (x^{T}, Y) an independent observation, which was calculated based on an independent test sample of size 10 000. The L_q loss for estimation is $\|\hat{\beta} - \beta_0\|_q$. A false positive means a selected covariate outside the true sparse model $\text{supp}(\beta_0)$, and a false negative means a missed covariate in $\text{supp}(\beta_0)$.

Table 1 lists the results under different performance measures. The combined L_1 and smoothly clipped absolute deviation, combined L_1 and hard-thresholding, and combined L_1 and smooth integration of counting and absolute deviation methods all performed similarly to the oracle procedure, outperforming the lasso. As the sample size increases, the performance of all methods tends to improve. Although theoretically the oracle inequalities for the L_1 penalty and the combined L_1 and concave penalty can have the same convergence rates, the constants in these oracle inequalities matter in finite samples. This explains the differences in prediction errors and other performance measures in Table 1 for various methods.

We also compared our method with the concave penalty alone. Simulation studies suggest that they have similar performance, except that our method is more stable. To illustrate this, we compared the smoothly clipped absolute deviation with the combined L_1 and smoothly clipped absolute deviation. Boxplots of different performance measures from the two methods showed that the latter reduces outliers and variability, and thus stabilizes the estimate. This result reveals that the same advantage as advocated in Liu & Wu (2007) remains true in high dimensions, with more general concave penalties.

Table 1. *Means and standard errors (in parentheses) of different performance measures*

	Lasso	$L_1 + \text{SCAD}$	$L_1 + \text{Hard}$	$L_1 + \text{SICA}$	Oracle
$n = 80$					
PE ($\times 10^{-2}$)	45·0 (1·7)	8·1 (0·2)	7·0 (0·1)	7·1 (0·1)	6·9 (0·0)
L_2 loss ($\times 10^{-2}$)	86·9 (1·9)	16·8 (1·0)	11·3 (0·4)	11·3 (0·5)	9·7 (0·3)
L_1 loss ($\times 10^{-1}$)	27·6 (0·6)	3·6 (0·2)	2·5 (0·1)	2·5 (0·1)	2·1 (0·1)
L_∞ loss ($\times 10^{-2}$)	48·2 (1·2)	12·1 (0·8)	7·5 (0·3)	7·5 (0·3)	6·6 (0·2)
FP	26·1 (0·5)	0·2 (0·0)	0 (0)	0 (0)	0 (0)
FN	1·0 (0·1)	0·1 (0·0)	0·0 (0·0)	0·0 (0·0)	0 (0)
$n = 160$					
PE ($\times 10^{-2}$)	16·9 (0·5)	6·7 (0·0)	7·0 (0·1)	7·0 (0·1)	6·6 (0·0)
L_2 loss ($\times 10^{-2}$)	45·3 (1·0)	7·7 (0·3)	9·2 (0·4)	9·2 (0·4)	6·6 (0·2)
L_1 loss ($\times 10^{-1}$)	16·2 (0·3)	1·7 (0·1)	2·1 (0·1)	2·1 (0·1)	1·4 (0·0)
L_∞ loss ($\times 10^{-2}$)	24·9 (0·6)	5·3 (0·2)	6·0 (0·2)	5·9 (0·2)	4·4 (0·1)
FP	52·8 (1·1)	0·1 (0·0)	0·7 (0·1)	0·7 (0·1)	0 (0)
FN	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)

$L_1 + \text{SCAD}$, combined L_1 and smoothly clipped absolute deviation; $L_1 + \text{Hard}$, combined L_1 and hard-thresholding; $L_1 + \text{SICA}$, combined L_1 and smooth integration of counting and absolute deviation; PE, prediction error; FP, number of false positives; FN, number of false negatives.

5. Real-data analysis

We applied our method to the lung cancer data originally studied in Gordon et al. (2002) and analysed in Fan & Fan (2008). This dataset consists of 181 tissue samples, with 31 from the malignant pleural mesothelioma of the lung, and 150 from the adenocarcinoma of the lung. Each sample tissue is described by 12 533 genes.

To better evaluate the suggested method, we randomly split the 181 samples into a training set and a test set such that the training set consists of 16 samples from the malignant pleural mesothelioma class and 75 samples from the adenocarcinoma class. Correspondingly, the test set has 15 samples from the malignant pleural mesothelioma class and 75 samples from the adenocarcinoma class. For each split, we employed the same methods as in § 4 to fit the logistic regression model to the training data, and then calculated the classification error using the test data. The tuning parameters were selected using crossvalidation. We repeated the random splitting 50 times, and the means and standard errors of classification errors were 2·960 (0·254) for the lasso, 3·080 (0·262) for combined L_1 and smoothly clipped absolute deviation, 2·960 (0·246) for combined L_1 and hard-thresholding, and 2·980 (0·228) for combined L_1 and smooth integration of counting and absolute deviation. We also calculated the median number of variables chosen by each method: 19 for the first one, 11 for the second one, 11 for the third one, and 12 for the fourth one; the mean model sizes are almost the same as the medians. For each method, we computed the percentage of times each gene was selected, and list the most frequently chosen m genes in the Supplementary Material, with m equal to the median model size by the method. The sets of genes selected by the combined L_1 and concave penalties are subsets of those selected by the lasso.

6. Discussion

Our theoretical analysis shows that the regularized estimate, as the global optimum, given by combined L_1 and concave regularization enjoys the same asymptotic properties as the lasso estimator, but with improved sparsity and false sign rate, in ultrahigh-dimensional linear regression models. These results may be extended to more general model settings and other convex penalties, such as the L_2 penalty. To quantify the stability of variable selection, one can use, for example,

the bootstrap method (Efron, 1979) to estimate the selection probabilities, significance, and estimation uncertainty of selected variables by the regularization method in practice.

ACKNOWLEDGEMENT

The authors sincerely thank the editor, an associate editor, and two referees for comments that significantly improved the paper. This work was supported by the U.S. National Science Foundation and the University of Southern California.

SUPPLEMENTARY MATERIAL

Supplementary material available at *Biometrika* online includes the proofs of Proposition 1 and Theorem 3, and further details for § 5.

APPENDIX

Proof of Theorem 1

Let $\delta = \hat{\beta} - \beta_0$ denote the estimation error with $\hat{\beta}$ the global minimizer of (2). By Condition 2, we see from Proposition 1 that each $\hat{\beta}_j$ is either 0 or of magnitude larger than $(1 - c_1)\lambda$. It follows from the global optimality of $\hat{\beta}$ that

$$(2n)^{-1}\|\varepsilon - X(\hat{\beta} - \beta_0)\|_2^2 + \lambda_0\|\hat{\beta}\|_1 + \|p_\lambda(\hat{\beta})\|_1 \leqslant (2n)^{-1}\|\varepsilon\|_2^2 + \lambda_0\|\beta_0\|_1 + \|p_\lambda(\beta_0)\|_1. \tag{A1}$$

With some simple algebra, (A1) becomes

$$(2n)^{-1}\|X\delta\|_2^2 - n^{-1}\varepsilon^{\mathrm{T}}X\delta + \lambda_0\|\beta_0 + \delta\|_1 + \|p_\lambda(\beta_0 + \delta)\|_1 \leqslant \lambda_0\|\beta_0\|_1 + \|p_\lambda(\beta_0)\|_1. \tag{A2}$$

For notational simplicity, we let $\tilde{a}_1$ and $\tilde{a}_2$ denote the subvectors of a p-vector a consisting of its first s components and remaining $p - s$ components, respectively. Since $\tilde{\beta}_{0,2} = 0$, we have $\tilde{\beta}_{0,2} + \tilde{\delta}_2 = \tilde{\delta}_2$. Thus we can rewrite (A2) as

$$(2n)^{-1}\|X\delta\|_2^2 - n^{-1}\varepsilon^{\mathrm{T}}X\delta + \lambda_0\|\tilde{\delta}_2\|_1 \leqslant \lambda_0\|\tilde{\beta}_{0,1}\|_1 - \lambda_0\|\tilde{\beta}_{0,1} + \tilde{\delta}_1\|_1 + \|p_\lambda(\beta_0)\|_1 - \|p_\lambda(\beta_0 + \delta)\|_1. \tag{A3}$$

The reverse triangle inequality $|\lambda_0\|\tilde{\beta}_{0,1}\|_1 - \lambda_0\|\tilde{\beta}_{0,1} + \tilde{\delta}_1\|_1| \leqslant \lambda_0\|\tilde{\delta}_1\|_1$, along with (A3), yields

$$(2n)^{-1}\|X\delta\|_2^2 - n^{-1}\varepsilon^{\mathrm{T}}X\delta + \lambda_0\|\tilde{\delta}_2\|_1 \leqslant \lambda_0\|\tilde{\delta}_1\|_1 + \|p_\lambda(\beta_0)\|_1 - \|p_\lambda(\beta_0 + \delta)\|_1, \tag{A4}$$

which is key to establishing bounds on prediction and variable selection losses.

To analyse the behaviour of δ, we need to use the concentration property of $n^{-1}X^{\mathrm{T}}\varepsilon$ around its mean zero, as given in the deviation probability bound (5). Condition on the event $\mathcal{E} = \{\|n^{-1}X^{\mathrm{T}}\varepsilon\|_\infty \leqslant \lambda_0/2\}$. On this event, we have

$$-n^{-1}\varepsilon^{\mathrm{T}}X\delta + \lambda_0\|\tilde{\delta}_2\|_1 - \lambda_0\|\tilde{\delta}_1\|_1 \geqslant -(\lambda_0/2)\|\delta\|_1 + \lambda_0\|\tilde{\delta}_2\|_1 - \lambda_0\|\tilde{\delta}_1\|_1 = (\lambda_0/2)\|\tilde{\delta}_2\|_1 - (3\lambda_0/2)\|\tilde{\delta}_1\|_1.$$

This inequality, together with (A4), gives

$$(2n)^{-1}\|X\delta\|_2^2 + (\lambda_0/2)\|\tilde{\delta}_2\|_1 \leqslant (3\lambda_0/2)\|\tilde{\delta}_1\|_1 + \|p_\lambda(\beta_0)\|_1 - \|p_\lambda(\beta_0 + \delta)\|_1. \tag{A5}$$

In order to proceed, we need to construct an upper bound for $\|p_\lambda(\beta_0)\|_1 - \|p_\lambda(\beta_0 + \delta)\|_1$. We claim that such an upper bound is $(4n)^{-1}\|X\delta\|_2^2 + 4^{-1}\lambda_0\|\delta\|_1$. To prove this, we consider two cases.

Case 1: $\|\hat{\beta}\|_0 \geqslant s$. Then by Condition 2, we have $|\beta_{0,j}| > (1 - c_1)\lambda$ $(j = 1, \ldots, s)$ and $p'_\lambda\{(1 - c_1)\lambda\} \leqslant \lambda_0/4$. For each $j = 1, \ldots, s$, if $\hat{\beta}_j \neq 0$, we must have $|\hat{\beta}_j| > (1 - c_1)\lambda$, and thus by the mean-value theorem, $|p_\lambda(|\beta_{0,j}|) - p_\lambda(|\hat{\beta}_j|)| = p'_\lambda(t)|(|\hat{\beta}_j| - |\beta_{0,j}|)| \leqslant p'_\lambda(t)|\delta_j|$, where t is between $|\beta_{0,j}|$ and $|\hat{\beta}_j|$, and δ_j is the jth component of δ. Clearly $t > (1 - c_1)\lambda$, which along with the concavity of $p_\lambda(t)$ leads to $p'_\lambda(t) \leqslant p'_\lambda\{(1 - c_1)\lambda\} \leqslant \lambda_0/4$. This shows that $|p_\lambda(|\beta_{0,j}|) - p_\lambda(|\hat{\beta}_j|)| \leqslant 4^{-1}\lambda_0|\delta_j|$ for each $j = 1, \ldots, s$

with $\hat{\beta}_j \neq 0$. We now consider $j = 1, \ldots, s$ with $\hat{\beta}_j = 0$. Since $\|\hat{\beta}\|_0 \geqslant s$, there exists some $j' > s$ such that $\hat{\beta}_{j'} \neq 0$ and the j' are distinct for different j. Similarly to above, we have that for some t_1 between $(1 - c_1)\lambda$ and $|\beta_{0,j}|$ and some t_2 between $(1 - c_1)\lambda$ and $|\hat{\beta}_{j'}|$,

$$\begin{aligned}|p_\lambda(|\beta_{0,j}|) - p_\lambda(|\hat{\beta}_{j'}|)| &\leqslant |p_\lambda(|\beta_{0,j}|) - p_\lambda\{(1-c_1)\lambda\}| + |p_\lambda(|\hat{\beta}_{j'}|) - p_\lambda\{(1-c_1)\lambda\}| \\ &= p'_\lambda(t_1)\{|\beta_{0,j}| - (1-c_1)\lambda\} + p'_\lambda(t_2)\{|\hat{\beta}_{j'}| - (1-c_1)\lambda\} \\ &\leqslant p'_\lambda(t_1)|\delta_j| + p'_\lambda(t_2)|\delta_{j'}| \leqslant (\lambda_0/4)(|\delta_j| + |\delta_{j'}|),\end{aligned}$$

since $\hat{\beta}_j = 0$ and $\beta_{0,j'} = 0$. Combining these two sets of inequalities yields the desired upper bound $\|p_\lambda(\beta_0)\|_1 - \|p_\lambda(\beta_0 + \delta)\|_1 \leqslant (\lambda_0/4)\|\delta\|_1 \leqslant (4n)^{-1}\|X\delta\|_2^2 + \lambda_0\|\delta\|_1/4$.

Case 2: $\|\hat{\beta}\|_0 = s - k$ for some $k \geqslant 1$. Then we have $\|\delta\|_0 \leqslant \|\hat{\beta}\|_0 + \|\beta_0\|_0 \leqslant s - k + s < 2s$ and $\|\delta\|_2 \geqslant k^{1/2} \min_{j=1,\ldots,s} |\beta_{0,j}|$, since there are at least k such j with $j = 1, \ldots, s$ and $\hat{\beta}_j = 0$. Thus it follows from the first part of Condition 1 and $\min_{j=1,\ldots,s} |\beta_{0,j}| > 2\kappa_0^{-1} p_\lambda^{1/2}(\infty)$ in Condition 2 that

$$(4n)^{-1}\|X\delta\|_2^2 \geqslant 4^{-1}\kappa_0^2\|\delta\|_2^2 \geqslant 4^{-1}\kappa_0^2 \left(k^{1/2} \min_{j=1,\ldots,s} |\beta_{0,j}|\right)^2 \geqslant k p_\lambda(\infty).$$

Since $p_\lambda(|\beta_{0,j}|) \leqslant p_\lambda(\infty)$ and there are $s - k$ nonzero quantities $\hat{\beta}_j$, applying the same arguments as in Case 1 gives our desired upper bound $\|p_\lambda(\beta_0)\|_1 - \|p_\lambda(\beta_0 + \delta)\|_1 \leqslant k p_\lambda(\infty) + (\lambda_0/4)\|\delta\|_1 \leqslant (4n)^{-1}\|X\delta\|_2^2 + (\lambda_0/4)\|\delta\|_1$.

Combining Cases 1 and 2 above along with (A5) and $\|\delta\|_1 = \|\tilde{\delta}_1\|_1 + \|\tilde{\delta}_2\|_1$ yields

$$n^{-1}\|X\delta\|_2^2 + \lambda_0\|\tilde{\delta}_2\|_1 \leqslant 7\lambda_0\|\tilde{\delta}_1\|_1, \tag{A6}$$

which entails a basic constraint

$$\|\tilde{\delta}_2\|_1 \leqslant 7\|\tilde{\delta}_1\|_1. \tag{A7}$$

With the aid of (A7), we will first establish a useful bound on $\|\tilde{\delta}_2\|_2$. In view of (A7), the restricted eigenvalue assumption in the second part of Condition 1 and (A6), as well as the Cauchy–Schwartz inequality, lead to

$$4^{-1}\kappa^2(s, 7)(\|\tilde{\delta}_1\|_2^2 \vee \|\tilde{\delta}_2'\|_2^2) \leqslant (4n)^{-1}\|X\delta\|_2^2 \leqslant (7/4)\lambda_0\|\tilde{\delta}_1\|_1 \leqslant (7/4)\lambda_0 s^{1/2}\|\tilde{\delta}_1\|_2. \tag{A8}$$

Solving this inequality gives

$$\|\tilde{\delta}_1\|_2 \leqslant 7\lambda_0 s^{1/2}/\kappa^2(s, 7), \quad \|\tilde{\delta}_1\|_1 \leqslant s^{1/2}\|\tilde{\delta}_1\|_2 \leqslant 7\lambda_0 s/\kappa^2(s, 7). \tag{A9}$$

Since the kth largest absolute component of $\tilde{\delta}_2$ is bounded from above by $\|\tilde{\delta}_2\|_1/k$, we have $\|\tilde{\delta}_3\|_2^2 \leqslant \sum_{k=s+1}^{p-s} \|\tilde{\delta}_2\|_1^2/k^2 \leqslant s^{-1}\|\tilde{\delta}_2\|_1^2$, where $\tilde{\delta}_3$ is a subvector of $\tilde{\delta}_2$ consisting of components excluding those with the s largest magnitudes. This inequality, (A7), and the Cauchy–Schwartz inequality imply that $\|\tilde{\delta}_3\|_2 \leqslant s^{-1/2}\|\tilde{\delta}_2\|_1 \leqslant 7s^{-1/2}\|\tilde{\delta}_1\|_1 \leqslant 7\|\tilde{\delta}_1\|_2$, and thus $\|\tilde{\delta}_2\|_2 \leqslant 7\|\tilde{\delta}_1\|_2 + \|\tilde{\delta}_2'\|_2$. By (A8), we have $\|\tilde{\delta}_2'\|_2 \leqslant 7^{1/2}\lambda_0^{1/2}s^{1/4}\|\tilde{\delta}_1\|_2^{1/2}/\kappa(s, 7)$. Combining these two inequalities with (A9) gives

$$\|\tilde{\delta}_2\|_2 \leqslant 7\|\tilde{\delta}_1\|_2 + 7^{1/2}\lambda_0^{1/2}s^{1/4}\|\tilde{\delta}_1\|_2^{1/2}/\kappa(s, 7) \leqslant 56\lambda_0 s^{1/2}/\kappa^2(s, 7). \tag{A10}$$

This bound enables us to conduct more delicate analysis on δ.

We proceed to prove the first part of Theorem 1. The inequality (8) on the prediction loss can be obtained by inserting (A9) into (A6):

$$n^{-1/2}\|X\delta\|_2 \leqslant 7\lambda_0 s^{1/2}/\kappa(s, 7). \tag{A11}$$

Combining (A9) with (A10) yields the following bound on the L_2 estimation loss,

$$\|\delta\|_2 \leqslant \|\tilde{\delta}_1\|_2 + \|\tilde{\delta}_2\|_2 \leqslant 63\lambda_0 s^{1/2}/\kappa^2(s, 7). \tag{A12}$$

For each $1 \leqslant q < 2$, an application of Hölder's inequality gives

$$\|\delta\|_q \leqslant \{s^{(2-q)/2}\|\tilde{\delta}_1\|_2^q\}^{1/q} = s^{(2-q)/(2q)}\|\delta\|_2 \leqslant 63\lambda_0 s^{1/q}/\kappa^2(s, 7). \tag{A13}$$

Now we bound the number of falsely discovered signs $\mathrm{FS}(\hat{\beta})$. If $\mathrm{sgn}(\hat{\beta}_j) \neq \mathrm{sgn}(\beta_{0,j})$, then by Proposition 1 and Condition 2, $|\delta_j| = |\hat{\beta}_j - \beta_{0,j}| \geqslant (1-c_1)\lambda$. Thus, it follows that $\|\delta\|_2 \geqslant \{\mathrm{FS}(\hat{\beta})\}^{1/2}(1-c_1)\lambda$. This, together with (A12), entails that

$$\mathrm{FS}(\hat{\beta}) \leqslant \{63/(1-c_1)\}^2(\lambda_0/\lambda)^2 s/\kappa^4(s,7). \tag{A14}$$

We finally note that all the above bounds for $\hat{\beta}$ are conditional on the event $\mathcal{E}$, and thus hold simultaneously with probability $1 - O(p^{-c_0})$, which concludes the proof of the first part of Theorem 1.

It remains to prove the second part of Theorem 1. Since $\lambda \geqslant 56(1-c_1)^{-1}\lambda_0 s^{1/2}/\kappa^2(s,7)$, we have by Condition 2 that $\min_{j=1,\ldots,s}|\beta_{0,j}| > 56\lambda_0 s^{1/2}/\kappa^2(s,7)$. This inequality, together with (A9), implies that for each $j = 1,\ldots,s$,

$$\mathrm{sgn}(\hat{\beta}_j) = \mathrm{sgn}(\beta_{0,j}), \tag{A15}$$

by a simple contradiction argument. In view of (A10) and the hard-thresholding feature of $\hat{\beta} = (\hat{\beta}_{0,1}^{\mathrm{T}}, \hat{\beta}_{0,2}^{\mathrm{T}})^{\mathrm{T}}$ with $\hat{\beta}_{0,1} = (\hat{\beta}_1,\ldots,\hat{\beta}_s)^{\mathrm{T}}$, a similar contradiction argument shows that $\hat{\beta}_{0,2} = 0$. Combining this result with (A15) leads to $\mathrm{sgn}(\hat{\beta}) = \mathrm{sgn}(\beta_0)$. With this strong result on the sign consistency of $\hat{\beta}$, we can derive tight bounds on the L_∞ loss. By Theorem 1 of Lv & Fan (2009), $\hat{\beta}_{0,1}$ solves the following equation for $\gamma \in \mathbb{R}^s$:

$$\gamma = \tilde{\beta}_{0,1} - (n^{-1}X_1^{\mathrm{T}}X_1)^{-1}b, \tag{A16}$$

where X_1 is an $n \times s$ submatrix of X corresponding to s nonzero regression coefficients $\beta_{0,j}$ and $b = \{\lambda_0 1_s + p'_\lambda(|\gamma|)\} \circ \mathrm{sgn}(\tilde{\beta}_{0,1}) - n^{-1}X_1^{\mathrm{T}}\varepsilon$, with the derivative taken componentwise and $\circ$ the Hadamard, componentwise, product. It follows from the concavity and monotonicity of $p_\lambda(t)$ and Condition 2 that for any $t > (1-c_1)\lambda$, we have $0 \leqslant p'_\lambda(t) \leqslant p'_\lambda\{(1-c_1)\lambda\} \leqslant \lambda_0/4$. In view of (A15) and the hard-thresholding feature of $\hat{\beta}$, each component of $\hat{\beta}_{0,1}$ has magnitude larger than $(1-c_1)\lambda$. Since $\|n^{-1}X_1^{\mathrm{T}}\varepsilon\|_\infty \leqslant \|n^{-1}X^{\mathrm{T}}\varepsilon\|_\infty \leqslant \lambda_0/2$ on the event $\mathcal{E}$, combining these results leads to

$$\mathrm{sgn}(b) = \mathrm{sgn}(\tilde{\beta}_{0,1}), \quad \lambda_0/2 \leqslant \|b\|_\infty \leqslant 7\lambda_0/4. \tag{A17}$$

Clearly $\tilde{\delta}_2 = \hat{\beta}_{0,2} = 0$. Thus it follows from (A16), (A17), and the first part of Condition 1 that

$$\|\delta\|_\infty \leqslant \|(n^{-1}X_1^{\mathrm{T}}X_1)^{-1}\|_\infty\|b\|_\infty \leqslant (7/4)\lambda_0\|(n^{-1}X_1^{\mathrm{T}}X_1)^{-1}\|_\infty,$$

which concludes the proof of the second part of Theorem 1.

Proof of Theorem 2

Let $\hat{\beta}$ be the global minimizer of (2) given in Theorem 1, with $\delta = \hat{\beta} - \beta_0$ denoting the estimation error. To calculate the risk of the regularized estimator $\hat{\beta}$ for different losses, we need to analyse its tail behaviour on the event $\mathcal{E}^c = \{\|n^{-1}X^{\mathrm{T}}\varepsilon\|_\infty > \lambda_0/2\}$. We work directly with inequality (A1). It follows easily from (A1) that

$$(2n)^{-1}\|X\delta - \varepsilon\|_2^2 + \lambda_0\|\delta\|_1 + \|p_\lambda(\hat{\beta})\|_1 \leqslant (2n)^{-1}\|\varepsilon\|_2^2 + 2\lambda_0\|\beta_0\|_1 + \|p_\lambda(\beta_0)\|_1. \tag{A18}$$

We need to bound the term $E\{(2n)^{-1}\|\varepsilon\|_2^2 1_{\mathcal{E}^c}\}$ from above, where $\varepsilon = (\varepsilon_1,\ldots,\varepsilon_n)^{\mathrm{T}}$. Consider the cases of bounded or unbounded error.

Case 1: Bounded error with range $[-b, b]$. Then in view of the deviation probability bound (5), we have

$$E\{(2n)^{-1}\|\varepsilon\|_2^2 1_{\mathcal{E}^c}\} \leqslant (b^2/2)\mathrm{pr}(\mathcal{E}^c) = O(p^{-c_0}). \tag{A19}$$

Case 2: Unbounded error. Then it follows from (5) that for each $i = 1,\ldots,n$ and any $\tau > 0$,

$$\begin{aligned} E\{\varepsilon_i^2 1_{\mathcal{E}^c}\} &\leqslant E\{\varepsilon_i^2 1_{\{|\varepsilon_i|\leqslant\tau\}\cap\mathcal{E}^c}\} + E\{\varepsilon_i^2 1_{\{|\varepsilon_i|>\tau\}}\} \leqslant \tau^2\mathrm{pr}(\mathcal{E}^c) + E\{\varepsilon_0^2 1_{\{|\varepsilon_0|>\tau\}}\} \\ &= O(\tau^2 p^{-c_0}) + E\{\varepsilon_0^2 1_{\{|\varepsilon_0|>\tau\}}\}. \end{aligned}$$

Thus we have

$$E\{(2n)^{-1}\|\varepsilon\|_2^2 1_{\mathcal{E}^c}\} = (2n)^{-1}\sum_{i=1}^{n} E\{\varepsilon_i^2 1_{\mathcal{E}^c}\} \leqslant 2^{-1}E\{\varepsilon_0^2 1_{\{|\varepsilon_0|>\tau\}}\} + O(\tau^2 p^{-c_0}). \tag{A20}$$

Clearly, the bound (A19) is a special case of the general bound (A20), with $\tau = b$.

We first consider the risks under the L_1 loss and prediction loss. Note that $\|p_\lambda(\beta_0)\|_1 \leqslant s p_\lambda(\infty)$. By (A18), (A20), and (5), we have

$$\begin{aligned} E\{\|\delta\|_1 1_{\mathcal{E}^c}\} &\leqslant \lambda_0^{-1}E\{(2n)^{-1}\|\varepsilon\|_2^2 1_{\mathcal{E}^c}\} + O[\{2\|\beta_0\|_1 + s\lambda_0^{-1}p_\lambda(\infty)\}p^{-c_0}] \\ &\leqslant (2\lambda_0)^{-1}E\{\varepsilon_0^2 1_{\{|\varepsilon_0|>\tau\}}\} + O(\gamma p^{-c_0}), \end{aligned} \tag{A21}$$

where $\gamma = \|\beta_0\|_1 + s\lambda_0^{-1}p_\lambda(\infty) + \tau^2\lambda_0^{-1}$. This inequality, along with (A13) on the event $\mathcal{E}$, yields that for any $\tau > 0$, $E\|\delta\|_1 \leqslant 63\lambda_0 s/\kappa^2(s,7) + (2\lambda_0)^{-1}E\{\varepsilon_0^2 1_{\{|\varepsilon_0|>\tau\}}\} + O(\gamma p^{-c_0})$. Note that $(2n)^{-1}\|X\delta - \varepsilon\|_2^2 \geqslant (4n)^{-1}\|X\delta\|_2^2 - (2n)^{-1}\|\varepsilon\|_2^2$. Thus in view of (A18), a similar argument to the one for (A21) applies to show that $E\{n^{-1}\|X\delta\|_2^2 1_{\mathcal{E}^c}\} \leqslant 4E\{\varepsilon_0^2 1_{\{|\varepsilon_0|>\tau\}}\} + O(\gamma\lambda_0 p^{-c_0})$. Combining this inequality with (A11) on the event $\mathcal{E}$ gives

$$E\{n^{-1}\|X\delta\|_2^2\} \leqslant 49\lambda_0^2 s/\kappa^2(s,7) + 4E\{\varepsilon_0^2 1_{\{|\varepsilon_0|>\tau\}}\} + O(\gamma\lambda_0 p^{-c_0}).$$

We now consider the risk under the variable selection loss. To this end, we need to bound $\|\hat\beta\|_0$ on the event $\mathcal{E}^c$. Since $\hat\beta$ always has the hard-thresholding property ensured by Proposition 1, it follows from the monotonicity of $p_\lambda(t)$ and Condition 2 that $\|p_\lambda(\hat\beta)\|_1 \geqslant \|\hat\beta\|_0 p_\lambda\{(1-c_1)\lambda\} \geqslant \|\hat\beta\|_0 p_{\mathrm{H},\lambda}\{(1-c_1)\lambda\} = \|\hat\beta\|_0 2^{-1}(1-c_1^2)\lambda^2$. This inequality, along with (A18), shows that

$$\|\hat\beta\|_0 \leqslant 2(1-c_1^2)^{-1}\lambda^{-2}\{(2n)^{-1}\|\varepsilon\|_2^2 + 2\lambda_0\|\beta_0\|_1 + \|p_\lambda(\beta_0)\|_1\}. \tag{A22}$$

Clearly, $\mathrm{FS}(\hat\beta) \leqslant \|\hat\beta\|_0 + s$. Thus by (A22), applying a similar argument to the one for (A21) gives

$$E\{\mathrm{FS}(\hat\beta)1_{\mathcal{E}^c}\} \leqslant (1-c_1^2)^{-1}\lambda^{-2}E\{\varepsilon_0^2 1_{\{|\varepsilon_0|>\tau\}}\} + O\{(\gamma\lambda_0/\lambda^2 + s)p^{-c_0}\}.$$

It follows from this bound and inequality (A14) on the event $\mathcal{E}$ that

$$E\{\mathrm{FS}(\hat\beta)\} \leqslant 63^2(1-c_1)^{-2}(\lambda_0/\lambda)^2 s/\kappa^4(s,7) + (1-c_1^2)^{-1}\lambda^{-2}E\{\varepsilon_0^2 1_{\{|\varepsilon_0|>\tau\}}\} + O\{(\gamma\lambda_0/\lambda^2 + s)p^{-c_0}\}.$$

We finally consider the risks under the L_q loss with $q \in (1,2]$. By (A18) and the norm inequality $\|\delta\|_2 \leqslant \|\delta\|_1$, we have

$$\begin{aligned} \|\delta\|_2^2 &\leqslant \lambda_0^{-2}\{(2n)^{-1}\|\varepsilon\|_2^2 + 2\lambda_0\|\beta_0\|_1 + \|p_\lambda(\beta_0)\|_1\}^2 \leqslant 3\lambda_0^{-2}\{(2n)^{-2}\|\varepsilon\|_2^4 + 4\lambda_0^2\|\beta_0\|_1^2 + \|p_\lambda(\beta_0)\|_1^2\} \\ &\leqslant 3\lambda_0^{-2}\left\{(4n)^{-1}\sum_{i=1}^{n}\varepsilon_i^4 + 4\lambda_0^2\|\beta_0\|_1^2 + s^2 p_\lambda^2(\infty)\right\}. \end{aligned}$$

With this inequality and (5), a similar argument to the one for (A20) applies to show that for any $\tau > 0$,

$$\begin{aligned} E\{\|\delta\|_2^2 1_{\mathcal{E}^c}\} &\leqslant 3\lambda_0^{-2}\left[(4n)^{-1}\sum_{i=1}^{n}E(\varepsilon_i^4 1_{\mathcal{E}^c}) + \{4\lambda_0^2\|\beta_0\|_1^2 + s^2 p_\lambda^2(\infty)\}\mathrm{pr}(\mathcal{E}^c)\right] \\ &\leqslant (3/4)\lambda_0^{-2}E(\varepsilon_0^4 1_{\{|\varepsilon_0|>\tau\}}) + O(\gamma^2 p^{-c_0}). \end{aligned} \tag{A23}$$

Combining (A23) with (A12) on the event $\mathcal{E}$ yields $E\|\delta\|_2^2 \leqslant 63^2\lambda_0^2 s/\kappa^4(s,7) + (3/4)\lambda_0^{-2}E\{\varepsilon_0^4 1_{\{|\varepsilon_0|>\tau\}}\} + O(\gamma^2 p^{-c_0})$. For the L_q loss with $q \in (1,2)$, an application of Hölder's inequality and Young's inequality

with (A21) and (A23) gives

$$E\{\|\delta\|_q^q 1_{\mathcal{E}^c}\} = E\left(\sum_{j=1}^{p} |\delta_j|^{2-q}|\delta_j|^{2q-2}1_{\mathcal{E}^c}\right) \leqslant \{E(\|\delta\|_1 1_{\mathcal{E}^c})\}^{2-q}\{E(\|\delta\|_2^2 1_{\mathcal{E}^c})\}^{q-1}$$
$$\leqslant (2-q)E\{\|\delta\|_1 1_{\mathcal{E}^c}\} + (q-1)E\{\|\delta\|_2^2 1_{\mathcal{E}^c}\}$$
$$\leqslant (2-q)(2\lambda_0)^{-1}E\{\varepsilon_0^2 1_{\{|\varepsilon_0|>\tau\}}\}$$
$$+ (q-1)(3/4)\lambda_0^{-2}E\{\varepsilon_0^4 1_{\{|\varepsilon_0|>\tau\}}\} + O[\{(2-q)\gamma + (q-1)\gamma^2\}p^{-c_0}],$$

where $\delta = (\delta_1, \ldots, \delta_p)^{\mathrm{T}}$. It follows from this inequality and (A13) on the event $\mathcal{E}$ that

$$E(\|\delta\|_q^q) \leqslant 63^q\lambda_0^q s\kappa^{-2q}(s,7) + (2-q)(2\lambda_0)^{-1}E\{\varepsilon_0^2 1_{\{|\varepsilon_0|>\tau\}}\} + (q-1)(3/4)\lambda_0^{-2}E\{\varepsilon_0^4 1_{\{|\varepsilon_0|>\tau\}}\}$$
$$+ O[\{(2-q)\gamma + (q-1)\gamma^2\}p^{-c_0}],$$

which completes the proof of the first part of Theorem 2.

The second part of Theorem 2 can be proved by noting that $\mathrm{sgn}(\hat{\beta}) = \mathrm{sgn}(\beta_0)$ under the additional condition and using similar arguments to the above.

REFERENCES

BICKEL, P. J., RITOV, Y. & TSYBAKOV, A. B. (2009). Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.* **37**, 1705–32.

CANDÈS, E. & TAO, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n (with Discussion). *Ann. Statist.* **35**, 2313–404.

CHEN, S. S., DONOHO, D. L. & SAUNDERS, M. A. (1999). Atomic decomposition by basis pursuit. *SIAM J. Sci. Comp.* **20**, 33–61.

EFRON, B. (1979). Bootstrap methods: Another look at the jackknife. *Ann. Statist.* **7**, 1–26.

EFRON, B., HASTIE, T. J., JOHNSTONE, I. M. & TIBSHIRANI, R. J. (2004). Least angle regression (with Discussion). *Ann. Statist.* **32**, 407–99.

FAN, J. & FAN, Y. (2008). High-dimensional classification using features annealed independence rules. *Ann. Statist.* **36**, 2605–37.

FAN, J. & LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Statist. Assoc.* **96**, 1348–60.

FAN, J. & LV, J. (2010). A selective overview of variable selection in high dimensional feature space (invited review article). *Statist. Sinica* **20**, 101–48.

FAN, J. & LV, J. (2011). Nonconcave penalized likelihood with NP-dimensionality. *IEEE Trans. Info. Theory* **57**, 5467–84.

GORDON, G. J., JENSEN, R. V., HSIAO, L. L., GULLANS, S. R., BLUMENSTOCK, J. E., RAMASWAMY, S., RICHARDS, W. G., SUGARBAKER, D. J. & BUENO, R. (2002). Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma. *Cancer Res.* **62**, 4963–7.

LIN, W & LV, J. (2013). High-dimensional sparse additive hazards regression. *J. Am. Statist. Assoc.* **108**, 247–64.

LIU, Y. & WU, Y. (2007). Variable selection via a combination of the L_0 and L_1 penalties. *J. Comp. Graph. Statist.* **16**, 782–98.

LV, J. & FAN, Y. (2009). A unified approach to model selection and sparse recovery using regularized least squares. *Ann. Statist.* **37**, 3498–528.

MAZUMDER, R., FRIEDMAN, J. H. & HASTIE, T. J. (2011). SparseNet: Coordinate descent with nonconvex penalties. *J. Am. Statist. Assoc.* **106**, 1125–38.

ROSSET, S. & ZHU, J. (2007). Piecewise linear regularized solution paths. *Ann. Statist.* **35**, 1012–30.

TIBSHIRANI, R. J. (1996). Regression shrinkage and selection via the lasso. *J. R. Statist. Soc.* B **58**, 267–88.

ZHANG, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.* **38**, 894–942.

ZHANG, C.-H. & ZHANG, T. (2012). A general theory of concave regularization for high-dimensional sparse estimation problems. *Statist. Sci.* **27**, 576–93.

ZOU, H. (2006). The adaptive lasso and its oracle properties. *J. Am. Statist. Assoc.* **101**, 1418–29.

ZOU, H. & ZHANG, H. H. (2009). On the adaptive elastic-net with a diverging number of parameters. *Ann. Statist.* **37**, 1733–51.

[*Received June* 2012. *Revised August* 2013]