SCIENCE CHINA
Information Sciences

• **RESEARCH PAPERS** •

June 2010 Vol. 53 No. 6: 1159–1169
doi: 10.1007/s11432-010-0090-0

$L_{1/2}$ regularization

XU ZongBen[1], ZHANG Hai[1,2*], WANG Yao[1], CHANG XiangYu[1] & LIANG Yong[3]

[1] *Institute of Information and System Science, Xi'an Jiaotong University, Xi'an* 710049, *China;*
[2] *Department of Mathematics, Northwest University, Xi'an* 710069, *China;*
[3] *University of Science and Technology, Macau* 999078, *China*

Received December 22, 2008; accepted February 26, 2009; published online May 8, 2010

Abstract In this paper we propose an $L_{1/2}$ regularizer which has a nonconvex penalty. The $L_{1/2}$ regularizer is shown to have many promising properties such as unbiasedness, sparsity and oracle properties. A reweighed iterative algorithm is proposed so that the solution of the $L_{1/2}$ regularizer can be solved through transforming it into the solution of a series of L_1 regularizers. The solution of the $L_{1/2}$ regularizer is more sparse than that of the L_1 regularizer, while solving the $L_{1/2}$ regularizer is much simpler than solving the L_0 regularizer. The experiments show that the $L_{1/2}$ regularizer is very useful and efficient, and can be taken as a representative of the $L_p (0 < p < 1)$ regularizer.

Keywords machine learning, variable selection, regularizer, compressed sensing

Citation Xu Z B, Zhang H, Wang Y, et al. $L_{1/2}$ regularization. Sci China Inf Sci, 2010, 53: 1159–1169, doi: 10.1007/s11432-010-0090-0

1 Introduction

It is well known that variable selection and feature extraction are basic problems in high-dimensional and massive data analysis. The traditional variable selection criteria such as AIC, BIC and Cp [1–3] involve solving an NP hard optimization problem so they are infeasible for high dimensional data. Consequently, innovative variable selection procedure is expected to cope with very high dimensionality, which is one of the hot topics in machine learning. The regularization methods are recently used as feasible approaches to solve the problem. In general, the regularization methods have the form

$$\min \left\{ \frac{1}{n} \sum_{i=1}^{n} l(y_i, f(x_i)) + \lambda \|f\|_k \right\}, \tag{1}$$

where $l(.,.)$ is a loss function, $(x_i, y_i)_{i=1}^{n}$ is a data set, and λ is the regularization parameter. When f is in the linear form and the loss function is square loss, $\|f\|_k$ is normally taken as the norm of the coefficient of linear model. Almost all the existing learning algorithms can be considered as a special form of this regularization framework. For example, when $k = 0$, it is AIC or BIC, which is referred to as the L_0 regularizer in this paper. When $k = 1$, it is the Lasso, which is called the L_1 regularizer in this paper. When $k = 2$, it is the ridge regression, which is called the L_2 regularizer. And when $k = \infty$, it is the L_∞ regularizer.

*Corresponding author (email: zhanghai@nwu.edu.cn)

The L_0 regularizer is the earliest regularization method applied to variable selection and feature extraction. Constrained by the number of coefficients, the L_0 regularizer yields the most sparse solutions, but it faces the problem of combinatory optimization. The L_1 regularizer (Lasso) proposed by Tibshirani [4] provides an alterative for variable selection and feature extraction, which just needs to solve a quadratic programming problem but is less sparse than the L_0 regularization. At the same time, Donoho [5–7] proposed Basis Pursuit when studying the signal sparsity recovery problem. They proved that under some conditions the solutions of the L_0 regularizer are equivalent to those of the L_1 regularizer for the sparsity problem, so the hard NP optimization problem can be avoided in the L_1 regularizer. Based on the work of the above mentioned scholars, the L_1 regularizer and the L_1 type regularizers, including SCAD [8], Adaptive Lasso [9], Elastic net [10], Stagewise Lasso [11] and Dantzig selector [12], have become the dominantly used tools for data analysis since then.

When the model has redundant irrelevant predictors, the variable selection and feature extraction are to distinguish them from the true predictors with any amount of data and any amount of regularization. A model is a sparse model if it has abundant irrelevant predictors. Clearly, the L_0 regularizer is ideal for variable selection in the sense of yielding the most sparse variables, but it is a combinatory optimization problem which is difficult to be solved. While the L_1 regularizer leads to a convex optimization problem easy to be solved, but it does not yield sufficiently sparse solution. The solutions of the L_2 regularizer have the properties of being smooth, but they do not possess the sparse property. The solutions of L_∞ regularizer do not have sparse property either. To our knowledge, the properties of the L_∞ regularizer are still unclear.

In recent years, there has been an explosion of researches on the properties of the L_1 regularizer. However, for many practical applications, the solutions of the L_1 regularizer are often less sparse than those of the L_0 regularizer. To find more sparse solutions than L_1 regularizer is, however, imperative and required for many variable selection applications. Also, the L_1 regularizer is inefficient when the errors in data have heavy tail distribution [4]. A question then arises: whether we can find a new regularizer which is more sparse than the L_1 regularizer while it is still easier to be solved than the L_0 regularizer? A natural choice is to try the $L_p(0 < p < 1)$ regularizer. But in so doing we have to answer the subsequent two questions: (i) Which p is the best and should be chosen? (ii) Is there an efficient algorithm for solving the nonconvex optimization problem deduced from the $L_p(0 < p < 1)$ regularizer?

In this paper, our aim is to provide a satisfactory answer to the above questions. We propose the $L_{1/2}(0 < p < 1)$ regularizer and show that the $L_{1/2}$ regularizer can be taken as a representative of the $L_p(0 < p < 1)$ regularizers for the sparsity problem. Therefore what we need to do is to focus on the situation when $p = 1/2$ for the $L_p(0 < p < 1)$ regularizers. A reweighted iteration algorithm is proposed so that the $L_{1/2}$ regularizer can be efficiently solved through transforming it into a series of weighted L_1 regularizer problems. We also present three application examples, a variable selection example, a prostate cancer example and a compressive sensing example, to demonstrate the effectiveness and powerfulness of the $L_{1/2}$ regularizer. The variable selection example shows that the $L_{1/2}$ regularizer is more efficient and robust than Lasso when the errors have heavy tail distribution. The prostate cancer application shows that the solutions of the $L_{1/2}$ regularizer are not only more sparse than those of Lasso but they bring about also the lower prediction error. In the compressive sensing example, it is shown that the $L_{1/2}$ regularizer can significantly reduce the necessary sampling number for sparse signal exact recovery and substantially require less measurements. Our research reveals that when $1/2 \leqslant p < 1$, the $L_{1/2}$ regularizer is the most sparse and robust among the L_p regularizers, and when $0 < p < 1/2$, the L_p regularizers have similar properties to the $L_{1/2}$ regularizer. So we conclude that the $L_{1/2}$ regularizer can be taken as the representative of the $L_p(0 < p < 1)$ regularizers.

2 Regularization framework and $L_{1/2}$ regularizer

To make things more clear, some notations used throughout the paper are introduced first. Then we introduce the framework of regularization and discuss the differences among the existing regularizers. Finally, we propose the $L_{1/2}$ regularizer and present its theoretical properties.

Let X be a set, and Y a subset of a Euclid space. $Z = X \times Y$. Y is identified as the input space and Y as the output space. Denote by $F(X, Y)$ an unknown probability distribution on $Z = X \times Y$. The set

$$S = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\} \in Z^n$$

of size n in Z drawn i.i.d. from P is called a training set. It is supposed that there exists a definite but unknown function $f^*(x) : X \to Y$ (the true model). The goal is to select the real variables to predict the sample data based on the training set by minimizing the expected loss (risk)

$$\min_{\beta} L(\beta) = E_{x,y} l(y, f(x, \beta)). \tag{2}$$

Unfortunately, the distribution F is unknown and this quantity cannot be computed. So a common practice is to substitute $L(\beta)$ by an empirical loss $L_n(\beta)$ and solve the problem via

$$\min_{\beta} L_n(\beta) = \frac{1}{n} \sum_{i=1}^{n} l(y_i, f(x_i, \beta)). \tag{3}$$

In general, (3) is known as an ill-posed problem. A direct computation based on this scheme very often leads to overfitting; that is, the empirical error is minimized, but the performance of prediction for new sample is poor. A common remedy for this is to replace it with

$$\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} l(y_i, f(x_i, \beta)) \quad \text{s.t.} \quad p(\beta) \leqslant t, \tag{4}$$

where $p(\beta)$ is a nonnegative function, reflecting the expection of the solutions to be found. Different p and t here are in correspondence with different constraints to the model, so different solutions will be obtained respectively. The constraint is the strongest when $t = 0$ and becomes weaker as t becomes larger. Denote $\hat{\beta} = \beta(t)$. Generally, the same procedure can be obtained through the penalized form of (4)

$$\min_{\beta} \left\{ \frac{1}{n} \sum_{i=1}^{n} l(y_i, f(x_i, \beta)) + \lambda P(\beta) \right\}, \tag{5}$$

where λ is a tuning parameter controlling the complexity of the model. (5) is the general framework of regularization. Note that setting $\lambda = \infty$ in (5) results in the totally constrained solution ($t = 0$) whereas $\lambda = 0$ yields the unconstrained solution ($t = \infty$). Denote $\hat{\beta} = \beta(\lambda)$. Obviously, (5) is determined by two elements—the loss function (data term) and the penalty form (penalty term). The different loss functions and different penalties will result in different regularization algorithms. For example, let the loss function be square loss and the penalty be L_1 norm of coefficients. It is the Lasso. When the loss function is hinge loss and the penalty is the L_2 norm of coefficients, it is the SVM. In this paper, we study the case in which the loss function is square loss but the penalty is a nonconvex function.

Consider the sparse linear model,

$$Y = X^{\mathrm{T}} \beta + \epsilon, \quad E\epsilon = 0, \quad \mathrm{Cov}(\epsilon) = \sigma^2 I, \tag{6}$$

where $Y = (Y_1, \ldots, Y_n)^{\mathrm{T}}$ is an $n \times 1$ response vector, $X = (X_1, X_2, \ldots, X_n)$ ($X_i^{\mathrm{T}} = (x_{i1}, \ldots, x_{ip})$, $i = 1, \ldots, n$) and $\beta = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$ is a vector of $p \times 1$ unknown parameters. ϵ is random error and σ^2 is a positive constant. We suppose that the true model is $f^*(x, \beta) = \beta_1^* x_1 + \beta_2^* x_2 + \cdots + \beta_{p_0}^* x_{p_0}$, where $p_0 \leqslant p$. Let $A = \{j : \beta_j^* \neq 0\}$. Then the true model depends only on a subset of the predictors. That is to say, Y is relevant to p_0 predictors, while the others are irrelevant predictors. Without loss generality, we assume that the data are normalized. Just as stated above, the L_0 regularizer defined by

$$\hat{\beta}_{L_0} = \arg\min_{\beta} \left\{ \frac{1}{n} \sum_{i=1}^{n} (Y_i - X_i^{\mathrm{T}} \beta)^2 + \lambda \sum_{i=1}^{p} I_{\beta_i \neq 0} \right\} \tag{7}$$

is an ideal method for variable selection. But unfortunately it is NP hard to solve, which is infeasible for high dimensional and huge data. Recently, researchers have shifted their interests to the L_1 regularizer

$$\hat{\beta}_{L1} = \arg\min_{\beta} \left\{ \frac{1}{n} \sum_{i=1}^{n} (Y_i - X_i^{\mathrm{T}}\beta)^2 + \lambda \sum_{i=1}^{p} |\beta_i| \right\}, \tag{8}$$

which is referred to as Lasso or Basis Pursuit. A series of L_1 type regularizers [8–12] have been proposed. A natural question is how to select a proper regularizer among all those regularizers. Some criteria have been suggested by Fan [8]. He proposed that a good regularizer should possess the sparsity property; that is, the resulting estimator should automatically set the irrelevant variables to zero; a good regularizer should be unbiased, i.e., the resulting estimator should have low bias; a good regularizer should also be continuous: the resulting estimator is continuous such that instability in model selection is reduced. And, he believed that a good regularizer should have the Oracle property: we can identify the right subset model exactly whenever we have it. Let $\hat{\beta}$ denote the estimation of the parameters. Then a regularizer has the Oracle property if and only if the following hold:

$$\begin{aligned} &(\mathrm{A1}) : \{j : \hat{\beta}_j \neq 0\} = A, \\ &(\mathrm{A2}) : \sqrt{n}(\hat{\beta} - \beta^*) \to_d N(0, \Sigma^*), \end{aligned}$$

where Σ^* is the covariance matrix of the known true subset model. All of those criteria have become the rule to determine a good regularizer.

In this paper, we propose the following $L_{1/2}$ regularizer:

$$\hat{\beta}_{L_{\frac{1}{2}}} = \arg\min_{\beta} \left\{ \frac{1}{n} \sum_{i=1}^{n} (Y_i - X_i^{\mathrm{T}}\beta)^2 + \lambda \sum_{i=1}^{p} |\beta_i|^{\frac{1}{2}} \right\}, \tag{9}$$

where λ is the tuning parameter. Different from the L_1 regularizer, the penalty in the $L_{1/2}$ regularizer is nonconvex. To show the value of the $L_{1/2}$ regularizer, we explain the relation between the $L_{1/2}$ regularizer and the exiting regularizers below. Under the transformation $p \to \frac{1}{p}$, the L_0 regularizer corresponds to the L_∞ regularizer, both of which have some extreme properties. The L_1 regularizer is at the center with the properties of sparsity and continuity. The $L_{1/2}$ regularizer clearly corresponds to the L_2 regularizer which yields the smooth solutions. So the $L_{1/2}$ regularizer most probably has special properties, which inspire us to explore it further.

We further show the sparsity property of the $L_{1/2}$ regularizer from the aspect of geometry. Figure 1 shows the graphics of the penalty of the $L_{1/2}, L_1, L_2$ and L_∞ regularizers. As shown in Figure 1, the constraint region of the L_1 regularizer is a rotated square. The Lasso solution is the first place at which the contours touch the square, and this will concur at a corner corresponding to a zero coefficient. The graphs for L_2 and L_∞ are shown in Figure 1 too. There are no corners for the contours to hit and hence zero solutions will rarely appear. It is obvious that the solution of the $L_{1/2}$ regularizer occurs at a corner with a higher possibility, which hints that it is more sparse than the L_1 regularizer.

The following theorem shows the theoretical properties of the $L_{1/2}$ regularizer.

Theorem 1. The $L_{1/2}$ regularizer possesses sparsity, unbiasedness and Oracle properties.

Proof. Fan [8] has already proved the properties of sparsity and unbiasedness of the $L_{1/2}$ regularizer. In [13], Knight studied the asymptotic normal property of the L_1 and the L_1 type regularizers, and he in essence has shown that the $L_p(0 < p < 1)$ regularizer has the Oracle property. So Theorem 1 follows.

For the $L_p(p > 1)$ regularizers, researcher have mainly focused on the L_2 (ridge regression or SVM) regularizer. Similarly, we will show that for the $L_p(0 < p < 1)$ regularizers, only the $L_{1/2}$ regularizer should be worth considering. In the next section, we will present an algorithm for solving the $L_{1/2}$ regularizer.

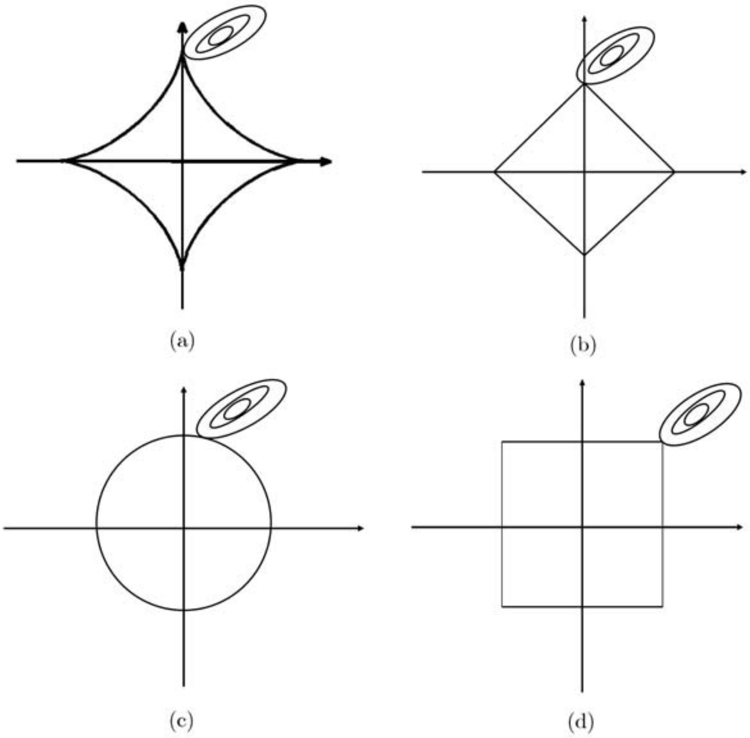

Figure 1 Estimation pictures for (a) $L_{1/2}$, (b) L_1, (c) L_2 and (d) L_∞ regularizers.

3 An algorithm for $L_{1/2}$ regularizer

In this section, we present an iteration algorithm to solve the $L_{1/2}$ regularizer. We show that the solution of the $L_{1/2}$ regularizer can be transformed into that of a series of convex weighted Lasso, to which the existing Lasso algorithms can be efficiently applied.

We first present the algorithm, and then analyze its convergence.

Algorithm.

Step 1. Set the initial value β^0 and the maximum iteration step K. Let $t = 0$.

Step 2. Solve

$$\beta^{t+1} = \arg\min \left\{ \frac{1}{n} \sum_{i=1}^{n} (Y_i - X_i^{\mathrm{T}}\beta)^2 + \lambda \sum_{i=1}^{p} \frac{1}{\sqrt{|\beta_i^t|}} |\beta_i| \right\},$$

with an existing L_1 regularizer algorithm, and let $t := t + 1$.

Step 3. If $t < K$, go to Step 2, otherwise, output β^t.

In the above algorithm, we have used K, the maximally allowable iteration step, as the termination criterion. The initial value β^0 normally can be taken as $\beta^0 = (1, 1, \ldots, 1)$ though not necessary it should be. However, with such a setting, the first iteration ($t = 0$) in Step 2 is exactly corresponding to solving an L_1 regularizer problem (thus, leading to a Lasso solution). When $t = 1$, Step 2 is to solve a reweighted L_1 regularizer problem, which can be transformed into an L_1 regularizer via linear transformation. It is possible that when $t \geqslant 1$, some β_i are zero. So to guarantee the feasibly, we replace $\frac{1}{\sqrt{|\beta_i^t|}}$ with $\frac{1}{\sqrt{|\beta_i^t|+\epsilon}}$ in Step 2 when implementing, where ϵ is any fixed positive real number.

Remark 1. The existing algorithms to solve the L_1 regularizer include the gradient boosting [14], quadratic programming [4], lars [15], piecewise linear [16] and interior point methods [17].

Below, we analyze the convergence of the algorithm. Let $R_n(\beta) = L_n(\beta) + \lambda \sum_{i=1}^{p} \frac{1}{\sqrt{|\beta_i^t|}} |\beta|$ and

$R_n^*(\beta) = L_n(\beta) + \lambda\sum_{i=1}^p \sqrt{|\beta_i|}$. We rewrite (9) as

$$\hat{\beta}_{1/2} = \arg\min_{\beta^+,\ \beta^-}\left\{\frac{1}{n}\sum_{i=1}^n (Y_i - X_i^{\mathrm{T}}(\beta^+ - \beta^-))^2 + \lambda\sum_{i=1}^p \sqrt{\beta_i^+ + \beta_i^-}\right\},$$

where $\beta^+ = \max(\beta, 0)$ and $\beta^- = -\min(\beta, 0)$ are respectively the positive part and the negative part of β. Obviously, $\beta = \beta^+ - \beta^-$ and $|\beta| = \beta^+ + \beta^-$. So we get

$$R_n(\beta) = \frac{1}{n}\sum_{i=1}^n (y_i - X_i^{\mathrm{T}}(\beta^+ - \beta^-))^2 + \lambda\sum_{i=1}^p \frac{1}{\sqrt{\beta_i^{+t} + \beta_i^{-t}}}\{\beta_i^+ + \beta_i^-\}.$$

Let $\beta^t = (\beta^+, \beta^-)^t$. Then Step 2 of the algorithm equates to minimizing $R_n(\beta)$; thus β^{t+1} satisfies

$$\overrightarrow{\bigtriangledown} L_n(\beta^{t+1}) = -\lambda\overrightarrow{\bigtriangledown} P(\beta^t). \tag{10}$$

We have the following theorem.

Theorem 2. β^t converges to the stationary point set of $R_n^*(\beta)$ as $t \to \infty$.

Proof. The convexity of $L_n(\beta)$ implies

$$L_n(\beta^t) \geqslant L_n(\beta^{t+1}) + (\beta^t - \beta^{t+1})\overrightarrow{\bigtriangledown} L_n(\beta^{t+1}), \tag{11}$$

for all β^t, β^{t+1}. While $\lambda\sum_{i=1}^p \sqrt{\beta^+ + \beta^-}$ is a concave function about β^+, β^-, we have for all β^t, β^{t+1},

$$\lambda P(\beta^{t+1}) \leqslant \lambda P(\beta^t) + (\beta^{t+1} - \beta^t)\lambda\overrightarrow{\bigtriangledown} P(\beta^t). \tag{12}$$

Combining (11) with (12) and using (10), we obtain

$$R_n^*(\beta^{t+1}) \leqslant R_n^*(\beta^t).$$

So $R_n^*(\beta)$ is a bounded and monotonically decreasing function. By the well-known Lasalle invariance principle, β^t converges to the set of stationary points of $R_n^*(\beta)$ as $t \to \infty$.

Theorem 2 shows that the $L_{1/2}$ regularizer will always approach to the set of local minimizers of $R_n^*(\beta)$, and $R_n^*(\beta^t)$ will converge to one of its local minima. Since, for the first iteration, the algorithm degenerates to solving an L_1 regularizer problem, β^1 is just the solution of Lasso. Consequently, the solution yielded by the $L_{1/2}$ regularizer algorithm must be more optimal than those of Lasso. Note that the nonconvex optimization has been always a hot topic [18, 19]. The algorithm proposed in this paper is inspired by their work. For example, Candes [20] recently proposed a regularization iteration methods $\beta^{t+1} = \arg\min\{\frac{1}{n}\sum_{i=1}^n (Y_i - X_i^{\mathrm{T}}\beta)^2 + \lambda\sum_{i=1}^p \frac{1}{|\beta_i^t|+\epsilon}|\beta_i|\}$. The efficiency of this algorithm is shown by experiments. It is easy to see that their work is just to iteratively solve $\min\{\frac{1}{n}\sum_{i=1}^n (Y_i - X_i^{\mathrm{T}}\beta)^2 + \lambda\sum_{i=1}^p \{\log(|\beta_i|+\epsilon)\}\}$, the theoretical properties of which can then be analyzed in the framework of this section.

4 Experiments

In this section, we apply the $L_{1/2}$ regularizer to three application examples: a variable selection example, a prostate cancer example and a compressive sensing example.

Example 1 (Variable selection). We consider the following linear model used in Tibshirani [4] when studying the sparsity of Lasso:

$$Y = X^{\mathrm{T}}\beta + \sigma\varepsilon, \tag{13}$$

where $\beta = (3, 1.5, 0, 0, 2, 0, 0, 0), X^{\mathrm{T}} = (X_1, \ldots, X_8)$ and ε is random error. We assume that ϵ obeys the mixture of normal distribution and Cauchy distribution. We assume also that each x_i obeys normal distribution and the correlation between x_i and x_j satisfies $\rho^{|i-j|}$ with $\rho = 0.5$. We have simulated 100

Table 1 Results of Lasso and $L_{1/2}$ regularizer

Method	CAN of zero	ICAN of zero
Lasso	4.01	0.07
$L_{1/2}$	4.6	0.23

datasets consisting of 100 observations with the ε being drawn from the standard normal distribution plus 30% outliers from the standard Cauchy distribution. Then each data set was divided into two parts: a training set with 60 observations and a test set with 40 observations. We applied the Lasso and the $L_{1/2}$ regularizer algorithm to the 100 datasets with the tuning parameter being selected by minimizing mean square error (MSE) on the test data. We have used the gradient boosting to solve the L_1 regularizer. The average number of correctly identified zero coefficients (CAN of zero in brief) over the 100 tests, and the average number of incorrectly identified zero coefficients (ICAN of zero in brief, that is, the number of the coefficients whose value is zero in the resultant model but nonzero in the true model) are recorded. These are shown in Table 1.

From Table 1, we can see that CAN of zero is 4.6 when the $L_{1/2}$ regularizer is applied, while it is 4.01 when Lasso is applied. This shows that the $L_{1/2}$ regularizer is more efficient and robust than Lasso when the errors have heavy tail distribution.

Example 2 (Prostate cancer). The data set in this example is derived from a study of prostate cancer by Blake et al. [21]. The dataset consists of the medical records of 97 patients who were about to receive a radical prostatectomy. The predictors are eight clinical measures: log (cancer volume) (lcavol), log (prostate weight) (lweight), age, the logarithm of the amount of benign prostatic hyperplasia (lbph), seminal vesicle invasion (svi), log (capsular penetration) (lcp), Gleason score (gleason) and percentage Gleason score4 or 5 (pgg45). The response is the logarithm of prostate-specific antigen (lpsa). One of the main aims here is to identify which predictors are more important in predicting the response. The prostate cancer data were divided into two parts: a training set with 67 observations and a test set with 30 observations. The tuning parameter is selected again by minimizing the mean square error on the test data. In simulation the gradient boosting algorithm was used to solve the L_1 regularizer. The simulation results are shown in Figure 2. From Figure 2, we can see that the Lasso has seleted lcavol, svi, lweight, lbph and pgg45 as the variables in the final model, whereas the $L_{1/2}$ regularizer selects lcavol, svi, lweight and lbph. The prediction error of Lasso is 0.478 while that of the $L_{1/2}$ regularizer is 0.468. Comparing the results, we can conclude that the solutions of the $L_{1/2}$ regularizer are not only more sparse than those of Lasso but they have also lower prediction error.

Example 3 (Compressive sensing). The compressive sensing has been one of the hot topics of research in recent years [22–24]. Different from the traditional Shannon/Nyquist theory, the compressive sensing is a novel sampling paradigm that goes against the common wisdom in data acquisition. It brings the reality of recovering certain signals from far fewer samples or measurements than Shannon sampling method.

Consider a real-valued, finite-length signal x, viewed as an $N \times 1$ vector in R^N. x then can be represented in an orthonormal basis $\{\psi_i\}_{i=1}^N$ of R^N. Let $\Psi = [\psi_1, \ldots, \psi_N]$. Then x can be expressed as

$$x = \tilde{\Psi} s = \sum_{i=1}^{N} s_i \psi_i, \tag{14}$$

where s is the column vector of coefficients. If for the chosen basis, x is sparse, then many coefficients s_i are equal to zero. Let us suppose that the signal x is K-sparse, namely it is a linear combination of only K basis vectors; that is, only K of the s_i coefficients are nonzero and the others are zero.

The traditional signal reconstruction methods first take N measurements (samplings) of x, obtain a complete set of coefficients s_i (via $s_i = x^{\mathrm{T}} \phi_i$), and then, select the largest K nonzero coefficients s_j^* and get the reconstructed signal $x^* = \sum_{i=1}^{s} s_j^* \phi_i$. Thus, to reconstruct an N length signal, N samplings are needed. The compressive sensing addresses the problem in a different way: it directly takes the compressed measurements of the signal without going through the intermediate step of acquiring N samples. Given an

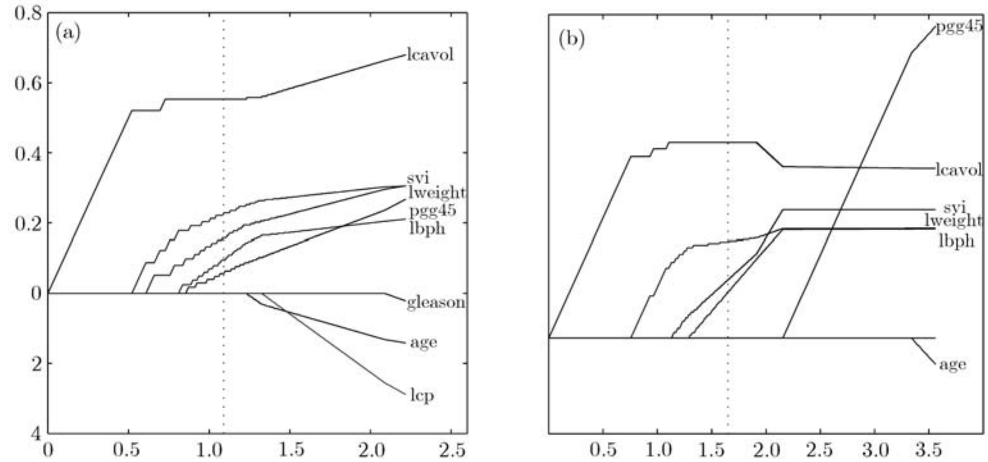

Figure 2 Comparison of variable selection results when Lasso (a) and $L_{1/2}$ (b) regularizer are applied to Example 2.

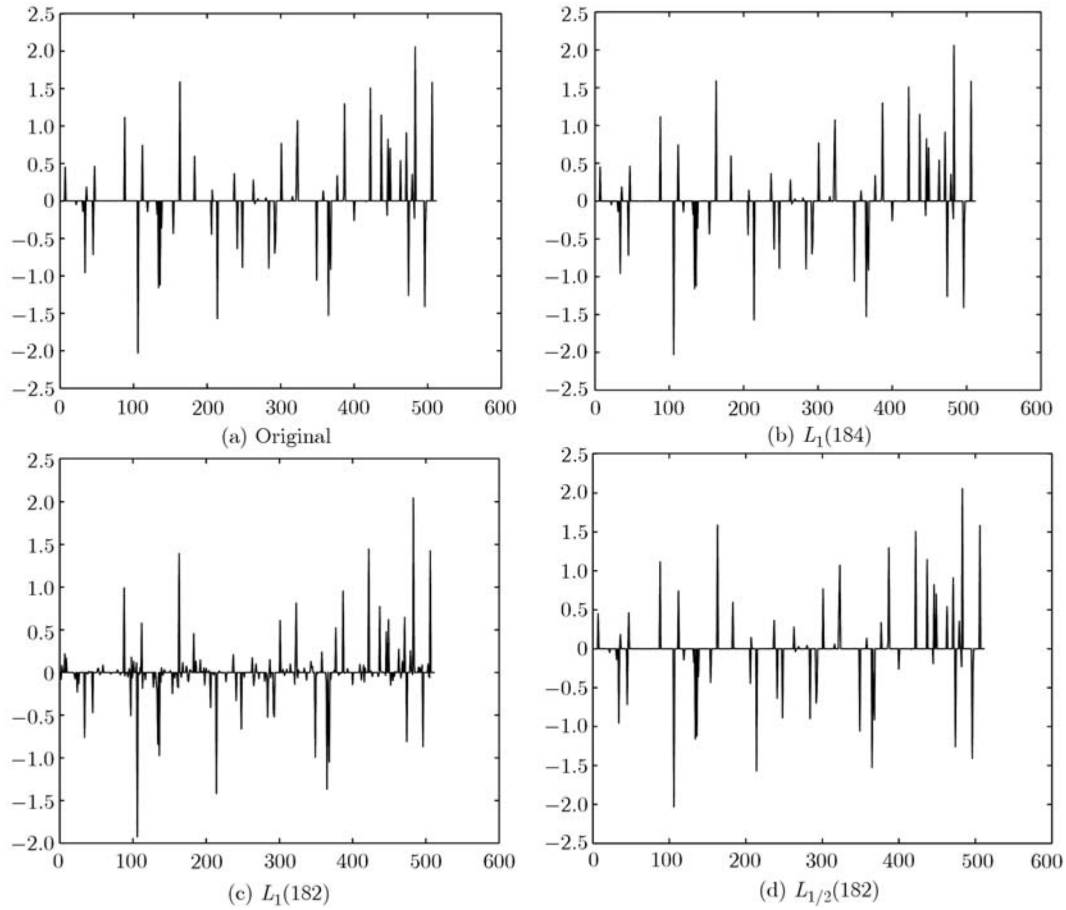

Figure 3 Sparse signal recovery by Lasso and $L_{1/2}$ regularizer.

$M \times N$ matrix $\Phi = [\phi_1, \ldots, \phi_M]$ (called a sensing matrix, understood as the composition of a compression matrix and an orthonormal basis matrix), we get M measurements y_i $(i = 1, \ldots, M)$ via the inner products $y_i = \langle x, \phi_i \rangle (K \leqslant M \leqslant N)$, and then we reconstruct x from the M measurements. For simplicity, we will consider only the problem $y = \Phi x$ where x is sparse. Since ϕ is a known basis, knowledge of x equivalent to knowledge of s. It is shown in [22–24] that the reconstruction of x can be modeled as finding the minimizer of the following L_0 problem:

$$\min_{x\in R^N}\sum_{i=1}^{N} I_{x_i\neq 0} \quad \text{s.t.} \quad y=\Phi x.$$

According to Donoho [23], the above L_0 problem can be replaced by the following simpler problem:

$$\min_{x\in R^N}\sum_{i=1}^{N} |x_i| \quad \text{s.t.} \quad y=\Phi x.$$

This is an L_1 problem. We propose to apply the $L_{1/2}$ regularizer to solve the problem, that is, to use the solution of

$$\min_{x\in R^N}\sum_{i=1}^{N} |x_i|^{1/2} \quad \text{s.t.} \quad y=\Phi x$$

to reconstruct the signal.

The following experiments were conducted to show the feasibility and powerfulness of the $L_{1/2}$ regularizer. We fixed x at a signal length of $N = 512$ which contains 60 nozero spikes. We took the sensing matrix Φ being Gaussian random, let sampling be uniform in $[0, 512]$, and then, applied the $L_{1/2}$ regularizer together with L_1 regularizer (L_1 magic algorithm was used) to reconstructe the signal. The error between the reconstructed signal and the original one, error=$\sum_{i=1}^{512}|x_i - x_i^*|$, was computed in the simulation. The simulation results are shown in Figure 3.

Figure 3(a) shows the original signal, and FIgure 3(b) shows the reconstructed signal when the L_1 regularizer is applied with 184 samplings. In this case, the reconstruction is perfect and error=2.6729e-04. Figure 3(c) shows that when sampling number becomes 182, the L_1 regularizer is very poor with error=19.3224. Nevertheless, when the $L_{1/2}$ regularizer is applied, as shown in Figure 3(d), the reconstruction is still perfect, with error=9.7304e-006, even based on the same measurements. This experiment shows that the lowest number of samplings for the L_1 regularizer is at least 184. When the sampling number is reduced, say, to 182, the L_1 regularizer cannot perfectly reconstruct the signal any more, but the $L_{1/2}$ regularizer can.

Another experiment was carried out to see whether the measurements required by the $L_{1/2}$ regularizer can be far less than the least measurements required by the L_1 regularizer (184). We have simulated the L_1 regularizer and the $L_{1/2}$ regularizer with many different measurements under $M \leqslant 184$. The simulation results are uniform: when sampling number is less than 184 (the sampling numbers are 160 and 150), the L_1 regularizer never can satisfactorily reconstruct the signal, as shown in Figure 4(a) and (c) (in these cases, error is 17.0238 and 14.1106 respectively). However, when sampling is less than 184, the $L_{1/2}$ regularizer can be sure to reach a perfect recovery of the signal, as demonstrated in Figure 4(b) and (d). In these cases, the $L_{1/2}$ reconstructed error respectively are 6.1918e-06 and 1.4769e-05. This experiment shows that for the perfect signal recovery of x, the lowest sampling number required by the $L_{1/2}$ regularizer is under 150, far less than 184, the number at least required by the L_1 regularizer. This proves that the capability of signal recovery of the $L_{1/2}$ regularizer is stronger than that of the L_1 regularizer.

The performance of the $L_p(0 < p < 1, p \neq 1/2)$ regularizers were also evaluated in this application. The evaluation shows that the $L_{1/2}$ regularizer is always best for $1/2 \leqslant p < 1$ and the L_p regularizers perform similarly when $0 < p \leqslant 1/2$.

5 Analysis of experiment results

From Experiment 1, we find that the $L_{1/2}$ regularizer is more sparse than the L_1 regularizer. At the same time, the $L_{1/2}$ regularizer is more efficient and effective for heavy tail datasets. From Experiment 2, we find that the $L_{1/2}$ regularizer is able to select less variables, which shows that the $L_{1/2}$ regularizer is good at the gene data analysis. In Experiment 3, we find that for the same original sparse signal and when the completely reconstruction condition is met, the $L_{1/2}$ regularizer requires far less samplings than the

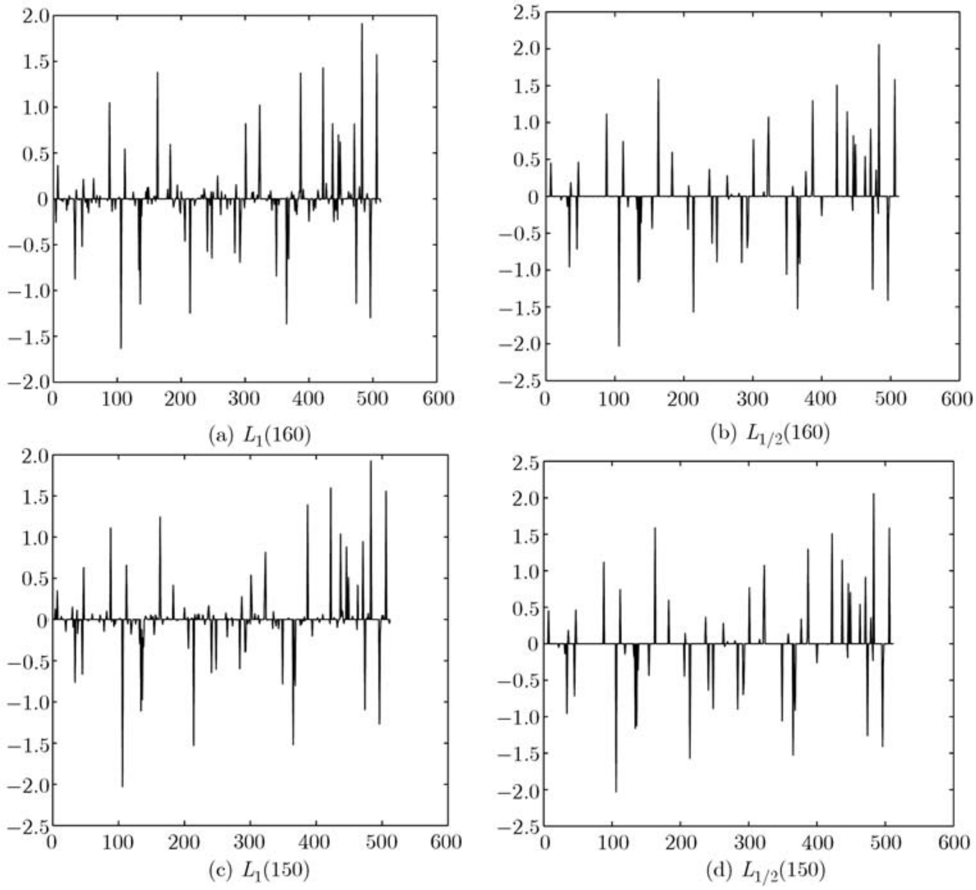

Figure 4 Capability comparison of signal recovery by Lasso and $L_{1/2}$ regularizer.

L_1 regularizer. Meanwhile, we find that in reconstructing the same original signal, the use of the $L_p(0 < p < 1/2)$ regularizer or the $L_{1/2}$ regularizer bears no significant difference.

Although sparsity is widely studied in recent years, no unified criterion is available to measure the sparsity of a problem. Our investigation into signal reconstruction in the compressive sensing example suggests that "the sampling number needed for a regularizer to exactly reconstruct a sparse signal" may serve as a measure, which might provide a feasible approach to analyzing the properties of an algorithm in the field of variable selection.

6 Conclusions

The $L_{1/2}$ regularizer proposed in this paper is easier to be solved than the L_0 regularizer and, meanwhile, more sparse and stable than the L_1 regularizer. Consequently, the $L_{1/2}$ regularizer can be more powerfully and widely used than the L_0 and L_1 regularizers. We have suggested an efficient algorithm to solve the $L_{1/2}$ regularizer which transforms a nonconvex problem into a series of L_1 regularizer problems to which the existing L_1 regularizer algorithms can be effectively applied.

Our experiments have shown that the solutions yielded from the $L_{1/2}$ regularizer are more sparse and stable than those of the L_1 regularizer. It is particularly more appropriate for heavy tail data. Furthermore, the variable selection application experiments have shown that the $L_p(0 < p < 1)$ regularizers can be represented by the $L_{1/2}$ regularizer because when $1/2 < p < 1$, the $L_{1/2}$ regularizer always yields the best sparse solution and when $0 < p < 1/2$, the $L_{1/2}$ regularizer has a sparse property similar to that of the L_p regularizers. All those properties show the great value of the $L_{1/2}$ regularizer. The results obtained in this work can be applied directly to other sparsity problems, such as blind source separation and sparse image representation. All these problems are under our current research.

Acknowledgements

This work was supported by the National Basic Research Program of China (Grant No. 2007CB311002), the National Natural Science Foundation of China (Grant No. 60975036), and the Scientific Research Plan Projects of Shaanxi Education Department (Grant No. 08jk473).

References

1 Akaike H. Information theory and an extension of the maximum likelihood principle. In: Petrov B N, Caki F, eds. Second International Symposium on Information Theory. Budapest: Akademiai Kiado, 1973. 267–281
2 Schwarz G. Estimating the dimension of a model. Ann Statist, 1978, 6: 461–464
3 Mallows C L. Some comments on Cp. Technometrics, 1973, 15: 661–675
4 Tibshirani R. Regression shrinkage and selection via the Lasso. J Royal Statist Soc B, 1996, 58: 267–288
5 Donoho D L, Huo X. Uncertainty principles and ideal atomic decomposition. IEEE Trans Inf Theory, 2001, 47: 2845–2862
6 Donoho D L, Elad E. Maximal sparsity representation via l_1 minimization. Proc Natl Acal Sci, 2003, 100: 2197–2202
7 Chen S, Donoho D L, Saunders M. Atomic decomposition by basis pursuit. SIAM Rev, 2001, 43: 129–159
8 Fan J, Heng P. Nonconcave penalty likelihood with a diverging number of parameters. Ann Statist, 2004, 32: 928–961
9 Zou H. The adaptive Lasso and its oracle properties. J Amer Statist Assoc, 2006, 101: 1418–1429
10 Zou H, Hastie T. Regularization and variable selection via the elastic net. J Royal Statist Soc B, 2005, 67: 301–320
11 Zhao P, Yu B. Stagewise Lasso. J Mach Learn Res, 2007, 8: 2701–2726
12 Candes E, Tao T. The Dantzig selector: Statistical estimation when p is much larger than n. Ann Statist, 2007, 35: 2313–2351
13 Knight K, Fu W J. Asymptotics for lasso-type estimators. Ann Statist, 2000, 28: 1356–1378
14 Friedman J, Hastie T, Tibshirani R. Additive logistic regression: a statistical view of boosting. Ann Statist, 2002, 28: 337–407
15 Efron B, Haistie T, Johnstone I, et al. Least angle regression. Ann Statist, 2004, 32: 407–499
16 Rosset S, Zhu J. Piecewise linear regularization solution paths. Ann Statist, 2007, 35: 1012–1030
17 Kim J, Koh K, Lustig M, et al. A method for large-scale l1-regularized least squares. IEEE J Se Top Signal Process, 2007, 1: 606–617
18 Horst R, Thoai N V. Dc programming: preview. J Optim Th, 1999, 103: 1–41
19 Yuille A, Rangarajan A. The concave convex procedure (CCCP). NIPS, 14. Cambridge, MA: MIT Press, 2002
20 Candes E, Wakin M, Boyd S. Enhancing sparsity by reweighted L1 minimization. J Fourier A, 2008, 14: 877–905
21 Blake C, Merz C. Repository of Machine Learning Databases [DB/OL]. Irvine, CA: University of California, Department of Information and Computer Science, 1998
22 Candes E, Romberg J, Tao T. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. IEEE Trans Inf Theory, 2006, 52: 489–509
23 Donoho D L. Compressed sensing. IEEE Trans Inf Theory, 2006, 52: 1289–1306
24 Candes E, Tao T. Near optimal signal recovery from random projections: universal encoding strategies. IEEE Trans Inf Theory, 2006, 52: 5406–5425