Research Article

Noisy Sparse Recovery Based on Parameterized Quadratic Programming by Thresholding

Jun Zhang,[1, 2] Yuanqing Li,[1] Zhuliang Yu,[1] and Zhenghui Gu[1]

[1] *Center for Brain-Computer Interfaces and Brain Information Processing, College of Automation Science and Engineering, South China University of Technology, Guangzhou 510640, China*
[2] *College of Information Engineering, Guangdong University of Technology, Guangzhou 510006, China*

Correspondence should be addressed to Jun Zhang, zhangjun7907@hotmail.com

Parameterized quadratic programming (Lasso) is a powerful tool for the recovery of sparse signals based on underdetermined observations contaminated by noise. In this paper, we study the problem of simultaneous sparsity pattern recovery and approximation recovery based on the Lasso. An extended Lasso method is proposed with the following main contributions: (1) we analyze the recovery accuracy of Lasso under the condition of guaranteeing the recovery of nonzero entries positions. Specifically, an upper bound of the tuning parameter h of Lasso is derived. If h exceeds this bound, the recovery error will increase with h; (2) an extended Lasso algorithm is developed by choosing the tuning parameter according to the bound and at the same time deriving a threshold to recover zero entries from the output of the Lasso. The simulation results validate that our method produces higher probability of sparsity pattern recovery and better approximation recovery compared to two state-of-the-art Lasso methods.

1. Introduction

The problem of recovering unknown sparse vector $\mathbf{S} \in \mathbb{R}^m$ based on the limited noisy observations $\mathbf{Y} = \mathbf{AS} + \mathbf{e}$ arises in many applications, including compressed sensing [1, 2], pattern recognition [3, 4], blind source separation [5, 6], signal reconstruction [7], and machine learning [8], where $\mathbf{A} \in \mathbb{R}^{n \times m}$ is referred to a measurement matrix with $n < m$ and $\mathbf{e} \in \mathbb{R}^n$ is an unknown vector of noise. In this paper, we suppose the positions and the signs of nonzero components of $\mathbf{S}$ are distributed uniformly at random, and their amplitudes follow an arbitrary distribution. We also assume that $\mathbf{e}$ follows zero-mean, independent, and identically distributed sub-Gaussian with parameter σ^2. Recently, many studies have advocated the use of the parameterized quadratic programming (*Lasso* [9, 10], also called *basis pursuit* [11]) to deal with the noisy sparse recovery problem through minimizing the following objective function which simultaneously executes approximation and stable recovery of the ideal sparse solution:

$$\min_{\mathbf{S}} \|\mathbf{Y} - \mathbf{AS}\|_2^2/2 + h\|\mathbf{S}\|_1. \tag{1}$$

Here and throughout, $\|\cdot\|_p$ denotes the L_p-norm ($0 \leq p \leq \infty$). Specially, $\|\mathbf{S}\|_0 = |\operatorname{supp}(\mathbf{S})|$, where $\operatorname{supp}(\mathbf{S}) := \{j \mid S_j \neq 0\}$, $|\mathbf{\Omega}|$ denotes the cardinality of a finite set $\mathbf{\Omega}$ and S_j denotes the jth component in the vector $\mathbf{S}$. In the optimization problem (1), the tuning parameter h is critical for deriving a satisfactory solution.

Up to date, many theoretical results have been obtained on Lasso to recover a sparse signal. The following two scenarios are usually of interest:

(1) *Sparsity pattern recovery:* given noisy observations $\mathbf{Y}$ of sparse signal $\mathbf{S}$, how to recover the positions and signs of $\mathbf{S}$'s nonzero entries.

(2) *Stable recovery:* analyzing the error bound between Lasso solution $\hat{\mathbf{S}}$ and true sparse vector $\mathbf{S}$.

About scenario (1), based on deterministic framework, Fuchs [12, 13] has provided a sufficient condition in mutual incoherence form. Tropp [14] and Donoho et al. [15, 16] have both discussed the sufficient conditions for the support of the solution to the Lasso method to be contained within the true support of $\mathbf{S}$. However, the sufficient condition

with the mutual incoherence form can easily be violated in applications due to the presence of highly correlated columns in measurement matrix **A**. In addition, the sufficient conditions derived in the deterministic framework are sometimes too strict to reflect the fact that the Lasso often finds the sparsity pattern of true signal **S**. Thereby, Wainwright [17] has investigated the sufficient conditions in the probabilistic framework. At the same time, in line with scenario (2), Donoho et al. [15], Tropp [14], Wainwright [17], Tseng [18], and Candès and Plan [19] have derived the error bounds under a noise tolerance constraint.

In fact, the sparsity pattern recovery requires simultaneous recovery of nonzero entries as well as zero entries. For this purpose, the Lasso utilizes the regulation of tuning parameter h to get a tradeoff between nonzero entry recovery and zero entry recovery [13]. To guarantee the zero entry recovery, a large h is always required. However, it will shrink the nonzero entries at the same time. Fuchs [12] has investigated the behavior of the solution of Lasso along with the h being increased: when the h becomes large enough, the solution of Lasso will shrink to zero. Hence, on the one hand, it is not easy to appropriately find the optimal tuning parameter h. On the other hand, to arrive at the sparsity pattern recovery, the Lasso requires high *signal-to-noise ratio* (*SNR*), and it may have poor performance in recovery accuracy which is also a basic goal in the recovery problem. In this paper, we use "approximation recovery" to reflect the performance in recovery accuracy, and its definition will be given in Section 2. It is worth noting that the achieved results of the stable recovery cannot guarantee the approximation recovery. Therefore, a significant problem in the sparse recovery is to achieve both the sparsity pattern recovery and the approximation recovery. This problem has not been previously discussed as far as we know.

To cope with this problem, we propose an extended Lasso method by utilizing thresholding. At first, we analyze the error of Lasso solution based on the L_2 norm criterion and get an upper bound of h associated with the power of noise. We also prove that, under the condition of guaranteeing the recovery of nonzero entries positions, the error of Lasso solution will increase when h exceeds the bound. Hence, for the purpose of both approximation recovery and sparsity pattern recovery, h has to be selected below this bound. Furthermore, we present a threshold estimation method and an algorithm utilizing the threshold to recover all the zero entries. The simulation results validate that this method not only recovers the sparsity pattern of S with high probability but also has a better approximation recovery than both the classical Lasso with any value of h and the basis pursuit denoising (BPDN) [11].

The rest of the paper is organized as follows. Section 2 describes the sparsity pattern recovery and the approximation recovery. In Section 3, the upper bound on h to achieve a "good" approximation recovery is found, and the threshold estimation method is presented. At the same time, an extended Lasso algorithm is proposed. Section 4 gives some simulation results to validate our algorithms. Section 5 concludes this paper.

2. Problem Formulation and Performance Analysis of Lasso

2.1. Sparsity Pattern Recovery.
The support and sign pattern (named sparsity pattern) of a sparse signal S are defined as

$$\mathbf{I_{supp}} = \left\{ j \mid S_j \neq 0 \right\},$$

$$\mathbf{I_{supp}^+} = \left\{ j \mid S_j > 0 \right\},\tag{2}$$

$$\mathbf{I_{supp}^-} = \left\{ j \mid S_j < 0 \right\},$$

where S_j is the jth entry of vector **S**. Furthermore, we denote $\mathbf{S}_{\min} = \min_{j \in \mathbf{I_{supp}}} |S_j|$ and $\mathbf{S_{supp}} = (S_j)_{j \in \mathbf{I_{supp}}}$. The vector $\mathbf{S_{supp}}$ is composed of the nonzero entries of **S**. Suppose $\hat{\mathbf{S}}$ be a recovery of **S** based on **Y**; we also denote $\hat{\mathbf{I}}_{\mathbf{supp}} = \{j \mid \hat{S}_j \neq 0\}$, $\hat{\mathbf{I}}_{\mathbf{supp}}^+ = \{j \mid \hat{S}_j > 0\}$ and $\hat{\mathbf{I}}_{\mathbf{supp}}^- = \{j \mid \hat{S}_j < 0\}$.

Based on the above notations, a definition of sparsity pattern recovery is given as follows.

Definition 1 (sparsity pattern recovery). A signal $\hat{\mathbf{S}}$ is a sparsity pattern recovery of **S** if and only if

$$\mathbf{N_m} + \mathbf{N_f} = 0,\tag{3}$$

where $\mathbf{N_m} = |\mathbf{I_{supp}}| - (|\mathbf{I_{supp}^+} \cap \hat{\mathbf{I}}_{\mathbf{supp}}^+| + |\mathbf{I_{supp}^-} \cap \hat{\mathbf{I}}_{\mathbf{supp}}^-|)$ and $\mathbf{N_f} = |\hat{\mathbf{I}}_{\mathbf{supp}}| - (|\mathbf{I_{supp}^+} \cap \hat{\mathbf{I}}_{\mathbf{supp}}^+| + |\mathbf{I_{supp}^-} \cap \hat{\mathbf{I}}_{\mathbf{supp}}^-|)$.

Remark 1. $\mathbf{N_m}$ and $\mathbf{N_f}$ denote the number of nonzero entries and zero entries of **S** that are failed in recovery, respectively. Therefore, for exact sparsity pattern recovery, the support and sign of the nonzero entries as well as the zero entries have to be recovered.

2.2. Approximation Recovery.
In noisy case, it is generally impossible to seek exact recovery of a sparse signal, and estimation error is inevitable in the process of recovery. In this paper, the squared error (SE) between the $\hat{\mathbf{S}}$ and **S** is defined as follows:

$$\text{SE} = \left(\hat{\mathbf{S}} - \mathbf{S}\right)^T \left(\hat{\mathbf{S}} - \mathbf{S}\right).\tag{4}$$

Using squared error as criterion, the problem of approximation recovery is defined as follows.

Definition 2 (approximation recovery). A recovery $\hat{\mathbf{S}}$ is a "good" approximation recovery to **S** if the SE is as small as possible.

2.3. Performance Analysis of Lasso.
Lasso utilizes the regulation of parameter h to reach both $\mathbf{N_m} = 0$ and $\mathbf{N_f} = 0$. Generally, as h increases, the probability that the solution of Lasso reaches $\mathbf{N_f} = 0$ increases, but the probability that the solution of Lasso reaches $\mathbf{N_m} = 0$ decreases. A brief analysis to support this conclusion is given as follows. Wainwright [17] established a sufficient condition for Lasso to reach sparsity pattern recovery: to guarantee $\mathbf{N_f} = 0$, h must satisfy

$$h > \frac{2}{\gamma}\sqrt{2n \log m}\sigma,\tag{5}$$

where $\gamma \in (0, 1]$ is called *incoherence parameter* of the matrix $\mathbf{A}$ such that

$$\left\| \left\| \left(A_{\mathbf{I_{supp}}} \right)^T A_{\mathbf{I_{supp}}} \left(\left(A_{\mathbf{I_{supp}}} \right)^T A_{\mathbf{I_{supp}}} \right)^{-1} \right\| \right\|_\infty \le 1 - \gamma, \qquad (6)$$

where $\mathbf{I_{supp}^c}$ denotes the supplementary set of $\mathbf{I_{supp}}$, $\mathbf{A}_\beta$ is a submatrix composed of the columns of $\mathbf{A}$ with their indices being in β, and, for a $n \times m$ matrix $\mathbf{M}$, $\|\|\mathbf{M}\|\|_\infty \triangleq \max_{i=1,\dots,n} \sum_{j=1}^m |M_{ij}|$.

In a necessary condition for Lasso to reach sparsity pattern recovery, Wainwright defined a quantity

$$\widetilde{g}_i(h) = h \mathbf{I_i}^T \left(\mathbf{S_{supp}}^T \mathbf{S_{supp}} \right)^{-1} \text{sign} \left(\mathbf{S_{supp}} \right), \qquad (7)$$

where $\mathbf{I_i}$ is a unit vector whose entries are zeros except that the ith entry equals one. Suppose the existence of inclusion $S_i \in (0, \widetilde{g}_i(h))$ or $S_i \in (\widetilde{g}_i(h), 0)$ for some $i \in \mathbf{I_{supp}}$, then the probability of sparsity pattern recovery is bounded away from one

$$\mathbf{P}(\mathbf{N_m} + \mathbf{N_f} = 0) \le \frac{1}{2}. \qquad (8)$$

According to the analysis in [17], the quantity $\widetilde{g}_i(h)$ corresponds to the amount that is "shrunken" by the Lasso in position $i \in \mathbf{I_{supp}}$. As the quantity $\widetilde{g}_i(h)$ increases with h, the assumption of the existence of inclusion $S_i \in (0, \widetilde{g}_i(h))$ or $S_i \in (\widetilde{g}_i(h), 0)$ for some $i \in \mathbf{I_{supp}}$ holds with higher probability. Combined with (5) and (8), it indicates that as h increases, the probability of $\mathbf{N_m} = 0$ decreases.

From the above analysis, we know that the probabilities of $\mathbf{N_m} = 0$ and $\mathbf{N_f} = 0$ have contradicting trends with the changing of h. Therefore, the appropriate selection of parameter h is an important and open problem [16]. In practice, a large h is always used to reach $\mathbf{N_f} = 0$. In such case, not only the nonzero entries caused by the noise but also the nonzero entries in positions $i \in \mathbf{I_{supp}}$ are shrunk by Lasso with a large h. This operation increases the probability of $\mathbf{N_m} \ne 0$ and results in poor performance in approximation recovery. To mitigate this problem, in the following context, we propose a method to reach high probability on sparsity pattern recovery and high accuracy on approximation.

3. Extended Lasso Method Combining Parameter Selection and Thresholding

In this section, an extended Lasso method is proposed for simultaneous sparsity pattern recovery and approximation recovery. The overall algorithm is presented in Section 3.1. Two key techniques of the proposed method, including threshold estimation and parameter selection, are given in Sections 3.2 and 3.3.

3.1. Proposed Extended Lasso Algorithm. The algorithm of the proposed extended Lasso is summarized as follows.

Step 1. Solve (1) through appropriate selection of parameter h based on Section 3.3.

Step 2. Estimate the threshold by following (18) in Section 3.2.

Step 3. Apply pruning to make the entries below the estimated threshold of the solution obtained in Step 1. to be zero.

Remark 2. We do not try to reach the sparsity pattern recovery at Step 1 but to recover the nonzero entries of $\mathbf{S}$ and achieve the approximation recovery. The sparsity pattern recovery is realized by thresholding in Step 3.

In the proposed algorithm, we have to answer the following two questions. The first one is how to select an appropriate tuning parameter h that guarantees a "good" approximation recovery at the nonzero entries of $\mathbf{S}$. The second is how to estimate the threshold. The answers of these questions are presented in the following context.

3.2. Threshold Estimation. The Kuhn-Tucker condition of the optimization problem (1) is considered as follows. We denote $\partial \|\widehat{\mathbf{S}}\|_1$ as the subdifferential of $\|\widehat{\mathbf{S}}\|_1$

$$\partial \left\| \widehat{\mathbf{S}} \right\|_1 = \left\{ \mathbf{u} \mid \mathbf{u}^T \widehat{\mathbf{S}} = \left\| \widehat{\mathbf{S}} \right\|_1, \|\mathbf{u}\|_\infty \le 1 \right\}$$
$$= \left\{ \mathbf{u} \left| \begin{array}{ll} u_i = \text{sign}\left(\widehat{S}_i\right), & \text{if } \widehat{S}_i \ne 0 \\ |u_i| \le 1, & \text{otherwise} \end{array} \right. \right\}. \qquad (9)$$

The recovery $\widehat{\mathbf{S}}$ is the optimum of (1), if and only if

$$\exists \mathbf{u} \in \partial \left\| \widehat{\mathbf{S}} \right\|_1, \quad \text{such that } \mathbf{A}^T \left(\mathbf{A}\widehat{\mathbf{S}} - \mathbf{Y} \right) + h\mathbf{u} = 0. \qquad (10)$$

Since the optimization problem (1) is convex, the Kuhn-Tucker condition is a sufficient condition. To distinguish the nonzero and zero components of $\widehat{\mathbf{S}}$, we denote $\widehat{\mathbf{S}}_{\mathbf{supp}}$ as the reduced dimensional vector built upon the nonzero components of $\widehat{\mathbf{S}}$. Similarly, $\overline{\mathbf{A}}$ denotes the associated columns in $\mathbf{A}$, and a_j is the jth column in $\mathbf{A}$. In terms of nonzero and zero entries of $\widehat{\mathbf{S}}$, the necessary and sufficient condition (10) can be expressed as

$$\overline{\mathbf{A}}^T \left(\mathbf{Y} - \overline{\mathbf{A}} \widehat{\mathbf{S}}_{\mathbf{supp}} \right) = h \times \text{sign} \left(\widehat{\mathbf{S}}_{\mathbf{supp}} \right), \qquad (11)$$

$$\left| a_j^T \left(\mathbf{Y} - \overline{\mathbf{A}} \widehat{\mathbf{S}}_{\mathbf{supp}} \right) \right| \le h, \quad \text{for } a_j \notin \overline{\mathbf{A}}. \qquad (12)$$

Since $\overline{\mathbf{A}}$ is always full rank, (11) can be given as

$$\widehat{\mathbf{S}}_{\mathbf{supp}} = \overline{\mathbf{A}}^+ \mathbf{Y} - h \left(\overline{\mathbf{A}}^T \overline{\mathbf{A}} \right)^{-1} \text{sign} \left(\widehat{\mathbf{S}}_{\mathbf{supp}} \right), \qquad (13)$$

where $\overline{\mathbf{A}}^+$ is the Moore-Penrose inverse of $\overline{\mathbf{A}}$. When $\mathbf{N_m} = 0$, (13) becomes

$$\widehat{\mathbf{S}}_{\mathbf{supp}} = \mathbf{S}_{\mathbf{supp}} + \overline{\mathbf{A}}^+ \mathbf{e} - h \left(\overline{\mathbf{A}}^T \overline{\mathbf{A}} \right)^{-1} \text{sign} \left(\widehat{\mathbf{S}}_{\mathbf{supp}} \right). \qquad (14)$$

If the maximum amplitudes of the second and third terms in the right-hand side of (14) are bounded, we can reach $\mathbf{N_f} = 0$ through pruning by setting a threshold. Furthermore, the sparsity pattern recovery can be achieved.

In fact, the third term in the right-hand side of (14) is a deterministic quantity. We only need to bound the second term. Suppose that the singular value decomposition $\overline{\mathbf{A}}^+ = \mathbf{V}\mathbf{\Sigma}^{-1}\mathbf{U}^T$ and $\lambda_{\min}$ is the minimal singular value of $\overline{\mathbf{A}}^+$. Since the elements of e are zero mean and i.i.d sub-Gaussian with parameter σ^2, it follows from property of sub-Gaussian [17] that $\overline{\mathbf{A}}^+e$ is zero mean and sub-Gaussian with a variance satisfying

$$\mathbf{V}\mathbf{\Sigma}^{-1}\mathbf{U}^T\mathbf{e} \le \frac{\sigma^2}{\lambda_{\min}}. \tag{15}$$

Consequently, using the sub-Gaussian tail bound and the union bound, we have

$$\mathbf{P}\left(\left\|\overline{\mathbf{A}}^+\mathbf{e}\right\|_\infty > t\right) \le 2\exp\left(-\frac{t^2\lambda_{\min}}{2\sigma^2} + \log\left(\left|\operatorname{supp}(\hat{\mathbf{S}})\right|\right)\right)$$
$$\triangleq 1 - T, \tag{16}$$

where T represents confidence probability. With a given T, the bound of the second term in (14) can be derived as

$$t = \sqrt{\frac{2\log\left(2\left|\operatorname{supp}(\hat{\mathbf{S}})\right|/(1 - T)\right)}{\lambda_{\min}}}\sigma. \tag{17}$$

Hence, according to (14), the threshold can be set as

$$\operatorname{Thresh} = \sqrt{\frac{2\log\left(2\left|\operatorname{supp}(\hat{\mathbf{S}})\right|/(1 - T)\right)}{\lambda_{\min}}}\sigma$$
$$+ h\left(\overline{\mathbf{A}}^T\overline{\mathbf{A}}\right)^{-1}\operatorname{sign}\left(\hat{\mathbf{S}}_{\mathbf{supp}}\right). \tag{18}$$

3.3. Selection of h for Approximation Recovery. Denoting $\hat{\mathbf{u}}^* \triangleq \operatorname{sign}(\hat{\mathbf{S}}_{\mathbf{supp}})$ and combining (4) and (14), we have

$$\operatorname{SE}(h) = \left(\overline{\mathbf{A}}^+\mathbf{e} - h\left(\overline{\mathbf{A}}^T\overline{\mathbf{A}}\right)^{-1}\hat{\mathbf{u}}^*\right)^T\left(\overline{\mathbf{A}}^+\mathbf{e} - h\left(\overline{\mathbf{A}}^T\overline{\mathbf{A}}\right)^{-1}\hat{\mathbf{u}}^*\right)$$
$$= \left[\hat{\mathbf{u}}^{*T}\left(\overline{\mathbf{A}}^T\overline{\mathbf{A}}\right)^{-1}\left(\overline{\mathbf{A}}^T\overline{\mathbf{A}}\right)^{-1}\hat{\mathbf{u}}^*\right]h^2$$
$$- \left[2\hat{\mathbf{u}}^{*T}\left(\overline{\mathbf{A}}^T\overline{\mathbf{A}}\right)^{-1}\left(\overline{\mathbf{A}}^T\overline{\mathbf{A}}\right)^{-1}\overline{\mathbf{A}}^+\mathbf{e}\right]h + \mathbf{e}^T\left(\overline{\mathbf{A}}^+\right)^T\overline{\mathbf{A}}^+\mathbf{e}. \tag{19}$$

It is obvious that $[\hat{\mathbf{u}}^{*T}(\overline{\mathbf{A}}^T\overline{\mathbf{A}})^{-1}(\overline{\mathbf{A}}^T\overline{\mathbf{A}})^{-1}\hat{\mathbf{u}}^*] > 0$. Fuchs [13] has checked the solution $\hat{\mathbf{S}}$ of (1) as h varies from 0 to $+\infty$. As h increases, $\hat{\mathbf{S}}_{\mathbf{supp}}$ is a continuous function of h, and $\|\hat{\mathbf{S}}_{\mathbf{supp}}\|_1$ is monotonically decreasing. Until $h > \|\mathbf{A}^T\mathbf{Y}\|_\infty$, $\hat{\mathbf{S}} = 0$. Thereby, the interval $]0+, \|\mathbf{A}^T\mathbf{Y}\|_\infty[$ can be divided into subintervals characterized by the fact that, within each such subinterval, the number of nonzero components of the optimum $\hat{\mathbf{S}}$ of Lasso is constant. According to the above discussion, in each of the subintervals, $\hat{\mathbf{u}}^*$ is a fixed vector, $\operatorname{SE}(h)$ is a quadratic function about h. Thereby, the following theorem holds.

Theorem 1. *In the general noisy case, let* $\mathbf{Y} = \mathbf{AS} + \mathbf{e}$ *with* $\|\mathbf{e}\|_2 \le \delta$; *one solves the optimization problem* (1) *with h. Then, if* $\mathbf{N_m} = 0$ *and* $h > \delta$, $SE(h)$ *increases with h.*

The proof of this theorem is given in the appendix.

Remark 3. Under the aim of sparsity pattern recovery, the selection of h has been discussed by Wainwright [17]. When $h > (2/\gamma)\sqrt{2n\log m}\sigma$ and $S_{\min} > [\|((\mathbf{A_{I_{supp}}})^T\mathbf{A_{I_{supp}}})^{-1}\|_\infty + 4\sigma/\sqrt{C_{\min}}]h$, where $C_{\min}$ denotes the minimal eigenvalue of matrix $((\mathbf{A_{I_{supp}}})^T\mathbf{A_{I_{supp}}})$, Lasso can recover the sparsity pattern of $\mathbf{S}$ with high probability. However, from (A.12), parameter h must satisfy $h < \|\mathbf{e}\|_2 \le \delta$ to achieve approximation recovery. Hence, the selection of the parameter h is a tradeoff when both sparsity pattern recovery and approximation recovery are simultaneously under consideration. This indicates that the selection of h may be a dilemma.

4. Simulation Results

In this section, we carried out two experiments to evaluate the performance of the proposed method. Firstly, we studied the distribution of the selected tuning parameter h against two aims, that is, sparsity pattern recovery and approximation recovery. Secondly, we compared the performance of the proposed method with BPDN and standard Lasso. The probabilities of sparsity pattern recovery and the recovery accuracies of these methods were compared numerically. In the standard Lasso, we search its optimal tuning parameter h in its feasible range with size $1e - 2$.

4.1. Histogram of Optimal h for Standard Lasso. In this experiment, we study the histograms of optimal h for the purpose of sparsity pattern recovery and approximation recovery respectively. The parameters used in this experiment are $n = 10$, $m = 30$, $\|\mathbf{S}\|_0 = 2, 3, 4$, and $\|\mathbf{e}\|_2 \le 0.1$. The simulation results are obtained by 10000 Monte Carlo experiments with randomly generated sparse signal $\mathbf{S}$ and matrix $\mathbf{A}$. Firstly, we applied the Lasso to achieve optimal approximation of true sparse signal $\mathbf{S}$. In this experiment, we searched for the optimal tuning parameter h in its feasible range with step size $1e - 2$ to achieve minimal mean-square error. The histograms of optimal h for approximation recovery with signals of different sparsity are shown in Figure 1. Secondly, we simulate the histograms of optimal h for sparsity pattern recovery. The results are also shown in Figure 1. From the histograms shown in Figure 1, it is clear that the optimal tuning parameter h for the purpose of sparsity pattern recovery and the h for approximation recovery cannot coincide generally. That is to say, we cannot achieve both purposes simply by tuning the parameter h. This motivates our research on new method to achieve sparsity pattern recovery and approximation recovery simultaneously.

4.2. Performance Comparison of the Proposed Method with BPDN and Lasso. In this experiment, we compare the probability of sparsity pattern recovery and the mean-square
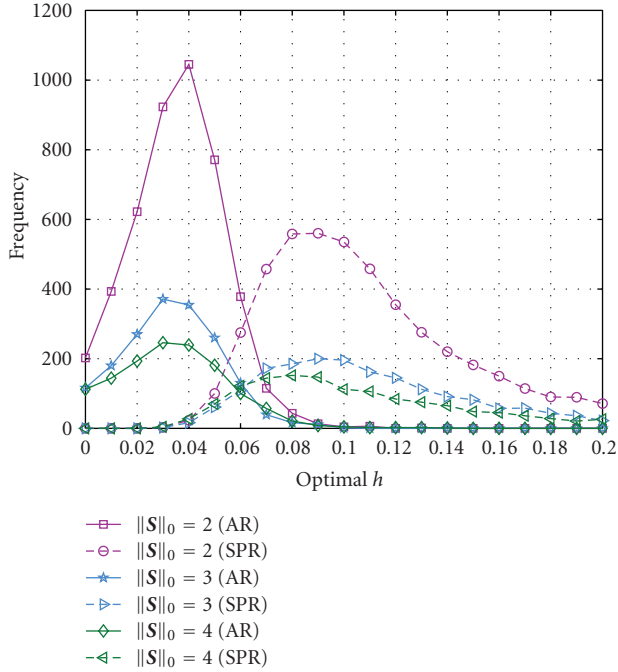
FIGURE 1: The histograms of h for the purposes of sparsity pattern recovery and approximation recovery (square-solid curves: the histogram of tuning parameter h for approximation recovery (AR), circinate-dash curves: the histogram of tuning parameter h for sparsity pattern recovery (SPR)).
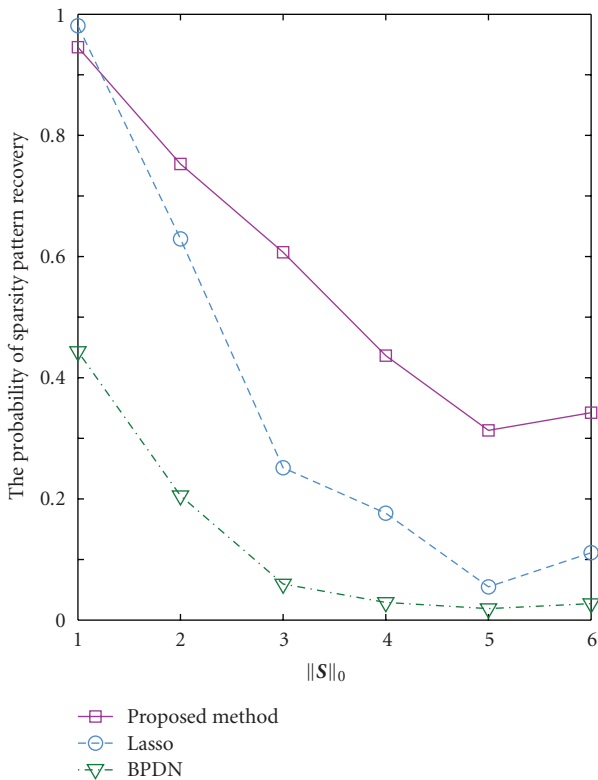

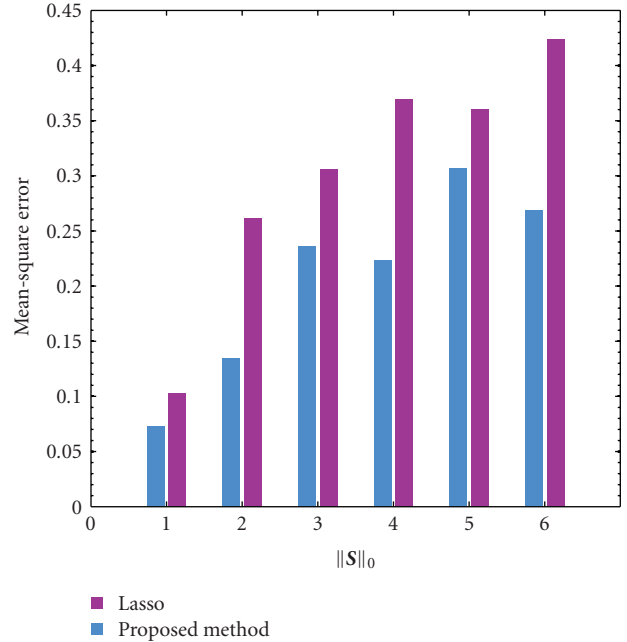FIGURE 2: Comparison on the probability of sparsity pattern recovery.

FIGURE 3: Comparison on the mean-square error.

error between the true sparse vector $\mathbf{S}$ and the solutions $\hat{\mathbf{S}}$ obtained from the proposed method, Lasso, and BPDN [11]. The parameters used in this experiment are $n = 10$, $m = 30$, $\|\mathbf{S}\|_0 = 1, 2, 3 \ldots, 6$, and $\|\mathbf{e}\|_2 \leq 0.1$. The results are also obtained from 10000 Monte Carlo experiments with randomly generated sparse signals S and matrices $\mathbf{A}$. In the sparsity pattern recovery experiments, supposing that n_p experiments can recover the sparsity pattern of $\mathbf{S}$, we use the ratio $n_p/10000$ to approximate the probability of sparsity pattern recovery. Based on this approach, we obtain the probabilities of sparsity pattern recovery of the Lasso, BPDN $(h = \sigma\sqrt{2\log(n)})$ [11], and the proposed method as shown in Figure 2. For the Lasso, it is worth noting that the optimal tuning parameter h is searched with step size $1e-2$ to achieve sparsity pattern recovery. Under the condition of sparsity pattern recovery, we further compared the mean-square error between the proposed method and Lasso. The results in Figures 2 and 3 reveal that the proposed method not only has higher probability of sparsity pattern recovery but also has better approximation performance. Since the performance of BPDN with $h = \sigma\sqrt{2\log(n)}$ is lower than that of the Lasso with optimal h, the approximation performance of BPDN is not shown in Figure 3.

5. Conclusion

In this paper, motivated by the fact that Lasso can hardly achieve simultaneous optimal sparsity pattern recovery and approximation recovery, we propose an extended Lasso method to achieve satisfactory performance in both aspects. In the proposed method, firstly, we select a tuning parameter h based on Theorem 1 and solve the Lasso

for approximation recovery. Then, a threshold estimation method is applied to prune the entries of the obtained solution to achieve sparsity pattern recovery. The simulation results demonstrate the good performance of the proposed method.

Appendix

Proof of Theorem 1

Proof. We first denote h_{optimal} the h value of vertex at each quadratic function $\mathbf{SE}(h)$. From the analysis of Section 3.3 and (19), the following inequality certainly holds:

$$h_{\text{optimal}}$$
$$\leq \max \left\{ h = \frac{\hat{\mathbf{u}}^{*\mathbf{T}} \left(\overline{\mathbf{A}}^{\mathbf{T}}\overline{\mathbf{A}}\right)^{-1} \left(\overline{\mathbf{A}}^{\mathbf{T}}\overline{\mathbf{A}}\right)^{-1} \overline{\mathbf{A}}^{\mathbf{T}} \mathbf{e}}{\hat{\mathbf{u}}^{*\mathbf{T}} \left(\overline{\mathbf{A}}^{\mathbf{T}}\overline{\mathbf{A}}\right)^{-1} \left(\overline{\mathbf{A}}^{\mathbf{T}}\overline{\mathbf{A}}\right)^{-1} \hat{\mathbf{u}}^{*}} \middle| \hat{u}^{*} \in \partial \left\|\hat{\mathbf{S}}\right\|_{1} \right\}. \tag{A.1}$$

Without loss of generality, the columns in $\mathbf{A}$ can be normalized to one in ℓ_2 norm. The term $\overline{\mathbf{A}}^{\mathbf{T}}\mathbf{e}$ can be rewritten as follows:

$$\overline{\mathbf{A}}^{\mathbf{T}} \mathbf{e} = \|\mathbf{e}\|_{2} \times \overline{\mathbf{A}}^{\mathbf{T}} \times \frac{\mathbf{e}}{\|\mathbf{e}\|_{2}} = \|\mathbf{e}\|_{2} \times \begin{bmatrix} s_1, 0, 0, \ldots \\ 0, s_2, 0, \ldots \\ \cdots \\ 0, \ldots, 0, s_q \end{bmatrix} \times \hat{\mathbf{u}}^{*}, \tag{A.2}$$

where $|s_i| \triangleq |\overline{\mathbf{A}_{\mathbf{i}}}^{T} \times (\mathbf{e}/\|\mathbf{e}\|_{2})|$ implies $|s_i| \leq 1$, $\text{sign}(s_i) \triangleq \hat{\mathbf{u}}_{\mathbf{i}}^{*} \times \text{sign}(\overline{\mathbf{A}_{\mathbf{i}}}^{T} \times (\mathbf{e}/\|\mathbf{e}\|_{2}))$.

Hence, according to (A.2), we have

$$\hat{\mathbf{u}}^{*\mathbf{T}} \left(\overline{\mathbf{A}}^{\mathbf{T}}\overline{\mathbf{A}}\right)^{-1} \left(\overline{\mathbf{A}}^{\mathbf{T}}\overline{\mathbf{A}}\right)^{-1} \overline{\mathbf{A}}^{\mathbf{T}} \mathbf{e}$$
$$= \|\mathbf{e}\|_{2} \times \hat{\mathbf{u}}^{*\mathbf{T}} \left(\overline{\mathbf{A}}^{\mathbf{T}}\overline{\mathbf{A}}\right)^{-1} \left(\overline{\mathbf{A}}^{\mathbf{T}}\overline{\mathbf{A}}\right)^{-1} \begin{bmatrix} s_1, 0, 0, \ldots \\ 0, s_2, 0, \ldots \\ \cdots \\ 0, \ldots, 0, s_q \end{bmatrix} \hat{\mathbf{u}}^{*}. \tag{A.3}$$

Let

$$\left(\overline{\mathbf{A}}^{\mathbf{T}}\overline{\mathbf{A}}\right)^{-1} \left(\overline{\mathbf{A}}^{\mathbf{T}}\overline{\mathbf{A}}\right)^{-1} = \mathbf{S}_1, \quad \begin{bmatrix} s_1, 0, 0, \ldots \\ 0, s_2, 0, \ldots \\ \cdots \\ 0, \ldots, 0, s_q \end{bmatrix} = \mathbf{S}_2, \tag{A.4}$$

we have

$$h_{\text{optimal}}$$
$$\leq \max \left\{ h = \frac{\hat{\mathbf{u}}^{*\mathbf{T}} \left(\overline{\mathbf{A}}^{\mathbf{T}}\overline{\mathbf{A}}\right)^{-1} \left(\overline{\mathbf{A}}^{\mathbf{T}}\overline{\mathbf{A}}\right)^{-1} \overline{\mathbf{A}}^{\mathbf{T}} \mathbf{e}}{\hat{\mathbf{u}}^{*\mathbf{T}} \left(\overline{\mathbf{A}}^{\mathbf{T}}\overline{\mathbf{A}}\right)^{-1} \left(\overline{\mathbf{A}}^{\mathbf{T}}\overline{\mathbf{A}}\right)^{-1} \hat{\mathbf{u}}^{*}} \middle| \hat{u}^{*} \in \partial \left\|\hat{\mathbf{S}}\right\|_{1} \right\}$$
$$= \max \left\{ h = \|\mathbf{e}\|_{2} \times \frac{\hat{\mathbf{u}}^{*\mathbf{T}} \mathbf{S}_1 \mathbf{S}_2 \hat{\mathbf{u}}^{*}}{\hat{\mathbf{u}}^{*\mathbf{T}} \mathbf{S}_1 \hat{\mathbf{u}}^{*}} \middle| \hat{u}^{*} \in \partial \left\|\hat{\mathbf{S}}\right\|_{1} \right\}. \tag{A.5}$$

According to the definitions of $\mathbf{S}_1$ and $\mathbf{S}_2$, it is easy to know that they are a positive definite matrix and a generalized positive definite matrix, respectively. To examine the size relationship between $\hat{\mathbf{u}}^{*\mathbf{T}} \mathbf{S}_1 \mathbf{S}_2 \hat{\mathbf{u}}^{*}$ and $\hat{\mathbf{u}}^{*\mathbf{T}} \mathbf{S}_1 \hat{\mathbf{u}}^{*}$, we construct the auxiliary function $f(\hat{\mathbf{u}}^{*})$ as follows:

$$f(\hat{\mathbf{u}}^{*}) = \hat{\mathbf{u}}^{*\mathbf{T}} \mathbf{S}_1 \mathbf{S}_2 \hat{\mathbf{u}}^{*} - \hat{\mathbf{u}}^{*\mathbf{T}} \mathbf{S}_1 \hat{\mathbf{u}}^{*}$$
$$= \hat{\mathbf{u}}^{*\mathbf{T}} \mathbf{S}_1 (\mathbf{S}_2 - \mathbf{I}) \hat{\mathbf{u}}^{*} \tag{A.6}$$
$$= -\hat{\mathbf{u}}^{*\mathbf{T}} \mathbf{S}_1 (\mathbf{I} - \mathbf{S}_2) \hat{\mathbf{u}}^{*},$$

where $\mathbf{I}$ denotes identity matrix. According to the definition of $\mathbf{S}_2$, matrix $(\mathbf{I} - \mathbf{S}_2)$ is a positive diagonal matrix.

Also, $\hat{\mathbf{u}}^{*}$ can be normalized to one in ℓ_2 norm. In the following, we discuss the extremum of function $f^{+}(\hat{\mathbf{u}}^{*}) = \hat{\mathbf{u}}^{*T} \mathbf{S}_1 (\mathbf{I} - \mathbf{S}_2) \hat{\mathbf{u}}^{*}$, where $\|\hat{\mathbf{u}}^{*}\|_{2} = 1$.

It is easy to know that $f^{+}(\hat{\mathbf{u}}^{*})$ is a real continuous function about $\hat{\mathbf{u}}^{*}$ and $\{\hat{\mathbf{u}}^{*} \mid \|\hat{\mathbf{u}}^{*}\|_{2} = 1\}$ is a compact set. Therefore, $f^{+}(\hat{\mathbf{u}}^{*})$ have the minimum and maximum in the set $\{\hat{\mathbf{u}}^{*} \mid \|\hat{\mathbf{u}}^{*}\|_{2} = 1\}$. It implies that the range of $f^{+}(\hat{\mathbf{u}}^{*})$ is a closed bound convex set. We construct the Lagrange function $G(\hat{\mathbf{u}}^{*})$ of $f^{+}(\hat{\mathbf{u}}^{*})$ as follow:

$$G(\hat{\mathbf{u}}^{*}) = \hat{\mathbf{u}}^{*\mathbf{T}} \mathbf{S}_1 (\mathbf{I} - \mathbf{S}_2) \hat{\mathbf{u}}^{*} - \lambda \left(\hat{\mathbf{u}}^{*T} \hat{\mathbf{u}}^{*} - 1 \right). \tag{A.7}$$

Let the derivative of $G(\hat{\mathbf{u}}^{*})$ equal zero, that is,

$$\mathbf{grad} \, G(\hat{\mathbf{u}}^{*}) = \mathbf{S}_1 (\mathbf{I} - \mathbf{S}_2) \hat{\mathbf{u}}^{*} + [\mathbf{S}_1 (\mathbf{I} - \mathbf{S}_2)]^{T} \hat{\mathbf{u}}^{*} - 2\lambda \hat{\mathbf{u}}^{*} = 0, \tag{A.8}$$

then we have

$$\frac{\mathbf{S}_1 (\mathbf{I} - \mathbf{S}_2) + [\mathbf{S}_1 (\mathbf{I} - \mathbf{S}_2)]^{T}}{2} \hat{\mathbf{u}}^{*} = \lambda \hat{\mathbf{u}}^{*}. \tag{A.9}$$

It implies that λ is the eigenvalue of matrix $(\mathbf{S}_1 (\mathbf{I} - \mathbf{S}_2) + [\mathbf{S}_1 (\mathbf{I} - \mathbf{S}_2)]^{T})/2$. So the extremum of $f^{+}(\hat{\mathbf{u}}^{*})$ is given by

$$f_{e}^{+}(\hat{\mathbf{u}}^{*}) = \hat{\mathbf{u}}^{*T} \left\{ \frac{\mathbf{S}_1 (\mathbf{I} - \mathbf{S}_2) + [\mathbf{S}_1 (\mathbf{I} - \mathbf{S}_2)]^{T}}{2} \right\} \hat{\mathbf{u}}^{*}$$
$$= \lambda \hat{\mathbf{u}}^{*T} \hat{\mathbf{u}}^{*} = \lambda \geq \lambda_{\min}, \tag{A.10}$$

and $\lambda_{\min}$ is the minimal eigenvalue of matrix $(\mathbf{S}_1 (\mathbf{I} - \mathbf{S}_2) + [\mathbf{S}_1 (\mathbf{I} - \mathbf{S}_2)]^{T})/2$.

Since matrix $(\mathbf{S}_1(\mathbf{I} - \mathbf{S}_2) + [\mathbf{S}_1(\mathbf{I} - \mathbf{S}_2)]^T)/2$ is a symmetric positive definite matrix, the eigenvalue is positive. We have

$$f(\hat{\mathbf{u}}^*) = -f^+(\hat{\mathbf{u}}^*) \leq -f_e^+(\hat{\mathbf{u}}^*) \leq -\lambda_{\min} < 0$$
$$\implies \hat{\mathbf{u}}^{*\mathbf{T}} \mathbf{S}_1 \mathbf{S}_2 \hat{\mathbf{u}}^* < \hat{\mathbf{u}}^{*\mathbf{T}} \mathbf{S}_1 \hat{\mathbf{u}}^*. \tag{A.11}$$

Therefore, applying the condition $\|\mathbf{e}\|_2 \leq \delta$,

$$h_{\text{optimal}} \leq \max \left\{ h = \|\mathbf{e}\|_2 \times \frac{\hat{\mathbf{u}}^{*\mathbf{T}} \mathbf{S}_1 \mathbf{S}_2 \hat{\mathbf{u}}^*}{\hat{\mathbf{u}}^{*\mathbf{T}} \mathbf{S}_1 \hat{\mathbf{u}}^*} \,\middle|\, \hat{\mathbf{u}}^* \in \partial \left\| \hat{\mathbf{S}} \right\|_1 \right\}$$

$$< \|\mathbf{e}\|_2 \leq \delta. \tag{A.12}$$

□

One can further check that $\hat{\mathbf{S}}$ is a continuous function of h. Then, with the above proof, the continuity of squared error function, and the property of the quadratic function, the theorem is proven.

Acknowledgments

References

[1] E. J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information," *IEEE Transactions on Information Theory*, vol. 52, no. 2, pp. 489–509, 2006.

[2] D. L. Donoho, "Compressed sensing," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[3] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210–227, 2009.

[4] Y. Li, P. Namburi, Z. Yu, C. Guan, J. Feng, and Z. Gu, "Voxel selection in FMRI data analysis based on sparse representation," *IEEE Transactions on Bio-Medical Engineering*, vol. 56, no. 10, pp. 2439–2451, 2009.

[5] Y. Li, S. I. Amari, A. Cichocki, D. W. C. Ho, and S. Xie, "Underdetermined blind source separation based on sparse representation," *IEEE Transactions on Signal Processing*, vol. 54, no. 2, pp. 423–437, 2006.

[6] Y. Li, A. Cichocki, and S. I. Amari, "Analysis of sparse representation and blind source separation," *Neural Computation*, vol. 16, no. 6, pp. 1193–1234, 2004.

[7] P. Xu, Y. Tian, H. Chen, and D. Yao, "Lp norm iterative sparse solution for EEG source localization," *IEEE Transactions on Biomedical Engineering*, vol. 54, no. 3, pp. 400–409, 2007.

[8] S. Agarwal, A. Awan, and D. Roth, "Learning to detect objects in images via a sparse, part-based representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 11, pp. 1475–1490, 2004.

[9] K. Knight and W. Fu, "Asymptotics for Lasso-type estimators," *Annals of Statistics*, vol. 28, no. 5, pp. 1356–1378, 2000.

[10] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B*, vol. 58, no. 1, pp. 267–288, 1996.

[11] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM Review*, vol. 43, no. 1, pp. 129–159, 2001.

[12] J. J. Fuchs, "Recovery of exact sparse representations in the presence of bounded noise," *IEEE Transactions on Information Theory*, vol. 51, no. 10, pp. 3601–3608, 2005.

[13] J. J. Fuchs, "On sparse representations in arbitrary redundant bases," *IEEE Transactions on Information Theory*, vol. 50, no. 6, pp. 1341–1344, 2004.

[14] J. A. Tropp, "Just relax: convex programming methods for identifying sparse signals in noise," *IEEE Transactions on Information Theory*, vol. 52, no. 3, pp. 1030–1051, 2006.

[15] D. L. Donoho, M. Elad, and V. N. Temlyakov, "Stable recovery of sparse overcomplete representations in the presence of noise," *IEEE Transactions on Information Theory*, vol. 52, no. 1, pp. 6–18, 2006.

[16] D. L. Donoho and M. Elad, "On the stability of the basis pursuit in the presence of noise," *Signal Processing*, vol. 86, no. 3, pp. 511–532, 2006.

[17] M. J. Wainwright, "Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso)," *IEEE Transactions on Information Theory*, vol. 55, no. 5, pp. 2183–2202, 2009.

[18] P. Tseng, "Further results on stable recovery of sparse overcomplete representations in the presence of noise," *IEEE Transactions on Information Theory*, vol. 55, no. 2, pp. 888–899, 2009.

[19] E. J. Candès and Y. Plan, "Near-ideal model selection by ℓ_1 minimization," *Annals of Statistics*, vol. 37, no. 5A, pp. 2145–2177, 2009.