SPARSE PARAMETER IDENTIFICATION FOR STOCHASTIC SYSTEMS BASED ON L_γ REGULARIZATION*

JIAN GUO†, YING WANG†, YANLONG ZHAO†, AND JI-FENG ZHANG†‡

Abstract. This paper is concerned with the reconstruction of the zero and non-zero elements of the sparse parameter vector of stochastic systems with general observation sequences. A sparse parameter identification algorithm based on L_γ penalty with $0 < \gamma < 1$ and the residual sum of squares is proposed. Without requiring independently and identically distributed (i.i.d) and stationary conditions on the observation sequences, the proposed algorithm is proved that not only the contributing variable corresponding to the non-zero parameters can be selected out with probability converging to one, but also the estimates of the non-zero parameters have the asymptotic normality property. In order to improve the performance of the L_γ regularization method, a two-step algorithm based on adaptively weighted L_γ penalty with $0 < \gamma \leq 1$ is designed, whose set and parameter almost sure convergence are established with non-i.i.d and non-stationary observation sequences. The proposed methods are applied to the structure selection of the nonlinear autoregressive models with exogenous variables and the sparse parameter identification of the linear feedback control systems. Finally, three numerical examples are given to verify the efficiency of the theoretical results.

Key word. Stochastic system, sparse identification, L_γ penalty, asymptotic normality, strong consistency.

MSC codes. 93E03, 93E12, 93E24

1. Introduction.

The sparsity problems are occurring in many areas of scientific research and engineering practice and have attracted considerable attention in recent years. Exemplary applications involve image processing [25, 30], wireless communication [13, 27], biometrics [31], compressed sampling [3], and so on. One of the most interesting issues is the exact reconstruction of zero and non-zero elements of sparse parameter vectors. This is of great importance in engineering applications as it provides a way to implement a parsimonious model with better predictive performance and can reduce the curse of dimensionality.

Classical parameter identification is a rapidly developing field for the reconstruction of system elements and has achieved a great success in both theoretical research and practical applications [6, 26]. A series of prestigious methods have been developed, including stochastic gradient descent, stochastic approximation, least-squares (LS), least mean square, and so on. These methods are usually obtained by minimizing some criteria such as the square error between the predicted and observed signals, and have some theoretical properties, such as consistency, convergence rate, asymptotic normality, etc. However, for sparse systems, since they tend to be high-dimensional or have a limited number of samples, these classical theories and methods will no longer be valid.

In the field of statistics, a number of effective and widely used methods have emerged for sparse problems [9]. For instance, there are several classical criteria to implement variable selection, such as Akaike's information criterion (AIC) [1] and

*Submitted to the editors on September 7, 2023.

Funding: The work is supported by National Key R&D Program of China under Grant 2018YFA0703800, National Natural Science Foundation of China under Grant T2293770, 62025306, 62303452 and 12226305, CAS Project for Young Scientists in Basic Research under Grant YSBR-008 and China Postdoctoral Science Foundation under Grant 2022M720159.

†Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China and School of Mathematics Sciences, University of Chinese Academy of Sciences, Bejing 100149, China (j.guo@amss.ac.cn, wangying96@amss.ac.cn, ylzhao@amss.ac.cn, jif@iss.ac.cn).

‡Corresponding author.

Bayesian information criterion (BIC) [32]. However, they are not applicable to high-dimensional data as they may involve solving NP-Hard optimization problems. Subsequently, regularization methods are proposed and widely used as a solution to sparse problem. Typically, regularization is designed by adding a penalty term to the LS objective, where the penalty term is generally defined as a norm over the parameter space. L_0 regularization is the first regularization method applied to variable selection, which can produce the sparsest solution, but requires solving a combinatorial optimization problem, whose complexity grows exponentially with dimension. [33] proposed an alternative method called LASSO which converts the combinatorial optimization problem of variable selection into an easily solvable quadratic programming problem, but is not as sparse as L_0 regularization. Thereafter, various regularization methods such as smoothly clipped absolute deviation [9], adaptive LASSO [43], elastic net [44], etc., have become the main tools for data analysis. In addition, [18] considers the asymptotic behavior of regression estimates that minimize the sum of squared residuals plus the L_γ penalty. [37] and [38] addressed the particular importance of $L_{1/2}$ regularization in sparse modeling and obtained promising practical results in image processing, matrix filling, etc.

With the rapid development of variable selection in statistics, some of these ideas and methods have been applied to stochastic systems and control. For instance, [34] used the L_0 regularization to obtain the sparsest estimate of the parameter vector. [24] utilized L_1 regularization to identify the system parameters and predict future signals assuming that the output noise components exhibited strong seriality and cross-sectional correlation. [42] introduced an LS sparse parameter identification algorithm based on L_1 penalty with adaptive weights and proved its convergence with general observation sequences, and then [12] generalized this approach to Multivariate ARMA Systems with Exogenous Inputs. In addition, some non-convex regularization methods are also employed for stochastic systems. [11] suggested a simple numerical scheme to compute solutions with minimal L_γ norm and studied its convergence. [29] proposed a new sparse signal reconstruction algorithm based on the minimization of the squared error of a smooth L_γ $(\gamma < 1)$ norm regularization, which provided better signal reconstruction performance. [36] presented generalized shrinkage penalties with explicit proximal mappings and thus gave iterative γ-shrinkage iterative algorithms that could be implemented to accurately recover a given sparse data with a given measurement matrix. However, these papers about non-convex penalized methods, do not give theoretical results like that in [42]: whether the solutions obtained by non-convex regularization methods are still convergent in the non-stationary and non-independently and identically distributed (i.i.d) situation.

Motivation of this work. As known in the literature, the L_1 regularization method has led to remarkable progress in sparse problems. However, L_1 regularization suffers from bias, leading to a heavily biased estimate and not achieving reliable recovery with the least observations [4]. Besides, L_1 regularization may produce inconsistent selections when applied to some situations [43]. In contrast, the non-convex penalty such as L_γ $(0 < \gamma < 1)$ regularization has the advantage of improving the bias problem and has led to significant performance improvements in many applications. For instance, [19] demonstrated the very high efficiency of applying $L_{1/2}$ and $L_{2/3}$ regularization to image deconvolution. This motivates us to investigate non-convex penalties in the fields of systems and control. However, in the existing literature on the non-convex regularization, the noise is usually required to be i.i.d or there is prior knowledge of the sample probability distribution, or the observed sequences are deterministic [10]. These conditions are difficult to satisfy for stochastic systems, especially feedback control systems. Besides, it is not clear whether the estimates obtained by utilizing such an approach in sparse system identification still have the theoretical asymptotic properties.

Thus, this paper sets out to investigate the non-convex $L_\gamma(0 < \gamma < 1)$ regularization method in sparse identification problems of stochastic dynamic systems with general observation sequences and non-i.i.d noise. The main contributions of this paper are as follows:

- This paper proposes a sparse parameter identification algorithm based on the L_γ $(0<\gamma<1)$ penalty and the residual sum of squares for stochastic sparse systems with non-i.i.d and non-stationary observation sequences and non-i.i.d noise. This algorithm yields significantly better performance in terms of sparsity induction and efficiency compared to the convex penalty. In addition, the theoretical properties of this algorithm are established. Specifically, the almost sure convergence of the estimates is proven. Besides, the set convergence in probability is shown, i.e., the probability that the proposed algorithm correctly selects the non-zero elements of the unknown sparse parameter vector converges to one. Moreover, the asymptotic normality of the parameter estimates is obtained. These results incorporate the results of bridge estimate [18] and do not require additional strong irrepresentable conditions compared with LASSO [40].
- In order to improve the performance of the L_γ regularization method, motivated by [42] and [43], a two-step algorithm based on the adaptively weighted $L_\gamma(0 < \gamma \leq 1)$ penalty and the residual sum of squares is proposed. For the case of non-i.i.d and non-stationary observation sequences and non- i.i.d noise, not only is almost sure parameter convergence established, but also almost sure set convergence is achieved, i.e., this algorithm correctly selects the non-zero elements of the unknown sparse parameter vector with probability one using a finite number of observations. Moreover, this algorithm is more efficient in sparsity induction than the adaptive LASSO and the algorithm in [42] and covers their results when $\gamma = 1$.
- The proposed sparse identification algorithms in this paper are applied to two kinds of typical scenes in stochastic sparse systems with non-i.i.d observation sequences. Specifically, the proposed algorithms can efficiently select the contributing basis functions out for the Nonlinear AutoRegressive models with eXogenous variables (NARX). Furthermore, the proposed algorithm is able to accurately reconstruct the sparse parameters of the linear feedback control systems with non-i.i.d and non-stationary observation sequences and non-i.i.d noise.

The rest of this paper is organized as follows: Section 2 gives the problem formulation. Section 3 proposes the $L_\gamma(0 < \gamma < 1)$ regularization algorithm, establishes its theoretical results and compares it with related works. Section 4 gives an adaptively weighted two-step algorithm and investigates its properties. In Section 5, the proposed algorithm is applied to accomplish the structure selection of the NARX model and the sparse identification of the linear feedback control systems. In Section 6, three typical simulation examples are given to illustrate the algorithms' performance. And in Section 7, some concluding remarks and further works are provided.

Notation: Let $(\Omega, \mathcal{F}, \mathrm{P})$ be the probability space, $\omega \in \Omega$ be the sample points, and $E(\cdot)$ be the expectation operator. $\|\cdot\|_1$ and $\|\cdot\|$ denote 1-norm and 2-norm for vectors or matrices, respectively. By $\mathbb{R}$ and $\mathbb{N}_+$, we denote the sets of real numbers and positive integers, respectively. $\mathbf{I}_p$ represents a unit matrix of order p and $0_p = [0, ..., 0]^T \in \mathbb{R}^p$. Moreover, $\operatorname{sign}(\cdot)$ is defined as $\operatorname{sign}(x) = 1$, when $x \geq 0$, and $\operatorname{sign}(x) = -1$, when $x < 0$, $\operatorname{vec}(x_j)|_{j=1}^q$ means $[x_1, x_2, \dots, x_q]^T$, and for a set A, by A^c, we denote the complement of A. For any two positive sequences $\{a_k\}_{k\geq 1}$ and $\{b_k\}_{k\geq 1}$,

$a_k = O(b_k)$ means there are $c > 0$ and $k_0 \in \mathbb{N}_+$ such that $a_k \le cb_k$ for all $k \ge k_0$; $a_k = o(b_k)$ means $a_k/b_k \to 0$ as $k \to \infty$. For two random sequences $\{x_k\}$ and $\{y_k\}$, we give the following two frequently-used definitions in this paper

- $x_k = O_p(y_k)$ means that for any $\epsilon > 0$, there is a finite $M > 0$ and a finite $N > 0$ such that $\mathrm{P}\{|x_k| \ge M|y_k|\} < \epsilon$ for all $k \ge N$;
- $x_k = o_p(y_k)$ means $x_k/y_k \xrightarrow{P} 0$ as $k \to \infty$, where $\xrightarrow{P}$ means convergence in probability.

2. Problem formulation.

Consider the stochastic sparse system

$$y_{k+1} = \theta^T \varphi_k + w_{k+1}, \quad k \ge 0, \tag{2.1}$$

where $\theta = [\theta(1), \dots, \theta(p)]^T \in \mathbb{R}^p$ is the unknown p-dimensional parameter vector containing many zero values, $\varphi_k \in \mathbb{R}^p$ consisting of possibly current and past inputs and outputs, is the stochastic regressor vector, y_{k+1} and w_{k+1} are the system output and noise, respectively. Denote the zero elements set of the unknown parameter θ by $A^* = \{j : \theta(j) = 0,\ j \in \{1, \dots, p\}\}$. Suppose that there are q non-zero elements in the vector θ. Without loss of generality, we assume that $\theta(j) = 0$ for $j = q+1, \dots, p$.

Problem. The identification problem of the stochastic sparse system is to infer the zero elements A^* and to estimate the non-zero elements of the unknown parameter vector θ by using the observed data $\{y_{k+1}, \varphi_k\}_{k=1}^n$.

Before giving the sparse identification algorithm, the following assumptions are introduced.

Assumptions. Denote the family of the σ-algebras $\{\mathcal{F}_k\}$ as

$$\mathcal{F}_k = \sigma\{\varphi_0, \dots, \varphi_k, w_1, \dots, w_k\}, \quad k \ge 1,$$

the maximum and minimum eigenvalues of $\sum_{k=1}^n \varphi_k \varphi_k^T$ as $\lambda_{\max}(n)$ and $\lambda_{\min}(n)$, respectively, and the maximum eigenvalue of $E\sum_{k=1}^n \varphi_k \varphi_k^T$ as $\lambda_{E,\max}(n)$.

(A1) The noise $\{w_k, \mathcal{F}_k\}_{k \ge 1}$ is a martingale difference sequence and there is $\delta > 0$ such that $\sup_k E\left[|w_{k+1}|^{2+\delta} \mid \mathcal{F}_k\right] < \infty$, a.s.

(A2) (a) For the maximal and minimal eigenvalues of $\sum_{k=1}^n \varphi_k \varphi_k^T$, it holds

$$\frac{\log \lambda_{\max}(n)}{\lambda_{\min}(n)} \xrightarrow[n\to\infty]{} 0 \text{ a.s.}$$

(b) For each n, there is a positive number d_n such that

$$d_n \lambda_{\min}(n)^{-1} = O_p(1) \text{ and } \frac{\sqrt{\lambda_{E,\max}(n)}}{d_n} \xrightarrow[n\to\infty]{} 0.$$

Remark 2.1. In Assumption (A1), a sequence of martingale differences is broader than a sequence of independent variables, which implies a much milder restriction on sequence memory than independence and allows w_{k+1} to depend on $\mathcal{F}_k$. Many random variables, such as Gaussian random variables, uniformly distributed random variables, and so on, all satisfy this assumption. Assumptions (A2) is about the system observation sequences. Assumption (A2)(a) is the classical weakest strong convergence condition for LS [22].

3. L_γ regularization algorithm and its properties.

This section constructs a sparse identification algorithm based on $L_\gamma(0 < \gamma < 1)$ regularization and gives the corresponding theoretical properties.

3.1. L_γ regularization algorithm. We start by giving the objective function based on L_γ penalty with $0 < \gamma < 1$ and residual sum of squares:

$$J_n(\beta) = \sum_{k=1}^{n} \left(y_{k+1} - \beta^T \varphi_k\right)^2 + \lambda_n \sum_{l=1}^{p} |\beta(l)|^\gamma, \tag{3.1}$$

where λ_n is a penalty parameter and $\beta = [\beta(1), \ldots, \beta(p)]^T$.

Assumption. We first give the following assumption about the parameter λ_n.

(A3) The penalty parameter $\{\lambda_n\}$ of (3.1) satisfies that

(a) $\frac{\lambda_n}{\lambda_{\min}(n)} \xrightarrow[n\to\infty]{} 0$, a.s., (b) $\frac{\lambda_n}{\lambda_{E,\max}(n)^{1/2}} \xrightarrow[n\to\infty]{} 0$, (c) $\frac{\lambda_n d_n^{2-\gamma}}{\lambda_{E,\max}(n)^{2-\frac{1}{2}\gamma}} \xrightarrow[n\to\infty]{} \infty$.

Remark 3.1. Assumptions (A3) is about the penalty parameter λ_n. It is able to be satisfied and cover the classical persistent excitation condition as a special case, i.e., $C_1 n \le \lambda_{\min}(n) \le \lambda_{\max}(n) \le C_2 n$ for some constants C_1 and C_2. Specifically, d_n in (A2)(b) can be n and for any given $0 < \gamma < 1$, λ_n can be chosen as n^α with $\frac{1}{2}\gamma < \alpha < \frac{1}{2}$ to meet Assumption (A3).

Algorithm. The sparse identification algorithm based on L_γ penalty is designed in Algorithm 3.1. This algorithm provides a method for combining variable selection and parameter estimation in a single step.

Algorithm 3.1 L_γ regularization.

Step 0 (Initialization). For given $0 < \gamma < 1$, choose a positive sequence $\{\lambda_n\}_{n\ge 1}$ satisfying (A3).

Step 1 (Sparse Optimization with L_γ penalty) With γ and λ_n, optimize the objective function

$$J_n(\beta) = \sum_{k=1}^{n} \left(y_{k+1} - \beta^T \varphi_k\right)^2 + \lambda_n \sum_{l=1}^{p} |\beta(l)|^\gamma, \tag{3.2}$$

and obtain

$$\beta_n = [\beta_n(1), \ldots, \beta_n(p)]^T = \underset{\beta}{\operatorname{argmin}}\ J_n(\beta), \tag{3.3}$$

$$A_n^* = \{j : \beta_n(j) = 0, j \in \{1, \ldots, p\}\}. \tag{3.4}$$

Remark 3.2. We now discuss the feasibility of (3.3). First, the global minimum point of non-convex function $J_n(\beta)$ exists (not infinity). This is because $J_n(\beta)$ is continuous, there exists a minimum point on any compact set; and since $\|\beta\| \to \infty$, $J_n(\beta) \to \infty$, the point that minimizes $J_n(\beta)$ must be finite. Thus, (3.3) is a well-defined estimator. Second, we present the computation methods of (3.3). It is worth noting that the standard gradient-based method fails to solve this problem, because the penalty objective function $J_n(\beta)$ is non-differentiable when β has zero components. While, a large number of approximate algorithms and nonconvex optimization solvers have emerged to solve this problem. For instance, [37] proposed an iterative half thresholding algorithm for fast solution of $L_{1/2}$ regularization, and [20] and [29] designed solving algorithms by approximating the L_γ penalty with a function that has finite gradient at zero. In addition, genetic algorithms, particle swarm algorithms, simulated annealing algorithms, etc. can be used to solve non-convex optimization problems as well as solvers such as IPOPT [35]. However, none of the above methods provide sufficient theoretical support. Thus, the focus of this paper is not on the discussion of the solution method of (3.3), but on the properties of its solution.

Remark 3.3. The currently existing papers on the sparse identification of L_γ penalty either lack theory, as in the papers [11, 29, 36], or discuss its properties only under the i.i.d and stationary condition, as in the papers [9, 18, 38]. However, in the fields of system and control, the regressor φ_k is generally non-stationary and non-independent because any real feedback controller depends essentially on the system output and hence the driven noise [17]. The main point of interest in this paper is whether the estimates (3.3) and (3.4) remain parameter convergence, set convergence and asymptotically normality under non-stationary and non-independent conditions.

Remark 3.4. [42] proved the convergence of L_1 penalty with adaptive weights under non-stationary and non-independent assumption. While, L_γ penalty is more efficient in sparsity induction than L_1 penalty. We give an example to explain. Consider the Auto Regression with eXtra input (ARX) system: $y_{k+1} = \theta_1 y_k + \theta_2 u_k + w_{k+1}$ with the true parameters $\theta_1 = 1$ and $\theta_2 = 0$. Let $\beta = [\beta_1, \beta_2]^T$. By Lagrange's multiplier method, the regularized LS problem (3.3) is equivalent to solving:

$$\min_{\beta} \quad J(\beta) = \sum_{k=1}^{n} (y_{k+1} - \beta_1 y_k - \beta_2 u_k)^2 \qquad \text{s.t.} \qquad |\beta_1|^\gamma + |\beta_2|^\gamma \le s,$$

for some $s > 0$. Fig. 1 shows the objective function equivalence graphs of L_1 and $L_{1/2}$ penalties. The constraint region of the L_1 penalty is a square after rotation, and the constraint region of the $L_{1/2}$ penalty is a graph concave inward. The solution to this problem occurs when the contour $J(\beta)$ is first tangent to the constraint region. It can be seen that the solutions of both L_1 and $L_{1/2}$ penalties may appear at the corners, which leads to a sparse solution. This geometrically demonstrates the sparsity of $L_\gamma (0 < \gamma \le 1)$ regularization. Moreover, the solution of the $L_{1/2}$ regularized LS problem is more likely to appear at the corners, which implies that the solution of the $L_{1/2}$ regularized LS problem is sparser than L_1.

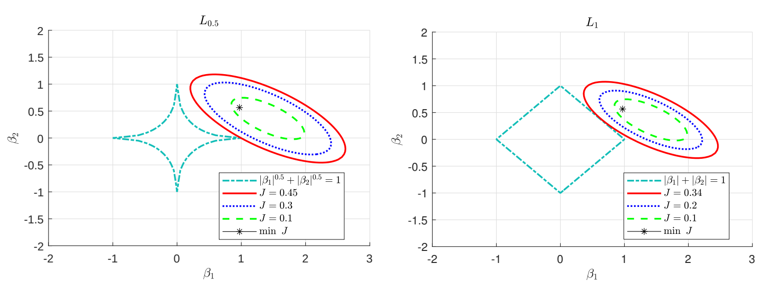

FIG. 1. *L_1 penalty v.s. $L_{1/2}$ penalty*

Remark 3.5. Now we give a way of choosing penalty parameter λ_n in Algorithm 3.1 for general cases. If $\frac{\lambda_{E,\max}(n)^{3/2-\gamma/2}}{d_n^{2-\gamma}} \xrightarrow[n\to\infty]{} 0$ and $\frac{\sqrt{\lambda_{E,\max}(n)}}{\lambda_{\min}(n)} = O(1)$ a.s., then, for any $0 < \beta < 1$, $\lambda_n = \frac{\lambda_{E,\max}(n)^{(\frac{3}{2}-\frac{1}{2}\gamma)\beta+1/2}}{d_n^{(2-\gamma)\beta}}$ satisfies Assumption (A3).

3.2. Theoretical properties. This section will give the theoretical properties of Algorithm 3.1. To prove these properties, we first give the following proposition.

PROPOSITION 3.6. [23] *For the system (2.1), if Assumptions (A1) and (A2) hold, then* $\left\|\left(\sum_{k=1}^{n}\varphi_k\varphi_k^T\right)^{-\frac{1}{2}}\sum_{k=1}^{n}\varphi_k w_{k+1}\right\| = O\left(\sqrt{\log\lambda_{\max}(n)}\right)$, a.s.

For the estimate β_n and A_n^* generated by Algorithm 3.1, the following theorem shows the almost sure convergence of the estimates.

THEOREM 3.7. *Under Assumptions (A1), (A2)(a) and (A3)(a), the estimate given by Algorithm 3.1 is almost surely convergent, i.e.,* $\lim_{n\to\infty}\beta_n = \theta$, a.s.

Proof. Noting that β_n is the minimizer of $J_n(\beta)$ in Algorithm 3.1, we have $J_n(\beta_n) \le J_n(\theta)$. Since $\lambda_n > 0$, $|\beta_n(j)|^\gamma \ge 0$, by (3.2) and a direct calculation, we have

$$\begin{aligned}\lambda_n\sum_{j=1}^{p}|\theta(j)|^\gamma &\ge \sum_{i=1}^{n}\left(y_{i+1}-\varphi_i^T\beta_n\right)^2-\sum_{i=1}^{n}\left(y_{i+1}-\varphi_i^T\theta\right)^2\\ &=(\beta_n-\theta)^T\sum_{i=1}^{n}(\varphi_i\varphi_i^T)(\beta_n-\theta)+2\sum_{i=1}^{n}\varphi_i^T(\theta-\beta_n)w_{i+1}.\end{aligned}\tag{3.5}$$

Let $P_n = \sum_{i=1}^{n}\varphi_i\varphi_i^T$, $\delta_n = P_n^{1/2}(\beta_n-\theta)$ and $Q_n = \left(\sum_{k=1}^{n}\varphi_k\varphi_k^T\right)^{-\frac{1}{2}}\sum_{k=1}^{n}\varphi_k w_{k+1}$. Then, (3.5) becomes

$$\begin{aligned}&(\beta_n-\theta)^T\sum_{i=1}^{n}(\varphi_i\varphi_i^T)(\beta_n-\theta)+2\sum_{i=1}^{n}\varphi_i^T(\theta-\beta_n)w_{i+1}\\ &= \delta_n^T\delta_n - 2\left[(\sum_{i=1}^{n}\varphi_i\varphi_i^T)^{-1/2}\sum_{i=1}^{n}\varphi_i w_{i+1}\right]^T\delta_n = \delta_n^T\delta_n - 2Q_n^T\delta_n.\end{aligned}\tag{3.6}$$

From (3.5) and (3.6) it follows that $\delta_n^T\delta_n - 2Q_n^T\delta_n - \lambda_n\sum_{j=1}^{p}|\theta(j)|^\gamma \le 0$, which implies $\|\delta_n - Q_n\|^2 - \|Q_n\|^2 - \lambda_n\sum_{j=1}^{p}|\theta(j)|^\gamma \le 0$. Hence, we have

$$\begin{aligned}\|\delta_n - Q_n\| &\le \sqrt{\|Q_n\|^2+\lambda_n\sum_{j=1}^{p}|\theta(j)|^\gamma}\\ &\le\sqrt{\|Q_n\|^2+\lambda_n\sum_{j=1}^{p}|\theta(j)|^\gamma+2\|Q_n\|(\lambda_n\sum_{j=1}^{p}|\theta(j)|^\gamma)^{1/2}} = \|Q_n\|+(\lambda_n\sum_{j=1}^{p}|\theta(j)|^\gamma)^{1/2}.\end{aligned}$$

Then, by the triangular inequality we have $\|\delta_n\| \le \|\delta_n - Q_n\| + \|Q_n\| \le 2\|Q_n\| + (\lambda_n\sum_{j=1}^{p}|\theta(j)|^\gamma)^{1/2}$. Noting Proposition 3.6 and $\lambda_n\sum_{j=1}^{p}|\theta(j)|^\gamma = O(\lambda_n)$, it follows that $\|\beta_n-\theta\| = O\left(\sqrt{\frac{\log\lambda_{\max}(n)}{\lambda_{\min}(n)}}+\sqrt{\frac{\lambda_n}{\lambda_{\min}(n)}}\right)$ a.s. By Assumptions (A2)(a) and (A3)(a), the proof is completed. ∎

Next, we discuss the set convergence in probability of the estimates, starting with the following lemma to illustrate the convergence properties in probability of the estimation error.

LEMMA 3.8. *If Assumptions (A1), (A2) and (A3)(a)-(b) hold, then*

$$\|\beta_n-\theta\| = O_p\left(\frac{\sqrt{q\lambda_{E,\max}(n)}}{d_n}\right).\tag{3.7}$$

Proof. To prove $\|\beta_n-\theta\| = O_p(\sqrt{q\lambda_{E,\max}(n)}/d_n)$, it is sufficient to prove that for any $\epsilon_1>0$, there exists a finite $\tilde{M}>0$ and N such that for any $n>N$, $\mathrm{P}(\|\beta_n-\theta\|> \tilde{M}\sqrt{q\lambda_{E,\max}(n)}/d_n)\le \epsilon_1$. Let $h_n=d_n/\sqrt{q\lambda_{E,\max}(n)}$. By the fact that $\forall\epsilon>0$,

$$\mathrm{P}\left(h_n\|\beta_n-\theta\|>\tilde{M}\right)\le \mathrm{P}\left(\|\beta_n-\theta\|\ge\epsilon/2\right)+\mathrm{P}\left(\tilde{M}/h_n<\|\beta_n-\theta\|<\epsilon/2\right), \tag{3.8}$$

we divide the proof into two steps: one is to prove $\mathrm{P}(\|\beta_n-\theta\|\ge\epsilon/2)\le\frac{\epsilon_1}{3}$ and the other is to prove $\mathrm{P}\left(\tilde{M}/h_n<\|\beta_n-\theta\|<\epsilon/2\right)\le\frac{2\epsilon_1}{3}$. Specifics are as follows.

Step 1: By Theorem 3.7, the probability $\mathrm{P}\left(\|\beta_n-\theta\|\ge\epsilon/2\right)$ converges to zero, which means for any given $\epsilon_1>0$, there is a finite $N_1\in\mathbb{N}_+$ such that for all $n>N_1$,

$$\mathrm{P}\left(\|\beta_n-\theta\|\ge\epsilon/2\right)\le\epsilon_1/3. \tag{3.9}$$

Step 2: This step is to prove $\mathrm{P}\left(\tilde{M}/h_n<\|\beta_n-\theta\|<\epsilon/2\right)\le\frac{2\epsilon_1}{3}$. For each $n\in\mathbb{N}_+$, denote $S_{j,n}=\left\{\beta:2^{j-1}<h_n\|\beta-\theta\|<2^j\right\}$ for $j\in\mathbb{Z}$. By Assumption (A2)(b), there is a finite $M_1>0$ and a finite $N_2\in\mathbb{N}_+$ such that for all $n>N_2$,

$$\mathrm{P}(\lambda_{\min}(n)\le M_1 d_n)=\mathrm{P}(d_n\lambda_{\min}(n)^{-1}\ge M_1^{-1})\le\frac{\epsilon_1}{3}. \tag{3.10}$$

Denote $A_n=\{\omega:\lambda_{\min}(n)\le M_1 d_n\}$. By the definition of $S_{j,n}$ and (3.10), we have

$$\begin{aligned}
&\mathrm{P}\left(2^M/h_n<\|\beta_n-\theta\|<\epsilon/2\right)\\
\le\ &\mathrm{P}\left(\{\omega:\beta_n\in S_{j,n},\ \forall j\ge M+1,\ 2^{j+1}\le\epsilon h_n\}\cap A_n^c\right)+\mathrm{P}(A_n)\\
\le &\sum_{j\ge M+1,2^{j+1}\le\epsilon h_n}\mathrm{P}(\{\omega:\beta_n\in S_{j,n}\}\cap A_n^c)+\frac{\epsilon_1}{3}.
\end{aligned}\tag{3.11}$$

Since β_n is the minimum of $J_n(\beta)$, for any set A containing the point β_n, we have $\inf\limits_{\beta\in A}(J_n(\beta)-J_n(\theta))\le 0$, which implies

$$\{\omega:\ \beta_n\in A\}\subset\{\omega:\ \inf_{\beta\in A}(J_n(\beta)-J_n(\theta))\le 0\}. \tag{3.12}$$

Thus, by (3.12) and (3.11) we have

$$\begin{aligned}
&\mathrm{P}\left(2^M/h_n<\|\beta_n-\theta\|<\epsilon/2\right)\\
\le\ &\frac{\epsilon_1}{3}+\sum_{j\ge M,2^j\le\epsilon h_n}\mathrm{P}\left(\{\inf_{\beta\in S_{j,n}}(J_n(\beta)-J_n(\theta))\le 0\}\cap A_n^c\right).
\end{aligned}\tag{3.13}$$

Next we consider the right hand of (3.13). Let $\beta=[\beta(1),\dots,\beta(p)]^T\in S_{j,n}$. Since $|\beta(j)|^\gamma>0$ and $\theta(j)=0$ for $j=q,q+1,\dots,p$, similar to (3.5), we have

$$\begin{aligned}
J_n(\beta)-J_n(\theta)&=(\beta-\theta)^T\sum_{i=1}^n(\varphi_i\varphi_i^T)(\beta-\theta)\\
&+2\sum_{i=1}^n\varphi_i^T(\theta-\beta)w_{i+1}+\lambda_n\sum_{j=1}^q\left[|\beta(j)|^\gamma-|\theta(j)|^\gamma\right].
\end{aligned}\tag{3.14}$$

For the first term on the right hand of (3.14), by noting $\beta\in S_{j,n}$ and (3.10), for any $\omega\in A_n^c$, it follows that

$$\begin{aligned}
&(\beta-\theta)^T\sum_{i=1}^n(\varphi_i\varphi_i^T)(\beta-\theta)\ge\lambda_{\min}(n)\|\beta-\theta\|^2\\
&\ge\lambda_{\min}(n)2^{2j-2}h_n^{-2}\ge M_1 d_n 2^{2j-2}h_n^{-2}.
\end{aligned}\tag{3.15}$$

For the third term on the right hand of (3.14), by the mean value theorem, there exists ξ_j between $\beta(j)$ and $\theta(j)$ such that

$$\lambda_n \sum_{j=1}^{q} \left||\beta(j)|^{\gamma} - |\theta(j)|^{\gamma}\right| = \lambda_n \gamma \sum_{j=1}^{q} |\xi_j|^{\gamma-1} \left|\beta(j) - \theta(j)\right|.$$

Since $\|\beta - \theta\| < \epsilon/2$, there is some constant $C_1 > 0$ such that $|\xi_j|^{\gamma-1} < C_1$. Thus,

$$\begin{aligned} \lambda_n \sum_{j=1}^{q} \left||\beta(j)|^{\gamma} - |\theta(j)|^{\gamma}\right| &\leq C_1 \lambda_n \gamma \sum_{j=1}^{q} |\beta(j) - \theta(j)| \\ &\leq C_1 \lambda_n \gamma \sqrt{q} \|\beta - \theta\| \leq C_1 \lambda_n \gamma \sqrt{q} 2^j h_n^{-1}. \end{aligned} \tag{3.16}$$

Then, it follows from (3.16) that

$$\lambda_n \sum_{j=1}^{q} \left[|\beta(j)|^{\gamma} - |\theta(j)|^{\gamma}\right] \geq -C_1 \lambda_n \gamma \sqrt{q} 2^j h_n^{-1}. \tag{3.17}$$

Hence, by (3.14), (3.15) and (3.17), we have for any $\omega \in A_n^c$,

$$J_n(\beta) - J_n(\theta) \geq M_1 d_n 2^{2j-2} h_n^{-2} - C_1 \lambda_n \gamma 2^j h_n^{-1} - \sup_{\beta \in S_{j,n}} 2 \left| \sum_{i=1}^{n} \varphi_i^T (\theta - \beta) w_{i+1} \right|. \tag{3.18}$$

When $\inf_{\beta \in S_{j,n}} (J_n(\beta) - J_n(\theta)) \leq 0$, by (3.18), for any $\omega \in A_n^c$, the following inequality holds

$$\sup_{\beta \in S_{j,n}} 2 \left| \sum_{i=1}^{n} \varphi_i^T (\theta - \beta) w_{i+1} \right| \geq M_1 d_n 2^{2j-2} h_n^{-2} - C_1 \lambda_n \gamma \sqrt{q} 2^j h_n^{-1}. \tag{3.19}$$

By Assumption (A3)(b), we have $\frac{\lambda_n \sqrt{q} 2^j h_n^{-1}}{d_n 2^{2j-2} h_n^{-2}} = \frac{\lambda_n}{2^{j-2}\sqrt{\lambda_{E,\max}(n)}} \xrightarrow[n\to\infty]{} 0$. Then, it follows that $M_1 d_n 2^{2j-2} h_n^{-2} > C_1 \lambda_n \gamma \sqrt{q} 2^j h_n^{-1}$ for all $n > N_3$ with N_3 being some positive integer. Therefore, by (3.19) and Markov inequality, we have

$$\begin{aligned} &\mathrm{P}\left(\{ \inf_{\beta \in S_{j,n}} (J_n(\beta) - J_n(\theta)) \leq 0 \} \cap A_n^c \right) \\ &\leq \mathrm{P}\left(\sup_{\beta \in S_{j,n}} 2 \left| \sum_{i=1}^{n} \varphi_i^T (\theta - \beta) w_{i+1} \right| \geq M_1 d_n 2^{2j-2} h_n^{-2} - C_1 \lambda_n \gamma \sqrt{q} 2^j h_n^{-1} \right) \\ &\leq \frac{E \sup_{\beta \in S_{j,n}} 2 \left| \sum_{i=1}^{n} \varphi_i^T (\theta - \beta) w_{i+1} \right|}{M_1 d_n 2^{2j-2} h_n^{-2} - C_1 \lambda_n \gamma \sqrt{q} 2^j h_n^{-1}}. \end{aligned} \tag{3.20}$$

In addition, by Assumption (A1), we further assume that $E\left(w_{k+1}^2 | \mathcal{F}_k\right) = \sigma_k^2 \leq \bar{\sigma}^2$ with $\bar{\sigma}$ being some constant. Then, by the definition of $S_{j,n}$, Jensen's inequality, and Cauchy-Schwarz inequality, we have

$$\begin{aligned} E \sup_{\beta \in S_{j,n}} 2 \left| \sum_{i=1}^{n} \varphi_i^T (\theta - \beta) w_{i+1} \right| &\leq 2 \sqrt{E \sup_{\beta \in S_{j,n}} \|\beta - \theta\|^2 \left\| \sum_{i=1}^{n} \varphi_i^T w_{i+1} \right\|^2} \\ &\leq 2^{j+1} h_n^{-1} \sqrt{E \left[\sum_{i=1}^{n} \varphi_i^T w_{i+1} \sum_{i=1}^{n} \varphi_i w_{i+1} \right]}. \end{aligned} \tag{3.21}$$

Noting Assumption (A1), we have

$$E\left[\sum_{i=1}^{n}\varphi_i^T w_{i+1}\sum_{i=1}^{n}\varphi_i w_{i+1}\right] = E\left[\sum_{i=1}^{n}\varphi_i^T\varphi_i w_{i+1}^2\right] = E\left[\sum_{i=1}^{n}E(\left[\varphi_i^T\varphi_i w_{i+1}^2\right]\mid\mathcal{F}_i)\right]$$

$$\le \bar{\sigma}^2 E\sum_{i=1}^{n}\varphi_i^T\varphi_i \le \bar{\sigma}^2\mathrm{tr}\left(E\sum_{i=1}^{n}\varphi_i\varphi_i^T\right) \le \bar{\sigma}^2 p\lambda_{E,\max}(n). \tag{3.22}$$

Therefore, from (3.20) and (3.22) it follows that

$$\mathrm{P}\left(\{\inf_{\beta\in S_{j,n}}(J_n(\beta)-J_n(\theta))\le 0\}\cap A_n^c\right)$$

$$\le \frac{2^{j+1}\bar{\sigma}\sqrt{p}h_n^{-1}\lambda_{E,\max}(n)^{1/2}}{M_1 d_n 2^{2j-2}h_n^{-2}-C_1\lambda_n\gamma\sqrt{q}2^j h_n^{-1}} \le \frac{2\bar{\sigma}}{M_1 2^{j-2}-\frac{C_1\lambda_n\gamma}{\lambda_{E,\max}(n)^{1/2}}}.$$

By Assumption (A3)(b), there is a finite $N_4\in\mathbb{N}_+$ such that for all $n>N_4$,

$$\mathrm{P}\left(\{\inf_{\beta\in S_{j,n}}(J_n(\beta)-J_n(\theta))\le 0\}\cap A_n^c\right)\le\frac{\bar{\sigma}}{M_1 2^{j-4}},$$

which leads to

$$\sum_{j\ge M,2^j\le\epsilon h_n}\mathrm{P}\left(\{\inf_{\beta\in S_{j,n}}(J_n(\beta)-J_n(\theta))\le 0\}\cap A_n^c\right)\le\sum_{j\ge M}\frac{\bar{\sigma}}{M_1 2^{j-4}}\le\frac{\bar{\sigma}}{M_1}2^{-(M-5)}. \tag{3.23}$$

Therefore, for the given ϵ_1, there is a finite M_2 and N_5 such that for all $n>N_5$,

$$\sum_{j\ge M_2,2^j\le\epsilon h_n}\mathrm{P}\left(\{\inf_{\beta\in S_{j,n}}(J_n(\beta)-J_n(\theta))\le 0\}\cap A_n^c\right)\le\frac{\epsilon_1}{3}. \tag{3.24}$$

Thus, from (3.8), (3.9), (3.13) and (3.24), letting $\tilde{M}=2^{M_2}$ and $N=\max\{N_1,\dots,N_5\}$, we have for all $n>N$, $\mathrm{P}\left(h_n\|\beta_n-\theta\|>\tilde{M}\right)\le\epsilon_1$. This completes the proof. ∎

Remark 3.9. From Equation (3.7), it can be obtained that the smaller the number of non-zero elements of the parameter vector θ, the faster the convergence rate and thus the better the identification performance. This is further verified by the simulation Example 1 in Section 6.

Based on Lemma 3.8, we give the following theorem demonstrating the set convergence in probability. Different from Theorem 3.7, the following theorem further states that the probability that the proposed algorithm correctly selects the non-zero elements of the unknown sparse parameter vector converges to one.

THEOREM 3.10. *Let $\beta_n=(\beta_{1n}^T,\beta_{2n}^T)^T$ with $\beta_{1n}\in\mathbb{R}^q$ and $\beta_{2n}\in\mathbb{R}^{p-q}$ being the vectors composed by the first q elements and the last $p-q$ elements of β_n, and $\theta=(\theta_{10}^T,0_{p-q}^T)^T$ with $\theta_{10}\in\mathbb{R}^q$. If Assumptions (A1)-(A3) hold, then we have the set convergence of the estimates with probability tending to one, i.e., $\lim_{n\to\infty}P(\beta_{2n}=0_{p-q})=1$.*

Proof. Let $t_n=\frac{\sqrt{\lambda_{E,\max}(n)}}{d_n}$. Denote the estimate $\beta_n=(\beta_{1n}^T,\beta_{2n}^T)^T$ as $\beta_{1n}=\theta_{10}+t_n u_{1n}$, $\beta_{2n}=t_n u_{2n}$, where $u_{1n}\in\mathbb{R}^q$ and $u_{2n}\in\mathbb{R}^{p-q}$. In addition, denote

$$\sum_{k=1}^{n}\varphi_k\varphi_k^T=\begin{bmatrix}\Phi_n^{(11)} & \Phi_n^{(12)}\\ \Phi_n^{(21)} & \Phi_n^{(22)}\end{bmatrix}\text{ and }\varphi_k=\begin{bmatrix}\varphi_k^{(1)}\\ \varphi_k^{(2)}\end{bmatrix}, \tag{3.25}$$

where $\Phi_n^{(11)} \in \mathbb{R}^{q\times q}$, $\varphi_k^{(q)} \in \mathbb{R}^d$ and others are with compatible dimensions. For any given $C > 0$, define

$$B_n = \{\omega : \ \beta_n \in \{\beta : \|\beta - \theta\| \le t_n C\}\}, \quad D_n = \{\omega : \ \beta_{2n} = 0\}. \tag{3.26}$$

Then, by (3.26) we have $\|u_{1n}\| \le C$ and $\|u_{2n}\| \le C$ for all $\omega \in B_n$. Next, we prove that for any given $\epsilon > 0$, there is $N \in \mathbb{N}_+$ such that $\mathrm{P}(D_n) \ge 1 - \epsilon$ for all $n > N$. In the following, we consider the estimate sequence $\{\beta_n\}_{n\ge1}$ on a fixed sample path $\omega \in B_n$. Direct calculation for (3.2) leads to

$$\begin{aligned} J_n(\beta_n) = &\sum_{i=1}^n w_{i+1}^2 + (\beta_n - \theta)^T \sum_{i=1}^n (\varphi_i \varphi_i^T)(\beta_n - \theta) \\ &+ 2\sum_{i=1}^n \varphi_i^T(\theta - \beta_n) w_{i+1} + \lambda_n \sum_{j=1}^p |\beta_n(j)|^\gamma . \end{aligned}$$

Then, we can obtain

$$\begin{aligned} &J_n(\theta_{10} + t_n u_{1n}, t_n u_{2n}) - J_n(\theta_{10} + t_n u_{1n}, 0) \\ = \ &t_n^2 \sum_{i=1}^n \left(\varphi_i^{(2)T} u_{2n}\right)^2 + 2t_n^2 \sum_{i=1}^n \left(\varphi_i^{(1)T} u_{1n}\right)\left(\varphi_i^{(2)T} u_{2n}\right) \\ &-2t_n \sum_{i=1}^n w_{i+1}\left(\varphi_i^{(2)^T} u_{2n}\right) + \lambda_n t_n^\gamma \sum_{j=1}^{p-q} |u_{2n}(j)|^\gamma . \end{aligned} \tag{3.27}$$

For the first two terms on the right hand of (3.27), we have

$$\begin{aligned} &t_n^2 \sum_{i=1}^n \left(\varphi_i^{(2)T} u_{2n}\right)^2 + 2t_n^2 \sum_{i=1}^n \left(\varphi_i^{(1)T} u_{1n}\right)\left(\varphi_i^{(2)T} u_{2n}\right) \\ \ge \ &t_n^2 \sum_{i=1}^n \left(\varphi_i^{(2)T} u_{2n}\right)^2 - t_n^2 \sum_{i=1}^n \left[\left(\varphi_i^{(1)T} u_{1n}\right)^2 + \left(\varphi_i^{(2)T} u_{2n}\right)^2\right] = -t_n^2 \sum_{i=1}^n \left(\varphi_i^{(1)T} u_1\right)^2 . \end{aligned} \tag{3.28}$$

By Markov inequality and noting that $\lambda_{\max}\{E\Phi_n^{(11)}\} \le \lambda_{E,\max}(n)$, for the above given ϵ, letting $M_1 = \frac{3}{\epsilon}$, we have

$$\mathrm{P}\left(t_n^2 \sum_{i=1}^n \left(\varphi_i^{(1)T} u_{1n}\right)^2 \ge M_1 \lambda_{E,\max}(n) t_n^2 C^2\right) \le \frac{\epsilon E\left(t_n^2 \sum_{i=1}^n \left(\varphi_i^{(1)T} u_{1n}\right)^2\right)}{3\lambda_{E,\max}(n) t_n^2 C^2} \le \epsilon/3. \tag{3.29}$$

Hence, it follows $\mathrm{P}(E_n^c) \le \epsilon/3$, where E_n is denoted as

$$E_n = \left\{\omega : t_n^2 \sum_{i=1}^n \left(\varphi_i^{(1)T} u_1\right)^2 \le M_1 t_n^2 C^2 \lambda_{E,\max}(n)\right\}. \tag{3.30}$$

For the third term on the right hand of (3.27), similar to (3.21) and (3.22), noting that $\lambda_{\max}\{E\Phi_n^{(22)}\} \le \lambda_{E,\max}(n)$ and $\|u_{2n}\| \le C$, we can get

$$\begin{aligned} &E\left|\sum_{i=1}^n w_{i+1}\left(\varphi_i^{(2)T} u_{2n}\right)\right| \le \left(E\left|\sum_{i=1}^n w_{i+1}\left(\varphi_i^{(2)T} u_{2n}\right)\right|^2\right)^{1/2} \\ &\le C\bar{\sigma}\lambda_{\max}\{E\Phi_n^{(22)}\}^{1/2} \le C\bar{\sigma}\sqrt{\lambda_{E,\max}(n)}, \end{aligned} \tag{3.31}$$

where $\mathbb{E}\left(w_{k+1}^2|\mathcal{F}_k\right) \le \bar{\sigma}^2$ with $\bar{\sigma}$ being some constant by Assumption (A1). By Markov inequality, for the above given ϵ, letting $M_2 = \frac{3C\bar{\sigma}}{\epsilon}$, it follows from (3.31) that

$$\text{(3.32)}\quad \mathrm{P}\left(\left|\sum_{i=1}^{n} w_{i+1}\left(\varphi_i^{(2)T}u_{2n}\right)\right| \ge M_2\sqrt{\lambda_{E,\max}(n)}\right) \le \frac{E\left|\sum_{i=1}^{n} w_{i+1}\left(\varphi_i^{(2)T}u_{2n}\right)\right|}{M_2\sqrt{\lambda_{E,\max}(n)}} \le \epsilon/3.$$

Denote

$$F_n = \left\{\omega : -\sum_{i=1}^{n} w_{i+1}\left(\varphi_i^{(2)^T}u_{2n}\right) \ge -M_2\lambda_{E,\max}^{1/2}(n)\right\}. \tag{3.33}$$

Then, from (3.32) it follows $\mathrm{P}(F_n^c) \le \epsilon/3$. For the last term on the right hand of (3.27), noting that $\left[\sum_{j=1}^{p-q}|u_{2n}(j)|^\gamma\right]^{2/\gamma} \ge \sum_{j=1}^{p-q}|u_{2n}(j)|^2 = \|u_{2n}\|^2$, we have

$$\lambda_n t_n^\gamma \sum_{j=1}^{p-q}|u_{2n}(j)|^\gamma \ge \|u_{2n}\|^\gamma \lambda_n t_n^\gamma. \tag{3.34}$$

For all $\omega \in E_n \cap F_n$, from (3.28), (3.30), (3.33) and (3.34), we have

$$\begin{aligned} &J_n(\theta_{10}+t_nu_{1n}, t_nu_{2n}) - J_n(\theta_{10}+t_nu_{1n}, 0) \ge \\ &-M_1t_n^2C^2\lambda_{E,\max}(n) + \|u_{2n}\|^\gamma\lambda_nt_n^\gamma - 2t_nM_2\lambda_{E,\max}^{1/2}(n). \end{aligned} \tag{3.35}$$

By Assumption (A3)(c), and noting that $\lambda_{E,\max}(n)/d_n \nrightarrow 0$, we have

$$\begin{aligned} \frac{\lambda_n t_n^\gamma}{t_n^2\lambda_{E,\max}(n)} &= \frac{\lambda_n d_n^{2-\gamma}}{\lambda_{E,\max}(n)^{2-\frac{1}{2}\gamma}} \xrightarrow[n\to\infty]{} \infty, \\ \frac{\lambda_n t_n^\gamma}{t_n\sqrt{\lambda_{E,\max}(n)}} &= \frac{\lambda_n d_n^{2-\gamma}}{\lambda_{E,\max}(n)^{2-\frac{1}{2}\gamma}}\frac{\lambda_{E,\max}(n)}{d_n} \xrightarrow[n\to\infty]{} \infty. \end{aligned}$$

Therefore, from (3.35), if $\|u_{2n}\| > 0$, then there is a finite $\tilde{N} \in \mathbb{N}_+$ such that $J_n(\beta_n) - J_n(\theta_{10}+t_nu_{1n}, 0) > 0,\ \forall n > \tilde{N}$, which contradicts $\beta_n = \underset{\beta}{\operatorname{argmin}}\ J_n(\beta)$. Thus, for any $\omega \in E_n \cap F_n$, there is a finite N_1 such that $\beta_{2n} = t_nu_{2n} = 0,\ \forall n > N_1$. Therefore, from (3.26) it follows $B_n \cap E_n \cap F_n \subset D_n \cap E_n \cap F_n,\ \forall n > N_1$. In addition, by Lemma 3.8, for the above given ϵ, there is an $N_2 \in \mathbb{N}_+$ such that $P(B_n^c) \le \epsilon/3$ for all $n > N_2$. Hence, combing the results above (3.30) and below (3.33), and letting $N = \max\{N_1, N_2\}$, we have that for all $n > N$,

$$\begin{aligned} P(\beta_{2n} = 0) \;=\; P(D_n) &\ge P(D_n \cap E_n \cap F_n) = 1 - P(D_n^c \cup E_n^c \cup F_n^c) \\ &\ge 1 - P(B_n^c) - P(E_n^c) - P(F_n^c) \ge 1 - \epsilon. \end{aligned}$$

This completes the proof. ∎

Using the central limit theorem, we immediately give the asymptotic normality of the estimated non-zero parameters below.

THEOREM 3.11. *Assume for each n that there is a non-random positive definite symmetric matrix R_n such that*

$$R_n^{-1}\Phi_n^{(11)} \xrightarrow{P} \mathbf{I}_p,\ \max_{1\le k\le n}\|R_n^{-1/2}\varphi_k^{(1)}\| \xrightarrow{P} 0,\ \textit{and} \tag{3.36}$$

$$\lim_{k\to\infty}\mathrm{E}(w_{k+1}^2\,|\,\mathcal{F}_k) = \sigma^2,\ \text{a.s. } \textit{for some constant } \sigma, \tag{3.37}$$

where $\varphi_k^{(1)}$ and $\Phi_n^{(11)}$ are defined in (3.25). Denote the estimate $\beta_n = (\beta_{1n}^T, \beta_{2n}^T)^T$ and $\theta = [\theta_{10}^T, 0_{p-q}^T]^T$. For any non-random $\alpha_n \in \mathbb{R}^q$ satisfying $\|\alpha_n\| \le 1$, let $s_n^2 = \sigma^2 \alpha_n^T R_n^{-1} \alpha_n$. If Assumptions (A1)-(A3) hold, then

$$s_n^{-1}\alpha_n^T(\beta_{1n}-\theta_{10}) = s_n^{-1}\sum_{k=1}^{n}\alpha_n^T\left(\Phi_n^{(11)}\right)^{-1}\varphi_k^{(1)}w_{k+1} + o_p(1) \xrightarrow{\mathrm{d}} N(0,1), \tag{3.38}$$

where $\xrightarrow{\mathrm{d}}$ denotes convergence in distribution and $N(0,1)$ denotes the standard normal distribution.

Remark 3.12. The existence of a non-random matrix R_n satisfying conditions (3.36) in Theorem 3.11 can be regarded as an stability assumption of the matrix $\Phi_n^{(11)}$. Moreover, this assumption is necessary for asymptotic normality and one can refer to Example 3 in [21] in which asymptotic normality fails to hold in the absence of (3.36). Besides, R_n can be selected as $\Phi_n^{(11)}$ if $\{\varphi_k^{(1)}\}$ is determined sequence; R_n can be selected as $nE\varphi_n^{(1)}\varphi_n^{(1)T}$ if $\varphi_n^{(1)}\varphi_n^{(1)T}$ is a stationary and ergodic random sequence with positive covariance matrix [39].

Proof. Denote $J_n(\beta)$ in (3.2) as $J_n(\beta) = J_n(\beta_1, \beta_2)$ with $\beta_1 \in \mathbb{R}^q$. By Theorem 3.7, we have $\|\beta_n - \theta\| \xrightarrow[n\to\infty]{} 0$ a.s. Since each component of θ_{10} is not equal to zero, when n is sufficiently large, each element of β_{1n} stays away from zero. Noting that the estimate $\beta_n = (\beta_{1n}^T, \beta_{2n}^T)^T$ is the minimum of $J_n(\beta)$, when n is sufficiently large, we have $\frac{\partial}{\partial \beta_1} J_n(\beta_{1n}, \beta_{2n}) = 0$, which implies

$$-2\sum_{k=1}^{n}\left(y_{k+1}-\beta_{1n}^T\varphi_k^{(1)}-\beta_{2n}^T\varphi_k^{(2)}\right)\varphi_k^{(1)}+\lambda_n\gamma\mathrm{vec}\left(\mathrm{sign}(\beta_{1n}(j))|\beta_{1n}(j)|^{\gamma-1}\right)\Big|_{j=1}^{q}=0. \tag{3.39}$$

From (2.1) and noting that $\theta = [\theta_{10}^T, 0_{1\times(p-q)}]^T$, it follows $y_{k+1} - \theta_{10}^T\varphi_k^{(1)} = w_{k+1}$. Then, by (3.39) we get

$$\begin{aligned}\sum_{k=1}^{n}\varphi_k^{(1)}\varphi_k^{(1)T}(\beta_{1n}-\theta_{10}) \\ = -\sum_{k=1}^{n}\beta_{2n}^T\varphi_k^{(2)}\varphi_k^{(1)} + \sum_{k=1}^{n}\varphi_k^{(1)}w_{k+1} - \frac{1}{2}\lambda_n\gamma\mathrm{vec}\left(\mathrm{sign}(\beta_{1n}(j))|\beta_{1n}(j)|^{\gamma-1}\right)\Big|_{j=1}^{q}.\end{aligned} \tag{3.40}$$

Thus, direct calculation from (3.40) leads to

$$\begin{aligned}s_n^{-1}\alpha_n^T(\beta_{1n}-\theta_{10}) = -s_n^{-1}\alpha_n^T\left(\Phi_n^{(11)}\right)^{-1}\sum_{k=1}^{n}\beta_{2n}^T\varphi_k^{(2)}\varphi_k^{(1)} + s_n^{-1}\sum_{k=1}^{n}\alpha_n^T\left(\Phi_n^{(11)}\right)^{-1}\varphi_k^{(1)}w_{k+1} \\ -\frac{1}{2}\lambda_n\gamma s_n^{-1}\alpha_n^T\left(\Phi_n^{(11)}\right)^{-1}\mathrm{vec}\left(\mathrm{sign}(\beta_{1n}(j))|\beta_{1n}(j)|^{\gamma-1}\right)\Big|_{j=1}^{q}.\end{aligned} \tag{3.41}$$

For the first term on the right hand of (3.41), by Theorem 3.10 that $\lim_{n\to\infty} P(\beta_{2n} = 0) = 1$, we have

$$\lim_{n\to\infty}\mathrm{P}\left(s_n^{-1}\alpha_n^T\left(\Phi_n^{(11)}\right)^{-1}\sum_{k=1}^{n}\beta_{2n}^T\varphi_k^{(2)}\varphi_k^{(1)}=0\right)=1. \tag{3.42}$$

For the last term on the right hand of (3.41), since $\beta_{1n} \to \theta_{10}$, there is a constant C such that $|\beta_{1n}(j)| \le C$ for $j = 1, \ldots, q$. By Assumption (A3)(a), we have

$$\left| \lambda_n \alpha_n^T \left(\Phi_n^{(11)}\right)^{-1} \mathrm{vec}\left(\mathrm{sign}(\beta_{1n}(j))|\beta_{1n}(j)|^{\gamma-1}\right)\Big|_{j=1}^{q} \right|$$

$$\le \lambda_n \lambda_{\min}(n)^{-1} q^{1/2} C^{\gamma-1} = o_p(1), \tag{3.43}$$

which together with (3.41) and (3.42) gives

$$s_n^{-1}\alpha_n^T(\beta_{1n}-\theta_{10}) = s_n^{-1}\sum_{k=1}^{n}\alpha_n^T\left(\Phi_n^{(11)}\right)^{-1}\varphi_k^{(1)} w_{k+1} + o_p(1). \tag{3.44}$$

In view of (3.36) and (3.44), to prove (3.38), we need only to show that

$$s_n^{-1}\sum_{k=1}^{n}\alpha_n^T R_n^{-1}\varphi_k^{(1)} w_{k+1} \xrightarrow{\mathrm{d}} N(0,1). \tag{3.45}$$

Similar to [21], the desired conclusion (3.45) can be obtained by making use of a martingale central limit theorem of [7]. ∎

3.3. Comparison of Algorithm 3.1 with related methods. In this part, we compare the sparse identification Algorithm 3.1 with Information Criterion-based variable selection [1, 32], LASSO [33], and bridge estimate [18].

Comparison with variable selection and order estimation based on information criterion. The variable selection problem aims to select a subset of relevant variables used in model construction. The usual approach is to select the optimal one from a set of reasonable models under some importance criteria, many of which contain measures of accuracy and the penalized term by the number of selected variables, for instance, AIC [1] and BIC [32] for stationary time series. Algorithm 3.1 in this paper not only fulfills the task of variable selection but also estimates the parameters corresponding to the selected variables. Moreover, compared with order estimation methods for stochastic systems such as control information criterion (CIC) [5, 15], the algorithm in this paper solves the problem as well, and furthermore, non-contributing variables within the order can also be selected out.

Comparison with LASSO and bridge estimate. Compared with the LASSO, Algorithm 3.1 does not require additional conditions; and compared with the bridge estimate, Algorithm 3.1 can be applied to general observations. In a typical setup, the sparsity problem can be described as follows [37]: Given a $n \times p$ matrix Ψ_n, and a procedure of generating an observation such as

$$Y = \Psi_n \theta + W$$

with $Y = [y_1, \ldots, y_n]^T$, $\Psi_n = [\varphi_0, \ldots, \varphi_{n-1}]^T$ and $W = [w_1, \ldots, w_n]$, we are asked to recover θ from the observation Y such that θ is of the sparsest structure. The problem can be solved by the following regularization method:

$$\min_{\theta\in\mathbb{R}^p}\left\{\|Y-\Psi_n\theta\|^2 + \lambda_n\|\theta\|_\nu^\nu\right\},$$

where $\nu > 0$ and $\|x\|_\nu$ is defined by $\|\theta\|_\nu = \sqrt[\nu]{\sum_{i=1}^{p}|\theta(i)|^\nu}$. The LASSO (for $\nu = 1$), the bridge estimate (for $\nu > 0$), and the Algorithm 3.1 (for $0 < \nu < 1$) in this paper all fall into this category, but Ψ_n in Algorithm 3.1 can be stochastic, whereas the others are deterministic. We then compare the application scope of these three algorithms. For LASSO, denote

$$\Phi_n = \sum_{k=1}^{n} \varphi_k \varphi_k^T = \begin{bmatrix} \Phi_n^{11} & \Phi_n^{12} \\ \Phi_n^{21} & \Phi_n^{22} \end{bmatrix}$$

with $\Phi_n^{11} \in \mathbb{R}^{q\times q}$ and $\beta_n = (\beta_{1n}^T, \beta_{2n}^T)^T$. [40] gave sufficient conditions for the set convergence in probability of the LASSO estimate: (a) $\frac{1}{n}\Phi_n \to \Phi$ with Φ being a positive definite matrix; (b) the following strong irrepresentable condition holds:

$$\left|\Phi_n^{21}\left(\Phi_n^{11}\right)^{-1} \operatorname{sign}\left(\beta_{1n}\right)\right| \leq \mathbf{1}_{p-q} - \eta,$$

with $\mathbf{1}_{p-q}$ being a $(p-q)\times 1$ vector of 1's, $\eta > 0$ and the inequality holding element-wise; and (c) λ_n is chosen as $\lambda_n = n^{\alpha}$ with $\frac{1}{2} < \alpha < 1$. Algorithm 3.1 of this paper can also reach set convergence while covering condition (a) as a special case without requiring the strong irrepresentable condition (b).

For bridge estimate, the conditions for the consistency of the estimates given by [18] are: (a) $\frac{1}{n}\Phi_n \to \Phi$ with Φ being a positive definite matrix; (b) $\lambda_n n^{-1/2} \to 0$ and $\lambda_n^2 n^{-\gamma} \to 0$. This result is consistent with the result of Algorithm 3.1 when $C_1 n \leq \lambda_{\min}(n) \leq \lambda_{\max}(n) \leq C_2 n$ for some constants C_1 and C_2. In addition, the theoretical results of Algorithm 3.1 go further and can be adapted to non-persistent excitation cases, in particular, the data series $\{\varphi_k, y_{k+1}\}_{k\geq 1}$ can be generated by feedback control where φ_k may be stochastic.

4. Weighted L_γ regularization algorithm and its properties. LASSO is a popular technique for simultaneous estimation and variable selection. However, in some cases, LASSO is inconsistent for variable selection. [28] showed the conflict between the optimal prediction and consistent variable selection in LASSO. To address this issue, [43] proposed a new version of the LASSO, the adaptive LASSO, in which adaptive weights were used to penalize different parameters in the L_1 penalty. [42] extended this result to general observation cases. Inspired by the improvement of the convergence properties of LASSO with this technique, in order to extend the scope of application and improve the performance of the L_γ penalty, in this section, we present a two-step algorithm with adaptively weighted $L_\gamma (0 < \gamma \leq 1)$ penalty term. The algorithm is more broadly applicable and has better convergence properties.

4.1. Weighted L_γ regularization algorithm.

Assumption. Given constants γ and μ satisfying $0 < \gamma \leq 1$ and $\mu > 0$. To proceed, we first introduce the assumptions to be used for the theoretical analysis of the weighted L_γ regularization algorithm.

(B1) For the maximal and minimal eigenvalues of $\sum_{k=1}^{n}\varphi_k\varphi_k^T$ and the positive sequence $\{\lambda_n\}_{n\geq 1}$, it holds,

$$(a) \left(\frac{\log\lambda_{\max}(n)}{\lambda_{\min}(n)}\right)^{1-\frac{\gamma}{2}+\frac{\mu}{2}} \frac{\lambda_{\max}(n)}{\lambda_n} \xrightarrow[n\to\infty]{} 0 \text{ a.s.}$$

$$(b) \frac{\log\lambda_{\max}(n)^{\frac{\mu}{2}}}{\lambda_{\min}(n)^{1-\frac{\gamma}{2}+\frac{\mu}{2}}} \frac{\lambda_{\max}(n)}{\lambda_n^{\frac{\gamma}{2}}} \xrightarrow[n\to\infty]{} 0 \text{ a.s. } (c) \frac{\log\lambda_{\max}(n)^{\frac{1}{2}+\frac{\mu}{2}}}{\lambda_{\min}(n)^{\frac{1}{2}-\frac{\gamma}{2}+\frac{\mu}{2}}} \frac{\lambda_{\max}(n)^{\frac{1}{2}}}{\lambda_n^{\frac{1}{2}+\frac{\gamma}{2}}} \xrightarrow[n\to\infty]{} 0 \text{ a.s.}$$

The adaptive sparse identification algorithm is proposed in Algorithm 4.1.

Algorithm 4.1 Weighted L_γ regularization.

Step 0 (Initialization). For given $0 < \gamma \leq 1$ and $\mu > 0$, choose a positive sequence $\{\lambda_n\}_{n\geq 1}$ satisfying Assumption (B1).

Step 1 (LS Estimation). Based on $\{y_{k+1}, \varphi_k\}_{k=1}^n$, compute the estimator:

$$\theta_{n+1} = \left(\sum_{k=1}^n \varphi_k \varphi_k^T\right)^{-1} \left(\sum_{k=1}^n \varphi_k y_{k+1}\right).$$

Let $\theta_{n+1} = [\theta_{n+1}(1), \ldots, \theta_{n+1}(p)]^T$, and for $1 \leq j \leq p$, define

$$\widehat{\theta}_{n+1}(j) = \theta_{n+1}(j) + \operatorname{sign}(\theta_{n+1}(j))\sqrt{\frac{\log \lambda_{\max}(n)}{\lambda_{\min}(n)}}.$$

Step 2 (Sparse Optimization with L_γ penalty). With λ_n and $\widehat{\theta}_{n+1}$, optimize the objective function $\widehat{J}_n(\beta) = \sum_{k=1}^n \left(y_{k+1} - \beta^T \varphi_k\right)^2 + \lambda_n \sum_{j=1}^p \frac{1}{|\widehat{\theta}_{n+1}(j)|^\mu} |\beta(j)|^\gamma$ and obtain

$$\widehat{\beta}_n = \left[\widehat{\beta}_n(1), \ldots, \widehat{\beta}_n(p)\right]^T = \underset{\beta}{\operatorname{argmin}}\ \widehat{J}_n(\beta) \tag{4.1}$$

$$\widehat{A}_n^* = \left\{j = 1, \ldots, p \mid \widehat{\beta}_n(j) = 0\right\}. \tag{4.2}$$

Remark 4.1. We discuss the choice of λ_n in the Algorithm 4.1. If we assume $\frac{\lambda_{\max}(n)}{\lambda_{\min}(n)} \left(\frac{\log \lambda_{\max}(n)}{\lambda_{\min}(n)}\right)^{\mu/2} \to 0$, a.s., then Assumption (B1) can be simplified to $\lambda_n = o(\lambda_{\min}(n))$ and $\lambda_{\max}(n)\left(\frac{\log \lambda_{\max}(n)}{\lambda_{\min}(n)}\right)^{\frac{\mu}{2}} = o(\lambda_n)$. Denote $a_n = \lambda_{\max}(n)\left(\frac{\log \lambda_{\max}(n)}{\lambda_{\min}(n)}\right)^{\frac{\mu}{2}}$ and $b_n = \lambda_{\min}(n)$. Then, λ_n can be chosen as $\lambda_n = a_n^\eta b_n^{1-\eta}$ for any fixed $\eta \in (0, 1)$ satisfying Assumption (B1). Specifically, by noting that $\frac{a_n}{b_n} = \frac{\lambda_{\max}(n)}{\lambda_{\min}(n)} \left(\frac{\log \lambda_{\max}(n)}{\lambda_{\min}(n)}\right)^{\mu/2} \to 0$ a.s., it follows that $\frac{\lambda_n}{b_n} = \left(\frac{a_n}{b_n}\right)^\eta \to 0$, and $\frac{a_n}{\lambda_n} = \left(\frac{a_n}{b_n}\right)^{1-\eta} \to 0$ a.s.

4.2. Theoretical properties. Recall that the parameter vector is assumed $\theta = [\theta(1), \ldots, \theta(q), \theta(q+1), \ldots, \theta(p)]^T$ with $\theta(i) \neq 0$ for $i = 1, \ldots, q$, and $\theta(j) = 0$ for $j = q+1, \ldots, p$. For the estimate $\widehat{\beta}_n$ and $\widehat{A}_n^*$ generated by Algorithm 4.1, the almost sure convergence of $\widehat{\beta}_n$ and the almost sure set convergence of $\widehat{A}_n^*$ are given in the following theorems.

THEOREM 4.2. *If Assumptions (A1), (A2)(a) and (A3)(a) hold, then*

$$\lim_{n\to\infty} \widehat{\beta}_n(j) = \theta(j),\ j = 1, \ldots, q,\ \text{a.s.}$$

Proof. The proof is similar to that of Theorem 3.7, and so, omitted here.

THEOREM 4.3. *If Assumptions (A1), (A3)(a) and (B1) hold, then there is an ω-space Ω_0 satisfying $P(\Omega_0) = 1$ and for any $\omega \in \Omega_0$, there is an integer $N_0(\omega)$ such that $\widehat{A}_n^* = A^*$ for all $n \geq N_0(\omega)$.*

Proof. Combining the proof of Theorem 3.10 with the proof of Lemma 4 in [42] yields the theorem. ∎

4.3. Comparison of Algorithms 4.1 with related methods. Noting Remark 3.4, Algorithm 4.1 is more likely to produce sparse solutions than the algorithm in [42] and adaptive LASSO [43]. Moreover, Algorithm 4.1 covers the results of the adaptive LASSO and the algorithm in [42]. Specifically, when $\mu = \gamma = 1$, by Remark 4.1, Assumption (B1) is degenerated to $\frac{\lambda_{\max}(n)}{\lambda_{\min}(n)}\sqrt{\frac{\log \lambda_{\max}(n)}{\lambda_{\min}(n)}} \to 0$ a.s., which is consistent with Assumption (A3) in [42]. If one further assumes that $\lambda_{\max}(n) = O(n)$ and $\lambda_{\min}(n) = O(n)$ a.s., then the result is consistent with the adaptive LASSO.

5. Application to typical scences.

5.1. Structure selection for a class of NARX models. This section applies Algorithm 3.1 to the structure selection of the NARX models with finite basis functions. One class of NARX models [41] is the kernel regression model:

$$y_{k+1} = \theta_N^T \varphi_{N,k} + w_{k+1}, \tag{5.1}$$

where y_{k+1} is the output, $\varphi_{N,k} = [\varphi_1(x(k)), \ldots, \varphi_m(x(k))]^T$ is the non-linear basis functions, $x(k)$ contains all past and current variables, $\theta_N = [c_1, \ldots, c_m]^T$ is the corresponding coefficient, $w_{k+1} \in \mathbb{R}$ is the noise and m is the number of basis functions. The objective of the NARX model structure selection is to select the contributing components from a large number of non-linear basis functions. Algorithm 3.1 can be applied directly to the model (5.1), and is more efficient in reducing the model size. Now we consider a special class of the NARX model: Hammerstein system as an example and give the corresponding theoretical results. The Hammerstein model consists of a static single-valued nonlinear element followed by a linear dynamic element, and can be described as:

$$\begin{aligned} y_{k+1} &= a_1 y_k + \cdots + a_{n_y} y_{k+1-n_y} + b_1 f(u_k) + \cdots + b_{n_u} f(u_{k+1-n_u}) + w_{k+1}, \\ f(u_k) &= \sum_{j=1}^{s} d_j g_j(u_k), \end{aligned} \tag{5.2}$$

where $\{g_j(\cdot)\}_{j=1}^s$ are the basis functions. The system (5.2) can be rewritten as the form (5.1) by denoting

$$\begin{aligned} \theta_N &= \left[a_1, \ldots, a_{n_y}, (b_1 d_1), \ldots, (b_1 d_s), \ldots, (b_{n_u} d_1), \ldots, (b_{n_u} d_s)\right]^T, \\ \varphi_{N,k} &= \left[y_k, \ldots, y_{k+1-n_y}, g_1(u_k), \ldots, g_s(u_k), \ldots, g_1(u_{k+1-n_u}), \ldots, g_s(u_{k+1-n_u})\right]^T. \end{aligned}$$

Problem. The structure selection problem of the Hammerstein system (5.2) is to select the contributing basis functions from the candidate full basis functions $\{g_j(\cdot)\}_{j=1}^s$ using the observed data $\{y_{k+1}, \varphi_{N,k}\}_{k=1}^n$.

Before presenting the results, we first give the following assumptions and the corresponding proposition.

(C1) $A(z) = 1 - a_1 z - \cdots - a_{n_y} z^{n_y}$ is stable and $b_1^2 + \cdots + b_{n_u}^2 \neq 0$;

(C2) There is an interval $[a, b]$ such that $\{1, g_1(x), \ldots, g_s(x)\}$ is linearly independent;

(C3) The sequence $\{u_k\}_{k\geq 1}$ is i.i.d, independent of the noise $\{w_k\}_{k\geq 1}$, whose density function is positive and continuous on $[a, b]$ and $0 < \mathbb{E} g_i^2(u_k) < \infty$ for $1 \leq i \leq s$.

PROPOSITION 5.1. [41] *If the Hammerstein system (5.2) satisfies Assumptions (A1) and (C1)-(C3), then with* $0 < c_1 < c_2,\ 0 < c_3 < c_4$*, we have*

$$c_1 n \le \lambda_{\max}\left\{\sum_{k=1}^{n} \varphi_{N,k}\varphi_{N,k}^T\right\} \le c_2 n,\ \ c_3 n \le \lambda_{\min}\left\{\sum_{k=1}^{n} \varphi_{N,k}\varphi_{N,k}^T\right\} \le c_4 n, \tag{5.3}$$

By use of Algorithm 3.1, we give the sparse estimate $\beta_{N,n}$ for the parameters in the system (5.2). Denote $\beta_{N n} = [a_{1 n},...,a_{n_y n}, (b_1 d_1)_n,...,(b_1 d_s)_n,...,(b_{n_u} d_1)_n,...,(b_{n_u} d_s)_n]^T$ and define $\chi = [\chi(1), \dots, \chi(s)]^T$ with $\chi(l) = \sum_{i=1}^{n_u} (b_i d_l)^2$. Then, the estimate of χ can be obtained by $\chi_n = [\chi_n(1), \dots, \chi_n(s)]^T$ with $\chi_n(l) = \sum_{i=1}^{n_u} (b_i d_l)_n^2$. Moreover, denote $D^* = \{l : d_l = 0,\ \text{for } l = 1, \dots, s\}$, $D_n^* = \{l : \chi_n(l) = 0, \text{ for } l = 1, \dots, s\}$. Assumption (C1) guarantees that $\{l : \chi(l) = 0\} = D^*$, which implies D_n^* can be regraded as an estimate of D^*. Then, we give the theoretical results for the structure selection of the contributing basis functions in $\{g_j(\cdot)\}_{j=1}^s$.

THEOREM 5.2. *Take* $\lambda_n = n^\alpha$ *with* $\frac{1}{2}\gamma < \alpha < \frac{1}{2}$. *If Assumptions (A1) and (C1)-(C3) hold for the Hammerstein system (5.2), then* $\lim_{n\to\infty} P(D_n^* = D^*) = 1$.

Proof. From (5.3) in Proposition 5.1, we have that $\lambda_{\max}\left\{\sum_{k=1}^{n} \varphi_{N,k}\varphi_{N,k}^T\right\} = O(n)$, $\lambda_{\min}\left\{\sum_{k=1}^{n} \varphi_{N,k}\varphi_{N,k}^T\right\} = O(n)$ and $E\lambda_{\max}\left\{\sum_{k=1}^{n} \varphi_{N,k}\varphi_{N,k}^T\right\} = O(n)$. Moreover, we can choose $d_n = c_5 n$ with $c_5 > 0$. Thus, noticing $\lambda_n = n^\alpha$ with $\frac{1}{2}\gamma < \alpha < \frac{1}{2}$, we can verify that (A2)-(A3) hold for the regression model (5.1)-(5.2). Thus, by Theorem 3.10, the results follow directly. ∎

5.2. Sparse identification of linear feedback control systems. This section applies Algorithm 3.1 to the sparse identification of the closed-loop systems using the self-tuning regulator (STR) control. Recall that the regressor is generally non-stationary and non-independent for linear feedback control systems [17]. The classical STR control, first proposed in [2], consists of an LS estimation algorithm for a linear stochastic dynamic system coupled online with a "least variance" control law. The goal of STR is to minimize the tracking error of the system with unknown parameters. Consider the following sparse ARX system:

$$y_{k+1} = a_1 y_k + \cdots + a_{n_y} y_{k+1-n_y} + b_1 u_k + \cdots + b_{n_u} u_{k+1-n_u} + w_{k+1}. \tag{5.4}$$

where $y_{k+1} \in \mathbb{R}$ is the system output, $w_{k+1} \in \mathbb{R}$ is the system noise, $u_k \in \mathbb{R}$ is the feedback control, and $a_1, \dots, a_{n_y}$ and $b_1, \dots, b_{n_u}$ are the unknown sparse parameters. Denote

$$\begin{aligned} &A(z) = 1 - a_1 z - \cdots - a_{n_y} z^{n_y},\ B(z) = b_1 + b_2 z + \cdots + b_{n_u} z^{n_u - 1}, \\ &\theta = \left[a_1, \dots, a_{n_y}, b_1, \dots, b_{n_u}\right]^T,\ \varphi_k = \left[y_k, \dots, y_{k+1-n_y}, u_k, \dots, u_{k+1-n_u}\right]^T. \end{aligned}$$

Let $\{y_k^*\}$ be the deterministic bounded reference signal or regulation signal. For the system (5.4), two problems need to be solved: first, to use the STR control to make the closed-loop system track the reference signal $\{y_k^*\}$; second, to select the zero parameters accurately and estimate the non-zero parameters asymptotically under the STR control.

For the control step, let the LS parameter estimate for the system be $\theta_{L,n} = [a_{1,n}, \dots, a_{n_y,n}, b_{1,n}, \dots, b_{n_u,n}]^T$. The Certainty Equivalence Principle [2] suggests an adaptive control defined as

$$u_k^0 = \frac{1}{b_{1,k}}\left\{y_{k+1}^* + \left(b_{1,k} u_k - \theta_{L,k}^T \varphi_k\right)\right\}. \tag{5.5}$$

For the identification step, it is generally necessary to impose excitation conditions on the system, and thus, the control design (5.5) needs to be modified. Specifically, in order not to make the system worse after applying the excitation, the diminishing excitation technique is introduced and a zero-trending perturbation [14] is added to the control (5.5), i.e.,

$$u_k = u_k^0 + \frac{\nu_k}{r_{k-1}^{\bar{\varepsilon}/2}}, \quad k \geq 1, \tag{5.6}$$

where $\{\nu_k\}$ is an i.i.d and bounded stochastic sequence satisfying $E(\nu_k) = 0$, $E(\nu_k^2) = 1$, $r_{k-1} = 1 + \sum_{i=1}^{k-1} \|\varphi_i\|^2$, $\bar{\varepsilon} \in \left(0, \frac{1}{2(\bar{n}_{yu}+1)}\right)$ and $\bar{n}_{yu} = \max\{n_y, n_u\} + n_y - 1$. Next, we give the assumptions for (5.4) and the stability and optimality in Proposition 5.3.

(D1) The noise $\{w_k\}$ satisfies $\lim\limits_{k\to\infty} \frac{1}{k}\sum_{j=1}^k w_j^2 = R > 0$ a.s.;

(D2) The system is of minimum phase, i.e., $B(z) \neq 0, \forall |z| \leq 1$;

(D3) $|a_{n_y}| + |b_{n_u}| \neq 0$.

PROPOSITION 5.3. [14] *If Assumptions (A1) and (D1)-(D3) hold, then the model (5.4) with the attenuating excitation control (5.5) based on the LS parameter estimate and (5.6) satisfies*

$$\limsup_{k\to\infty} \frac{1}{k}\sum_{i=1}^k \left(\|u_i\|^2 + \|y_i\|^2\right) < \infty \quad \text{a.s. } \textit{and } \lim_{k\to\infty} \frac{1}{k}\sum_{i=1}^k (y_i - y_i^*)^2 = R \quad \text{a.s.},$$

and the regressor φ_k satisfies the following excitation:

$$\lambda_{\max}(n) \triangleq \lambda_{\max}\left\{\sum_{k=1}^n \varphi_k\varphi_k^T\right\} = O(n), \lambda_{\min}(n) \triangleq \lambda_{\min}\left\{\sum_{k=1}^n \varphi_k\varphi_k^T\right\} \geq cn^{1-\bar{\varepsilon}(t+1)} \text{ a.s.} \tag{5.7}$$

for some $c > 0$, which may depend on sample paths and the $\bar{\varepsilon}$ defined below (5.6).

For the input and output signals generated by the system (5.4), by minimizing (3.2) in Algorithm 3.1, we can obtain the estimate of the sparse system parameters in (5.4). Denote the estimate as $\beta_{L,n} = [\beta_{L,n}(1), \ldots, \beta_{L,n}(n_y + n_u)]^T$, and set

$$H^* = \{i : \ a_i = 0 \text{ for } 1 \leq i \leq n_y; \text{ and } b_{i-n_y} = 0 \text{ for } n_y + 1 \leq i \leq n_y + n_u\},$$
$$H_n^* = \{i : \beta_{L,n}(i) = 0 \text{ for } 1 \leq i \leq n_y + n_u\}.$$

Then, for the estimate $\beta_{L,n}$ obtained by Algorithm 3.1 with data $\{y_{k+1}, \varphi_k\}_{k=1}^n$ generated from (5.4)-(5.6), the following theorem demonstrates the set convergence of the estimate in probability.

THEOREM 5.4. *If Assumptions (A1) and (D1)-(D3) hold, then*

$$\lim_{n\to\infty} P(H_n^* = H^*) = 1, \tag{5.8}$$

where $\lambda_n = n^\tau$ in Algorithm 3.1 with $\tau \in (\frac{1}{2}\gamma + \frac{(1-\gamma)(2-\gamma)}{8-2\gamma}, \frac{1}{2})$ and $\bar{\varepsilon} = \frac{1-\gamma}{8-2\gamma}\frac{1}{\bar{n}_{yu}+1}$ in the controller (5.6).

Proof. First, τ is well-defined, which can be verified by the following inequality:

$$\frac{1}{2} - \left(\frac{1}{2}\gamma + \frac{(1-\gamma)(2-\gamma)}{8-2\gamma}\right) = \frac{1}{2}(1-\gamma) - \frac{(1-\gamma)(2-\gamma)}{8-2\gamma} = (1-\gamma)\frac{1}{4-\gamma} > 0.$$

Denote $\lambda_{E,\max}(n) = \lambda_{\max}\left\{E\sum_{k=1}^n \varphi_k\varphi_k^T\right\}$. From (5.7) in Proposition 5.3, we have $\lambda_{E,\max}(n) = O(n)$. Moreover, we can choose $d_n = c_1 n^{1-\bar{\varepsilon}(t+1)}$ with $c_1 \leq c$. By the specification of $\bar{\varepsilon}$ below (5.6) and noting that $\tau < \frac{1}{2}$, we have $0 < \tau < \frac{1}{2} < 1-\bar{\varepsilon}(t+1) < 1$. Then, it follows

$$\frac{\log \lambda_{\max}(n)}{\lambda_{\min}(n)} = O\left(\frac{\log n}{n^{1-\bar{\varepsilon}(t+1)}}\right) \xrightarrow[n\to\infty]{} 0, \quad \frac{\lambda_n}{\lambda_{\min}(n)} = O\left(\frac{1}{n^{1-\bar{\varepsilon}(t+1)-\tau}}\right) \xrightarrow[n\to\infty]{} 0,$$
$$\frac{\lambda_n}{\lambda_{E,\max}(n)^{1/2}} = O\left(\frac{1}{n^{\frac{1}{2}-\tau}}\right) \xrightarrow[n\to\infty]{} 0, \quad \frac{\sqrt{\lambda_{E,\max}(n)}}{d_n} = O\left(\frac{1}{n^{(1-\bar{\varepsilon}(t+1))-1/2}}\right) \xrightarrow[n\to\infty]{} 0.$$

Moreover, by noting $\bar{\varepsilon} = \frac{1-\gamma}{8-2\gamma}\frac{1}{t+1}$ and $\tau > \frac{1}{2}\gamma + \frac{(1-\gamma)(2-\gamma)}{8-2\gamma}$, we have $0 < \tau - \left(\frac{1}{2}\gamma + \frac{(1-\gamma)(2-\gamma)}{8-2\gamma}\right) = \tau - \left(\frac{1}{2}\gamma + \bar{\varepsilon}(t+1)(2-\gamma)\right)$, which implies

$$\begin{aligned}\frac{\lambda_n d_n^{2-\gamma}}{\lambda_{E,\max}(n)^{2-\frac{1}{2}\gamma}} &= O\left(n^{\tau+(1-\bar{\varepsilon}(t+1))(2-\gamma)-(2-\frac{1}{2}\gamma)}\right)\\ &= O\left(n^{\tau-\left(\frac{1}{2}\gamma+\bar{\varepsilon}(t+1)(2-\gamma)\right)}\right) \xrightarrow[n\to\infty]{} \infty.\end{aligned}$$

By applying Theorem 3.10, the conclusion holds. □

Remark 5.5. The weighted L_γ regularization Algorithm 4.1 can also be applied to these two typical problems, and the analyses are similar, and hence, omitted here.

6. Simulation study. This section sets up four simulations to validate the sparse identification performance of the proposed algorithms in this paper, including two finite impulse response (FIR) systems, a polynomial expansion NARX system and a linear feedback control system. In this paper, we use the particle swarm algorithm to solve (3.3) and (4.1).

Example 1. For the simulation of sparsity and estimation performance, consider the following FIR system: $y_{k+1} = \theta^T\varphi_k + w_{k+1}$, where $\theta = (1_{q\times 1}, 0_{(30-q)\times 1})^T$ with $q = 5, 10, 15, 20, 25$, φ_k are randomly generated in the interval [-5, 5], and the noise sequence $\{w_k\}$ is i.i.d. with the Gaussian distribution $N(0, 0.1)$ and independent of $\{\varphi_k\}$. From Fig. 2, it can be seen that as the number of non-zero parameters q increases, the estimation error will be larger for the same number of samples, which also indicates that the smaller q is, the better the algorithm performs.

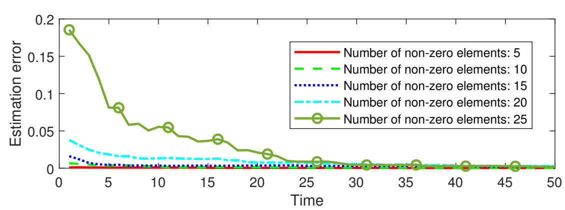

FIG. 2. *Estimation error with different number of non-zero elements*

Example 2. For the system (5.1), a common type of function expansion is the polynomial expansion [8], whose basis function is:

$$\varphi_j(x(k)) = y_{k-d_{j1}} \times \cdots \times y_{k-d_{ji}} \times u_{k-d_{j,i+1}} \times \cdots \times u_{k-d_{jl}}, \tag{6.1}$$

where $d_{j1}, \ldots, d_{jl} \in \mathbb{N}_+,\ l = 1, \ldots, M$ with M being the maximum order of the polynomial expansion. Consider such a polynomial expansion NARX model, where $M = n_u = n_y = 2$. Then, the regressor $\varphi(k)$ contains $\frac{(M+n_y+n_u)!}{M!(n_y+n_u)!} = 15$ of possible basis functions. Let the real system be $y_{k+1} = \theta^T\varphi_k + w_{k+1}$, where $\theta = [\theta_1, ..., \theta_{15}]$,

$$\begin{aligned}\varphi_k = \big[&u_k^2, u_k u_{k-1}, u_k y_k, u_k y_{k-1}, u_k, u_{k-1}^2, u_{k-1}y_k,\\ &u_{k-1}y_{k-1}, u_{k-1}, y_k^2, y_k y_{k-1}, y_k, y_{k-1}^2, y_{k-1}, 1\big]^T\end{aligned}$$

TABLE 1
The objective function and parameter settings corresponding to the algorithms

Algorithm	Objective function	Algorithm parameters
LS	$\lambda_n = 0$	
LASSO	$\gamma = 1,\ \rho = 1$	$\lambda_n = n^{0.25}$
Ridge regression	$\rho = 0$	$\lambda_n = n^{0.05}$
Elastic net	$\gamma = 1$	$\rho = 0.5, \lambda_n = n^{0.25}$
Algorithm 3.1	$\rho = 1$	$\gamma = 0.4,\ \lambda_n = n^{0.25}$

The real parameters are set as $\theta = [0, -0.5, 0.7, 0, 0.45, 0, 0, -0.006, -0.5, 0, 0.008, -0.2, 0, 1, 0]^T$. In this example, we use LS method ([6]), LASSO method ([33]), ridge regression method ([16]), elastic net method ([44]), and Algorithm 3.1 to identify the system parameters, respectively. A unified objective function of these methods takes the following form

$$J_{n+1}(\beta) = \sum_{k=1}^{n} \left(y_{k+1} - \beta^T \varphi_k\right)^2 + \lambda_n \rho \sum_{l=1}^{q} |\beta(l)|^\gamma + \lambda_n \frac{1-\rho}{2} \sum_{l=1}^{q} |\beta(l)|^2. \tag{6.2}$$

The form of the objective function corresponding to the algorithm and the parameter settings are given in Table 1. For this system, set the initial value to be i.i.d with the input $\{u_k\}$, obeying the uniform distribution $U(-1, 1)$ and the noise $\{w_k\}$, independent of $\{u_k\}$, obeying the normal distribution $N(0, 0.1)$.

Table 2 and Fig. 3 show the parameter estimation results of Algorithm 3.1, LS, LASSO, ridge regression, and elastic net with 200 observations, respectively. From Table 2 and Fig. 3, we can see that the Algorithm 3.1 has about the same accuracy in estimating the non-zero parameters as the rest of the algorithms, but at the same time, can significantly increase the accuracy of the selection of the zero parameters. When $n = 200$, the approximation solution of the estimates of the zero parameters are all less than 10^{-16}, indicating that Algorithm 3.1 performs better than the other algorithms in identifying the zero parameters. Table 2 also shows the running time of different methods. It is worth pointing out that the non-convex criterion adopted in this paper greatly improves the identification accuracy although it inevitably increases the computational complexity and the running time is relatively long.

TABLE 2
Comparison between Algorithm 3.1, LS, LASSO, Ridge regression and Elastic net under 200 observations.

Algorithms	$\theta_1 = 0$	$\theta_2 = -0.5$	$\theta_5 = 0.45$	$\theta_7 = 0$	Time
Algorithm 2.1	-2.2404×10^{-16}	-0.4950	0.4472	3.0363×10^{-17}	6.9296s
LS	-0.0011	-0.5010	0.4517	-0.0036	0.0228s
LASSO	-0.0015	-0.4958	0.4481	-7.6290×10^{-4}	0.5586s
Ridge regression	-6.9156×10^{-4}	-0.2805	0.3229	-0.0015	0.0348s
Elastic net	-0.0016	-0.4975	0.4493	-0.0022	0.6818s

Example 3. This example shows the application of the Algorithm 3.1 to the identification of a linear feedback control system. Let the ARX system be

$$y_{k+1} = \theta^T \varphi_k + w_{k+1} = \theta_1 y_k + \cdots + \theta_5 y_{k+1-5} + \theta_6 u_k + \cdots + \theta_{10} u_{k+1-5} + w_{k+1}, \tag{6.3}$$

where the true sparse parameters are $\theta = [0.5, 3, 0, -1, 0.5, 0, 0, 0, 0, 0]^T$. The noise $\{w_k\}$ is i.i.d, obeying the normal distribution $N(0, 0.025)$. The discrete reference

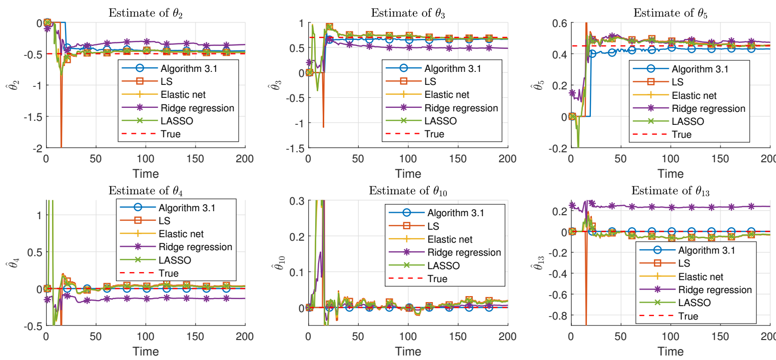

FIG. 3. *Comparison between Algorithm 3.1, LS, LASSO, Ridge regression and Elastic net.*

TABLE 3
Comparisons between Algorithm 3.1, LS, LASSO, Ridge regression and Elastic net at the 200th iteration.

	N=200		
Algorithm	$\theta_2 = 0$	$\theta_4 = 0$	$\theta_{10} = 0$
Algorithm 1	1.0433×10^{-17}	7.3991×10^{-17}	2.7050×10^{-17}
LS	0.0367	-0.0469	-0.1366
LASSO	0.0486	-0.0912	-0.1460
Ridge regression	0.0358	-0.0160	-0.0386
Elastic net	0.0481	-0.0210	-0.1476

signal is written as $y^*_{k+1} = \sin\left(\frac{1}{200}k\right),\ k \geq 0$. Let the LS estimate be $\theta_k = [\theta_k(1), \ldots, \theta_k(10)]^T$, then the self-tuning regulation control with diminishing excitation is

$$u_k = \frac{1}{\theta_k(6)}\left(y^*_{k+1} - \left(\theta_k(6)u_k - \theta_k^T\varphi_k\right)\right) + \frac{w'_k}{r_{k-1}^{\bar{\varepsilon}/2}}, \tag{6.4}$$

where $r_{k-1} = 1 + \sum_{l=1}^{k-1} \|\varphi_l\|^2$, $\bar{\varepsilon} = \frac{1}{20}$ and $\{w'_k\}$ are i.i.d with the uniform distribution $U(-0.1, 0.1)$. Fig. 4 plots the outputs of the closed-loop control system (6.3)-(6.4) and the reference signals.

For the identification problem of the closed-loop control system, Table 3 and Fig. 5 show that, as long as excitation conditions are satisfied, Algorithm 3.1 can accurately distinguish between zero and non-zero parameters, and has more precise estimates of the zero parameters than other algorithms.

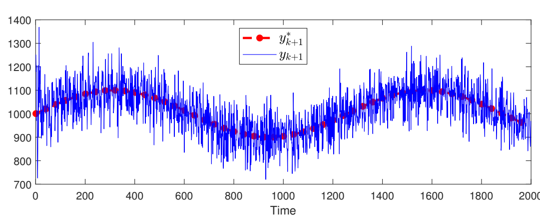

FIG. 4. *Trajectories of y_{k+1} v.s. y^*_{k+1} for Example 2.*

Example 4. This example aims to compare the performance of LS in [6], adaptive LASSO in [42] with Algorithm 4.1 in this paper. Consider the following FIR system:

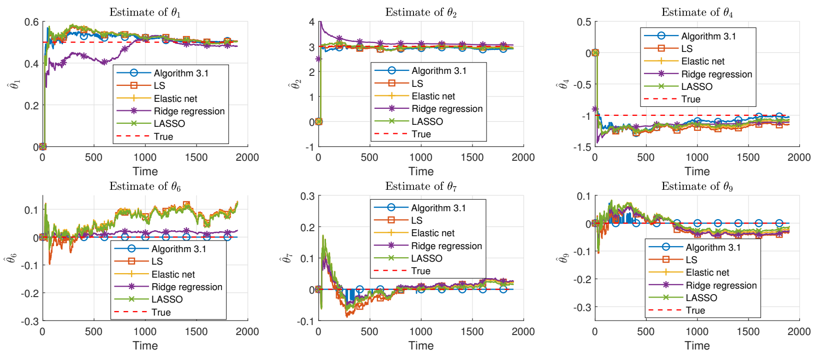

FIG. 5. *Comparisons between Algorithm 3.1, LS, LASSO, Ridge regression and Elastic net.*

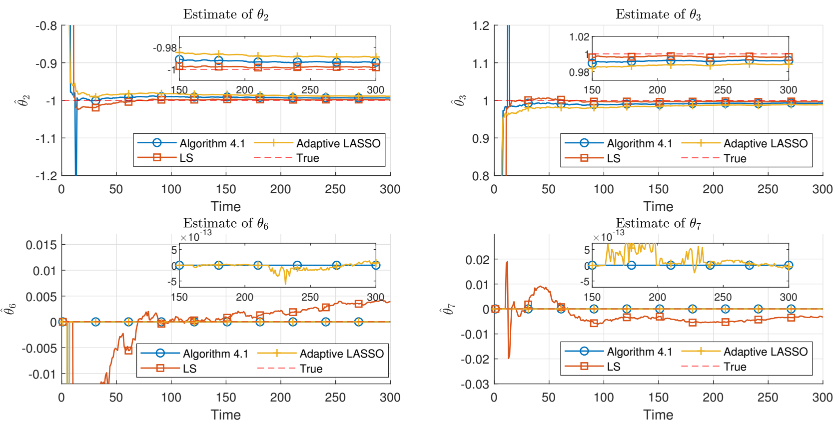

FIG. 6. *Comparison between Algorithm 4.1, LS and adaptive LASSO.*

$y_{k+1} = \theta^T \varphi_k + w_{k+1}$, where $\theta = [0, -1, 1, 2, 0.5, 0, 0, 0]^T$, φ_k are randomly generated in the interval [-5, 5], and the noise sequence $\{w_k\}$ is i.i.d. with the Gaussian distribution $N(0, 0.1)$ and independent of $\{\varphi_k\}$. Set $\lambda_n = n^{0.65}$ for the adaptive LASSO in [42] and Algorithm 4.1. It can be seen from Fig. 6, Algorithm 4.1 provides a more sparse estimate of the system parameters than LS and the algorithm in [42], and a more accurate estimate than the adaptive LASSO in [42].

7. Conclusion. This paper investigates two kinds of sparse identification algorithms based on the non-convex L_γ penalty for the stochastic systems with non-i.i.d and non-stationary observation sequences and non-i.i.d noise. First, a one-step sparse parameter identification algorithm is proposed based on the $L_\gamma(0 < \gamma < 1)$ penalty and the residual sum of squares. The almost sure convergence, the set convergence in probability, and the asymptotic normality property of the estimates generated by the proposed algorithm are established. Moreover, to improve the performance of the L_γ regularization method, a two-step algorithm based on the adaptively weighted $L_\gamma(0 < \gamma \leq 1)$ penalty is provided. Not only is the almost sure parameter convergence of the estimates established, but also the almost sure set convergence is achieved.

Compared with existing literature, the theoretical results of the algorithms in this paper are applicable to the stochastic sparse system with non-i.i.d and non-stationary observation sequences and non-i.i.d noise and the algorithms are more efficient in sparsity induction. Furthermore, these algorithms are successfully implemented in the structure selection of the NARX models and the sparse parameter identification of the linear feedback control systems.

In the future, since sparsity is often accompanied by high dimensionality, it is interesting to consider the identification of stochastic sparse systems in high dimensional settings, i.e., $p = p(n)$. Moreover, it is essential to propose a recursive algorithm for the sparse system identification, and consequently, to design controls.

REFERENCES

[1] H. Akaike, *Information theory and an extension of the maximum likelihood principle*, Selected papers of hirotugu akaike, (1998), pp. 199–213.

[2] K. J. Åström and B. Wittenmark, *On self tuning regulators*, Automatica, 9 (1973), pp. 185–199.

[3] E. J. Candès and M. B. Wakin, *An introduction to compressive sampling*, IEEE Signal Processing Magazine, 25 (2008), pp. 21–30.

[4] R. Chartrand and V. Staneva, *Restricted isometry properties and nonconvex compressive sensing*, Inverse Problems, 24 (2008), p. 035020.

[5] H. F. Chen and L. Guo, *Consistent estimation of the order of stochastic control systems*, IEEE Transactions on Automatic Control, 32 (1987), pp. 531–535.

[6] H. F. Chen and L. Guo, *Identification and stochastic adaptive control*, Springer Science & Business Media, 2012.

[7] A. Dvoretzky, *Asymptotic normality for sums of dependent random variables*, in Proceedings of the sixth Berkeley symposium on mathematical statistics and probability, vol. 2, University of California Press Berkeley, 1972, pp. 513–535.

[8] A. Falsone, L. Piroddi, and M. Prandini, *A randomized algorithm for nonlinear model structure selection*, Automatica, 60 (2015), pp. 227–238.

[9] J. Q. Fan and R. Z. Li, *Variable selection via nonconcave penalized likelihood and its oracle properties*, Journal of the American Statistical Association, 96 (2001), pp. 1348–1360.

[10] J. Q. Fan and J. C. Lv, *Nonconcave penalized likelihood with NP-dimensionality*, IEEE Transactions on Information Theory, 57 (2011), pp. 5467–5484.

[11] S. Foucart and M. J. Lai, *Sparsest solutions of underdetermined linear systems via l_q-minimization for $0 < q \leq 1$*, Applied and Computational Harmonic Analysis, 26 (2009), pp. 395–407.

[12] Y. X. Fu and W. X. Zhao, *Support recovery and parameter identification of multivariate ARMA systems with Exogenous inputs*, SIAM Journal on Control and Optimization, 61 (2023), pp. 1835–1860.

[13] A. Goldsmith, *Wireless communications*, Cambridge university press, 2005.

[14] L. Guo and H. F. Chen, *The Astrom-Wittenmark self-tuning regulator revisited and ELS-based adaptive trackers*, IEEE Transactions on Automatic Control, 36 (1991), pp. 802–812.

[15] L. Guo, H. F. Chen, and J. F. Zhang, *Consistent order estimation for linear stochastic feedback control systems (CARMA model)*, Automatica, 25 (1989), pp. 147–151.

[16] A. E. Hoerl and R. W. Kennard, *Ridge regression: Biased estimation for nonorthogonal problems*, Technometrics, 12 (1970), pp. 55–67.

[17] D. W. Huang and L. Guo, *Estimation of nonstationary ARMAX models based on the hannan-rissanen method*, The Annals of Statistics, 18 (1990), pp. 1729–1756.

[18] K. Knight and W. J. Fu, *Asymptotics for LASSO-type estimators*, Annals of statistics, (2000), pp. 1356–1378.

[19] D. Krishnan and R. Fergus, *Fast image deconvolution using hyper-laplacian priors*, Advances in Neural Information Processing Systems, 22 (2009).

[20] M. J. Lai, Y. Y. Xu, and W. T. Yin, *Improved iteratively reweighted least squares for unconstrained smoothed l_q minimization*, SIAM Journal on Numerical Analysis, 51 (2013), pp. 927–957.

[21] T. L. Lai and H. Robbins, *Consistency and asymptotic efficiency of slope estimates in stochastic approximation schemes*, Z. Wahrsch. verw. Gebiete, 56 (1981), pp. 329–360.

[22] T. L. Lai and C. Z. Wei, *Least squares estimates in stochastic regression models with applica-*

tions to identification and control of dynamic systems, The Annals of Statistics, 10 (1982), pp. 154–166.

[23] T. L. Lai and C. Z. Wei, *On the concept of excitation in least squares identification and adaptive control*, Stochastics: An International Journal of Probability and Stochastic Processes, 16 (1986), pp. 227–254.

[24] J. H. Lin and G. Michailidis, *System identification of high-dimensional linear dynamical systems with serially correlated output noise components*, IEEE Transactions on Signal Processing, 68 (2020), pp. 5573–5587.

[25] N. Liu, W. Li, Y. J. Wang, R. Tao, Q. Du, and J. Chanussot, *A survey on hyperspectral image restoration: From the view of low-rank tensor approximation*, Science China Information Sciences, 66 (2023), pp. 1–31.

[26] L. Ljung, *System identification*, Wiley encyclopedia of electrical and electronics engineering, (1999), pp. 1–19.

[27] K. Lu, H. Liu, L. Zeng, J. Y. Wang, Z. S. Zhang, and J. P. An, *Applications and prospects of artificial intelligence in covert satellite communication: a review*, Science China Information Sciences, 66 (2023), pp. 1–31.

[28] N. Meinshausen and P. Bühlmann, *Variable selection and high-dimensional graphs with the lasso*, The Annals of Statistics, 34 (2006), pp. 1436–1462.

[29] J. K. Pant, W. S. Lu, and A. Antoniou, *New improved algorithms for compressive sensing based on l_p norm*, IEEE Transactions on Circuits and Systems II: Express Briefs, 61 (2014), pp. 198–202.

[30] M. M. Petrou and C. Petrou, *Image processing: the fundamentals*, John Wiley & Sons, 2010.

[31] A. Ross and A. Jain, *Information fusion in biometrics*, Pattern Recognition Letters, 24 (2003), pp. 2115–2125.

[32] G. Schwarz, *Estimating the dimension of a model*, The Annals of Statistics, 6 (1978), pp. 461–464.

[33] R. Tibshirani, *Regression shrinkage and selection via the lasso*, Journal of the Royal Statistical Society: Series B (Methodological), 58 (1996), pp. 267–288.

[34] R. Tóth, B. M. Sanandaji, K. Poolla, and T. L. Vincent, *Compressive system identification in the linear time-invariant framework*, in 2011 50th IEEE Conference on Decision and Control and European Control Conference, IEEE, 2011, pp. 783–790.

[35] A. Wächter and L. T. Biegler, *On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming*, Mathematical Programming, 106 (2006), pp. 25–57.

[36] J. Woodworth and R. Chartrand, *Compressed sensing recovery via nonconvex shrinkage penalties*, Inverse Problems, 32 (2016), p. 075004.

[37] Z. B. Xu, X. Y. Chang, F. M. Xu, and H. Zhang, *$L_{1/2}$ regularization: A thresholding representation theory and a fast solver*, IEEE Transactions on Neural Networks and Learning Systems, 23 (2012), pp. 1013–1027.

[38] Z. B. Xu, H. Zhang, Y. Wang, X. Y. Chang, and Y. Liang, *$L_{1/2}$ regularization*, Science China Information Sciences, 53 (2010), pp. 1159–1169.

[39] L. T. Zhang and L. Guo, *Adaptive identification with guaranteed performance under saturated observation and nonpersistent excitation*, IEEE Transactions on Automatic Control, 69 (2024), pp. 1584–1599.

[40] P. Zhao and B. Yu, *On model selection consistency of Lasso*, The Journal of Machine Learning Research, 7 (2006), pp. 2541–2563.

[41] W. X. Zhao, *Parametric identification of Hammerstein systems with consistency results using stochastic inputs*, IEEE Transactions on Automatic Control, 55 (2010), pp. 474–480.

[42] W. X. Zhao, G. Yin, and E.-W. Bai, *Sparse system identification for stochastic systems with general observation sequences*, Automatica, 121 (2020), p. 109162.

[43] H. Zou, *The adaptive LASSO and its oracle properties*, Journal of the American Statistical Association, 101 (2006), pp. 1418–1429.

[44] H. Zou and T. Hastie, *Regularization and variable selection via the elastic net*, Journal of the Royal Statistical Society: series B (statistical methodology), 67 (2005), pp. 301–320.